GeoLS: an Intensity-based, Geodesic Soft Labeling for Image Segmentation

Sukesh Adiga Vasudeva sukesh.adiga-vasudeva.1@ens.etsmtl.ca
ETS Montreal, Canada and Polytechnique Montreal, Canada
MILA-Quebec AI Institute, Montreal, Canada

Jose Dolz jose.dolz@etsmtl.ca
ETS Montreal, Canada
International Laboratory on Learning Systems (ILLS), Canada

Herve Lombaert herve.lombaert@polymtl.ca
Polytechnique Montreal, Canada and ETS Montreal, Canada
MILA-Quebec AI Institute, Montreal, Canada

Abstract

Soft-label assignments have emerged as prominent strategies in training dense prediction problems, such as image segmentation. These approaches mitigate the limitations of hard labels, such as inter-class relationships in the data and spatial relationships between a given pixel and its neighbors. Nevertheless, most existing methods rely only on ground-truth masks and ignore the underlying image context associated with each label. For instance, image intensities convey information that could potentially clear ambiguities in the annotation. This paper, therefore, proposes a Geodesic Label Smoothing (GeoLS) approach that incorporates image intensity information within the soft labeling process. Specifically, we leverage the geodesic distance transform to capture the intensity variations between pixels. The generated maps geodesically modify the hard labels to obtain new intensity-based soft labels. The resulting geodesic soft labels better model spatial and class-wise relationships as they capture the variations of image gradients across classes and anatomy. The benefits of our intensity-based geodesic soft labels are assessed on three diverse sets of publicly accessible segmentation datasets. Our experimental results show that the proposed method consistently improves the segmentation accuracy compared to state-of-the-art soft-labeling techniques in terms of the Dice similarity and Hausdorff distance.

Keywords: Geodesic Distance, Soft Labeling, Label Smoothing, Image Segmentation.

1. Introduction

Image segmentation is a highly structured and dense prediction problem where pixels in an image are grouped into a set of target regions, such as organs or tumors (Pham et al., 2000; Suetens, 2017). It plays a pivotal role in clinical decision systems, notably in computer-assisted prognosis and diagnosis, treatment planning, and intervention support (Duncan and Ayache, 2000; Zhou et al., 2019). Recent advancements in segmentation methods are primarily due to the ability of deep learning techniques to solve such complex predictive tasks (Litjens et al., 2017; Hesamian et al., 2019). Training these approaches involves minimizing the deviation of the network predictions from the given ground-truth annotations

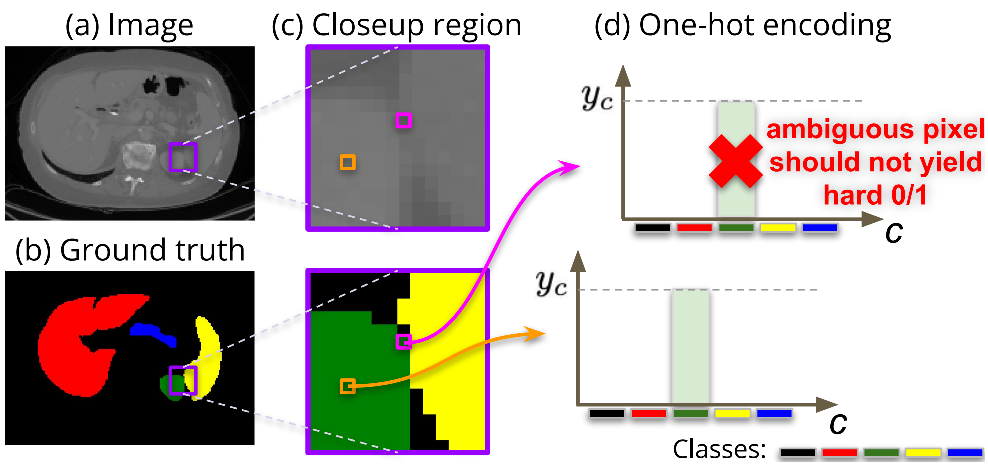


Figure 1: **Limitation of one-hot label assignments.** (a) A sample image and (b) its corresponding ground-truth mask, (c) a closeup image around the boundary region (purple), and (d) the one-hot (OH) encoding for two pixels (orange and pink in closeup images). The OH encoding of a pixel (orange) inside the kidney region (green label) may represent the true class distribution (y_c) since the label is spatially consistent with neighboring pixels. Conversely, the OH label assignment of a pixel (pink) near the boundary region may not reflect the true class distribution as it does not capture the underlying spatial ambiguities in the image. Different colors denote the class labels c.

using various objective functions (Rubinstein and Kroese, 2004; Sudre et al., 2017; Lin et al., 2017).

A common strategy to measure this deviation is to employ the cross-entropy function with the ground-truth mask represented as one-hot encoded vectors. This learning objective exhibits remarkable performance in problems needing predictions of independent classes, such as in whole-image classification (Baum and Wilczek, 1987; He et al., 2016; Szegedy et al., 2017). Nevertheless, the use of standard one-hot encoding in segmentation tasks can be sub-optimal since class predictions at each pixel are inherently conditioned with surrounding pixels. Such encoding indeed fails to capture the spatial relationships across neighborhoods as well as inter-class relationships within an image. These relationships, however, are crucial for the segmentation of medical images. For instance, labels can be similar for pixels within a homogeneous region, but vary near object boundaries due to various image ambiguities (Fig. 1). Such ambiguity can be attributed to partial volume effect, motion artifacts, or image acquisition, among other reasons. Moreover, the one-hot label assignments are solely based on the provided ground-truth masks, where the underlying spatial and inter-class relationships are disregarded. Therefore, explicitly modeling spatial and inter-class relationships in the label assignments is sought to improve the performance of the segmentation model.

Recent attempts to incorporate the inter-class relationships in the labels (Szegedy et al., 2016; Galdran et al., 2020) generally modify the hard one-hot encoding into a softer version. For instance, Label Smoothing (LS) (Szegedy et al., 2016) uniformly redistributes a portion of the target-class probability into all non-target classes to obtain a new soft label assignment for training a deep model. In (Galdran et al., 2020), a non-uniform label smoothing approach is proposed to capture the underlying structure within annotations. This method uses a Gaussian smoothing on each target class to redistribute probability over other classes. It is particularly suitable for datasets featuring ordered class labels, such as tumor or disease grading. These label-smoothing approaches, however, disregard the spatial relationships in their soft-label assignments.

To capture the spatial relationships, a few approaches alter the target segmentation mask to obtain softer labels in the boundary regions (Kats et al., 2019; Gros et al., 2021). For instance, Kats et al. (2019) generates the soft labels in the dilated regions of the target masks by adding granularity in the object boundaries. Furthermore, a Spatially-Varying Label Smoothing (SVLS) approach models the annotation ambiguity around object boundaries in target masks (Islam and Glocker, 2021). Its soft labels capture the local structural variations by applying a Gaussian-smoothing operation on the target masks. However, the annotation ambiguities of object boundaries stem from poorly defined image intensities caused by imaging techniques or existing pathologies, which inherently leads to labeling inaccuracies (Joskowicz et al., 2019; Hayward et al., 2008). These ambiguities are not captured in these soft-labeling methods, as they solely rely on the given ground-truth masks.

One solution is to incorporate image-based metrics in the soft-label assignments process. More specifically, a geodesic distance transform captures intensity variations and spatial distances within an image (Toivanen, 1996; Criminisi et al., 2008). Our approach, therefore, leverages the geodesic distance in order to capture inter-pixel and inter-class relationships during the label smoothing process. The generated soft labels thus become intensity-aware, capturing image gradient information across object boundaries. Incorporating our geodesic soft labels in model training is found to improve the segmentation performance, as they model the underlying intensity variations across objects and labels.

Our contributions: This work introduces a novel Geodesic Label Smoothing (GeoLS) approach to enhance image segmentation. Specifically, our originality lies in leveraging the geodesic distance transform to embed intensity variations in the soft-labeling process. Unlike existing soft-labeling strategies, our proposed method utilizes geodesic maps to smooth the hard labels, thus capturing the essential intensity information that is crucial for medical image segmentation. The resulting intensity-based soft labels capture class-wise relationships by considering image gradient information between two or more object categories. Furthermore, the geodesic distance between pixels captures the spatial relationships, integrating richer information than the Euclidean distance. Our GeoLS method is extensively validated across three distinct medical image segmentation benchmarks: the brain tumor dataset (Bakas et al., 2017, 2018), the abdominal organ dataset (Ma et al., 2022), and the prostatic zone dataset (Litjens et al., 2014). The findings in our experiments demonstrate the merit of GeoLS over existing soft-labeling methods.

