Computers in Biology and Medicine 174 (2024) 108414

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A new family of instance-level loss functions for improving instance-level segmentation and detection of white matter hyperintensities in routine clinical brain MRI

Muhammad Febrian Rachmadi [a,b,*], Michal Byra [a,c], Henrik Skibbe [a]

[a] Brain Image Analysis Unit, RIKEN Center for Brain Science, Wako-shi, Japan
[b] Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[c] Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland

ARTICLE INFO

Keywords:
Instance-level segmentation loss
Instance-level detection loss
White matter hyperintensities
Brain lesions
Ensemble inference

ABSTRACT

In this study, we introduce "instance loss functions", a new family of loss functions designed to enhance the training of neural networks in the instance-level segmentation and detection of objects in biomedical image data, particularly those of varied numbers and sizes. Intended to be utilized conjointly with traditional loss functions, these proposed functions, prioritize object instances over pixel-by-pixel comparisons. The specific functions, the instance segmentation loss ($\mathcal{L}_{\text{instance}}$), the instance center loss ($\mathcal{L}_{\text{center}}$), the false instance rate loss ($\mathcal{L}_{\text{false}}$), and the instance proximity loss ($\mathcal{L}_{\text{proximity}}$), serve distinct purposes. Specifically, $\mathcal{L}_{\text{instance}}$ improves instance-wise segmentation quality, $\mathcal{L}_{\text{center}}$ enhances segmentation quality of small instances, $\mathcal{L}_{\text{false}}$ minimizes the rate of false and missed detections across varied instance sizes, and $\mathcal{L}_{\text{proximity}}$ improves detection quality by pulling predicted instances towards the ground truth instances. Through the task of segmenting white matter hyperintensities (WMH) in brain MRI, we benchmarked our proposed instance loss functions, both individually and in combination via an ensemble inference models approach, against traditional pixel-level loss functions. Data were sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the WMH Segmentation Challenge datasets, which exhibit significant variation in WMH instance sizes. Empirical evaluations demonstrate that combining two instance-level loss functions through ensemble inference models outperforms models using other loss function on both the ADNI and WMH Segmentation Challenge datasets for the segmentation and detection of WMH instances. Further, applying these functions to the segmentation of nuclei in histopathology images demonstrated their effectiveness and generalizability beyond WMH, improving performance even in contexts with less severe instance imbalance.

1. Introduction

Semantic segmentation is a common task in biomedical image analysis, yet challenges such as class imbalance and instance imbalance problems remain unsolved. Class imbalance in images can be seen when the number of pixels in one class is much higher than the other classes. On the other hand, instance imbalance can be seen in images where larger objects, or instances, dominate over smaller instances of the same class, often resulting in failure to detect/segment the smaller instances. Both class and instance imbalance problems are commonly observed in tasks such as segmentation of white matter hyperintensities [1,2], vascular lesions [3,4], ischemic stroke lesions [5,6], and multiple sclerosis lesions [7,8], where lesions are small compared to the background class and vary in size.

Instance imbalance is a pervasive problem that directly affects on the detection quality of small instances during segmentation tasks. In the presence of this issue, employing a pixel-level segmentation loss, such as Dice loss [9], frequently causes under-segmentation (i.e., missed detections) or over-segmentation (i.e., false detections) of small instances. This typically occurs because small instances contribute less to the Dice similarity coefficient (DSC) score than larger instances [10,11]. Fig. 1 graphically demonstrates this dilemma: while all predictions (depicted in magenta) maintain identical DSCs, the detection quality varies on an instance level, exhibiting different numbers of missed and false detections. Recognizing these nuances, our instance-level detection loss functions, which are sensitive to such disparities, can be effectively utilized to regularize object detection during the training of segmentation tasks.

* Corresponding author at: Brain Image Analysis Unit, RIKEN Center for Brain Science, Wako-shi, Japan.
E-mail addresses: febrian.rachmadi@riken.jp, febrianrachmadi@cs.ui.ac.id (M.F. Rachmadi).

https://doi.org/10.1016/j.compbiomed.2024.108414
Received 17 October 2023; Received in revised form 16 March 2024; Accepted 3 April 2024
Available online 8 April 2024

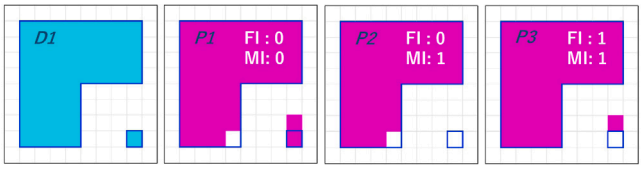

Fig. 1. Toy images (with size of 10 × 10 pixels) of a ground truth (D1 in cyan) and three different predictions (P1, P2, and P3 in magenta). The ground truth image has two instances of the same class. On pixel-level, all predictions have the same DSC score, which is 0.9796. However, on instance-level, P1 is better than P2 and P3 because P1 does not have false instances (FI) or missed instances (MI). Whereas, P3 is worse than P1 and P2 because it has the most FI and MI.

Previous deep learning segmentation methods have often employed pixel-level loss functions to evaluate the quality of the predicted segmentations against ground truth masks. Common examples of such loss functions are the cross-entropy (CE) loss [12,13], Dice loss [9], Focal loss [14], Generalized Dice loss [15], and Unified Focal loss [16]. The more recent losses, such as Unified Focal and Generalized Dice losses, have demonstrated superior performance compared to the classical CE and Dice losses when segmenting difficult-to-segment pixels. Reviews of various segmentation losses for medical image segmentation are discussed in [17,18]. However, all pixel-level loss functions tend to under-perform in segmenting small objects/instances due to their focus on individual pixels/voxels, rather than considering the context of objects-of-interest in images.

To address the issue of instance imbalance, recent studies have introduced a novel approach called the instance-level segmentation loss function. This approach assigns a loss value to each object/instance rather than to each individual pixel. An instance typically refers to a connected component in a ground truth mask. Examples of such losses include the blob loss [11] and the Instance-level and Center-of-Instance (ICI) loss [19]. However, both the blob loss and ICI loss still operate by optimizing instance-level segmentation, rather than detection. Consequently, despite the benefits of these approaches in improving the quality of segmentation results, they have shown limited improvement in object detection, particularly when object sizes vary by a wide margin.

This study provides **three main contributions** listed below.

- We introduce **"instance loss functions"**, a new family of loss functions designed to enhance the training of neural networks in the instance-level segmentation and detection of objects in biomedical image data, particularly those of varied numbers and sizes.
- We have formalized and computed ***instance-level detection loss functions from segmentation masks***, enabling them to enhance detection capability of various deep segmentation architectures when used in tandem with any segmentation loss. To the best of our knowledge, this innovative approach is unprecedented in the field.
- We demonstrate that an ensemble inference model approach, which ***allocates different models to various ranges of instance volumes***, effectively enhances the segmentation and detection quality across multiple instances-of-interest of varying sizes.

Preliminary results involving aspects of these loss functions have previously been presented in a conference paper, under the name ICI loss [19].

We evaluated our proposed loss functions on the segmentation of white matter hyperintensities (WMH) from T2-FLAIR brain MRI. WMH, presumed of vascular origin, have been identified as a predictor of stroke [20] and are linked with cognitive decline [21,22] as well as dementia progression [23]. The segmentation of WMH poses a notable challenge, particularly in the early stages when WMH manifest as small lesions, thereby complicating differentiation from normal brain tissues due to sharing similar image intensity characteristics [24,25]. If normal brain tissues were mistaken for WMH, it could negatively impact the design of clinical research trials [26]. Consequently, WMH typically present both class- and instance-imbalance problems, making it a fitting task for evaluating instance-level loss functions.

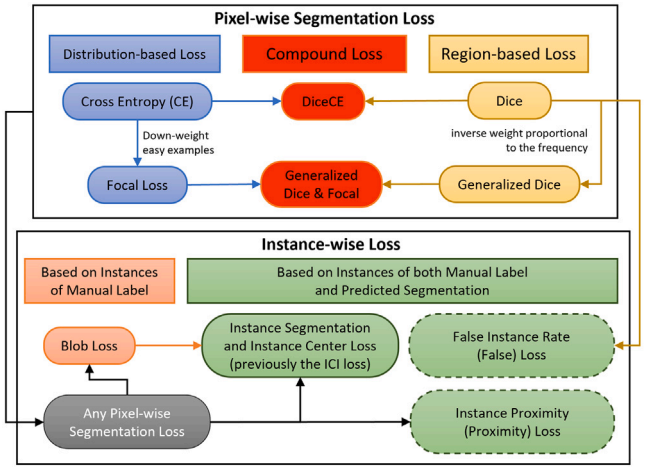

Fig. 2. Relationship between loss functions in biomedical image segmentation inspired by Ma et al. [18]. The newly proposed instance loss functions are highlighted in green color, where the instance segmentation and instance center losses were previously proposed as ICI loss [19].

2. Instance loss functions

We introduce a set of four distinct instance-level loss functions as follows:

- instance segmentation loss ($\mathcal{L}_{\text{instance}}$),
- instance center loss ($\mathcal{L}_{\text{center}}$),
- false instance rate loss ($\mathcal{L}_{\text{false}}$), and
- instance proximity loss ($\mathcal{L}_{\text{proximity}}$).

The instance-level segmentation losses $\mathcal{L}_{\text{instance}}$ and $\mathcal{L}_{\text{center}}$ were presented in our earlier conference paper [19]. Conversely, $\mathcal{L}_{\text{false}}$ and $\mathcal{L}_{\text{proximity}}$ present our newly proposed loss functions, designed to optimize instance-level detection quality, and are calculated from segmentation masks.

The relationship between the newly proposed family of instance loss functions and other loss functions is depicted in Fig. 2, inspired by a similar relationship chart of loss functions found in [18]. For clarity, DSC is defined as in Eq. (1), where TP, FP and FN represent the values of (pixel-level) true positive, false positive, and false negative, respectively.

$$DSC = \frac{2 \times \text{TP}}{\text{FP} + 2 \times \text{TP} + \text{FN}} \tag{1}$$

2.1. Formalism of instance loss functions

Let Ω be the image domain, and let $y_c : \Omega \rightarrow \{0, 1\}$ be a binary mask indicating pixels from a categorical class c. Let $\hat{y_c} : \Omega \rightarrow [0, 1]$ be a continuous predicted segmentation of a segmentation network that predicts the binary mask y_c from data. To extract individual instances of class c from both the binary mask and predicted segmentation, we perform connected component analysis (CCA). Specifically, we identify each connected component in y_c and $\hat{y_c}$ with $\mathcal{I}_{y_c,n}$ and $\mathcal{I}_{\hat{y_c},m}$, respectively, where n and m are the component numbers.

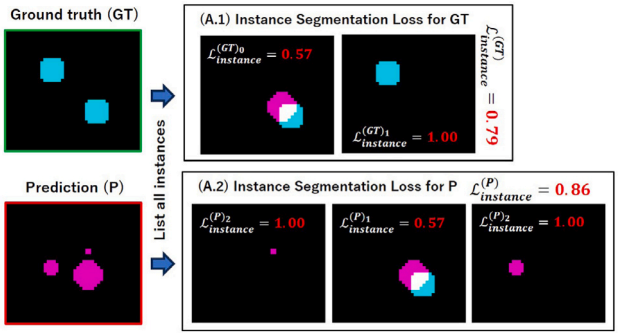

Fig. 3. Illustration of instance segmentation loss ($\mathcal{L}_{\text{instance}}$).

