Medical Image Analysis 99 (2025) 103355

Contents lists available at ScienceDirect

Medical Image Analysis

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

Label refinement network from synthetic error augmentation for medical image segmentation

Shuai Chen [a,b,1], Antonio Garcia-Uceda [b,1], Jiahang Su [b,1], Gijs van Tulder [c], Lennard Wolff [d], Theo van Walsum [b], Marleen de Bruijne [b,e,*]

[a] *China Electric Power Research Institute Co., Ltd, Beijing, China*
[b] *Biomedical Imaging Group Rotterdam, Department of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands*
[c] *Data Science group, Faculty of Science, Radboud University, Nijmegen, The Netherlands*
[d] *Department of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands*
[e] *Department of Computer Science, University of Copenhagen, DK-2110 Copenhagen, Denmark*

ARTICLE INFO

Keywords:
Label refinement
Medical images
Segmentation
Synthetic error
Tree-structure shape

ABSTRACT

Deep convolutional neural networks for image segmentation do not learn the label structure explicitly and may produce segmentations with an incorrect structure, e.g., with disconnected cylindrical structures in the segmentation of tree-like structures such as airways or blood vessels. In this paper, we propose a novel label refinement method to correct such errors from an initial segmentation, implicitly incorporating information about label structure. This method features two novel parts: (1) a model that generates synthetic structural errors, and (2) a label appearance simulation network that produces segmentations with synthetic errors that are similar in appearance to the real initial segmentations. Using these segmentations with synthetic errors and the original images, the label refinement network is trained to correct errors and improve the initial segmentations. The proposed method is validated on two segmentation tasks: airway segmentation from chest computed tomography (CT) scans and brain vessel segmentation from 3D CT angiography (CTA) images of the brain. In both applications, our method significantly outperformed a standard 3D U-Net, four previous label refinement methods, and a U-Net trained with a loss tailored for tubular structures. Improvements are even larger when additional unlabeled data is used for model training. In an ablation study, we demonstrate the value of the different components of the proposed method.

1. Introduction

Convolutional neural networks (CNNs) are the state-of-the-art for many biomedical imaging segmentation tasks. Many CNN segmentation architectures have been proposed, such as fully connected networks (Long et al., 2015), Dense-Net (Huang et al., 2017), and the U-Net (Ronneberger et al., 2015). The U-Net has become the most popular network for biomedical image segmentation, due to its efficient structural design featuring skip-connections, showing superior accuracy and robustness in various segmentation tasks (Isensee et al., 2021; Siddique et al., 2021). Most CNN-based segmentation methods including the U-Net do not fully exploit and encode the structural information of the objects to be segmented. Consequently, these methods may produce segmentations with errors that become obvious when looking at the full segmented structure. Examples of such errors are discontinuities in the segmentations of elongated tubular structures, such as airways in the lungs, as shown in Fig. 1. Using label structural knowledge such as continuity in the branches of the airway tree can help prevent these errors. However, it is not trivial to explicitly encode this global information in CNNs.

In this paper, we propose a framework to implicitly encode the label structural information into CNNs by formulating this as a label refinement step. Specifically, we generate structural synthetic errors in segmentations (ground truth or baseline) and train a label refinement network to correct these errors. The trained network is expected to generalize to the real errors in the initial segmentations produced by a baseline segmentation network and correct them. To enhance the generalizability of the label refinement network on the initial segmentations, a label appearance simulation network is applied to reduce the appearance difference between the segmentations with synthetic errors and the initial segmentations. With either these segmentations

[*] Correspondence to: Room Na 26-20, Erasmus MC, Rotterdam, The Netherlands.
E-mail address: marleen.debruijne@erasmusmc.nl (M. de Bruijne).
[1] S. Chen, A. Garcia-Uceda, and J. Su contributed equally.

https://doi.org/10.1016/j.media.2024.103355
Received 1 March 2023; Received in revised form 25 May 2024; Accepted 20 September 2024
Available online 27 September 2024

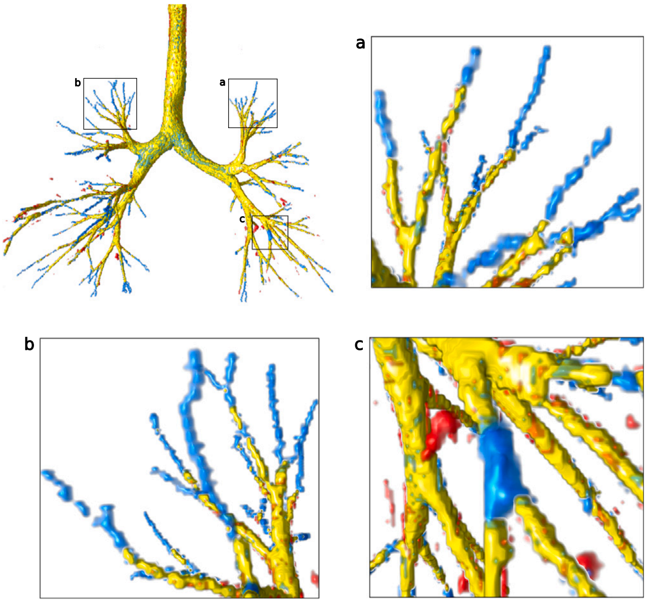

Fig. 1. Common structural errors in the segmentations obtained by a U-Net, trained to segment airways (Garcia-Uceda et al., 2021). True positives are displayed in yellow, false negatives in blue and false positives in red. Detailed views a–b show errors as missing terminal branches, and view c shows a discontinuity error in the branch.

with appearance-enhanced synthetic errors or the initial segmentations, together with the original images, as inputs and the ground truth segmentations as reference, the label refinement network can learn to correct those errors and incorporate this in its segmentation decisions.