This manuscript provides a significant extension upon our preliminary work (Adiga Vasudeva et al., 2023). Specifically, we conduct exhaustive experiments on a variety of datasets with thorough analyses to demonstrate the performance of our geodesic approach. Notably, our method is evaluated on a diversity of segmentation datasets, including tumors in brain MRIs (BraTS), multi-organs in abdominal CT scans (FLARE), and multiple zones in prostatic MRIs (ProstateX). Moreover, our experiments include comprehensive ablation studies to further highlight the effectiveness of our geodesic soft labels for image segmentation. In particular, we investigate the parameters influencing the generation of geodesic soft labels, such as studying the impact of intensity variation and different seeding strategies in obtaining our soft labels. Additionally, we conduct experiments focusing on the combination of our proposed loss with a Dice loss, a boundary loss, and a focal loss, which aim to assess the synergies in combining these approaches.

2. Related Work

2.1 Soft labeling

Soft labeling has been actively investigated in the machine learning community (Szegedy et al., 2016; Müller et al., 2019; Zhang et al., 2021). The early methods often leverage the nearest-neighbor points to obtain a soft label (Keller et al., 1985; Seo et al., 2003). Such a labeling scheme captures multiple class characteristics in the dataset, which are later used to train a classifier (El Gayar et al., 2006). More recently, Szegedy et al. (2016) proposed a label smoothing strategy for training deep neural networks. This smoothing strategy uniformly redistributes the portion of the one-hot label of a given class to all other classes. The model trained with these soft labels has been shown to improve the performance in classification tasks in both computer vision (Szegedy et al., 2016; Müller et al., 2019) and medical imaging domains (Galdran et al., 2020; He et al., 2020; Islam et al., 2020). It is also shown to be effective in handling noisy labels (Lukasik et al., 2020; Lukov et al., 2022).

In the context of image segmentation tasks, the label smoothing strategy (Szegedy et al., 2016) captures inter-class relationships within an image. However, It is also essential to consider the spatial relationships within neighboring regions. Recent approaches (Kats et al., 2019; Gros et al., 2021; Islam and Glocker, 2021) attempt to capture such relationships with spatially-varying smooth labels, improving segmentation performance. For instance, Kats et al. (2019) obtains soft labels by expanding the original binary mask using a dilation operation and subsequently assigns a soft value in the extended region. In (Gros et al., 2021), non-binary pre-processing and data augmentation techniques are employed on the target mask to obtain soft labels around the boundaries. These strategies are designed for binary segmentation tasks, where they disregard the probability distribution in the label assignments. Therefore, adopting them directly to multi-class segmentation is not trivial. A SVLS approach generates the soft labels by redistributing the class probabilities based on Gaussian filtering (Islam and Glocker, 2021). Nevertheless, these soft-labeling methods are entirely based on ground-truth masks while ignoring the ambiguities arising from image intensities. Alternately, soft labels can also be generated using multi-rater annotations (Lourenço-Silva and Oliveira, 2021). Although having multiple annotations for soft labels is ideal, it is even more expensive to obtain in practice since it requires multiple independent annotators. Furthermore, a few methods also utilize uncertainty maps for soft segmentation (Tang et al., 2022; Wang et al., 2023). Nevertheless, these methods require multiple segmentation predictions to compute uncertainty maps, which are computationally expensive. Compared to these approaches, our method leverages the geodesic distance transform (Toivanen, 1996) to capture the intensity variations in the label smoothing process. The resulting intensity-based soft labels capture spatial and class-wise relationships through the geodesic maps. Moreover, the generated soft labels are computed once and incorporated into the learning objective to train a segmentation model. Also, our method generates new soft labels from a single annotation and can be seamlessly integrated into the segmentation network.

2.2 Geodesic Distance Transform (GDT)

The GDT is commonly used for smooth and contrast-sensitive image segmentation (Criminisi et al., 2008; Protiere and Sapiro, 2007; Toivanen, 1996), as it captures the local contrast and structural information within an image. The seminal work, GeoS (Criminisi et al., 2008), proposes a generalized geodesic distance (GGD) method for segmentation tasks in an energy-based model. The effectiveness of GeoS has led to various segmentation approaches (Kontschieder et al., 2013; Wang et al., 2014; Qiu et al., 2015). For instance, Wang et al. (2014) utilizes GGDs to bring the spatial context between object boundaries in an atlas-based label propagation method. Recent approaches have leveraged GGDs in deep learning techniques to improve image segmentation (Wang et al., 2018; Bui et al., 2019; Hammoumi et al., 2021; Wei et al., 2022). For instance, Bui et al. (2019) proposes a regression of the geodesic distance maps to regularize the segmentation network through an additional prediction branch. Similarly, Ying et al. (2023) regularizes geodesic distance maps in a dual-branch network to enhance edge details for weakly supervised segmentation. To improve initial segmentation, the geodesic distance from user interactions (Wang et al., 2018) or initial network predictions (Wei et al., 2022) are employed to provide the contextual information. The resulting geodesic maps are subsequently used as additional inputs to the refinement network. These existing approaches require an extra prediction branch or refinement network to integrate the geodesic maps. In contrast, our method leverages the geodesic distance to embed underlying image context information into the label smoothing process. The generated soft labels are computed once and consequently incorporated into the learning objective to train the segmentation model. Our geodesic soft-labels, therefore, can be directly plugged into any segmentation network.

3. Method

An outline of the proposed approach comparing hard labels (OH) and existing soft labels (LS and SVLS) is shown in Fig. 2. Consider two closeup regions with the same masks but differing image intensities as in Fig. 2. The existing methods rely only on ground-truth masks to generate the soft labels. Therefore, they have the same class probability maps in both closeup regions. In contrast, our approach adds image context by leveraging geodesic distance transform in the soft-labeling process. The resulting intensity-based soft labels capture the underlying image ambiguities through geodesic maps. Thus, our method produces different class probability maps in the two closeups. The following subsections describe the label smoothing formulation and proposed geodesic soft-labeling process.

3.1 Preliminaries

Let $\{(x_i, y_i)\}_{i=1}^{N}$ indicate the training dataset with N samples, where $x_i \in \mathbb{R}^{S\times H\times W}$ represents a 3D input volume of size $S \times H \times W$, and $y_i \in \{0,1\}^{C\times S\times H\times W}$ denotes the corresponding ground truth in OH representation with C number of classes. The Cross-Entropy