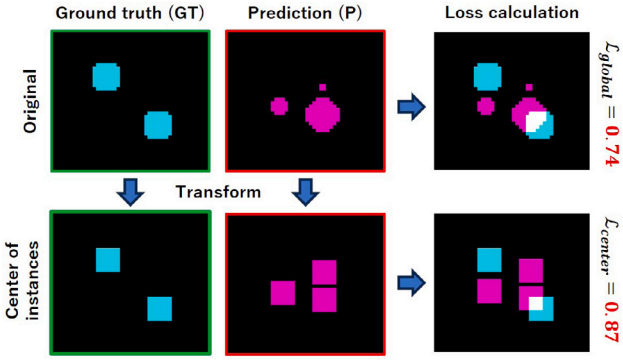

Fig. 4. Illustration of the instance center loss ($\mathcal{L}_{\text{center}}$).

2.1.1. Instance segmentation loss

The instance segmentation loss ($\mathcal{L}_{\text{instance}}$) applies a traditional loss function to each instance separately, contrasting with approaches that apply it to the entire image. The $\mathcal{L}_{\text{instance}}$ is illustrated in Fig. 3, where $\mathcal{L}^{(\text{GT})}_{\text{instance}}$ (Eq. (2)) assess instance-level segmentation quality of instances from the ground truth while $\mathcal{L}^{(\text{P})}_{\text{instance}}$ (Eq. (3)) assess instance-level segmentation quality of instances from the prediction. In the $\mathcal{L}^{(\text{GT})}_{\text{instance}}$, segmentation quality is assessed individually for each ground truth instance, by comparing it to all intersecting predicted segmentation instances while masking out any other ground truth instances. This process is formalized as the first term of function $\mathcal{L}_{seg}$ in Eq. (2), where each ground truth instance is denoted as $\mathcal{I}_{y_c,n}$ and each predicted segmentation instance is denoted as $\mathcal{I}_{\hat{y}_c,m}$. Notations c indicates the class of instance-of-interest, N is the total number of ground truth instances, M is the total number of predicted instances, Z is the total number of all ground truth instances N, and $\mathcal{L}_{seg}$ is any pixel-level segmentation loss function. A more detailed visualization of this process on toy images can be seen in Fig. 7(A.1).

$$\mathcal{L}^{(\text{GT})}_{\text{instance}} = \frac{1}{Z}\sum_{n=1}^{N}\mathcal{L}_{seg}\Big(\Big\{\mathcal{I}_{\hat{y}_c,m} \mid \Big\{\mathcal{I}_{\hat{y}_c,m}\cap\mathcal{I}_{y_c,n}\Big\}\neq\emptyset\Big\}_{m=1}^{M} \setminus \Big\{\mathcal{I}_{y_c,k} \mid k\neq n\Big\}_{k=1}^{N}, \mathcal{I}_{y_c,n}\Big) \tag{2}$$

Similarly, segmentation quality in $\mathcal{L}^{(\text{P})}_{\text{instance}}$ is assessed individually for each predicted segmentation instance, by comparing it to all intersecting ground truth instances. Similar notations as Eq. (2) are used in Eq. (3). A more detailed visualization of this process on toy images can be seen in Fig. 7(A.2).

$$\mathcal{L}^{(\text{P})}_{\text{instance}} = \frac{1}{Z}\sum_{m=1}^{M}\mathcal{L}_{seg}\Big(\mathcal{I}_{\hat{y}_c,m}, \Big\{\mathcal{I}_{y_c,n} \mid \Big\{\mathcal{I}_{y_c,n}\cap\mathcal{I}_{\hat{y}_c,m}\Big\}\neq\emptyset\Big\}_{n=1}^{N}\Big) \tag{3}$$

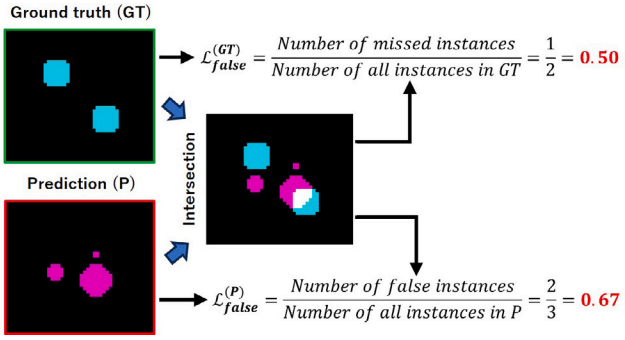

Fig. 5. Illustration of false instance rate loss ($\mathcal{L}_{\text{false}}$).

2.1.2. Instance center loss

The instance center loss $\mathcal{L}_{\text{center}}$ facilitates precise alignment by focusing on accurately matching the centers of detected instances with those in the ground truth. The $\mathcal{L}_{\text{center}}$ is illustrated in Fig. 4 and formalized in Eq. (4), which measures the segmentation quality of normalized instances where the size and shape of each instance are normalized into a square/cube (2D/3D) based on its center-of-mass. The transformation into the normalized square/cube is denoted as $C(\mathcal{I},\phi)$, where $\mathcal{I}$ is an instance to be normalized and parameter ϕ is used to control the normalized size of the center-of-mass. For example, if $\phi = 3$ then the normalized size of center-of-mass will be 3×3 in 2D or $3\times3\times3$ in 3D. A more detailed visualization of this process on toy images can be seen in Fig. 7(B).

$$\mathcal{L}_{\text{center}} = \mathcal{L}_{seg}\left(C\left(\left\{\mathcal{I}_{\hat{y}_c,m}\right\}_{m=1}^{M},\phi\right), C\left(\left\{\mathcal{I}_{y_c,n}\right\}_{n=1}^{N},\phi\right)\right) \tag{4}$$

2.1.3. False instance rate loss

The false instance rate loss ($\mathcal{L}_{\text{false}}$) aims to enhance detection quality by minimizing the rate of both false and missed detections across instances of varied sizes. The $\mathcal{L}_{\text{false}}$ is divided into two losses, which are $\mathcal{L}^{(\text{GT})}_{\text{false}}$ and $\mathcal{L}^{(\text{P})}_{\text{false}}$ where they can be calculated by counting the total numbers of missed instances (MI) and false instances (FI) divided by the total number of ground truth and predicted instances, respectively, as illustrated in Fig. 5. Both $\mathcal{L}^{(\text{GT})}_{\text{false}}$ and $\mathcal{L}^{(\text{P})}_{\text{false}}$ are formalized by Eq. (5) and Eq. (6), respectively.

$$\mathcal{L}^{(\text{GT})}_{\text{false}} = \frac{1}{N}\sum_{n=1}^{N}\left[\mathcal{L}^{\text{DL}}_{seg}\left(\left\{\mathcal{I}_{\hat{y}_c,m} \mid \left\{\mathcal{I}_{y_c,n}\cap\mathcal{I}_{\hat{y}_c,m}\right\}\neq\emptyset\right\}_{m=1}^{M}, \mathcal{I}_{y_c,n}\right)\right] \tag{5}$$

$$\mathcal{L}^{(\text{P})}_{\text{false}} = \frac{1}{M}\sum_{m=1}^{M}\left[\mathcal{L}^{\text{DL}}_{seg}\left(\mathcal{I}_{\hat{y}_c,m}, \left\{\mathcal{I}_{y_c,n} \mid \left\{\mathcal{I}_{y_c,n}\cap\mathcal{I}_{\hat{y}_c,m}\right\}\neq\emptyset\right\}_{n=1}^{N}\right)\right] \tag{6}$$

In this study, MI and FI are defined as ground truth and predicted instances, respectively, that have an "instance-level Dice segmentation loss of 1". It means that, for MI, no intersection with any predicted instances, and, for FI, no intersection with any ground truth instances, which are clearly indicated by having "instance-level Dice segmentation loss of 1". Thus, notation $\mathcal{L}^{\text{DL}}_{seg}$ in Eqs. (5) and (6) denotes the Dice segmentation loss function. Furthermore, there is a small constant number of ϵ in the implementation (not shown in Eqs. (5) and (6)) for numerical reasons:

1. avoid division by zero when N or M is equal to 0 and
2. ensure that the output of floor function applied to $\mathcal{L}^{\text{DL}}_{seg}$ for MI and FI is always 1.

Other notations follow Eq. (3). A more detailed visualization of this process on toy images can be seen in Fig. 7(C). Furthermore, it is also worth mentioning that floor function is necessary in the computation of $\mathcal{L}^{(\text{GT})}_{\text{false}}$ and $\mathcal{L}^{(\text{P})}_{\text{false}}$ in Eqs. (5) and (6). Without the floor function, $\mathcal{L}^{(\text{GT})}_{\text{false}}$ and $\mathcal{L}^{(\text{P})}_{\text{false}}$ would be the same as $\mathcal{L}^{(\text{GT})}_{\text{instance}}$ and $\mathcal{L}^{(\text{P})}_{\text{instance}}$ in Eqs. (2) and (3).

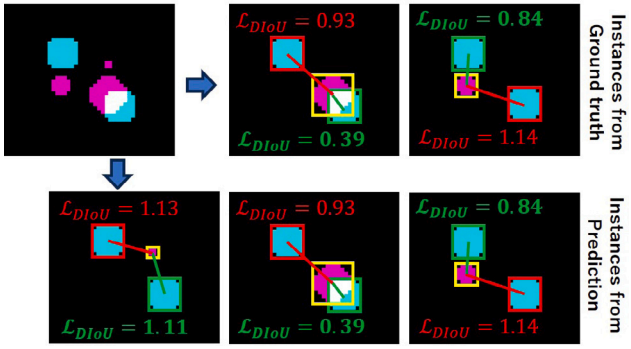

Fig. 6. Illustration of the use of Distance-IoU loss ($\mathcal{L}_{\text{DIoU}}$) in the newly proposed instance proximity loss ($\mathcal{L}_{\text{proximity}}$). $\mathcal{L}_{\text{DIoU}}$ is computed by using Eq. (7). Yellow color indicates the instance-of-interest, green color indicates the closest instance from the instance-of-interest, and the red color indicate other irrelevant instances.

2.1.4. Instance proximity loss

The instance proximity loss ($\mathcal{L}_{\text{proximity}}$) is crafted to refine detection quality of deep segmentation networks by pulling predicted segmentation instances towards the ground truth instances. The main limitation of segmentation loss functions is largely due to their inability to tell whether the predicted segmentation instances are located close to or far away from any ground truth instances. Even instance-level segmentation loss functions, such as blob loss [11] and ICI loss [19], are unable to do this based on their formulations.

To improve the detection quality of the segmentation results, the $\mathcal{L}_{\text{proximity}}$ utilizes an object detection loss named Distance-IoU loss ($\mathcal{L}_{\text{DIoU}}$) that was proposed for enhancing object detection models such as YOLOv3 and Faster R-CNN [27]. $\mathcal{L}_{\text{proximity}}$ works by calculating the distance value between the predicted instance and the ground truth instance as illustrated by using toy images in Fig. 6. The distance value produced by the $\mathcal{L}_{\text{DIoU}}$ is then used to weigh the instance segmentation loss value produced by the $\mathcal{L}_{\text{instance}}$ (Eqs. (2) and (3)) for calculating the $\mathcal{L}_{\text{proximity}}$ (Eqs. (8), (9), and (10)).