We validated the proposed label refinement method on two segmentation tasks: airway segmentation from chest computed tomography (CT) scans (Garcia-Uceda et al., 2021) and brain vessel segmentation from 3D CT angiography (CTA) images of the brain (Su et al., 2020). We compared our method with a U-Net baseline, four other label refinement methods: DoubleU-Net (Jha et al., 2020), SCAN (Dai et al., 2018), Post-DAE (Larrazabal et al., 2020) and DVAE (Araújo et al., 2019); and a U-Net trained with the clDice loss (Shit et al., 2021). Moreover, we conducted an ablation study to show the contribution of each component of the label refinement method. Finally, we performed experiments in a semi-supervised setting to train our method using additional unlabeled data.

2. Related work

2.1. Label refinement

In this work, we apply a refinement network on the initial segmentation from a baseline segmentation network together with the original image, intending to correct errors in the initial segmentation. Similar approaches have been used in other previous papers. Yang et al. (2019) refined low-quality manual annotations made by non-experts by training their method with added noise to reduce the inter-observer inconsistency of the annotations. Unlike our method, Yang et al. (2019) does not focus on refining an initial automatic segmentation and therefore the label appearance simulation network is not needed. Dai et al. (2018) refined the segmentations from a fully convolutional network by using adversarial training to reduce the domain gap between the target predictions and the ground truth segmentations on training data. Araújo et al. (2019) attached a variational auto-encoder after a U-Net network to encode the label structure of the ground truth segmentations for a better label reconstruction. Larrazabal et al. (2020) applied a denoising autoencoder after a U-Net as a post-processing method to recover the label structure of the segmentation. Jha et al. (2020) attached a second U-Net network to a baseline U-Net, using as inputs the original image multiplied with the output of the first U-Net. Different from these works, our method does not focus on encoding (Araújo et al., 2019; Larrazabal et al., 2020) or discriminating (Dai et al., 2018; Jha et al., 2020) the overall label structure, but instead on learning to correct the most common errors in the segmentations.

2.2. Airway segmentation

The airway tree in the lungs forms a complex 3D tree-like branching network, with many branches of different sizes and orientations. The peripheral branches of smaller size are challenging to segment from chest CT scans, as they have obscured borders due to partial volume effects. Many classical methods for airway tree extraction are based on a region growing algorithm (Graham et al., 2010; Lo et al., 2009, 2010). However, their accuracy is limited, and they typically miss a large number of the smaller peripheral airways (Lo et al., 2012). Many state-of-the-art airway segmentation methods are based on CNNs, and especially the U-Net (Cheng et al., 2021; Garcia-Uceda et al., 2021; Qin et al., 2021; Zheng et al., 2021). CNN-based methods can obtain more accurate and complete segmentations than previous intensity-based methods. However, even the latest U-Net-based methods usually miss several terminal branches and make errors in continuity around the smaller segmented branches.

2.3. Brain vessel segmentation

The brain vessels form a complex 3D branching network that consists of veins and arteries. In 3D CTA images of the brain, many seemingly isolated vessel structures can be present due to the image acquisition and vascular diseases, such as ischemic large vessel occlusions. State-of-the-art vessel segmentation methods have been applied to 3D time-of-flight (TOF) magnetic resonance angiography (MRA) images (Hilbert et al., 2020; Livne et al., 2019; Sanchesa et al., 2019), and to 3D and 4D CTA images (Meijs et al., 2017) using U-Nets. Su et al. (2020) used a U-Net-based method to extract a dilated vessel centerline approximation. Compared to previous vessel segmentation methods (Hilbert et al., 2020; Livne et al., 2019; Meijs et al., 2017; Sanchesa et al., 2019), centerline extraction recovers the topology of the vessel structure more accurately (e.g., “kissing vessels” appear connected in the full segmentations but are disconnected in centerline extraction). However, the U-Net still makes other structural errors such as local connectivity gaps in vessel branches.

3. Method

The proposed method consists of four steps, schematically shown in Fig. 2. Firstly, a baseline segmentation network generates the initial segmentations (Section 3.1). Secondly, synthetic errors are generated and added to every ground truth segmentation, to create segmentations with synthetic errors to train the label refinement network (Section 3.2). Thirdly, a label appearance simulation network (LASN) based on adversarial learning is used to reduce the appearance difference between the segmentations with synthetic errors and the initial segmentations (Section 3.3). Steps 2–3 constitute a realistic data augmentation (error augmentation) technique to generate training samples for the label refinement network, with a much larger variety of errors than in the initial segmentations. Finally, a label refinement network is trained to predict the final refined segmentations, using the segmentations with appearance-enhanced synthetic errors or the initial segmentations, together with the original images, as inputs and the ground truth segmentations as reference (Section 3.4).