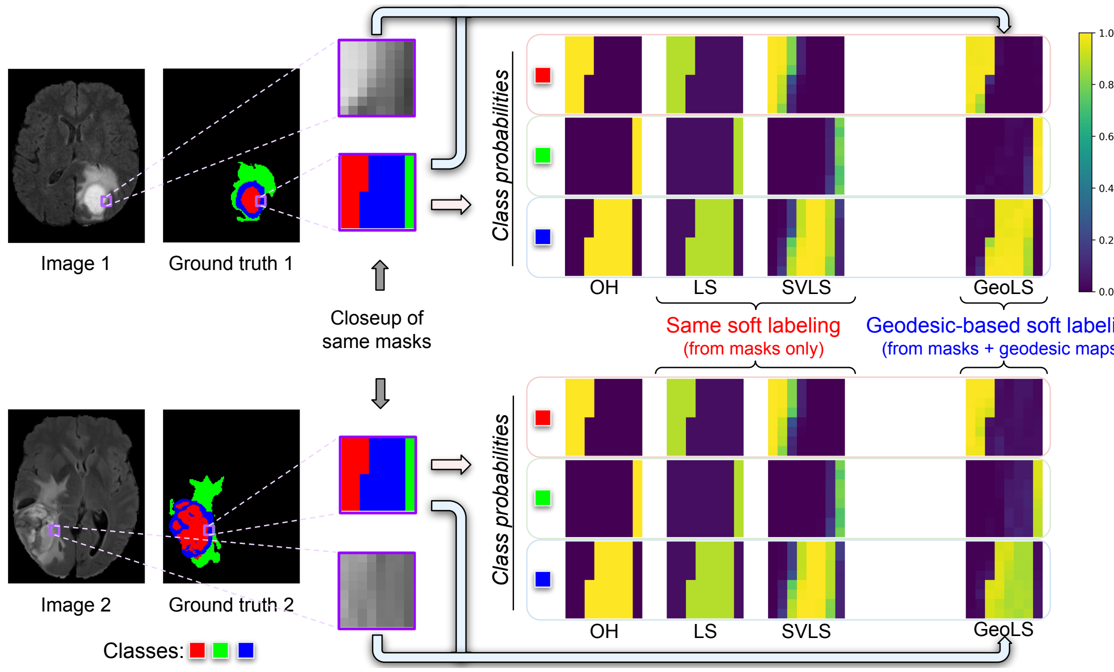


Figure 2: **Visualization of different soft labelings**. Left side: Two samples, their corresponding ground-truth masks, and closeup images having the same ground-truth masks around tumor regions. Right side: The probabilities of each class (in red, blue, and green colors) for the same closeup images from One-Hot (OH) encoding, Label Smoothing (LS), Spatially-Varying LS (SVLS), and ours (GeoLS). Since OH, LS, and SVLS are solely obtained from ground-truth masks, they have the same class probabilities maps for both closeup regions (compare top vs bottom). In contrast, our proposed method employs geodesic maps to smooth the hard labels, thus capturing intensity variations across object boundaries. Best viewed in color.

(CE) loss function for a given voxel is defined as:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log(p_c), \tag{1}$$

where p_c is the predicted softmax probability from the segmentation network. For simplicity, we use i and c notations wherever necessary and assume that the cardinality of the training set normalizes the loss function.

The OH label encoding, y_c, assigns a probability of '1' for the target class and '0' for the non-target classes. Such assignments fail to provide the model with annotation ambiguity since they do not capture the underlying inter-class relationships within the image. One way to model these relationships is by softening the hard OH encoding during the training process. For instance, the LS method (Szegedy et al., 2016) reduces the probability of the target class by a factor α and evenly distributes it across all classes. The resulting soft label for a given voxel is:

$$y_c^{LS} = (1-\alpha)y_c + \frac{\alpha}{C} \tag{2}$$

These soft labels are subsequently used in training a segmentation network by replacing the original OH label in Eq 1. This strategy has been shown to improve performance in classification tasks (Szegedy et al., 2016; He et al., 2020; Islam et al., 2020). Nevertheless, LS ignores the intrinsic spatial structure that is essential for the segmentation tasks.

3.2 Geodesic Label Smoothing (GeoLS)

Existing soft-labeling approaches modify the segmentation masks to capture the spatial relationships (Kats et al., 2019; Gros et al., 2021; Islam and Glocker, 2021), thereby accounting for the annotation ambiguities around the object boundaries. Nevertheless, they largely overlook the annotation ambiguities coming from the image intensities, being prone to annotation mistakes. To consider such image ambiguities, we integrate the geodesic distance transform (Toivanen, 1996) directly in the soft labeling of pixels. This addition captures the intensity variations as well as the spatial distance between pixels in an image. The following subsections elaborate on our geodesic label-smoothing method.

3.2.1 Generalized Geodesic Distance (GGD) Transform

The GGD transform (Criminisi et al., 2008) computes the shortest geodesic distance between a set of reference points, known as seed points, and each pixel in an image. This transform produces a distance map derived from a spatial distance and image gradient combination. The seed points can be either a single point or a set of points selected from the object of interest. Let $\mathcal{S}_c$ represent a set of seed points upon the target class c. The generalized geodesic distance of each voxel v to the set $\mathcal{S}_c$ of a target class is described as:

$$D_c(v; \mathcal{S}_c, x_i) = \min_{v' \in \mathcal{S}_c} d(v, v', x_i), \tag{3}$$

with:

$$d(v, v', x_i) = \min_{\mathbf{p} \in P_{v,v'}} \int \sqrt{||\mathbf{p}'(s)||^2 + \gamma^2(\nabla x_i \cdot \mathbf{u}(s))^2} ds, \tag{4}$$

where $P_{v,v'}$ represents the set of all paths between voxels v and v', and $\mathbf{p}(s)$ denotes one such path parameterized by $s \in [0,1]$. We define a unit vector $\mathbf{u}(s) = \frac{\mathbf{p}'(s)}{||\mathbf{p}'(s)||}$, which is tangent in the direction of the path, and whose spatial derivative is $\mathbf{p}'(s) = \frac{\partial \mathbf{p}(s)}{\partial s}$.

In Eq. 4, the first term, $\mathbf{p}'(s)$, accounts for the Euclidean distance, while the second term captures the image gradient information (∇x_i). The parameter γ, termed the geodesic factor, balances the contribution of the image gradient, and the Euclidean distance between the seed set $\mathcal{S}_c$ and each voxel in the image. When $\gamma = 0$, Eq. 4 simplifies to the Euclidean Distance, whereas setting γ to 1 facilitates computation of the geodesic distance as described in (Criminisi et al., 2008). In practice, the geodesic distance transform is optimally estimated using the raster scan algorithm (Toivanen, 1996; Criminisi et al., 2008).

An example of generating a geodesic map is shown in Fig. 3. The seed points are chosen by the skeletonization operation on a target mask. The GGD map is subsequently obtained

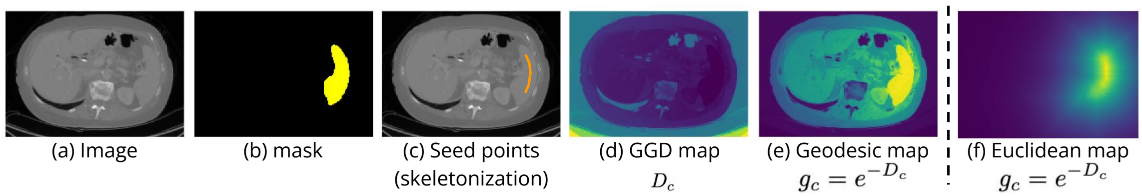


Figure 3: **Geodesic map generation.** (a) A sample image and (b) a corresponding segmentation mask of a spleen organ. (c) Seed points (orange, overlaid on the image) are derived by skeletonization of the segmentation mask. (d) The GGD map is generated from seed sets to each pixel in the image. (e) Our final geodesic map is obtained by inverting the GGD map. (f) An Euclidean map is similarly obtained for the same seed points. Notice that the Euclidean map spreads uniformly from seed points in all directions. Whereas our geodesic map spreads based on both spatial distance and gradient information, capturing the underlying intensity similarities.

using Eq. 4. To highlight the object of interest, we invert the GGD map to get the final geodesic map for each target class as follows:

$$g_c = e^{-D_c} \tag{5}$$

The resulting maps are thus in the range $[0, 1]$. The geodesic map of the background class is obtained by inverting the average of foreground geodesic maps, also in the range $[0, 1]$. In Fig. 3, we have also added an Euclidean distance map for comparison with a geodesic map. The Euclidean map spreads uniformly from seed points in all directions. In contrast, our geodesic map propagates based on both spatial distance and gradient information, capturing the underlying intensity similarities.

3.2.2 Geodesic Soft Labels

The geodesic maps encode image gradient details as a function of distance from the target objects. Such maps account for the intensity variations across object boundaries. Our approach, therefore, avails the geodesic maps for smoothing the hard labels. In order to accomplish this, we first normalize the geodesic map of each class as $\tilde{g}_c = \frac{g_c}{\sum_c g_c}$, such that it follows a probability distribution. Subsequently, the normalized geodesic maps are integrated with the original one-hot encoding to produce the new intensity-based soft labels, as defined below:

$$y_c^{GeoLS} = (1 - \alpha) y_c + \alpha \tilde{g}_c \tag{6}$$

These generated soft labels are thereafter substituted in Eq. 1 to facilitate the training of the segmentation network. The generation of our proposed geodesic soft labels is demonstrated in Fig 4. As our approach incorporates intensity variations into the target label assignments through geodesic maps, it effectively guides the network toward better segmentation.