The $\mathcal{L}_{\text{DIoU}}$ is formalized in Eq. (7), which follows its original study, where m and n are indices for the predicted and ground truth instances, $\mathcal{I}_{\hat{y}_c,m}$ and $\mathcal{I}_{y_c,n}$ denote the predicted and ground truth instances, $\mathfrak{C}_{\mathcal{I}_{\hat{y}_c,m}}$ and $\mathfrak{C}_{\mathcal{I}_{y_c,n}}$ denote the centers of the bounding boxes of instances $\mathcal{I}_{\hat{y}_c,m}$ and $\mathcal{I}_{y_c,n}$, $\rho(\cdot)$ is the Euclidean distance function, and τ is the diagonal length of the smallest enclosing box covering the two boxes. If performed to all predicted and ground truth instances, $\mathcal{L}_{\text{DIoU}}$ produces an $N \times M$ matrix (i.e., denoted as $\mathcal{L}_{\text{DIoU}}^{(\mathcal{I}_{\hat{y}_c},\mathcal{I}_{y_c})}$) which describes the closeness between all N ground truth instances and all M predicted instances. In the matrix $\mathcal{L}_{\text{DIoU}}^{(\mathcal{I}_{\hat{y}_c},\mathcal{I}_{y_c})}$, value 0 indicating maximal closeness (i.e., the two bounding boxes are perfectly intersected with 0 distance between the two centers).

$$\mathcal{L}_{\text{DIoU}}\left(\mathcal{I}_{\hat{y}_c,m}, \mathcal{I}_{y_c,n}\right) = 1 - \text{IoU}(\mathcal{I}_{\hat{y}_c,m}, \mathcal{I}_{y_c,n}) + \frac{\rho^2\left(\mathfrak{C}_{\mathcal{I}_{\hat{y}_c,m}}, \mathfrak{C}_{\mathcal{I}_{y_c,n}}\right)}{\tau^2} \tag{7}$$

$$\mathcal{L}_{\text{proximity}} = \text{MSE}\left(\mathcal{L}_{\text{DIoU}}^{(\text{P})}, \mathcal{L}_{\text{DIoU}}^{(\text{GT})}\right) \tag{8}$$

$$\mathcal{L}_{\text{DIoU}}^{(\text{GT})} = \sum_{n}^{N} \min_{m \in M}\left(\mathcal{L}_{\text{DIoU}}^{(\mathcal{I}_{\hat{y}_c},\mathcal{I}_{y_c})}(n,m)\right) \cdot \mathcal{L}_{instance}^{(\text{GT})}(n) \tag{9}$$

$$\mathcal{L}_{\text{DIoU}}^{(\text{P})} = \sum_{m}^{M} \min_{n \in N}\left(\mathcal{L}_{\text{DIoU}}^{(\mathcal{I}_{\hat{y}_c},\mathcal{I}_{y_c})}(n,m)\right) \cdot \mathcal{L}_{instance}^{(\text{P})}(m) \tag{10}$$

The newly proposed $\mathcal{L}_{\text{proximity}}$ itself is formalized in Eq. (8), where it can be optimized by minimizing a mean square error (MSE) between weighted summations of minimum values for each row and column of matrix $\mathcal{L}_{\text{DIoU}}^{(\mathcal{I}_{\hat{y}_c},\mathcal{I}_{y_c})}$ as formalized by Eqs. (9) and (10), respectively. Based on Eqs. (8), (9), and (10), the newly proposed $\mathcal{L}_{\text{proximity}}$ practically calculates instance-wise segmentation loss values for all ground truth and predicted segmentation instances (i.e., $\mathcal{L}_{instance}^{(\text{GT})}$ and $\mathcal{L}_{instance}^{(\text{P})}$, respectively) and weighs them with distance values of the closest predicted segmentation and ground truth instances (i.e., $\min_{m \in M}\left(\mathcal{L}_{\text{DIoU}}^{(\mathcal{I}_{\hat{y}_c},\mathcal{I}_{y_c})}(n,m)\right)$ and $\min_{n \in N}\left(\mathcal{L}_{\text{DIoU}}^{(\mathcal{I}_{\hat{y}_c},\mathcal{I}_{y_c})}(n,m)\right)$, respectively). A more detailed visualization of this process on toy images can be seen in Fig. 7(D).

In the computation of $\mathcal{L}_{\text{proximity}}$, three edge cases arises when no instances are present in an image/patch. To navigate these, a small artificial instance of the size 3 × 3 pixels (3 × 3 × 3 pixels in case of 3D) is centrally positioned within the image/patch. The three edge cases are delineated below.

1. **Correct true negative**: No predicted segmentation instances are produced in an image/patch with no ground truth instances. In this case, $\mathcal{L}_{\text{DIoU}}$ always produces a value equals to 0.
2. **False negative**: No predicted segmentation instances are produced in an image/patch where ground truth instances are available. In this case, $\mathcal{L}_{\text{DIoU}}$ produces values that are always bigger than 0.
3. **False positive**: Predicted segmentation instances are produced in an image/patch where no ground truth instances are available. In this case, $\mathcal{L}_{\text{DIoU}}$ produces values that are always bigger than 0.

2.2. Gradients of the instance loss functions

The newly proposed detection loss functions are built upon existing pixel-level segmentation loss functions, meaning gradients are propagated according to the underlying pixel-wise segmentation loss functions. To isolate individual instances, connected component analysis (CCA) is performed to mask out instances that are not relevant to specific computations. Masking out irrelevant instances using CCA mask is based on fundamental mathematical operations, such as multiplication and addition, thereby eliminating the need for the explicit definition of new back-propagation rules for CCA and all proposed instance loss functions. In our GPU implementation, we extended the *connected_components* function from the Kornia library [28] to 3D images. Details regarding our CCA implementation on the GPU can be found in Appendix B. Fig. 7 illustrates the computational flow of all proposed instance loss functions in graph. The gradient flow is indicated in red.

Specifically, for the newly introduced detection loss functions $\mathcal{L}_{\text{false}}$ and $\mathcal{L}_{\text{proximity}}$, the gradients are derived from the instance-wise segmentation loss ($\mathcal{L}_{\text{instance}}$) as seen in Fig. 7. In scenarios where instances are not available, such as in edge cases of $\mathcal{L}_{\text{proximity}}$ loss, gradients necessary for loss value computation are provided by the pixel-wise segmentation loss ($\mathcal{L}_{\text{seg}}$).

2.3. The use of instance loss functions

In this study, we utilized instance loss functions independently, applying weights to modulate the influence of various terms in the loss computation, as formalized in Eq. (11). In total, there are 7 weights: $\alpha, \beta, \gamma, \lambda, \sigma, \omega, \delta \in \mathcal{R}^+$, each controlling the influence of its respective term in Eq. (11).

$$\begin{aligned}\mathcal{L} = {} & \alpha \cdot \mathcal{L}_{seg} + \beta \cdot \mathcal{L}_{\text{instance}}^{(\text{GT})} + \gamma \cdot \mathcal{L}_{\text{instance}}^{(\text{P})} + \lambda \cdot \mathcal{L}_{\text{center}} \\ & + \sigma \cdot \mathcal{L}_{\text{false}}^{(\text{GT})} + \omega \cdot \mathcal{L}_{\text{false}}^{(\text{P})} + \delta \cdot \mathcal{L}_{\text{proximity}}\end{aligned} \tag{11}$$

In this work, we evaluate three new compound losses derived from Eq. (11):

ICI loss: The ICI loss is a combination of a pixel-level segmentation loss, the terms $\mathcal{L}_{\text{instance}}^{(\text{GT})}$ and the instance center loss term $\mathcal{L}_{\text{center}}$. We

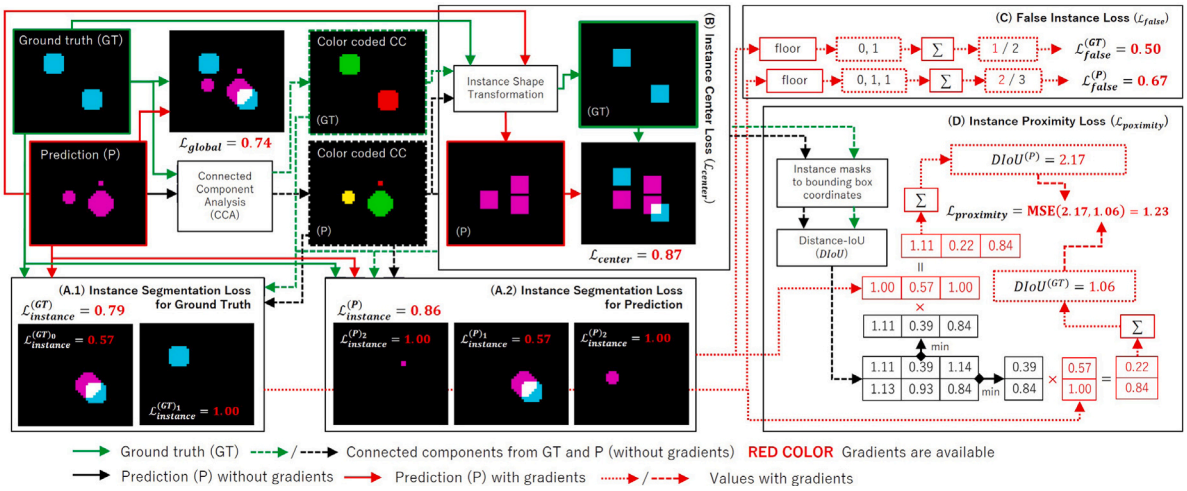

Fig. 7. Flowchart on how the proposed instance-level "Instance" family loss functions are calculated. "Instance" family loss functions consists of (A.1) and (A.2) for instance segmentation losses ($\mathcal{L}_{instance}$) for ground truth and prediction masks, respectively, (B) for instance center loss ($\mathcal{L}_{center}$), (C) for false instance loss ($\mathcal{L}_{false}$), and (D) for instance proximity loss ($\mathcal{L}_{proximity}$). Everything illustrated in red color means that the gradients are available for computation. Note that (A.1), (A.2), and (B) were first introduced as the Instance-level and Center-of-Instance (ICI) loss [19].

adhered to the advised weights $\alpha = 0.25$, $\beta = 0.5$ and $\lambda = 0.25$ [19], with

$$\text{ICI} := 0.25 \cdot \mathcal{L}_{seg} + 0.5 \cdot \mathcal{L}^{(\text{GT})}_{\text{instance}} + 0.25 \cdot \mathcal{L}_{\text{center}}. \tag{12}$$

False loss: The False loss function combines a pixel-level segmentation loss with the loss terms $\mathcal{L}^{(\text{GT})}_{\text{false}}$ and $\mathcal{L}^{(\text{P})}_{\text{false}}$ with weights equal to 1:

$$\text{False} := \mathcal{L}_{seg} + \mathcal{L}^{(\text{P})}_{\text{false}} + \mathcal{L}^{(\text{GT})}_{\text{false}}. \tag{13}$$

Proximity loss: The Proximity loss function combines a pixel-level segmentation loss term with the proximity loss term $\mathcal{L}_{\text{proximity}}$, with $\alpha = 1$ and $\delta = 0.1$:

$$\text{Proximity} := \mathcal{L}_{seg} + 0.1 \cdot \mathcal{L}_{\text{proximity}} \tag{14}$$

We determined the weights for False and Proximity through parameter tuning; see Section 3.4.

3. Experimental settings

3.1. Training dataset

The dataset used for training in this study was sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) public database [29,30]. ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The investigators within the ADNI[1] contributed to the design and implementation of ADNI and/or provided data but did neither participate in the analysis nor the writing of this paper.

The dataset contains MRI data from 20 ADNI-GO participants (12 men and 8 women, mean age at baseline 71.7(SD 7.18) years), imaged in 3 consecutive years, resulting in data from a total of 60 MRI scans. Three subjects were cognitively normal (CN), 12 had early mild cognitive impairment (EMCI), and 5 had late mild cognitive impairment (LMCI). Only T2-FLAIR MRI sequence was used as recommended by previous computational studies for WMH segmentation. The image size is 256 × 256 × 35 voxels with an anisotropic spacing of 0.8594 × 0.8594 × 5 mm^3. Ground truth was produced semi-automatically by an expert in medical image analysis using the region-growing algorithm in the Object Extractor tool in Analyze™ software. Furthermore, skull stripping was performed by using optiBET [31]. This dataset has been used in previous clinical [32,33] and computational [2,25,34] studies, which generated reference segmentations. The details of the data acquisition information are described in [2,35] while the manual WMH segmentations of the dataset are available on the data-share page [36].