3.1. Base segmentation network

We use a base segmentation network f_1 to predict the initial segmentations. Given a medical imaging dataset that contains an image

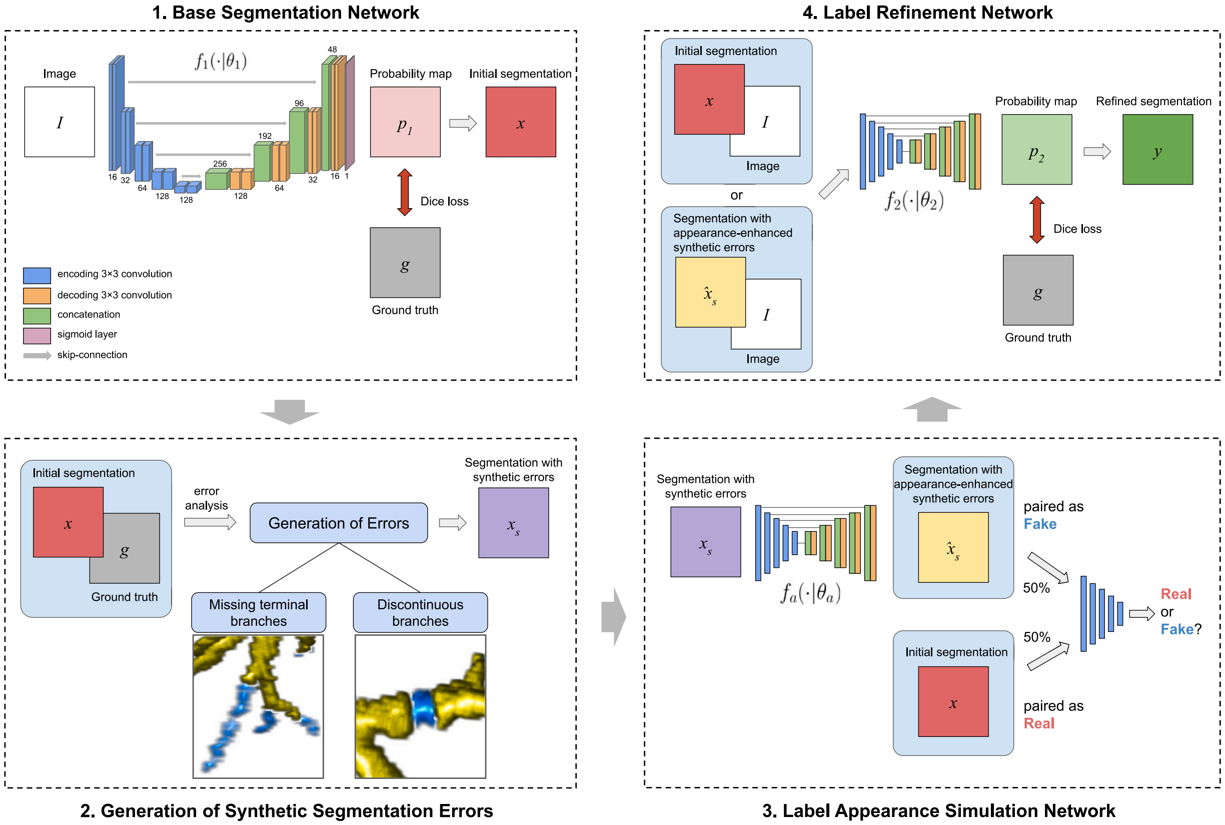

Fig. 2. Schematics of the proposed label refinement method. First, a base segmentation network f_1 is trained to obtain the initial segmentations x. Second, we create segmentations with synthetic errors x_s that are similar to the errors in x. Third, a label appearance improvement network f_a (together with a discriminator D) is trained to obtain segmentations with appearance-enhanced synthetic errors $\hat{x}_s$. Finally, the label refinement network f_2 is trained to correct these synthetic errors, with either $\hat{x}_s$ or x together with the image I as inputs.

I and the ground truth segmentation g for each subject, the model $f_1(I|\theta_1)$, with θ_1 the trainable parameters, is trained by minimizing the Dice loss $\mathcal{L}_1 = \mathcal{L}_{\text{dc}}(f_1(I), g)$:

$$\mathcal{L}_{\text{dc}}(p, g) = -\frac{2\sum_{i\in I} p_i g_i}{\sum_{i\in I} p_i + \sum_{i\in I} g_i} \tag{1}$$

where p_i and g_i are the ith voxel values of the probability maps output by the model (in this case $p = f_1(I)$), and the ground truth segmentation, respectively.

The initial predicted segmentation is x, obtained by thresholding the output probability maps p_1 of the network with a value of 0.5. x may contain label structural errors, such as discontinuous branches in a tree-like structure. Next, we show how to design synthetic errors similar to those in x that can be used to train the label refinement network.

3.2. Generation of synthetic segmentation errors

We use segmentations with synthetic errors x_s to train the label refinement network. The synthetic errors are generated to resemble those present in the initial segmentations x, based on our initial analysis of common errors. In this paper, we focus on two structures: airways in the lungs and vessels in the brain. Airways and vessels share several characteristics: they both form 3D branching networks, with branches of cylindrical shape and various sizes and orientations. We use this prior shape knowledge to generate synthetic errors, as described below for each structure.