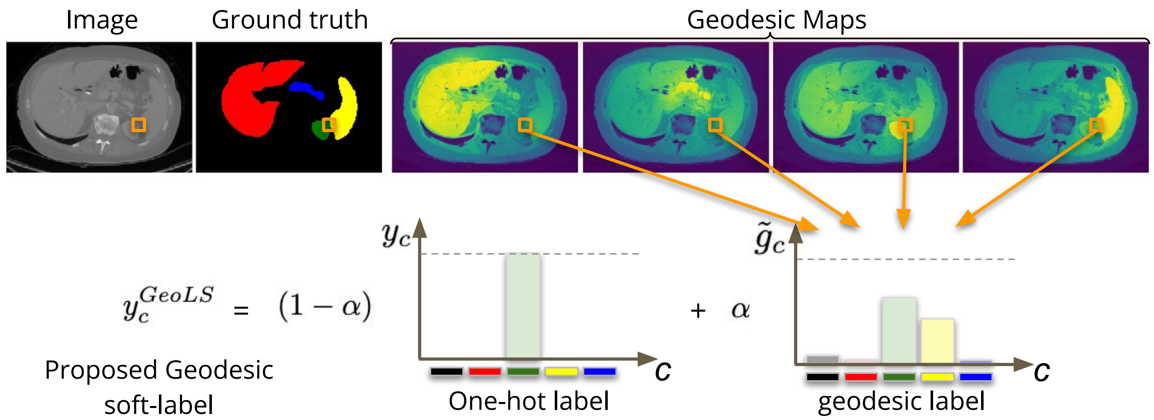


Figure 4: **Illustration of our proposed Geodesic Label Smoothing (GeoLS).** The geodesic maps for all target labels are combined to form a probability distribution. The generated geodesic label is subsequently used to modify the one-hot encoding to obtain the proposed intensity-based soft label. Our soft label captures the underlying intensity variation, thus it can better guide the segmentation network in ambiguous regions.

4. Experiments and Results

4.1 Datasets

In order to validate our geodesic label-smoothing method, we utilize three publicly accessible segmentation datasets. These datasets include: a) the Brain Tumor Segmentation dataset obtained from the 2019 BraTS challenge (Bakas et al., 2017, 2018), b) the multi-organ abdominal segmentation dataset from the 2021 FLARE challenge (Ma et al., 2022), and c) the prostatic zone segmentation dataset from the ProstateX challenge (Litjens et al., 2014). A detailed description of these datasets and our experimental settings are presented next.

a) BraTS: This dataset comprises 335 multimodal MRI volumes of the brain, containing T1, T2, FLAIR, and T1ce sequences. These volumes are preprocessed with skull-striped, co-registered to a fixed template, and resampled to an isotropic resolution of 1 mm^3. The dataset contains corresponding annotations of glioma tumors, including delineations of the necrotic and non-enhancing core, edema, and enhancing tumor regions. These regions are converted into Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) for evaluation purposes. The dataset is partitioned into 235 for training, 32 for validation, and 68 for testing across all our experiments.

b) FLARE: The dataset consists of 361 CT volumes of abdominal regions with segmentation masks of four organs: liver, kidney, spleen, and pancreas. These volumes have variable resolutions, which are standardized by resampling to a consistent resolution of $2 \times 2 \times 2.5$ mm^3. Subsequently, they are intensity normalized by retaining values within the percentile range of $[0.5, 0.95]$, as followed in the literature (Isensee et al., 2021). We employ a predefined dataset split for all experiments, allocating 260 volumes for training, 26 for validation, and the remaining 75 for testing.

c) ProstateX: The dataset includes 98 prostatic T2 MRI scans and corresponding segmentation labels of four anatomical zones, including the peripheral zone (PZ), transition zone (TZ), distal prostatic urethra (DPU), and anterior fibromuscular stroma (AFS). All volumes are resampled into a fixed resolution of $3 \times 0.5 \times 0.5\ mm^3$ as followed in (Islam and Glocker, 2021). For all our experiments, the dataset is split into 68 for training, 10 for validation, and the remaining 20 for testing.

4.2 Training and implementation details.

To assess the contribution of our geodesic soft labeling, we utilize a 3D U-net (Çiçek et al., 2016) architecture for the segmentation network. This model is trained using Adam optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-4} and weight decay of 10^{-4}. The input size of $128 \times 192 \times 192$ in BraTS, $112 \times 160 \times 208$ in FLARE, and $24 \times 320 \times 320$ in ProstateX experiments are fed into the network. The data augmentations such as random flipping and rotation are utilized, as in (Islam and Glocker, 2021). The network is trained for 200 epochs with a batch size of 4. For inference, the model with the best dice score on the validation set is selected for testing. Our evaluation includes experiments with CE, Focal Loss (FL) (Lin et al., 2017), LS (Szegedy et al., 2016), and SVLS (Islam and Glocker, 2021) losses as training objectives. Following the literature, commonly utilized hyperparameter values are considered for each baseline approach, and the result is reported for a value with the best dice score on the validation set. In particular, the focusing parameter γ in FL is set to $\{1, 2, 3\}$. In the case of LS, $\alpha \in \{0.1, 0.2, 0.3\}$ are used, whereas $\sigma \in \{0.5, 1, 2\}$ values are employed in SVLS with a kernel size of 3. In our method, the geodesic factor γ is explored for $\{0.5, 0.75, 1\}$ values with a fixed smoothing factor of $\alpha = 0.1$. To obtain the geodesic maps, an open-source library, *GeodisTK* [1], is employed with a skeletonization of a segmentation mask as seed points. Note that our soft labels are computed offline, requiring virtually no additional computation during the training process. The only additional cost is loading the geodesic maps, whose computational burden is negligible. The geodesic maps are not needed during the inference step, resulting in exactly the same computation cost as existing approaches. All our experiments were executed on an NVIDIA RTX A6000 GPU with PyTorch 1.8.0. Our GeoLS implementation is available at: `https://github.com/adigasu/GeoLS`.

4.3 Evaluation Metrics

The segmentation performance is evaluated with standard and widely used evaluation measures, such as the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD). The former measure estimates the overlap between ground truth labels and predictions, whereas the latter measures the distance between ground truth and predicted segmentation boundaries. To ensure a fair comparison, we conducted all experiments three times with fixed seed sets on identical machines, presenting results with mean and standard deviation values.

1. `https://github.com/taigw/GeodisTK`

4.4 Comparison with the state-of-the-art

The performance of the proposed geodesic soft-labeling approach is first compared with CE, FL, and state-of-the-art soft-labeling methods (LS (Szegedy et al., 2016) and SVLS (Islam and Glocker, 2021)), and their discriminative results are reported in Tables 1-3 for all three datasets. The table also includes the hyperparameter value corresponding to the best-performing model for each method.

The performance of various methods on multi-class brain tumor segmentation dataset is shown in Table 1. The results show that employing soft labels improves the segmentation performance compared to models trained with a CE loss on hard labels in both scores. Among soft-labeling baselines, FL and SVLS achieve the best DSC and HD scores, respectively. Our approach outperforms these best-performing baselines in both DSC and HD scores in all tumor categories. Notably, we observe that the proposed GeoLS indeed benefits in the enhancing tumor (ET) region. Such a region is often irregular and poorly defined, which leads to imprecise annotation (Menze et al., 2014). Our method improves this challenging region by 1.06% in DSC score and 0.45 mm in HD, highlighting the advantage of combining the intensity information in our soft labels. These results demonstrate the merit of using our geodesic soft-labeling over hard-labeling and existing soft-labeling approaches.

Table 2 presents the results of the multi-organ abdominal segmentation on the FLARE test set. A similar pattern is observable in the LS, SVLS, and GeoLS results compared to those obtained from the BraTS dataset (Table 1). Nevertheless, there is an apparent performance gap in FL compared to CE results, which may be attributed to the over-emphasis on mislabeled pixels present in the data. Overall, our GeoLS yields the best segmentation performance corresponding to the baselines, notably enhancing the segmentation in the challenging pancreas and spleen regions.

The results of the multi-class prostatic zone segmentation on the ProstateX dataset are reported in Table 3. A similar trend in FL, LS, and GeoLS results is observed as in Table 1. However, SVLS produces a drop in performance compared to CE results (HD), possibly due to the over-suppression of original one-hot encoding in the boundaries. Moreover,

Table 1: **Segmentation results on the BraTS test set.** In all tumor structures (ET, TC, WT), our method yields the best DSC and HD scores. For each tumor structure, bold and underlined indicate the best and second-best methods.