The instance-imbalance problem within the dataset is demonstrated in Fig. 8, which displays the distributions of WMH instances in the ADNI dataset. Every subject exhibits multiple WMH instances of varied sizes/volumes, highlighting a clear instance-imbalance issue. Additionally, Fig. 8(C) reveals a dominance of small, or even very small, WMH instances when compared to the average volume of the adult human brain (approximately 1130 mL for women and 1260 mL for men, with some variations [37]), indicating a class-imbalance problem. The grouping of instance-wise WMH volumes in Fig. 8(C) and (D) itself was done heuristically based on our experience in previous studies [2,35], where we grouped all WMH instances based on their volumes from smaller than 0.05 mL (0.05) to bigger than 1 mL (>1) with a step of 0.05 mL between each group.

3.2. 5-Fold nested cross validation

In this study, we performed 2D experiments with full axial images of T2-FLAIR brain MRI from the ADNI dataset with image size of 256 × 256. After bias field correction [38], volume-wise intensity normalization of zero-mean unit-variance was performed, and various random data augmentation were also performed including rotation, axis flipping, and intensity scaling and shifting. No other pre-processing methods were used. We performed 5-fold nested cross validation [39] on subject-level, where 12 subjects were used in training (i.e., 48 MRI), 4 subjects were used in validation (i.e., 12 MRI), and 4 subjects were used in testing (i.e., 12 MRI). We trained *state-of-the-art* 2D SwinUNETR models [40] (with sigmoid function for binary segmentation of WMH) for 200 epochs by using the Adam optimizer [41]. A mini-batch of 2 subjects, where 24 axial slices were randomly sampled from each subject, was used for training in each optimization step. The trained 2D SwinUNETR models that produced the best Dice metric for WMH segmentation in the validation set was used in the testing. We conducted

[1] http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

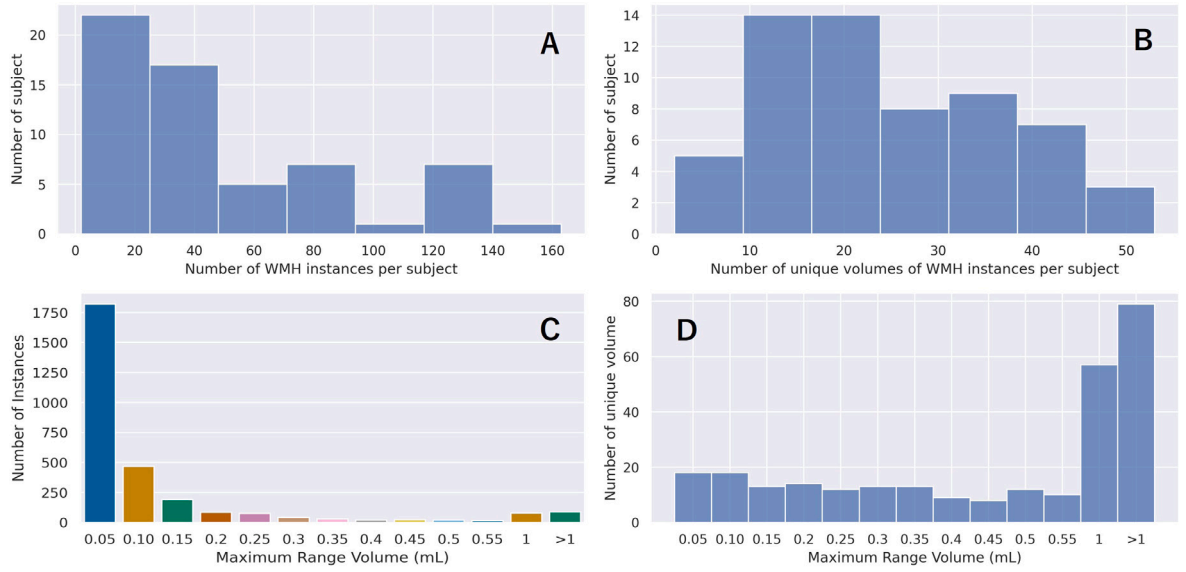

Fig. 8. Distributions of (A) number of WMH instances per subject, (B) number of unique volumes of WMH instances per subject, (C) number of WMH instances grouped by maximum range volume (in mL), and (D) WMH instances' unique volume variations per group in the ADNI dataset used in this study. For grouping: Group '0.10', for example, is for WMH instances with volumes that are bigger than or equal to 0.05 and less than 0.10 mL while Group '>1' is for WMH instances with volumes bigger than or equal to 1 mL. (D) shows that even though the majority of WMH instances are smaller than 0.05 mL (as shown in (C)), their unique volumes of WMH instances are similar to the other groups except for WMH instances in the 1 and >1 groups.

our experiments using an NVIDIA's a100 GPU 40 GB with CUDA version 11.7, Pytorch version 1.13.0, and MONAI version 1.1.0. The results on test folds are presented in Sections 4.1, 4.2, and 4.3.

3.3. Test dataset

We chose the WMH Segmentation Challenge[2] [42] to test the robustness of different loss functions on different datasets. This dataset contains data from three different institutions (i.e., Singapore, Amsterdam, and Utrecht), where each institution has 20 patients (the total is 60 MRI scans), with different data characteristics. MRI scans from Singapore, Amsterdam, and Utrecht have different dimension of $232 \times 256 \times 48$, $132 \times 256 \times 83$, and $240 \times 240 \times 48$ voxels with anisotropic spacings of $1 \times 1 \times 3$ mm^3, $1.2 \times 0.9766 \times 3$ mm^3, and $0.9583 \times 0.9583 \times 3$ mm^3, respectively. Akin to the training based on the ADNI dataset, we performed inference on 2D axial image slices. The robustness test results are presented in Section 4.4.

3.4. Loss functions for comparison

In preliminary experiments designed to validate the new losses, we explored the efficacy of various commonly-used pixel-level segmentation losses in the task of WMH segmentation and detection on the ADNI dataset namely the Dice loss [9], a compound of Dice and CE (DiceCE) losses [43], Generalized Dice loss [15], Focal loss [14], Tversky loss [44], and a compound of the Generalized Dice and Focal (GDF) losses with the recommended weights from their respective paper/documentation. Additionally, we tested two existing loss functions designed to address the instance imbalance problem; the inverse weighting Dice [45] and blob loss [11]. The results, detailed in Appendix A, revealed that the GDF loss, sourced from the MONAI library [46], yielded the most optimal results in both segmentation and detection of WMH. Additionally, when used as the primary segmentation loss for the blob loss, GDF outperformed the Dice loss. Consequently, we decided to use the GDF loss as the baseline and primary segmentation loss across all instance-level segmentation loss functions. We proceeded to test various combinations of GDF with our proposed instance-level segmentation loss functions.

We compared four different instance loss functions, the proposed ICI, False and Proximity loss functions, and the existing blob loss function. We adopted the recommended weights for the blob loss, utilizing $\alpha = 2$ for the pixel-level segmentation loss and $\beta = 1$ for the instance-level segmentation loss [11]. For the ICI loss, see Eq. (12), we adhered to the advised weights [19]. Post hyperparameter tuning for the newly proposed False and Proximity loss functions, optimal parameters emerged as $\alpha = 1$, $\sigma = 1$, and $\omega = 1$ for the former (Eq. (13)), and $\alpha = 1$ and $\delta = 0.1$ for the latter (Eq. (14)), respectively. Interestingly, these parameter values were first selected heuristically as the default parameter values in the early development, and they turned out to be the best and most effective ones in this study. Specifically, for the Proximity loss, $\delta = 0.1$ was carefully chosen to ensure the MSE value of $\mathcal{L}_{\text{proximity}}$ did not overshadow the pixel-wise segmentation loss.

3.5. Performance measurements

We evaluated the quality of 2D SwinUNETR binary segmentation results (with threshold value of 0.5 for binarization) on a global level, and an instance-level. The importance of instance-level evaluation in biomedical image analysis has been discussed thoroughly in a recent study [47]. On global-level, we measured the Dice similarity coefficient (DSC) for the quality of WMH segmentation, and computed the volumetric difference between ground truth and predicted WMH volumes (Vol.Diff). On an instance-level, we quantified the quality of WMH detections using the False Discovery Rate (FDR), False Negative Rate (FNR), Positive Predictive Value or Precision (PPV), True Positive Rate or Sensitivity/Recall (SEN), and the F1-score (F1). Furthermore, we used the Panoptic Quality (PQ) measurement [48] from the *panoptica* library [49], a measure for the quality of both instance segmentation (based on DSC) and detection (based on instance-level true positive (TP), false positive (FP), and false negative (FN) measurements) in an instance segmentation task. It is defined by

$$PQ = \frac{\sum_{(\text{GT,P}) \in \text{TP}} DSC(\text{GT}, \text{P})}{|\text{TP}| + 0.5 \cdot |\text{FP}| + 0.5 \cdot |\text{FN}|}. \tag{15}$$

To determine which loss function produced the best overall results, a numeric rank (r) was assigned to each performance measurement,

[2] https://wmh.isi.uu.nl/.

Table 1
Quantitative results for the pixel-level GDF segmentation loss and the tested instance-level loss functions for WMH the segmentation calculated on a subject-level from the ADNI dataset. Alphanumeric characters written in bold green indicate the best values/rankings in each metric while the underlined blue ones indicate the second best values/rankings.

(A) No post-processing

Naming	RA ↓	RG ↓	RI ↓	Panoptic		Global-level				Instance-level									
				PQ ↑	r	DSC ↑	r	Vol.Diff	r	FDR ↓	r	FNR ↓	r	PPV ↑	r	SEN ↑	r	F1 ↑	r
GDF	3.4	3.0	3.2	0.1610	5	0.5635	3	−1.53	3	0.5576	3	0.4199	3	0.4038	3	0.5801	3	0.4436	4
blob-GDF	**2.3**	3.0	2.2	**0.1689**	**1**	**0.5740**	**1**	−2.14	5	**0.5084**	**1**	0.4285	4	**0.4337**	**1**	0.5715	4	**0.4653**	**1**
ICI	3.8	**1.5**	5.0	0.1667	2	0.5735	2	**−1.30**	**1**	0.5720	5	0.4312	5	0.3972	5	0.5688	5	0.4340	5
False	2.8	3.0	2.6	0.1653	3	0.5633	4	−1.36	2	0.5592	4	**0.3963**	**1**	0.4015	4	**0.6037**	**1**	0.4449	3
Proximity	2.9	4.5	**2.0**	0.1626	4	0.5573	5	−1.90	4	0.5275	2	0.4190	2	0.4336	2	0.5810	2	0.4604	2

(B) Ensemble inference where blob-GDF was used to segment WMH instances that are smaller than 0.05 mL

Naming	RA ↓	RG ↓	RI ↓	Panoptic		Global-level				Instance-level									
				PQ ↑	r	DSC ↑	r	Vol.Diff	r	FDR ↓	r	FNR ↓	r	PPV ↑	r	SEN ↑	r	F1 ↑	r
GDF (+ blob-GDF)	3.9	4.5	3.6	0.1874	4	0.5666	5	−2.78	4	0.3964	2	0.5112	5	0.5424	3	0.4888	5	0.4987	3
blob-GDF	3.3	**2.0**	3.4	0.1689	5	0.5740	3	**−2.14**	**1**	0.5084	5	**0.4285**	**1**	0.4337	5	**0.5715**	**1**	0.4653	5
ICI (+ blob-GDF)	2.9	**2.0**	3.6	**0.1965**	**1**	0.5817	2	−2.52	2	0.4240	4	0.4932	3	0.5288	4	0.5068	3	0.4946	4
False (+ blob-GDF)	**2.0**	**2.0**	**2.0**	0.1959	2	**0.5840**	**1**	−2.76	3	0.3998	3	0.4791	2	0.5468	2	0.5209	2	**0.5088**	**1**
Proximity (+ blob-GDF)	3.0	4.5	2.4	0.1924	3	0.5696	4	−3.03	5	**0.3894**	**1**	0.4977	4	**0.5603**	**1**	0.5023	4	0.5002	2

such that mean ranks could be calculated (i.e., Rank All-level (RA), Rank Global-level (RG), and Rank Instance-level (RI)).