3.2.1. Synthetic errors for airways

Most of the errors in airway segmentations can be grouped into two types: (1) missing terminal branches, partially or totally, and (2) discontinuity in the segmented branches, which occurs more frequently in smaller (thinner) branches. Examples of errors in airway segmentations obtained by the baseline segmentation network in Section 3.1 are shown in Fig. 1. To generate similar synthetic errors, we select a random subset of branches in the ground-truth segmentation of the airway tree and remove a section of each branch by masking it at a random position and with a random length. Branches are identified using the airway centerline tree, extracted from the airway segmentation (Lo et al., 2012). Single branches are defined as the segments between two bifurcation points or between the last bifurcation and the end of terminal branches. The applied masking is defined differently for each type of error:

Missing terminal branches: The subset of branches in which to synthesize this type of errors is randomly sampled from all the terminal branches in the airway tree, defined as branches with no further bifurcations downstream. A mask of cylindrical shape is applied to (partially) remove the selected branch. The mask is defined by (1) a starting point, which is a random position along the branch centerline between the branch start and middle points; (2) a length, which is the distance between the mask start point and branch end; and (3) a width, which is three times the branch diameter. An example of this error type is schematically shown in Fig. 3.

Discontinuity in branches: The subset of branches in which to synthesize this type of errors is randomly sampled from all the branches in the airway tree, excluding the trachea, the two main bronchi and the 2nd generation airways. Since the more peripheral and thinner airway branches have a higher chance of being incomplete, we assign a higher sampling probability to branches that are further away from the root of the airway tree (the trachea). Each branch i is assigned an airway generation g_i that is defined as the number of bifurcations counted in the path from the trachea to the given branch. The sampling probability p_i for a candidate branch is then defined as $p_i = g_i / \sum_{k=1}^{N_c} g_k, \forall i =$

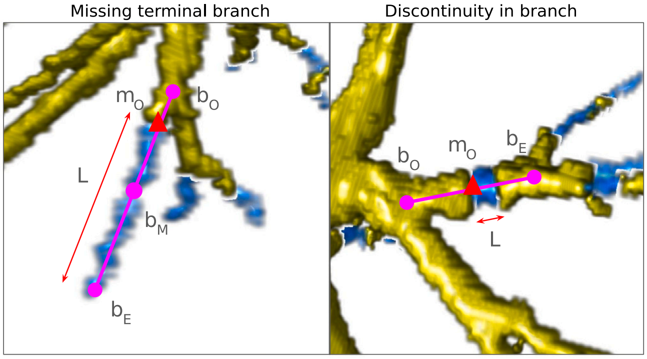

Fig. 3. Schematics of the synthetic segmentation errors defined for airways. Definitions are shown for a randomly selected terminal branch (left) and non-terminal branch (right). b_O: branch start point, b_E: branch end point, b_M: branch middle point, m_O: mask start point, L: mask length. The masked section of airway branches is displayed in blue (for the selected one as well as other nearby branches).

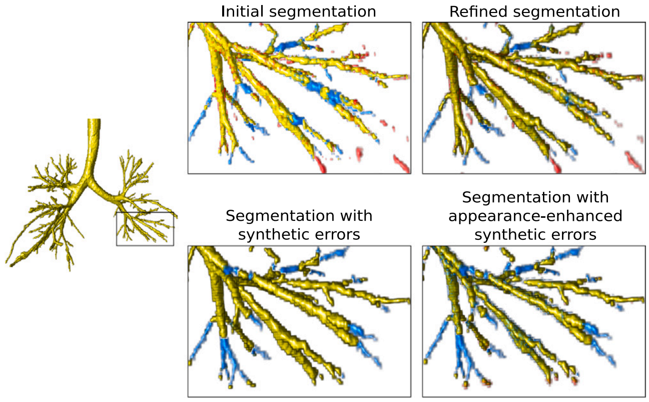
Initial segmentation
Refined segmentation
Segmentation with synthetic errors
Segmentation with appearance-enhanced synthetic errors

Fig. 4. Example of segmentation of airways obtained by the different components of the proposed method. In the detailed views, true positives are displayed in yellow, false negatives in blue and false positives in red.

$1 \dots N_c$, where g_i is the airway generation and N_c the number of candidate branches. A mask of cylindrical shape is applied to create a gap in the selected branch. The mask is defined by (1) a center, which is a random position along the branch centerline; (2) a length, which is a random distance between a minimum of 10 voxels and the total branch length; and (3) a width, which is three times the branch diameter. An example of this error type is schematically shown in Fig. 3.

Parameters: The extent of each type of errors in the airway segmentations with synthetic errors is determined by a separate parameter, denoted as p_1^a and p_2^a. p_1^a is the proportion of selected branches with errors of type "missing terminal branches", with respect to all the terminal branches. p_2^a is the proportion of selected branches with errors of type "discontinuity in branches", with respect to all the branches in the airway tree (excluding the trachea, the two main bronchi and the 2nd generation airways).