	Methods	ET	TC	WT	Average
DSC (%) $\uparrow$	CE	72.05 ± 2.14	82.38 ± 0.91	90.09 ± 0.39	81.51 ± 1.03
	FL ($\gamma = 1$)	73.55 ± 0.49	82.82 ± 0.20	90.37 ± 0.16	82.25 ± 0.20
	LS ($\alpha = 0.1$)	73.28 ± 0.85	82.65 ± 0.30	90.46 ± 0.08	82.13 ± 0.35
	SVLS ($\sigma = 1.0$)	73.15 ± 2.82	82.67 ± 1.96	90.43 ± 0.78	82.08 ± 1.81
	Ours ($\gamma = 0.75$)	**74.61 ± 0.79**	**83.51 ± 0.24**	**90.88 ± 0.12**	**83.00 ± 0.31**
HD (mm) $\downarrow$	CE	14.55 ± 1.61	7.64 ± 1.15	6.28 ± 0.86	9.49 ± 1.20
	FL ($\gamma = 1$)	12.81 ± 1.11	7.31 ± 0.32	5.96 ± 0.18	8.69 ± 0.31
	LS ($\alpha = 0.1$)	13.52 ± 0.35	7.23 ± 0.16	5.95 ± 0.16	8.90 ± 0.21
	SVLS ($\sigma = 1.0$)	12.83 ± 2.70	6.93 ± 1.37	5.72 ± 1.10	8.50 ± 1.70
	Ours ($\gamma = 0.75$)	**12.36 ± 0.56**	**6.08 ± 0.61**	**5.22 ± 0.52**	**7.89 ± 0.32**

Table 2: **Segmentation results on the FLARE test set.** Our method produces the best DSC and HD scores on average results as well as on a challenging pancreas organ. For each abdominal organ, bold and underlined indicate the best and second-best methods.

	Methods	Liver	Kidney	Spleen	Pancreas	Average
DSC (%) ↑	CE	94.88 ± 0.31	94.70 ± 0.33	95.46 ± 0.85	72.52 ± 0.61	89.39 ± 0.14
	FL ($\gamma = 1$)	94.84 ± 1.08	94.38 ± 0.35	95.56 ± 0.72	69.66 ± 2.02	88.61 ± 0.90
	LS ($\alpha = 0.1$)	**95.96 ± 1.11**	**94.89 ± 0.35**	95.61 ± 0.63	73.07 ± 1.35	89.88 ± 0.38
	SVLS ($\sigma = 0.5$)	95.76 ± 0.34	94.28 ± 0.34	95.01 ± 0.09	73.39 ± 0.16	89.61 ± 0.10
	Ours ($\gamma = 1.0$)	95.60 ± 0.87	94.80 ± 0.37	**96.52 ± 0.30**	**73.72 ± 1.02**	**90.16 ± 0.44**
HD (mm) ↓	CE	4.15 ± 1.10	2.94 ± 0.11	2.98 ± 1.06	6.72 ± 1.18	4.20 ± 0.19
	FL ($\gamma = 1$)	3.28 ± 1.28	3.22 ± 0.32	2.80 ± 1.08	8.03 ± 0.46	4.33 ± 0.61
	LS ($\alpha = 0.1$)	2.87 ± 1.14	2.93 ± 0.37	2.60 ± 0.24	6.37 ± 1.03	3.69 ± 0.26
	SVLS ($\sigma = 0.5$)	**2.61 ± 1.06**	3.17 ± 0.78	**1.42 ± 0.18**	6.26 ± 0.48	3.36 ± 0.20
	Ours ($\gamma = 1.0$)	3.01 ± 1.05	**2.40 ± 0.50**	1.49 ± 0.55	**5.59 ± 0.20**	**3.12 ± 0.21**

Table 3: **Segmentation results on the ProstateX test set.** Our method is competitive in most cases and achieves the best DSC score on average results. At the same time, baselines are ranked differently across prostatic zones (PZ, TZ, DPU, and AFS). For each prostatic zone, bold and underlined indicate the best and second-best methods.

	Methods	PZ	TZ	DPU	AFS	Average
DSC (%) ↑	CE	71.56 ± 0.55	86.34 ± 0.28	48.39 ± 2.46	38.27 ± 4.46	61.14 ± 1.21
	FL ($\gamma = 1$)	**72.18 ± 1.11**	86.38 ± 0.20	51.19 ± 2.73	35.50 ± 6.85	61.31 ± 1.96
	LS ($\alpha = 0.2$)	70.52 ± 0.31	86.34 ± 0.46	**53.31 ± 2.89**	35.16 ± 6.65	61.33 ± 1.29
	SVLS ($\sigma = 1.0$)	72.08 ± 1.89	85.89 ± 0.64	51.10 ± 4.14	35.67 ± 3.08	61.19 ± 2.12
	Ours ($\gamma = 1.0$)	70.86 ± 1.11	**86.51 ± 0.36**	51.50 ± 0.50	**39.50 ± 2.60**	**62.09 ± 0.75**
HD (mm) ↓	CE	6.51 ± 0.34	**3.22 ± 0.10**	11.28 ± 0.44	**9.58 ± 1.21**	7.65 ± 0.24
	FL ($\gamma = 1$)	**5.76 ± 0.97**	3.38 ± 0.39	7.89 ± 3.34	9.68 ± 0.59	**6.68 ± 1.05**
	LS ($\alpha = 0.2$)	6.64 ± 0.69	3.33 ± 0.15	7.28 ± 2.20	9.75 ± 1.14	6.75 ± 0.70
	SVLS ($\sigma = 1.0$)	7.04 ± 0.84	3.73 ± 0.24	10.94 ± 5.75	10.2 ± 1.26	7.98 ± 1.59
	Ours ($\gamma = 1.0$)	7.83 ± 2.72	**3.22 ± 0.06**	**6.50 ± 0.52**	9.78 ± 0.26	6.83 ± 0.78

existing methods are ranked differently across datasets and evaluation measures, indicating that these approaches are sensitive to datasets. In contrast, our GeoLS outperforms the state-of-the-art approaches in most cases. Based on these results, we can conclude that our method remains consistent across diverse datasets, highlighting the robustness of our intensity-based soft labels.

4.5 Qualitative Results

Figure 5 shows the visual comparison of different segmentation results on brain tumors from BraTS, abdominal organs from FLARE, and prostatic zones from ProstateX datasets. In brain tumor segmentations (top row), the results of existing approaches (OH, FL, SVLS) are predominantly over-segmenting in non-enhancing core regions (blue), whereas the LS and GeoLS reduce the segmentation errors. In the middle row of Fig. 5, the existing methods struggle to segment the challenging pancreas organ (yellow) organ. In contrast to these baselines, our GeoLS delivers a superior segmentation of the pancreas organ. The prostatic zone

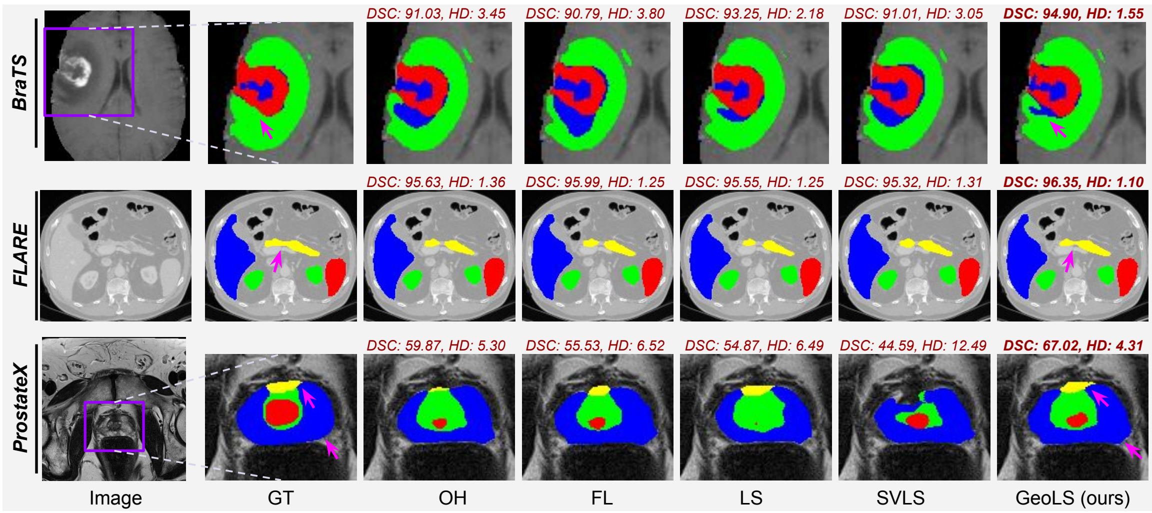


Figure 5: **Qualitative results across BraTS (top), FLARE (middle), and ProstateX (bottom) datasets.** For BraTS and ProstateX, segmentation results are shown from the region highlighted in the image (purple). Average DSC (%) and HD (mm) scores are mentioned at the top of each prediction. Our GeoLS minimizes classification errors in ambiguous regions, such as the non-enhancing core (blue) in BraTS, the pancreas (yellow) in FLARE, and PZ (blue) and AFS (yellow) zones in the ProstateX examples. Coloring denotes different tumor structures (top), abdominal organs (middle), and prostatic zones (bottom).