4. Results and discussions

4.1. Quantitative evaluation on the ADNI dataset

In this section, our main focus is on evaluating our new instance loss functions by using the test folds of the 5-fold nested cross validation on the ADNI dataset: ICI, False and Proximity, while also comparing them to established losses like the GDF loss and the blob-GDF loss. The latter two were selected based on their strong performance in our preliminary experiments, as detailed in Appendix A.

Table 1(A) presents the main results, revealing that there is no single loss function that excels across all performance measurements. From an overall perspective, measured by the rank average (RA), blob-GDF secured the best results with an RA = 2.3. However, our introduced False and Proximity functions performed commendably, each achieving RA = 2.8 and RA = 2.9, respectively.

Upon examining specific metrics, we observe that blob-GDF performed exceptionally well in global-level DSC and instance-level FDR, PPV, and F1 (all with $r = 1$), but fell short in instance-level FNR and SEN (each with $r = 4$), indicating it missed many ground truth instances. In contrast, the Proximity loss provided balanced results across various instance-level metrics (all with $r = 2$), but did not perform as well in global-level DSC ($r = 5$) and Vol.Diff ($r = 4$). Similarly, False achieved the best results in FNR and SEN (both with $r = 1$), indicating adept detection of ground truth instances.

Previous studies on WMH segmentation have shown that the presence of a large number of small WMH instances in a dataset can lead to lower DSC values [2,34,35,42,50]. This challenge has been attributed to the significantly more difficult task of segmenting small WMH instances compared to larger ones. In response, we have delved deeper by assessing the results in Table 1(A) at the instance level, organizing all WMH instances by their volumes heuristically, ranging from under 0.05 mL to 1.0 mL and above, as illustrated in Fig. 8(C). We found that the majority of WMH instances are smaller than 0.05 mL and exhibit distinct trends in evaluations (detailed in Appendix C). Thus, we conducted additional analyses based on the number of missed, detected, and false instances, as demonstrated in Fig. 9. Notably, while blob-GDF produced significantly fewer false instances, it also recorded the worst numbers of missed and correctly detected WMH instances. Nevertheless, due to its lower number of false instances, it managed to achieve the best global-level DSC and instance-level FDR, PPV, and F1 in Table 1(A).

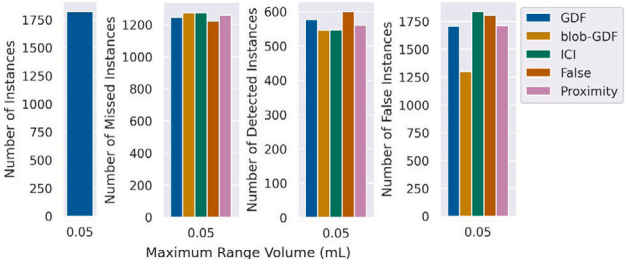

Fig. 9. Numbers of small WMH instances (lower than 0.05 mL) that are available, missed detected, correctly detected, and falsely detected by different segmentation models trained by using different instance-level loss functions from all subjects in the ADNI dataset.

4.2. Improvement on the ADNI dataset: Ensemble inference based on the volume of predicted WMH instance

The experiments in Section 4.1 on the ADNI dataset revealed that blob-GDF demonstrated superior performance in detecting WMH instances smaller than 0.05 mL, yet conversely, its performance diminished with larger instances. As Fig. 9 illustrates, a model trained with blob-GDF minimized the number of missed instances for smaller segments while maintaining comparable scores in other metrics relative to alternative approaches.

In light of this, we opted for an ensemble inference approach, pivoting on the predicted volume of WMH instances. Within this methodology, two segmentation models were independently utilized: one trained using blob-GDF, targeting WMH instances smaller than 0.05 mL, and another, trained with a different loss function, focused on larger WMH instances. The specific steps undertaken for each subject were as follows.

1. The model trained with blob-GDF loss was employed to perform the initial WMH segmentation.
2. The initial segmentation result was refined by removing WMH instances equal to or larger than 0.05 mL.
3. Subsequently, the second model, trained with an alternative loss function, was applied to produce another WMH segmentation.
4. This second segmentation result was then refined by excluding WMH instances smaller than 0.05 mL.
5. Lastly, the first and second segmentation results were merged to derive the final segmentation result.

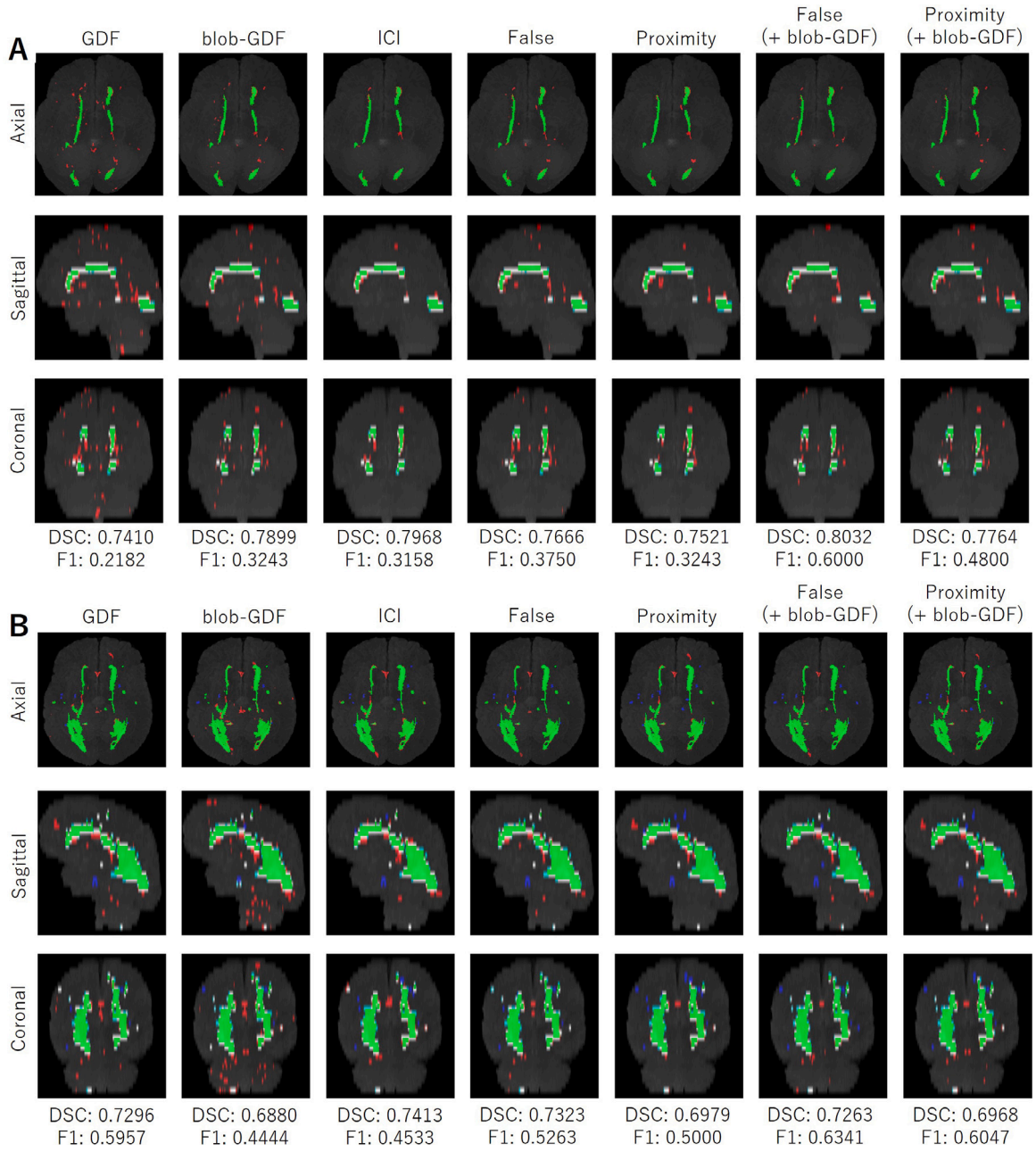

Fig. 10. Comparison of WMH detection performances of 2 different subjects from the ADNI dataset produced by different loss functions, where maximum intensity projections of T2-FLAIR images are overlaid by predicted segmentations. Green represents correct detection, blue represents missed detection, and red represent false detection. In subject (A), the performances of WMH detection (F1) improved by almost 3 folds when using the ensemble inference models of "False (+ blob-GDF)". In subject (B), only the ensemble inference models (i.e., "False (+ blob-GDF)" and "Proximity (+ blob-GDF)")) successfully improved the WMH detection performances (F1) while no significant improvements on WMH segmentation performances (DSC) were observed.

Table 1(B) presents the quantitative results, with ensemble models distinguished by a (+ blob-GDF) suffix. Notably, this table reflects an uplift in performance across most measurements for the ensemble inference models. All tested ensemble inference strategies enhanced the performance, including the pixel-level segmentation loss function of GDF. The model utilizing the False (+ blob-GDF) loss function emerged as the top performer, recording RA = RG = RI = 2.0. However, it is worth to note that the trade-off between FDR/PPV and FNR/SEN remained present, even when implementing ensemble inference models.

4.3. Qualitative results on the ADNI dataset

Fig. 10 shows qualitative results of WMH detection performances of 2 different subjects from the ADNI dataset produced by different loss functions. Because the locations of WMH instances, especially the predicted instances, could be anywhere in the brain, we purposely chose maximum intensity projections of T2-FLAIR MRI from different views (i.e., axial, sagittal, and coronal views) to show WMH detections in the whole brain. In Fig. 10, green represents correct detection, blue represents missed detection, and red represent false detection.

Table 2
Quantitative results of pixel-level GDF segmentation loss and all instance-level loss functions for WMH segmentation calculated on subject-level from the WMH Segmentation Challenge dataset. Alphanumeric characters written in bold green indicate the best values/rankings in each metric while the underlined blue ones indicate the second best values/rankings.