3.2.2. Synthetic errors for brain vessels

Most of the errors in brain vessel segmentation are in the form of partially missing vessel branches. To generate similar synthetic errors, we create random gaps in the ground-truth segmentation of each vessel by masking it at a random position and with a random length. Since the errors occur more frequently for long vessels than for short ones, we group all the vessels into three equal-sized groups: long, medium size and short, based on the relative centerline segment lengths in each subject. The distribution of vessel lengths (in voxels), using the median and interquartile range (IQR), is: for long segments 70 (49–106), for medium-size segments 29 (22–36), and for short segments 13 (9–17). For long segments, the maximum number of injected gaps is 6 (randomly sampled from a uniform distribution between 0 and 6 positions) with a gap length between 10–35 voxels. For medium-size segments, the maximum number of gaps is 4 with a gap length between 10–20 voxels. For the short segments, the maximum number of gaps is 2 with a gap length between 6–15 voxels. Those error injections are applied on the 1 voxel-wide ground truth centerlines, by dilating it with a 3 × 3 × 3 cubic structure element to generate the final vessel segmentation with synthetic errors.

Parameters: The extent of errors in the vessel segmentations with synthetic errors is determined by only one parameter, denoted as p^v. p^v is the proportion of selected branches with errors with respect to all the branches in the vessel network.

3.3. Label appearance simulation network

Although the segmentations with synthetic errors x_s are designed to have similar structural errors to the initial segmentations x, there may be an appearance difference between x_s and x (see an example in Fig. 4). The label refinement network trained on x_s may therefore generalize poorly to x. To prevent this, we use a label appearance simulation network $f_a(\cdot|\theta_a)$ to change the appearance of x_s to be more similar to that of x, while preserving the synthetic errors that we added to x_s.

The label appearance simulation network $f_a(\cdot|\theta_a)$, with θ_a the trainable parameters, is optimized by adversarial learning via a discriminator D:

$$f_a^* = \arg\min_{f_a}((\max_{D} \mathcal{L}_{adv}(f_a, D)) + \lambda \mathcal{L}_{dc}(\hat{x_s}, x_s)) \quad (2)$$

with the adversarial loss $\mathcal{L}_{adv}$ defined as:

$$\mathcal{L}_{adv}(f_a, D) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_{\hat{x_s}}[\log(1 - D(\hat{x_s}))] \quad (3)$$

where D is a classifier, discriminating the given segmentation with errors $\hat{x_s}$ and the initial segmentation x. It outputs a probability between 0.0 and 1.0. $\hat{x_s} = f_a(x_s)$ is the obtained segmentation with appearance-enhanced synthetic errors. We added a Dice-based identity loss $\mathcal{L}_{dc}(\hat{x_s}, x_s)$ to train $f_a(\cdot)$, in order to preserve the synthetic errors that we added in x_s. The hyperparameter λ controls the balance between the identity loss and the dissimilarity adversarial loss.

3.4. Label refinement network

Finally, we optimize a label refinement network f_2 to predict the ground truth segmentations, based on the segmentations with appearance-enhanced synthetic errors $\hat{x_s}$ together with the original images as inputs. This way, f_2 learns to correct segmentation errors and can be used to improve the initial segmentations x. The model $f_2((I, \tilde{x})|\theta_2)$, with θ_2 the trainable parameters, is trained by minimizing the Dice loss between the model output and ground truth segmentations $\mathcal{L}_2 = \mathcal{L}_{dc}(f_2(I, \tilde{x}), g)$, given by Eq. (1). The final refined segmentation result is y, obtained by thresholding the output probability maps p_2 of the refinement network with value 0.5.

The main difference between the base segmentation network f_1 and the proposed combination of base network followed by label refinement network f_2 is that the synthetic errors allow the label refinement network f_2 to incorporate prior knowledge about likely mistakes by the base segmentation network f_1. For example, in airway and brain vessel segmentation, one type of prior knowledge we use is that airway trees and brain vessels should always be continuous structures, while the base segmentation network f_1 sometimes makes discontinuity errors. By adding the synthetic errors to the ground truth segmentations, the label refinement network f_2 learns to correct the mistakes that the base segmentation network f_1 is likely to make.

4. Experiments

4.1. Datasets

We validated the proposed method on two biomedical imaging segmentation tasks: segmenting airways from chest CT scans and brain vessels from CTA images of the brain.

4.1.1. Chest CT data

The dataset of chest CT scans is from a retrospective study of pediatric patients (6 to 17 years old) with cystic fibrosis lung disease, acquired routinely at the hospital Erasmus MC-Sophia Rotterdam (Bouma et al., 2020). The CT scans show noticeable structural airway abnormalities resulting from the disease. In our study, we used 178 low-dose CT scans acquired at full inspiration breath-hold. All CT scans have slice dimensions 512 × 512, with a variable number of slices between 200–1000. Each CT scan has an in-plane voxel size in the range 0.35–0.65 mm, with slice thickness between 0.75–1.0 mm, and slice spacing between 0.3–0.8 mm. A random subset of 65 CT scans from the total 178 scans have annotations of the airway lumen. To obtain these annotations, Thirona's lung quantification software LungQ (Thirona, Nijmegen, the Netherlands) was used to automatically extract the airway lumen from the CT scan. Then, these segmentations were visually checked by trained data analysts for accuracy, and corrected as needed.

For our experiments, we used as testing data 41 random CT scans from the subset of 65 CT scans with ground truth segmentations. From the remaining 24 CT scans with annotations, we used three different random data splits with 20 CT scans for training the networks and 4 CT scans for validation. The remaining 113 CT scans without ground truth segmentations were used as unlabeled training data for the experiments with semi-supervised learning.