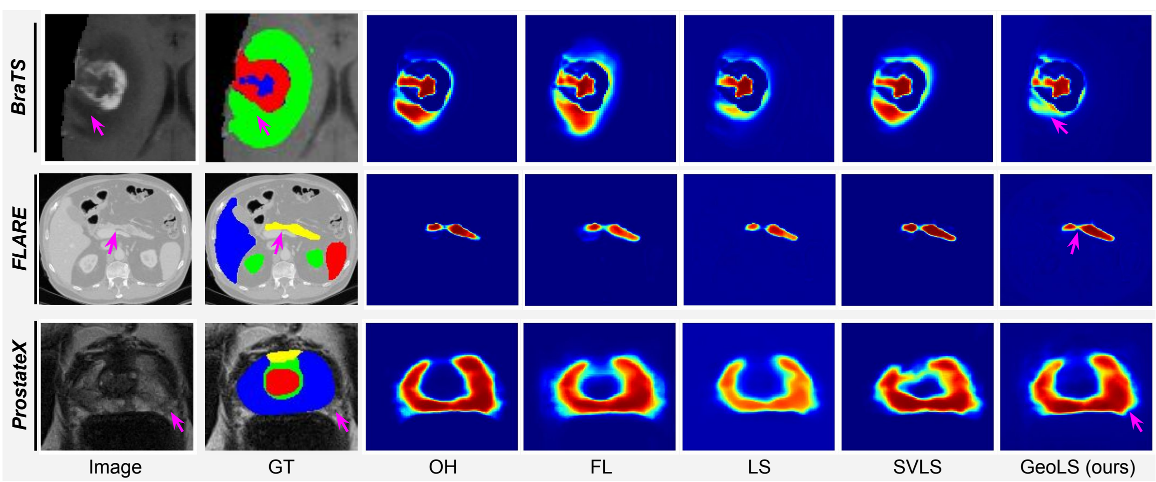
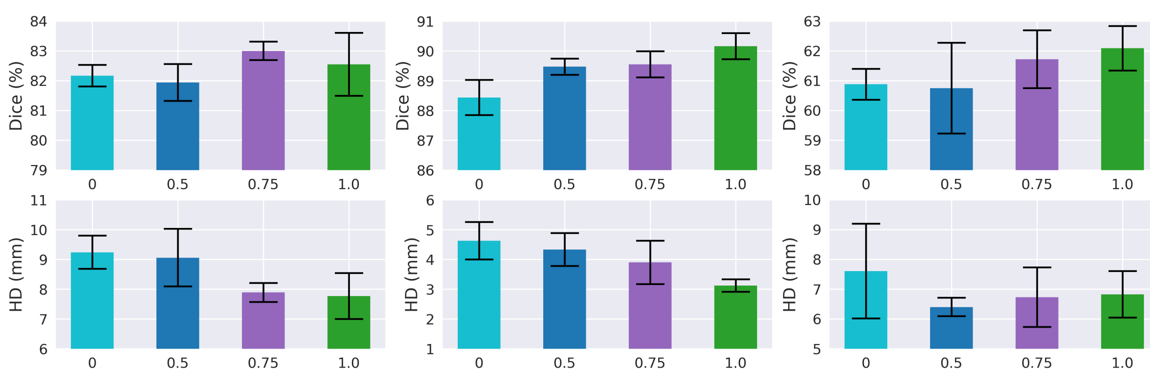


Figure 6: **Predicted probability maps.** The probability maps indicate a non-enhancing core (blue) in BraTS (top), a pancreas (yellow) in FLARE (middle), and a PZ (blue) in ProstateX (bottom), corresponding to the examples shown in the qualitative results. Our GeoLS yields reasonably low probabilities in poorly defined image intensities and misclassified regions while maintaining high probabilities in non-ambiguous regions.

segmentations are arguably challenging due to imprecise boundaries between different zones. In the bottom row, the results of prostatic zone segmentations are poor in all approaches. Our method produces reasonable segmentation results, notably in the AFS prostatic zone (yellow). In addition, the prediction probability maps of baselines and our method for the same examples are shown in Fig. 6. Our GeoLS produces reasonably low probabilities in poorly defined image intensities and misclassified regions, ensuring segmentation accuracy even in challenging areas. At the same time, it consistently maintains high probabilities in well-defined image intensity regions. Furthermore, the quantitative results presented in Sec. 4.4 support these visual results. These results indicate that supplying image gradient information through geodesic maps in our intensity-based soft-labeling approach enhances the segmentation performance.

4.6 Sensitivity to γ

The hyperparameter γ in Eq. 4 plays a crucial role in balancing between the Geodesic Distance and the Euclidean Distance. Since the intensity variations and spatial distance can influence the generalized geodesic distance transform, we investigate the segmentation performance by varying the γ parameter and report their results in Fig. 7, across all datasets. Additionally, we include the segmentation result obtained from a model trained with $\gamma = 0$, i.e., utilizing only the Euclidean Distance for soft labels. The results demonstrate that the segmentation performance is better for higher γ values compared to the models solely relying on Euclidean distance maps. This indicates that incorporating geodesic information based on image gradients in our soft labels positively impacts the performance of segmentation tasks.



(a) BraTS dataset (b) FLARE dataset (c) ProstateX dataset

Figure 7: **Sensitivity of geodesic factor γ on segmentation performance** - Each bar indicates the average DSC $\uparrow$ (top) and HD $\downarrow$ (bottom) scores for BraTS, FLARE, and ProstateX datasets. $\gamma = 0$ here uses only using Euclidean Distance. Segmentation accuracy improves when the γ value is increased towards 1, indicating a higher emphasis on Geodesic Distance in soft labels.

Table 4: **Performance under different seed sets $\mathcal{S}$** . Average DSC and HD scores on BraTS, FLARE, and ProstateX datasets are reported. Segmentation accuracy is consistent across datasets for skeleton-based seed points. The bold and underlined indicate the best and second-best results.