(A) No post-processing

Naming	RA ↓	RG ↓	RI ↓	Panoptic		Global-level				Instance-level									
				PQ ↑	r	DSC ↑	r	Vol.Diff	r	FDR ↓	r	FNR ↓	r	PPV ↑	r	SEN ↑	r	F1 ↑	r
GDF	**2.4**	2.0	**2.4**	0.1046	3	0.5694	2	1.68	2	0.4641	3	0.5162	3	**0.4258**	**1**	0.4838	3	0.4566	2
blob-GDF	2.6	**1.0**	3.6	**0.1228**	**1**	**0.5886**	**1**	**−0.78**	**1**	0.4560	2	0.5320	5	0.4242	3	0.4680	5	0.4501	3
ICI	3.1	3.0	3.4	0.1096	2	0.5654	3	2.73	3	0.4960	4	0.5118	2	0.4105	4	0.4882	2	0.4447	5
False	3.9	5.0	3.2	0.0966	5	0.5600	5	3.54	5	0.4980	5	**0.5017**	**1**	0.4002	5	**0.4983**	**1**	0.4450	4
Proximity	3.0	4.0	**2.4**	0.1001	4	0.5641	4	2.91	4	**0.4515**	**1**	0.5214	4	0.4253	2	0.4786	4	**0.4580**	**1**

(B) Ensemble inference where blob-GDF was used to segment WMH instances that are smaller than 0.05 mL

Naming	RA ↓	RG ↓	RI ↓	Panoptic		Global-level				Instance-level									
				PQ ↑	r	DSC ↑	r	Vol.Diff	r	FDR ↓	r	FNR ↓	r	PPV ↑	r	SEN ↑	r	F1 ↑	r
GDF (+ blob-GDF)	3.3	4.0	3.0	0.1479	3	0.6123	3	−2.95	5	0.3977	2	0.5692	5	**0.4772**	**1**	0.4308	5	0.4557	2
blob-GDF	3.5	3.0	3.4	0.1228	5	0.5886	5	**−0.78**	**1**	0.4560	5	**0.5320**	**1**	0.4242	5	**0.4680**	**1**	0.4501	5
ICI (+ blob-GDF)	3.1	3.0	3.0	0.1466	4	0.6071	4	−2.09	2	0.4150	4	0.5582	2	0.4681	4	0.4418	2	0.4557	3
False (+ blob-GDF)	2.9	**2.5**	3.2	0.1493	2	0.6131	2	−2.25	3	0.4057	3	0.5640	3	0.4713	3	0.4360	3	0.4544	4
Proximity (+ blob-GDF)	**2.3**	**2.5**	**2.4**	**0.1535**	**1**	**0.6148**	**1**	−2.74	4	**0.3896**	**1**	0.5660	4	0.4744	2	0.4340	4	**0.4599**	**1**

Overall, improvements on WMH instances detection (based on F1-score) were largely due to omission of false detections of small WMH instances as shown in Fig. 10. Furthermore, improvements of WMH instances detection based on F1-score did not always translate to the improvements of WMH segmentation based on DSC, as shown in Fig. 10(B). As previously discussed in Section 4.1 and Appendix C, the blob-GDF did not perform well except for small WMH instances smaller than 0.05 mL which are hard to detect even by experts. Because of that, blob-GDF produced the worsed WMH segmentation and detection results for WMH instances that are bigger than 0.05 mL as shown in Fig. 10(B). However, combining the blob-GDF with other loss functions through ensemble inference models approach showed significant improvements on both WMH segmentation and detection results.

4.4. Robustness test on the WMH challenge dataset

We conducted a robustness test employing five distinct models trained on different folds of the 5-fold nested cross-validation process, where each model was used to segment WMH in every data sample within the WMH Challenge dataset. Upon obtaining predictions from five different models, we implemented a threshold value of 0.5, and subsequently, the five values for a voxel were converged through a majority vote, labeling it as WMH if at least three different models concurred. Table 2(A) presents the results for the conventional approach (analogous to Section 4.1), whereas Table 2(B) depicts the outcomes utilizing the ensemble inference models approach (akin to Section 4.2).

As per Table 2(A), the GDF loss function surfaced as the most proficient, yielding the optimal overall score RA = 2.4, signaling balanced performance results at both global and instance levels. In contrast, the blob-GDF achieved the peak in global-level WMH segmentation and volume prediction, and the Proximity loss function achieved the best results for detecting WMH instances as per the F1-score. Notably, these robustness test outcomes varied from the quantitative results on the ADNI dataset found in Table 1(A), particularly regarding the GDF loss, attributable to the divergent characteristics inherent to the two datasets.

Conversely, Table 2(B) mirrors the results from previous experiments shown in Table 1(B) which underscores that the ensemble inference models approach was effective across different datasets, enhancing WMH segmentation and detection quality in a generalized manner. Proximity (+ blob-GDF) surfaced with the most impressive overall results, achieving RA = 2.4, RG = 2.5, and RI = 2.4, while False (+ blob-GDF) produced the second-best overall results with RA = 2.9, RG = 2.5, and RI = 3.2. It is worth to note that False (+ blob-GDF) also emerged as the best performer in the original ADNI dataset in Table 1(B), so both results in Tables 1(B) and 2(B) show that models trained by using False (+ blob-GDF) are the most robust models amongst all tested models in this study.

4.5. Generalization to other biomedical images

In order to test the potential generalizability of the proposed loss functions to other kinds of biomedical data, we trained a neural network for the segmentation of nuclei in histopathology images from the Triple Negative Breast Cancer (TNBC) dataset [51]. The TNBC dataset consists of 50 images with a resolution of 512 × 512 pixels and is publicly available at https://zenodo.org/records/1175282. Similar to the WMH experiments, we utilized the 2D SwinUNETR models with a sigmoid activation function for binary segmentation, conducted a 5-fold nested cross validation, trained for 200 epochs in each fold, and employed the Adam optimizer (detailed in Section 3.2). We also applied the same set of loss functions, which include Generalized Dice and Focal (GDF) loss as the pixel-level segmentation loss, along with blob-GDF, ICI, False, and Proximity losses with their respective weights (detailed in Section 3.4). The Proximity loss emerged as a clear winner where it produced the best performance values in all performance measurements, except for PQ, with RA = 1.1. The detailed results have been shifted to Table 5 in the Appendix. Although the instance imbalance problem in the TNBC dataset was less severe compared to the WMH experiments, the new losses clearly improved the performance of both segmentation and detection tasks.

4.6. Discussion on the limitations

While the results showed clear benefits for the proposed instance loss functions, they demanded longer training times than other loss functions. Table 3 shows quantitative measurements of training time per one mini-batch from the ADNI dataset (detailed in Section 3.2).

We identified three key factors that prolonged the computation time:

Table 3
Training time for each tested loss function per each mini-batch in millisecond (ms).

Loss	Train time per mini-batch (in ms)	
	No CCA pre-computing	With CCA pre-computing
Dice [9]	1.58	–
DiceCE [43]	5.99	–
GenDice [15]	2.55	–
Focal [14]	2.23	–
Tversky [44]	0.62	–
GDF	20.95	–
blob-Dice [11]	1625.74	660.62
blob-GDF	1734.02	771.01
ICI [19]	18,312.64	17,668.82
False	18,475.02	17,970.53
Proximity	24,592.12	23,638.50

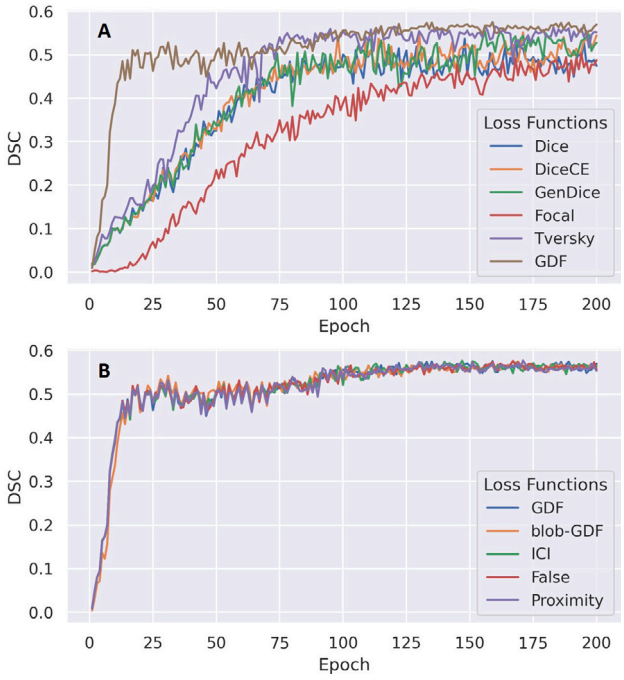

Fig. 11. Validation curves based on Dice segmentation score values produced in training processes (averaged over all 5 folds) for pixel-level segmentation loss functions in (A) and instance-level loss functions with GDF pixel-level segmentation loss function in (B).

1. **Connected component analysis (CCA)**: In level-instance loss functions, the CCA has to be done for both the ground truth and the predicted masks. Fortunately, pre-computation of CCA for ground truth images can be done for all loss functions to alleviate half of this problem, as shown in the third column of Table 3.
2. **Iterations over all ground truth and predicted instances** are required for our proposed instance loss functions. Contrary, the related blob loss [11] only requires to iterate over the ground truth instances, which gives it a computational advantage. Based on our observation, iterations over predicted instances usually take longer, especially in the earlier epochs, due to the presence of many small false instances.
3. **Other computations** are required for $\mathcal{L}_{\text{proximity}}$, such as computations for the Distance-IoU loss (i.e., $\mathcal{L}_{\text{DIoU}}$).

Another limitation of instance-level loss functions observed in this study is their validation curves (which are usually based on DSC) are similar to the pixel-instance segmentation loss function used as the main segmentation loss. For example, Fig. 11(B) shows that all instance-level loss functions produced DSC validation curves that are similar to the GDF pixel-level segmentation loss. On the other hand, Fig. 11(A) shows that some pixel-level segmentation losses, such as the Focal and GDF segmentation losses, have distinctive DSC validation curves.

It is worth mentioning that DSC is not a reliable measurement for the semantic segmentation quality of small WMH instances when they are dominated by large WMH instances, as per instance-imbalance problem definition [10,11]. Conclusively, different performance measurements are needed for validation curves to track the training progress in the presence of instance-imbalance problems in a semantic segmentation task, especially in biomedical images.

Another limitation of instance loss functions is a lot of hyperparameter tuning might be needed to get the best result for a specific task. In the newly proposed instance loss functions, there are 7 parameters to modulate the influence of various terms in the loss computation as shown in Eq. (11). In this study, we presented the best parameters that can be used for other studies, but careful hyperparameter tuning is beneficial to get the best semantic segmentation and detection results at the same time.

5. Conclusion

We introduced a novel family of loss functions tailored for addressing instance imbalance problems in biomedical imaging, enhancing both segmentation and detection tasks. Our evaluations on the ADNI [29] and WMH Segmentation Challenge [42] datasets demonstrated significant improvements in semantic segmentation and object detection. Moreover, the versatility of our loss functions was proven on the TNBC dataset [51] for nuclei segmentation tasks, where their application improved performance, even in contexts with a less pronounced instance imbalance problem.

Despite these advancements, we observed limitations including extended computational times and the complexity of managing numerous hyper-parameters. Additionally, our experiments revealed that these loss functions could not resolve the ambiguity problem in biomedical image segmentation tasks. Most falsely segmented and detected WMH instances were early-stage and often indistinguishable from normal brain tissues, a segmentation challenge observed by various studies [2,52]. Probabilistic and diffusion models, such as the Probabilistic U-Net [53] and Collectively Intelligent Medical Diffusion [54] models, have shown promising capability of dealing with ambiguous biomedical image segmentation tasks in recent years.

The introduced instance loss functions are designed for seamless integration into any deep segmentation networks allowing existing segmentation tasks to be seamlessly optimized for object detection at the same time without any architectural modifications, thereby providing potential benefits to any segmentation task that addresses the instance imbalance problem. The optimization of computational strategies and exploration of other combinations of the individual terms of the instance loss function remains a subject for future studies. For public and further development use, the instance loss functions are written in PyTorch and are available for both 2D and 3D images on GitHub (https://github.com/BrainImageAnalysis/instance-loss.git).

Table 4
Quantitative results of baseline loss functions, which are pixel-level segmentation loss functions and blob loss functions, for WMH segmentation calculated on subject-level from the ADNI dataset. Alphanumeric characters written in bold green indicate the best values/rankings in each metric while the underlined blue ones indicate the second best values/rankings.