4.1.2. CTA data of the brain

The dataset of CTA images of the brain is from the MR CLEAN Registry (Jansen et al., 2018), an ongoing registry for patients who underwent endovascular treatment for acute ischemic stroke in one of 19 hospitals in the Netherlands since March 2014. The data was collected during clinical practice, and we applied the following data inclusion criteria: (1) slice thickness $\leq$1.5 mm, (2) slice spacing $\leq$1.5 mm, (3) the contrast acquisition phase has to be peak arterial phase, equilibrium or early venous phase (Rodriguez-Luna et al., 2014), and (4) the image should cover at least half of the brain. In our study, we used 69 CTA images from 69 different subjects used in Su et al. (2020). All CTA images were skull-stripped with an atlas-based registration method (Peter et al., 2017). 20 CTA images had no vessel annotations, 9 CTA images had a complete brain vessel centerline annotation, and the remaining 40 CTA images (randomly sampled from the whole dataset) had vessel centerline annotations in a randomly sampled subvolume of 140 × 140 × 140 voxels. The centerline annotations were dilated with a 3 × 3 × 3 cubic structure element to obtain the ground truth segmentations. Each CTA image has an in-plane voxel size in the range 0.4–0.68 mm, with slice thickness between 0.5–1.5 mm, and slice spacing between 0.3–1.0 mm.

For our experiments, we used as testing data 2 random full-volume CTA scans and 20 random CTA cubes from the set of 9 full-volume, annotated CTA scans and 40 CTA cubes, respectively. From the remaining data with annotations, we used three different random data splits with 7 full-volume CTA scans and 14 CTA cubes for training the networks, and 6 CTA cubes for validation. The remaining 20 full-volume CTA scans without manual annotations were used as unlabeled training data for the experiments with semi-supervised learning.

4.2. Parameters for generating synthetic errors

The generation of synthetic errors depends on the parameters p_1^{a} and p_2^{a} for airways, and p^{v} for vessels, described in Sections 3.2.1 and 3.2.2 respectively. In the rest of the paper we refer to these parameters as "synthetic error rate", for each type of error. For each training sample, the synthetic error rate is randomly sampled from a uniform distribution between 0.0 and the upper bound, or maximum synthetic error rate. These upper bounds are hyperparameters for the proposed method, denoted as P_1^{a} and P_2^{a} for airways, and P^{v} for vessels.

We conducted experiments varying the hyperparameters for the error generation in the proposed method, i.e., the maximum synthetic error rates (P_1^{a} and P_2^{a} for airways, and P^{v} for vessels), to investigate their influence in the method performance. The results are shown in Section 5.3 below.

In our further experiments, the optimal hyperparameters were determined on the validation set for each of the three random data splits that we used, for both applications. Each hyperparameter was searched independently, from 0.0 to 1.0, while fixing the parameters for other error types to 0.0.

4.3. Network architecture

The baseline segmentation network f_1 is a 3D U-Net (Çiçek et al., 2016), shown in Fig. 2. The label refinement network f_2 and the label appearance simulation network f_{a} use a similar U-Net layout, with the discriminator D in f_{a} using the same layout as the U-Net encoder. The U-Net consists of an encoding path followed by a decoding path, with skip-connections linking the two paths. The network has 5 levels of depth, 16 feature channels in the first layer, and an input image size of 128 × 128 × 128. Each level of the encoding/decoding paths consists of two 3 × 3 × 3 convolutional layers followed by a 2 × 2 × 2 pooling or upsampling layer, respectively. Each convolutional layer consists of 3 × 3 × 3 convolution with zero padding followed by instance normalization and leaky ReLU activation. The number of feature channels is doubled or halved after every pooling or upsampling layer, respectively. The last layer of the U-Net is a 1 × 1 × 1 convolution, combining the outputs into a single feature map, followed by a sigmoid activation. A training batch contains only one image due to GPU memory limits. The networks are implemented using PyTorch (Paszke et al., 2019). The source code is publicly available: https://github.com/ShuaiChenBIGR/Label-refinement-network.

4.4. Details of training and inference of networks

For training, we first apply random rigid transformations as data augmentation, in the form of (1) random 3D rotations up to 30 degrees for all axes, (2) random scaling with a factor between 0.7–1.4 and (3) random flipping in the three directions. Then, we generate samples by extracting random image patches of size 128 × 128 × 128 on the fly from the input training images and corresponding ground truth segmentations. For the airway segmentation experiments, a lung mask is applied to the output of the network and the ground truth patches before computing the training loss. For this operation, we use a precomputed lung mask that is easily obtained with a region growing algorithm (Lo et al., 2010). During training, we used the Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of $1{\times}10^{-2}$. To train the refinement network f_2, the segmentation $\tilde{x}$ in each training sample is randomly sampled with equal probability from either the initial segmentations x or the segmentations with appearance-enhanced synthetic errors $\hat{x_s}$ from the label appearance simulation network.

During inference on new images, the input patches are extracted in a sliding-window fashion, with an overlap of 50% in the three directions. Then, the patch-wise predicted output by the network is aggregated by stitching the patches together, to reconstruct the full-size segmentation result. For the airway segmentation experiments, we applied a lung

Table 1

Results for airway segmentation. Average performance (standard deviation) over the results obtained from three random data splits. LR: simple label refinement network. LR+Syn(init): label refinement method with synthetic errors on initial segmentations. LR+Syn: label refinement method with synthetic errors on ground truth segmentations. LR+Syn+LASN: label refinement method with label appearance simulation network. ↑: significantly better than the U-Net baseline ($p < 0.05$). ↓: significantly worse than the U-Net baseline ($p < 0.05$). P-values are calculated by the paired two-sided Student's T-test (on the average results from the three data splits). Boldface: best results, or not significantly different from the best results.