Datasets	BraTS		FLARE		ProstateX	
choice of $\mathcal{S}$	**DSC (%) ↑**	**HD (mm) ↓**	**DSC (%) ↑**	**HD (mm) ↓**	**DSC (%) ↑**	**HD (mm) ↓**
random-3	82.98 ± 0.68	8.10 ± 0.09	87.83 ± 1.02	4.79 ± 0.16	58.65 ± 3.73	7.41 ± 1.59
random-5	82.51 ± 0.80	9.00 ± 0.70	89.46 ± 1.00	4.20 ± 0.97	60.88 ± 0.85	7.07 ± 0.33
random-7	82.36 ± 0.48	8.89 ± 0.81	89.23 ± 0.21	4.41 ± 0.49	61.76 ± 2.62	6.84 ± 0.91
skeleton	**83.00 ± 0.31**	**7.89 ± 0.32**	**90.16 ± 0.44**	**3.12 ± 0.21**	**62.09 ± 0.75**	**6.83 ± 0.78**
erosion	81.93 ± 0.93	9.17 ± 0.68	89.56 ± 0.08	3.63 ± 0.27	61.72 ± 0.90	6.96 ± 0.55

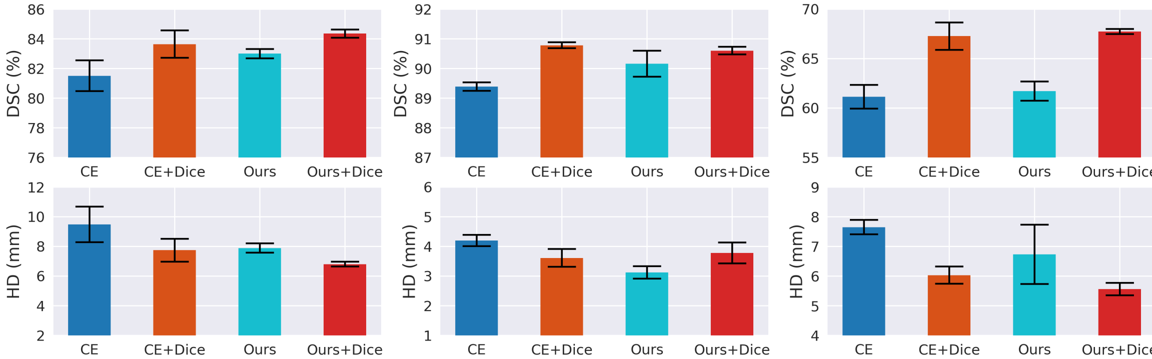
4.7 Choice of seed set $\mathcal{S}$

Our soft label relies on the geodesic maps, which vary with the different choices of seed set $\mathcal{S}$. Therefore, to validate the effectiveness of our seeding strategy on segmentation performance, we conduct experiments with different seed-set strategies. These strategies involve obtaining a random selection of pixels within each target class. For this, our experiments include 3, 5, and 7 randomly selected pixels as seed points. Such seed points are inadequate for large regions, such as the liver, or multiple instances of a class label, such as the kidney. To address this issue, seed sets are also obtained using the remainings of the skeletonization and erosion operations applied to each target class. The results of these experiments are reported in Table 4. It shows that the segmentation performances are comparable for different seed-set choices, which further demonstrates the strength of our geodesic soft labels. Furthermore, the results suggest that the skeleton-based seed strategy consistently yields favorable results across all datasets, which indicates that this seeding strategy could also be viable on new datasets.

4.8 Combination of loss functions

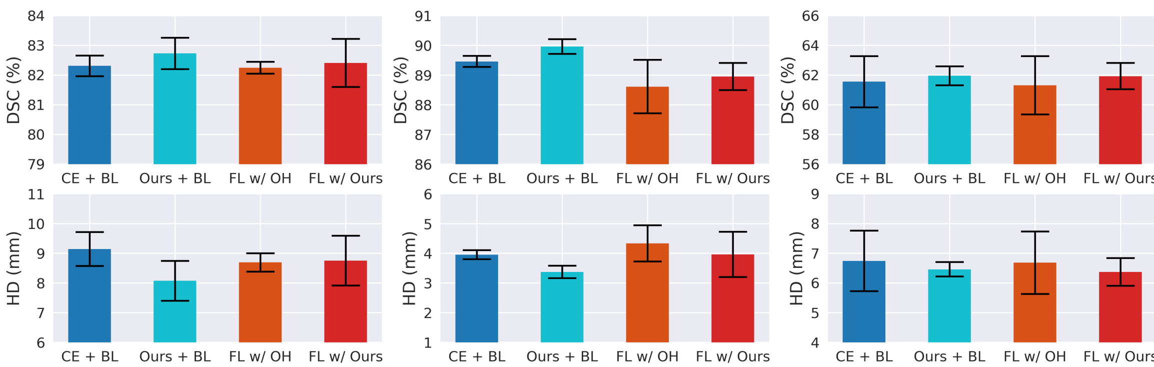
The main goal of this work is to provide an alternative to state-of-the-art soft labeling losses by leveraging geodesic distance transform. Nevertheless, the proposed approach is orthogonal to other types of segmentation losses, including widely used Dice loss (Sudre et al., 2017). Moreover, combined CE and Dice losses are often employed to train segmentation models for medical images (Ma et al., 2021; Taghanaki et al., 2019). Thus, we investigate whether the findings observed when comparing the CE loss hold when we combine the proposed GeoLS with the Dice loss. These results, depicted in Fig. 8, demonstrate that adding the Dice loss improves the segmentation performance of both CE and GeoLS across all datasets. Moreover, combining GeoLS and Dice losses achieves the best results in most cases, demonstrating the consistency of our geodesic label-smoothing approach.

Furthermore, we performed experiments by combining our GeoLS with a boundary loss (BL) first and then with a focal loss (FL), and their results are reported in Fig. 9. The results show a similar trend as with a combination of Dice loss. Combining our method with the BL and FL yields better segmentation results compared to the CE combined with the BL and FL across all three datasets, in most cases. These results demonstrate the robustness of the proposed GeoLS when combined with other loss functions.



(a) BraTS dataset (b) FLARE dataset (c) ProstateX dataset

Figure 8: **Segmentation results with a combination of Dice loss** - Each bar indicates the average DSC ↑ (top) and HD ↓ (bottom) scores on all three datasets. The performance of segmentation improves by adding Dice loss on both CE and our models. Combination of Dice loss with our yields consistently best in most cases.



(a) BraTS dataset (b) FLARE dataset (c) ProstateX dataset

Figure 9: **Segmentation results with a combination of Boundary loss (BL) and Focal loss (FL)** - Each bar indicates the average DSC ↑ (top) and HD ↓ (bottom) scores on all three datasets. Combining our method with BL and FL consistently provides better segmentation results compared to CE combined with BL and FL in most cases.

5. Discussion and Conclusion

Despite the growing popularity of contemporary soft-labeling approaches, the underlying image context information associated with the label is largely overlooked in the soft labels for image segmentation. This work demonstrates that incorporating such information into standard hard labels would improve the performance of deep segmentation networks. To that effect, our contribution, a Geodesic label smoothing (GeoLS), incorporates intensity variation details into the soft-labeling process through geodesic distance transforms. More specifically, our proposed approach generates new intensity-based soft labels that capture ambiguity between neighboring target regions. Employing our soft labels in the training of segmentation models has consequently demonstrated an improved segmentation performance. Our results have in fact shown that our geodesic-based smoothing consistently outperforms state-of-the-art approaches in soft-labeling, across three different datasets: multi-class tumor segmentation in brain MRIs, organ segmentation in abdominal CTs, and zone segmentation in prostatic MR volumes. Both quantitative and qualitative results indicate notable improvements in the segmentation of known challenging regions, such as of enhancing tumors, as well as the pancreas.

Furthermore, the ablation study conducted on the geodesic factor parameter indicates that our geodesic maps integrate richer intensity information in the yielded soft labels, effectively producing an improved segmentation performance than utilizing only Euclidean distance maps. Our experiments have also evaluated several key seeding strategies for generating soft labels. These results show that the skeleton-based strategy remains consistent across all datasets. The design of the seeding process can be further explored in order to better capture the intrinsic structures of target objects. This work provides, therefore, a valuable alternative to hard-labeling and existing soft-labeling losses. Nonetheless, our geodesic label smoothing loss can also be combined with other segmentation losses, such as the conventional Dice loss. The use of such loss has in fact shown further improvements in the segmentation accuracy within our experiments. As future work, our approach could also be potentially applicable to segmentation tasks under noisy annotations (Lukasik et al., 2020; Karimi et al., 2023). Overall, our proposed geodesic-based soft-labeling could be virtually leveraged in broader ranges of applications where annotation remains challenging due to ambiguities in image intensities across regions.

Acknowledgments

This work has been funded by the Canada Research Chair on Shape Analysis in Medical Imaging, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Fonds de Recherche du Quebec (FQRNT). Special thanks to Mustafa Chasmai and Kumar Devesh for their initial discussions on this work.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding the treatment of animals or human subjects.

Conflicts of Interest

The authors declare no known conflicts of interest in any personal relationship, financial or otherwise, to disclose. The authors are affiliated with the Department of Computer and Software Engineering at ETS Montreal, Montreal, Canada (etsmtl.ca), and Polytechnique Montreal, Canada (polymtl.ca). Fundings are from government agencies (Canada Research Chair, NSERC, FRQNT).

Data availability

All the datasets used in this research work are publicly available.