Naming	RA	Global-level				Instance-level									
		DSC ↑	r	Vol.Diff	r	FDR ↓	r	FNR ↓	r	PPV ↑	r	SEN ↑	r	F1 ↑	r
Dice [9]	5.9	0.5519	7	−1.89	3	0.5673	7	0.4252	4	0.3805	8	0.5748	4	0.4235	8
DiceCE [43]	6.0	0.5596	5	−1.95	4	0.5538	4	0.4500	8	0.3982	6	0.5500	8	0.4280	7
Generalized Dice [15]	4.4	0.5659	3	−1.88	2	0.5696	8	0.4219	3	0.3890	7	0.5781	3	0.4344	5
Focal [14]	5.7	0.5026	8	−4.04	8	**0.4065**	**1**	0.5481	9	**0.4618**	**1**	0.4519	9	0.4367	4
Generalized Dice and Focal (GDF)	**3.0**	0.5635	4	**−1.53**	**1**	0.5576	5	0.4199	2	0.4038	4	0.5801	2	0.4435	3
Tversky [44]	4.4	0.5739	2	−2.25	7	0.5228	3	0.4441	7	0.4288	3	0.5559	7	0.4553	2
Inverse Weighting Dice [45]	6.7	0.2854	9	24.63	9	0.9265	9	**0.0781**	**1**	0.0974	9	**0.9219**	**1**	0.1585	9
blob-Dice [11]	5.4	0.5521	6	−1.99	5	0.5608	6	0.4277	5	0.4007	5	0.5723	5	0.4304	6
blob-GDF	3.4	**0.5740**	**1**	−2.14	6	0.5084	2	0.4285	6	0.4337	2	0.5715	6	**0.4653**	**1**

Table 5
Quantitative results of different loss functions for cell segmentation calculated on image-level the TNBC dataset [51]. Alphanumeric characters written in bold green indicate the best values/rankings in each metric while the underlined blue ones indicate the second best values/rankings.

Naming	RA ↓	Global-level		Instance-level										Panoptic	
		DSC ↑	r	DSC ↑	r	FDR ↓	r	FNR ↓	r	PPV ↑	r	SEN ↑	r	F1 ↑	r	PQ ↑	r
GDF	4.4	0.7411	5	0.5030	5	0.3365	5	0.0224	3	0.7609	4	0.9776	3	0.7609	5	0.4324	5
blob-GDF	3.1	0.7411	4	0.5078	2	0.3250	2	0.0240	5	0.7671	3	0.9760	5	0.7691	3	**0.4370**	**1**
ICI	3.1	0.7428	2	0.5052	4	0.3278	3	0.0225	4	0.7731	2	0.9775	4	0.7693	2	0.4326	4
False	3.3	0.7424	3	0.5641	4	0.3298	4	0.0211	2	0.7580	5	0.9789	2	0.7676	4	0.4349	2
Proximity	**1.1**	**0.7443**	**1**	**0.5115**	**1**	**0.3189**	**1**	**0.0207**	**1**	**0.7758**	**1**	**0.9793**	**1**	**0.7750**	**1**	0.4350	2

CRediT authorship contribution statement

Muhammad Febrian Rachmadi: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Michal Byra:** Writing – review & editing, Validation, Methodology. **Henrik Skibbe:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Henrik Skibbe reports financial support was provided by Japan Agency for Medical Research and Development. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Funds from RIKEN's Special Postdoctoral Researchers (SPDR) program, Japan are gratefully acknowledged (MFR). This research was also supported by the program for Brain Mapping by Integrated Neurotechnologies for Disease Studies (Brain/MINDS) from the Japan Agency for Medical Research and Development AMED (JP15dm0207001). Library access provided by the Faculty of Computer Science, Universitas Indonesia is also gratefully acknowledged.

Appendix A. Preliminary experiments: Pixel-level and blob loss functions

Firstly, we tested various pixel-wise segmentation losses and blob loss as baseline loss functions. The results are listed in Table 4 where they are evaluated on subject-level. Looking at the results, it is clear that Generalized Dice and Focal (GDF) loss is the best pixel-level segmentation loss with Rank All (RA) of 3.0. Because of that, we tested another version of blob loss that used GDF loss as its main segmentation loss (denoted as blob-GDF), instead of Dice loss like in the original study [11] denoted as blob-Dice. The results show that GDF loss improved the performance of blob loss, where it produced the best performance values of DSC on global-level and F1 on instance-level detection (highlighted as bold green values in Table 4). On the other hand, the pixel-level GDF segmentation loss produced the best performance values on Vol.Diff on global-level while having better FNR and SEN on instance-level than the blob-GDF loss. These results suggested that the pixel-level GDF segmentation loss detected true WMH instances better than blob-GDF loss, but blob-GDF loss had better performance in segmentation and detection of WMH instances in general. On the other hand, inverse weighting Dice [45] failed to produced meaningful WMH segmentation because of the highest rate of FDR (0.9265).

Appendix B. Connected components analysis on GPU

Based on the implementation of the connected components analysis (CCA) function provided by the Kornia library [28], CCA can be performed on the GPU by iteratively applying the max-pooling operation. Assume that there exist M foreground (nonzero) pixel/voxels in a 2D/3D image. Then each foreground pixel/voxel is initialized with a unique natural number. Next, the maximum-pooling operation is applied N-time to determine the connected components. If N is sufficiently large, then, after the iterations, each connected component (instance) will have a unique integer ID. Based on our experiments, a good trade-of between the number of iterations (N) and the computation complexity is setting N equals to half the image's maximum resolution. For example, for an image with 512 × 128 resolution, N would be $512 \div 2 = 256$.

Appendix C. Instance-level analysis

Based on grouping explained in Fig. 8, we performed instance level quantitative analysis by using DSC, PPV (Precision), SEN (Sensitivity), and F1 (F1-score) shown in Fig. 12A, B, C, and D, respectively. Based on these figures, we can see that all tested loss functions produced the worst results for small WMH instances that are smaller than 0.05 mL in all performance measurements.

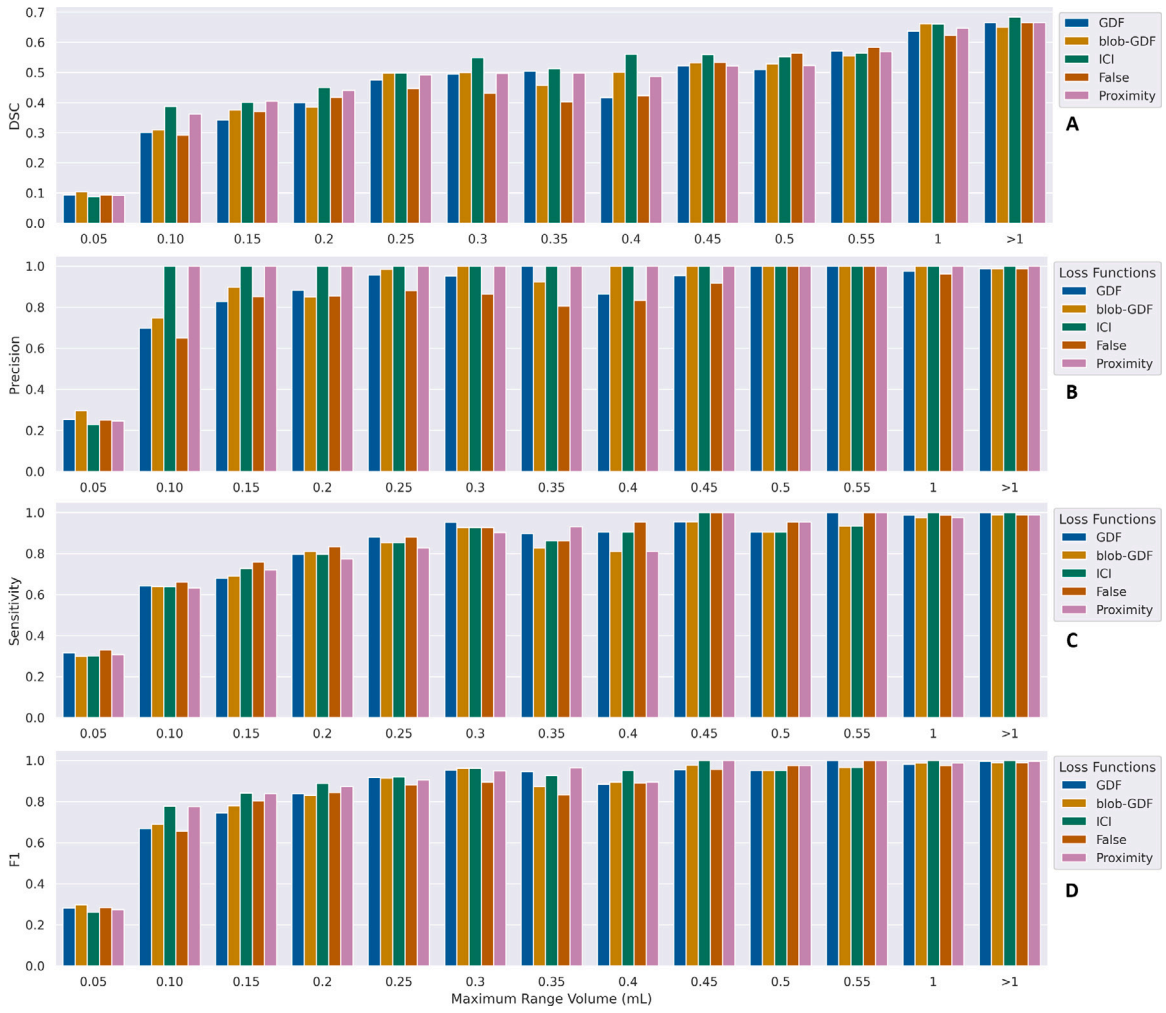

Fig. 12. Bar plots for (A) DSC, (B) precision, (C) sensitivity, and (D) F1-score performance measurements calculated on instance-level and grouped by instance's size (in mL) as per Fig. 8 from the ADNI dataset. The best average performances were produced by ICI loss (ICI) for DSC (with average of 0.4972), Proximity loss (Proximity) for precision (with average of 0.9421), False loss (False) for sensitivity (with average of 0.8562), and ICI loss (ICI) for F1 (with average of 0.8801).

References

[1] J.M. Wardlaw, E.E. Smith, G.J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R.I. Lindley, J.T. O'Brien, F. Barkhof, O.R. Benavente, S.E. Black, C. Brayne, M. Breteler, H. Chabriat, C. Decarli, F.-E. de Leeuw, F. Doubal, M. Duering, N.C. Fox, S. Greenberg, V. Hachinski, I. Kilimann, V. Mok, R.v. Oostenbrugge, L. Pantoni, O. Speck, B.C.M. Stephan, S. Teipel, A. Viswanathan, D. Werring, C. Chen, C. Smith, M. van Buchem, B. Norrving, P.B. Gorelick, M. Dichgans, STandards for ReportIng Vascular changes on nEuroimaging (STRIVE v1), Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration, Lancet Neurol. 12 (8) (2013) 822–838, http://dx.doi.org/10.1016/S1474-4422(13)70124-8.

[2] M.F. Rachmadi, M.d.C. Valdes-Hernandez, M.L.F. Agan, C. Di Perri, T. Komura, A.D.N. Initiative, et al., Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology, Comput. Med. Imaging Graph. 66 (2018) 28–43.

[3] C.H. Sudre, K. Van Wijnen, F. Dubost, H. Adams, D. Atkinson, F. Barkhof, M.A. Birhanu, E.E. Bron, R. Camarasa, N. Chaturvedi, et al., Where is VALDO? Vascular lesions detection and segmentation challenge at MICCAI 2021, Medical Image Analysis 91 (2024) 103029.