Method	Dice	Completeness	Leakage	Gaps
U-Net baseline (Garcia-Uceda et al., 2021)	0.76 (0.05)	0.74 (0.12)	0.23 (0.19)	95.73 (47.94)
DoubleU-Net (Jha et al., 2020)	0.77 (0.05)↑	0.73 (0.11)	0.21 (0.18)	99.93 (48.11)
SCAN (Dai et al., 2018)	0.77 (0.05)↑	0.75 (0.11)↑	0.31 (0.23)↓	98.83 (48.81)
Post-DAE (Larrazabal et al., 2020)	0.76 (0.06)	0.74 (0.12)	0.23 (0.19)	94.35 (48.17)
DVAE (Araújo et al., 2019)	0.75 (0.06)	0.72 (0.12)	0.18 (0.17)↑	**93.68** (49.69)↑
U-Net + clDice (Shit et al., 2021)	0.78 (0.05)↑	**0.75** (0.11)↑	0.25 (0.18)	95.96 (49.28)
LR	0.76 (0.05)	0.74 (0.11)↑	0.23 (0.17)	94.90 (47.66)
LR+Syn(init)	0.77 (0.06)	0.73 (0.12)	0.19 (0.17)	94.92 (50.14)
LR+Syn	**0.79** (0.05)↑	0.73 (0.12)	**0.17** (0.17)↑	**93.54** (50.83)↑
LR+Syn+LASN (proposed)	**0.79** (0.05)↑	**0.75** (0.11)↑	0.20 (0.16)↑	**91.63** (48.63)↑

mask to the final segmentation to remove any spurious noise prediction outside the lungs. For this operation, we use the same region growing algorithm as during training.

For the adversarial loss in Eq. (2), the weight λ is set to 0.01 for all experiments in this paper, based on visual inspection of the generated segmentations with appearance-enhanced synthetic errors $\hat{x_s}$.

4.5. Comparisons

We compared the results of our proposed method with the baseline 3D U-Net segmentation network described in Section 4.3. Also, we compared our method with four other label refinement methods: DoubleU-Net (Jha et al., 2020), SCAN (Dai et al., 2018), Post-DAE (Larrazabal et al., 2020) and DVAE (Araújo et al., 2019); and a U-Net trained with the clDice loss (Shit et al., 2021). In all cases, we reimplemented the methods from the original papers. DoubleU-Net consists of two consecutive U-Nets, with skip connections from the encoder of the first U-Net to the decoders of both U-Nets. For DoubleU-Net, no hyperparameters needed to be tuned. SCAN uses a U-Net with a discriminator and adversarial loss, discriminating between the segmentation results and the ground truth. We tuned the weight that balances the segmentation loss and the adversarial loss (low value on the adversarial term) between 0.001 and 0.1, on the validation sets for each application. For Post-DAE, we trained the denoising autoencoder (DAE) to map the segmentation with synthetic errors (x_s) to the ground truth segmentation (g). Then, we passed the initial segmentation (x) through the trained DAE to obtain the final segmentation (y). We used the DAE implementation for binary segmentation (Larrazabal et al., 2020). For DVAE, we trained the method with the class-weighted binary cross entropy loss with weights 0.7 and 0.3 for foreground and background classes, respectively (Araújo et al., 2019). For the U-Net trained with the clDice loss, we used the combination of clDice and Dice losses with parameter $\alpha = 0.5$ (Shit et al., 2021). Our implementations of all methods used the same 3D U-Net backbone as our proposed method and the U-Net baseline.

We also conducted an ablation study of the proposed method (LR+Syn+LASN) by removing some of the components. We evaluated (1) a simple label refinement method by inputting the original images and the initial segmentations without any synthetic errors (LR), (2) a label refinement method with synthetic errors added to the initial segmentations (LR+Syn(init)), and (3) a label refinement method with synthetic errors added to the ground truth segmentations but without the label appearance simulation network (LR+Syn).

4.6. Evaluation metrics

We evaluated the methods with the Dice coefficient to measure the overall segmentation quality, as well as with three metrics designed for tree-like structures: centerline completeness, centerline leakage, and number of gaps. For the airway segmentation experiments, the required centerlines were obtained by applying a skeletonization method (Lee et al., 1994) to the ground truth segmentation mask. For the vessel segmentation experiments, the ground truth centerlines were manually annotated. The evaluation metrics are defined below:

Dice coefficient measures the voxelwise overlap between the predicted mask Y and the ground truth mask G:

$$Dice = \frac{2|Y \cap G|}{|Y| + |G|} \tag{4}$$

Centerline completeness measures the proportion of the length of correctly detected centerlines (i.e., the intersection between the predicted mask Y and the ground truth centerlines G_{cl}) with respect to the length of ground truth centerlines G_{cl}:

$$Completeness = \frac{|Y \cap G_{\text{cl}}|}{|G_{\text{cl}}|} \tag{5}$$

Centerline leakage measures the proportion of the length of false positive centerlines (i.e., the intersection between the predicted centerlines Y_{cl} and the ground truth background $1 - G$) with respect to the length of ground truth centerlines G_{cl}:

$$Leakage = \frac{|Y_{\text{cl}} \cap (1 - G)|}{|G_{\text{cl}}|} \tag{6}$$

Gaps measures the number of continuity gaps in the correctly detected centerlines (i.e., the intersection between the predicted mask Y and the ground truth centerlines G_{cl}). It is calculated with connected component analysis (Fiorio and Gustedt, 1996) as follows:

$$Gaps = NCC(Y \cap G_{\text{cl}}) - NCC(G_{\text{cl}}) \tag{7}$$

with NCC counting the number of 26-neighbor-connected components in the input centerlines.