References

Sukesh Adiga Vasudeva, Jose Dolz, and Herve Lombaert. GeoLS: Geodesic label smoothing for image segmentation. In *Medical Imaging with Deep Learning*, 2023.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BraTS challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Eric Baum and Frank Wilczek. Supervised learning of probability distributions by neural networks. In *Neural information processing systems*, 1987.

Toan Duc Bui, Li Wang, Jian Chen, Weili Lin, Gang Li, and Dinggang Shen. Multi-task learning for neonatal brain segmentation using 3D Dense-UNet with dense attention guided by geodesic distance. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data - Medical Image Computing and Computer Assisted Intervention Workshop*, pages 243–251. Springer, 2019.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

Antonio Criminisi, Toby Sharp, and Andrew Blake. GeoS: Geodesic image Segmentation. In *European Conference on Computer Vision*, pages 99–112. Springer, 2008.

James S Duncan and Nicholas Ayache. Medical image analysis: Progress over two decades and the challenges ahead. *Transactions on Pattern Analysis and Machine Intelligence*, 22(1):85–106, 2000.

Neamat El Gayar, Friedhelm Schwenker, and Günther Palm. A study of the robustness of KNN classifiers trained using soft labels. In *Artificial Neural Networks in Pattern Recognition IAPR Workshop*, pages 67–80. Springer, 2006.

Adrian Galdran, Jihed Chelbi, Riadh Kobi, José Dolz, Hervé Lombaert, Ismail Ben Ayed, and Hadi Chakor. Non-uniform label smoothing for diabetic retinopathy grading from retinal fundus images with deep neural networks. *Translational Vision Science & Technology*, 9(2):34–34, 2020.

Charley Gros, Andreanne Lemay, and Julien Cohen-Adad. SoftSeg: advantages of soft versus binary training for image segmentation. *Medical Image Analysis*, 71:102038, 2021.

Adam Hammoumi, Maxime Moreaud, Christophe Ducottet, and Sylvain Desroziers. Adding geodesic information and stochastic patch-wise image prediction for small dataset learning. *Neurocomputing*, 456:481–491, 2021.

Robert M Hayward, Nicolas Patronas, Eva H Baker, Gilbert Vézina, Paul S Albert, and Katherine E Warren. Inter-observer variability in the measurement of diffuse intrinsic pontine gliomas. *Journal of neuro-oncology*, 90:57–61, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016.

Xingxin He, Leyuan Fang, Hossein Rabbani, Xiangdong Chen, and Zhimin Liu. Retinal Optical Coherence Tomography image classification with label smoothing generative adversarial network. *Neurocomputing*, 405:37–47, 2020.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of Digital Imaging*, 32(4):582–596, 2019.

Fabian Isensee, Paul F Jaeger, Simon AA 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, 2021.

Mobarakol Islam and Ben Glocker. Spatially varying label smoothing: Capturing uncertainty from expert annotations. In *Information Processing in Medical Imaging*, pages 677–688. Springer, 2021.

Mobarakol Islam, Lalithkumar Seenivasan, Lim Chwee Ming, and Hongliang Ren. Learning and reasoning with the graph structure representation in robotic surgery. In *Medical Image Computing and Computer-Assisted Intervention*, pages 627–636. Springer, 2020.

Leo Joskowicz, D Cohen, N Caplan, and Jacob Sosna. Inter-observer variability of manual contour delineation of structures in CT. *European Radiology*, 29(3):1391–1399, 2019.

Davood Karimi, Caitlin K Rollins, Clemente Velasco-Annis, Abdelhakim Ouaalam, and Ali Gholipour. Learning to segment fetal brain tissue from noisy annotations. *Medical Image Analysis*, 85:102731, 2023.

Eytan Kats, Jacob Goldberger, and Hayit Greenspan. Soft labeling by distilling anatomical knowledge for improved MS lesion segmentation. In *International Symposium on Biomedical Imaging*, pages 1563–1566. IEEE, 2019.

James M Keller, Michael R Gray, and James A Givens. A fuzzy k-nearest neighbor algorithm. *Transactions on Systems, Man, and Cybernetics*, (4):580–585, 1985.

DP Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Peter Kontschieder, Pushmeet Kohli, Jamie Shotton, and Antonio Criminisi. GeoF: Geodesic Forests for learning coupled predictors. In *Computer Vision and Pattern Recognition*, pages 65–72. IEEE, 2013.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision*, pages 2980–2988. IEEE, 2017.

Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in MRI. *Transactions on Medical Imaging*, 33(5):1083–1092, 2014.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Joao Lourenço-Silva and Arlindo L Oliveira. Using soft labels to model uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention - Brainlesion Workshop*, pages 585–596. Springer, 2021.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020.

Tohar Lukov, Na Zhao, Gim Hee Lee, and Ser-Nam Lim. Teaching with soft label smoothing for mitigating noisy labels in facial expressions. In *European Conference on Computer Vision*, pages 648–665. Springer, 2022.

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, 2021.

Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and Low-GPU-memory Abdomen CT organ sEgmentation: The FLARE challenge. *Medical Image Analysis*, 82:102616, 2022.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BraTS). *Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in Neural Information Processing Systems*, 32, 2019.

Dzung L Pham, Chenyang Xu, and Jerry L Prince. Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2(1):315–337, 2000.

Alexis Protiere and Guillermo Sapiro. Interactive image segmentation via adaptive weighted distances. *Transactions on Image Processing*, 16(4):1046–1057, 2007.

Wu Qiu, Jing Yuan, Martin Rajchl, Jessica Kishimoto, Yimin Chen, Sandrine de Ribaupierre, Bernard Chiu, and Aaron Fenster. 3D MR ventricle segmentation in pre-term infants with post-hemorrhagic ventricle dilatation (PHVD) using multi-phase geodesic level-sets. *NeuroImage*, 118:13–25, 2015.

Reuven Y Rubinstein and Dirk P Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 133. Springer, 2004.

Sambu Seo, Mathias Bode, and Klaus Obermayer. Soft nearest prototype classification. *Transactions on Neural Networks*, 14(2):390–398, 2003.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and 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*, pages 240–248. Springer, 2017.

Paul Suetens. *Fundamentals of medical imaging*. Cambridge University Press, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition*, pages 2818–2826. IEEE, 2016.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Association for the Advancement of Artificial Intelligence*, volume 31, 2017.

Saeid Asgari Taghanaki, Yefeng Zheng, S Kevin Zhou, Bogdan Georgescu, Puneet Sharma, Daguang Xu, Dorin Comaniciu, and Ghassan Hamarneh. Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75:24–33, 2019.

Pin Tang, Pinli Yang, Dong Nie, Xi Wu, Jiliu Zhou, and Yan Wang. Unified medical image segmentation by learning from uncertainty in an end-to-end manner. *Knowledge-Based Systems*, 241:108215, 2022.

Pekka J Toivanen. New geodesic distance transforms for gray-scale images. *Pattern Recognition Letters*, 17(5):437–450, 1996.

Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1559–1572, 2018.

Lin Wang, Xiufen Ye, Lie Ju, Wanji He, Donghao Zhang, Xin Wang, Yelin Huang, Wei Feng, Kaimin Song, and Zongyuan Ge. Medical matting: Medical image segmentation with uncertainty from the matting perspective. *Computers in Biology and Medicine*, 158: 106714, 2023.

Zehan Wang, Kanwal K Bhatia, Ben Glocker, Antonio Marvao, Tim Dawes, Kazunari Misawa, Kensaku Mori, and Daniel Rueckert. Geodesic patch-based segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 666–673. Springer, 2014.

Jie Wei, Zhengwang Wu, Li Wang, Toan Duc Bui, Liangqiong Qu, Pew-Thian Yap, Yong Xia, Gang Li, and Dinggang Shen. A cascaded nested network for 3T brain MR image segmentation guided by 7T labeling. *Pattern Recognition*, 124:108420, 2022.

Jie Ying, Wei Huang, Le Fu, Haima Yang, and Jiangzihao Cheng. Weakly supervised segmentation of uterus by scribble labeling on endometrial cancer MR images. *Computers in Biology and Medicine*, 167:107582, 2023.

Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *Transactions on Image Processing*, 30: 5984–5996, 2021.

S Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger. *Handbook of medical image computing and computer assisted intervention*. Academic Press, 2019.