[4] R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M.d.C. Valdés-Hernández, D.A. Dickie, J. Wardlaw, et al., White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks, NeuroImage: Clin. 17 (2018) 918–934.

[5] Y. Kabir, M. Dojat, B. Scherrer, F. Forbes, C. Garbay, Multimodal MRI segmentation of ischemic stroke lesions, in: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2007, pp. 1595–1598.

[6] A. Clerigues, S. Valverde, J. Bernal, J. Freixenet, A. Oliver, X. Lladó, Acute ischemic stroke lesion core segmentation in CT perfusion images using fully convolutional neural networks, Comput. Biol. Med. 115 (2019) 103487.

[7] C. Zeng, L. Gu, Z. Liu, S. Zhao, Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI, Front. Neuroinform. 14 (2020) 610967.

[8] O. Commowick, M. Kain, R. Casey, R. Ameli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, S. Camarasu-Pop, T. Glatard, et al., Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset, Neuroimage 244 (2021) 118589.

[9] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision, 3DV, IEEE, 2016, pp. 565–571.

[10] A. Reinke, M.D. Tizabi, C.H. Sudre, M. Eisenmann, T. Rädsch, M. Baumgartner, L. Acion, M. Antonelli, T. Arbel, S. Bakas, et al., Common limitations of image processing metrics: A picture story, 2021, Eprint http://arxiv.org/abs/2104.05642. arXiv:2104.05642.

[11] F. Kofler, S. Shit, I. Ezhov, L. Fidon, I. Horvath, R. Al-Maskari, H.B. Li, H. Bhatia, T. Loehr, M. Piraud, et al., Blob loss: instance imbalance aware loss functions for semantic segmentation, in: International Conference on Information Processing in Medical Imaging, Springer, 2023, pp. 755–767.

[12] M. Yi-de, L. Qing, Q. Zhi-Bai, Automated image segmentation using improved PCNN model based on cross-entropy, in: Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004, IEEE, 2004, pp. 743–746.

[13] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[15] C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017, pp. 240–248.
[16] M. Yeung, E. Sala, C.-B. Schönlieb, L. Rundo, Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation, Comput. Med. Imaging Graph. 95 (2022) 102026.
[17] S. Jadon, A survey of loss functions for semantic segmentation, in: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB, IEEE, 2020, pp. 1–7.
[18] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, A.L. Martel, Loss odyssey in medical image segmentation, Med. Image Anal. 71 (2021) 102035.
[19] M.F. Rachmadi, C. Poon, H. Skibbe, Improving segmentation of objects with varying sizes in biomedical images using instance-wise and center-of-instance segmentation loss function, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2023.
[20] S.P. Rensma, T.T. van Sloten, L.J. Launer, C.D. Stehouwer, Cerebral small vessel disease and risk of incident stroke, dementia and depression, and all-cause mortality: A systematic review and meta-analysis, Neurosci. Biobehav. Rev. (2018).
[21] T. Pohjasvaara, R. Mäntylä, O. Salonen, H.J. Aronen, R. Ylikoski, M. Hietanen, M. Kaste, T. Erkinjuntti, How complex interactions of ischemic brain infarcts, white matter lesions, and atrophy relate to poststroke dementia, Arch. Neurol. 57 (9) (2000) 1295–1300.
[22] M.d.C. Valdés Hernández, T. Booth, C. Murray, A.J. Gow, L. Penke, Z. Morris, S.M. Maniega, N.A. Royle, B.S. Aribisala, M.E. Bastin, et al., Brain white matter damage in aging and cognitive ability in youth and older age, Neurobiol. Aging 34 (12) (2013) 2740–2747.
[23] J.M. Wardlaw, E.E. Smith, G.J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R.I. Lindley, J. T O'Brien, F. Barkhof, O.R. Benavente, et al., Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration, Lancet Neurol. 12 (8) (2013) 822–838.
[24] M.d.C. Valdés Hernández, R.J. Piper, X. Wang, I.J. Deary, J.M. Wardlaw, Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: a systematic review, J. Magn. Reson. Imaging 38 (4) (2013) 774–785.
[25] R. Maulana, M.F. Rachmadi, L. Rahadianti, Robustness of probabilistic U-net for automated segmentation of white matter hyperintensities in different datasets of brain MRI, in: 2021 International Conference on Advanced Computer Science and Information Systems, ICACSIS, IEEE, 2021, pp. 1–7.
[26] X. Wang, M.C.V. Hernández, F. Doubal, F.M. Chappell, J.M. Wardlaw, How much do focal infarcts distort white matter lesions and global cerebral atrophy measures? Cerebrovasc. Dis. 34 (5–6) (2012) 336–342.
[27] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-IoU loss: Faster and better learning for bounding box regression, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 12993–13000.
[28] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, G. Bradski, Kornia: an open source differentiable computer vision library for pytorch, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 3674–3683.
[29] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C. Jack, W. Jagust, J.Q. Trojanowski, A.W. Toga, L. Beckett, The Alzheimer's disease neuroimaging initiative, Neuroimaging Clin. North Am. 15 (4) (2005) 869–877.
[30] M.W. Weiner, D.P. Veitch, P.S. Aisen, L.A. Beckett, N.J. Cairns, R.C. Green, D. Harvey, C.R. Jack, W. Jagust, E. Liu, et al., The Alzheimer's disease neuroimaging initiative: A review of papers published since its inception, Alzheimer's Dementia 8 (1) (2012) S1–S68.
[31] E.S. Lutkenhoff, M. Rosenberg, J. Chiang, K. Zhang, J.D. Pickard, A.M. Owen, M.M. Monti, Optimized brain extraction for pathological brains (optiBET), PLoS One 9 (12) (2014) e115551.
[32] M.d.C. Valdés-Hernández, S. Reid, S. Mikhael, C. Pernet, A.D.N. Initiative, et al., Do 2-year changes in superior frontal gyrus and global brain atrophy affect cognition? Alzheimer's Dementia: Diagn. Assess. Dis. Monit. 10 (2018) 706–716.
[33] A.M. Harper, L. Clayson, J.M. Wardlaw, M.d.C. Valdés Hernández, A.D.N. Initiative, Considerations on accuracy, pattern and possible underlying factors of brain microbleed progression in older adults with absence or mild presence of vascular pathology, J. Int. Med. Res. 46 (9) (2018) 3518–3538.
[34] Y. Jeong, M.F. Rachmadi, M.d.C. Valdés-Hernández, T. Komura, Dilated saliency u-net for white matter hyperintensities segmentation using irregularity age map, Front. Aging Neurosci. 11 (2019) 150.
[35] M.F. Rachmadi, M.d.C. Valdés-Hernández, H. Li, R. Guerrero, R. Meijboom, S. Wiseman, A. Waldman, J. Zhang, D. Rueckert, J. Wardlaw, et al., Limited one-time sampling irregularity map (lots-im) for automatic unsupervised assessment of white matter hyperintensities and multiple sclerosis lesions in structural brain magnetic resonance images, Comput. Med. Imaging Graph. 79 (2020) 101685.
[36] M. Valdés Hernández, Reference Segmentations of White Matter Hyperintensities from a Subset of 20 Subjects Scanned Three Consecutive Years, 2010–2014 [dataset], University of Edinburgh, Centre for Clinical Brain Sciences, Edinburgh, 2016, http://dx.doi.org/10.7488/ds/1578.
[37] K.P. Cosgrove, C.M. Mazure, J.K. Staley, Evolving knowledge of sex differences in brain structure, function, and chemistry, Biol. Psychiatry 62 (8) (2007) 847–855.
[38] J. Sled, A. Zijdenbos, A. Evans, A nonparametric method for automatic correction of intensity nonuniformity in MRI data, IEEE Trans. Med. Imaging 17 (1) (1998) 87–97, http://dx.doi.org/10.1109/42.668698.
[39] M.P. Deisenroth, A.A. Faisal, C.S. Ong, Mathematics for Machine Learning, Cambridge University Press, 2020.
[40] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 574–584.
[41] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015, URL: http://arxiv.org/abs/1412.6980.
[42] H.J. Kuijf, J.M. Biesbroek, J. De Bresser, R. Heinen, S. Andermatt, M. Bento, M. Berseth, M. Belyaev, M.J. Cardoso, A. Casamitjana, et al., Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge, IEEE Trans. Med. Imaging 38 (11) (2019) 2556–2568.
[43] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.
[44] S.S.M. Salehi, D. Erdogmus, A. Gholipour, Tversky loss function for image segmentation using 3D fully convolutional deep networks, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2017, pp. 379–387.
[45] B. Shirokikh, A. Shevtsov, A. Kurmukov, A. Dalechina, E. Krivov, V. Kostjuchenko, A. Golanov, M. Belyaev, Universal loss reweighting to balance lesion size inequality in 3D medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 523–532.
[46] M.J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murray, A. Myronenko, C. Zhao, D. Yang, V. Nath, Y. He, Z. Xu, A. Hatamizadeh, W. Zhu, Y. Liu, M. Zheng, Y. Tang, I. Yang, M. Zephyr, B. Hashemian, S. Alle, M. Zalbagi Darestani, C. Budd, M. Modat, T. Vercauteren, G. Wang, Y. Li, Y. Hu, Y. Fu, B. Gorman, H. Johnson, B. Genereaux, B.S. Erdal, V. Gupta, A. Diaz-Pinto, A. Dourson, L. Maier-Hein, P.F. Jaeger, M. Baumgartner, J. Kalpathy-Cramer, M. Flores, J. Kirby, L.A. Cooper, H.R. Roth, D. Xu, D. Bericat, R. Floca, S.K. Zhou, H. Shuaib, K. Farahani, K.H. Maier-Hein, S. Aylward, P. Dogra, S. Ourselin, A. Feng, MONAI: An open-source framework for deep learning in healthcare, 2022, http://dx.doi.org/10.48550/arXiv.2211.02701, arXiv preprint arXiv:2211.02701.
[47] M. Eisenmann, A. Reinke, V. Weru, M.D. Tizabi, F. Isensee, T.J. Adler, S. Ali, V. Andrearczyk, M. Aubreville, U. Baid, et al., Why is the winner the best? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19955–19966.
[48] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár, Panoptic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9404–9413.
[49] F. Kofler, H. Möller, J.A. Buchner, E. de la Rosa, I. Ezhov, M. Rosier, I. Mekki, S. Shit, M. Negwer, R. Al-Maskari, et al., Panoptica–instance-wise evaluation of 3D semantic and instance segmentation maps, 2023, arXiv preprint arXiv:2312.02608.
[50] G. Park, J. Hong, B.A. Duffy, J.-M. Lee, H. Kim, White matter hyperintensities segmentation using the ensemble U-net with multi-scale highlighting foregrounds, Neuroimage 237 (2021) 118140.
[51] P. Naylor, M. Laé, F. Reyal, T. Walter, Segmentation of nuclei in histopathology images by deep regression of the distance map, IEEE Trans. Med. Imaging 38 (2) (2018) 448–459.
[52] R. Balakrishnan, M.d.C.V. Hernández, A.J. Farrall, Automatic segmentation of white matter hyperintensities from brain magnetic resonance images in the era of deep learning and big data–A systematic review, Comput. Med. Imaging Graph. 88 (2021) 101867.
[53] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J.R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, O. Ronneberger, A probabilistic u-net for segmentation of ambiguous images, Adv. Neural Inf. Process. Syst. 31 (2018).
[54] A. Rahman, J.M.J. Valanarasu, I. Hacihaliloglu, V.M. Patel, Ambiguous medical image segmentation using diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11536–11546.