5. Results

5.1. Segmentation results

The results of our experiments for airway and brain vessel segmentation are shown in Tables 1 and 2, respectively. In both applications, the proposed label refinement method achieves the highest Dice and completeness scores, the lowest number of gaps, with a moderate leakage compared to the other methods. This indicates that our method learns from the synthesized errors and succeeds in correcting errors in the real data. In both applications, the methods with the highest completeness (U-Net trained with clDice loss for airways, and DoubleU-Net for vessels) show both more leakage and more gaps than our method. This indicates that these methods may lack the ability to learn relevant label structural information, and over-segment branches to increase the completeness rather than correcting errors in continuity. For airway segmentation, DVAE shows a similar number of gaps to the

Table 2
Results for brain vessel segmentation. Average performance (standard deviation) over the results obtained from three random data splits. LR: simple label refinement network. LR+Syn(init): label refinement method with synthetic errors on initial segmentations. LR+Syn: label refinement method with synthetic errors on ground truth segmentations. LR+Syn+LASN: label refinement method with label appearance simulation network. ↑: significantly better than the U-Net baseline ($p < 0.05$). ↓: significantly worse than the U-Net baseline ($p < 0.05$). P-values are calculated by the paired two-sided Student's T-test (on the average results from the three data splits). Boldface: best results, or not significantly different from the best results.

Method	Dice	Completeness	Leakage	Gaps
U-Net baseline (Su et al., 2020)	0.57 (0.10)	0.70 (0.18)	0.19 (0.18)	106.68 (161.41)
DoubleU-Net (Jha et al., 2020)	0.59 (0.09)↑	**0.73** (0.18)↑	0.18 (0.16)	92.41 (151.27)↑
SCAN (Dai et al., 2018)	0.57 (0.09)	0.70 (0.18)	0.17 (0.15)	104.05 (160.91)
Post-DAE (Larrazabal et al., 2020)	0.58 (0.10)	0.69 (0.19)	0.13 (0.14)↑	97.36 (172.81)
DVAE (Araújo et al., 2019)	0.57 (0.10)	0.68 (0.18)	0.15 (0.14)↑	77.44 (143.21)↑
U-Net + clDice (Shit et al., 2021)	**0.61** (0.11)↑	0.72 (0.18)↑	0.16 (0.16)↑	67.52 (122.44)↑
LR	0.57 (0.10)	0.70 (0.18)	0.16 (0.16)↑	82.05 (139.45)↑
LR+Syn(init)	0.58 (0.11)	0.71 (0.19)	0.18 (0.15)	69.91 (126.89)↑
LR+Syn	0.60 (0.11)↑	0.71 (0.19)	**0.12** (0.11)↑	64.86 (115.21)↑
LR+Syn+LASN (proposed)	**0.62** (0.10)↑	**0.74** (0.20)↑	0.14 (0.11)↑	**46.64** (76.57)↑

Table 3
Results with semi-supervised learning for airway segmentation. Average performance (standard deviation) over the results obtained from three random data splits. LR+Syn+LASN: proposed method trained only with labeled data. LR+Syn+LASN+Unlabeled: proposed method trained with both labeled and unlabeled data. Boldface: significantly better than the supervised results ($p < 0.05$). P-values are calculated by the paired two-sided Student's T-test (on the average results from the three data splits).

Method	Dice	Completeness	Leakage	Gaps
LR+Syn+LASN	0.79 (0.05)	0.75 (0.11)	0.20 (0.16)	91.63 (48.63)
LR+Syn+LASN+Unlabeled	**0.81** (0.04)	**0.77** (0.10)	0.19 (0.16)	90.53 (48.80)

Table 4
Results with semi-supervised learning for brain vessel segmentation. Average performance (standard deviation) over the results obtained from three random data splits. LR+Syn+LASN: proposed method trained only with labeled data. LR+Syn+LASN+Unlabeled: proposed method trained with both labeled and unlabeled data. Boldface: significantly better than the supervised results ($p < 0.05$). P-values are calculated by the paired two-sided Student's T-test (on the average results from the three data splits).

Method	Dice	Completeness	Leakage	Gaps
LR+Syn+LASN	0.62 (0.10)	0.74 (0.20)	0.14 (0.11)	46.64 (76.57)
LR+Syn+LASN+Unlabeled	**0.63** (0.09)	0.75 (0.18)	0.13 (0.11)	**42.45** (71.26)

proposed method, while suffering from lower Dice and completeness. For vessel segmentation, both Post-DAE and DVAE show lower Dice and completeness than the proposed method, with a moderate improvement in the number of gaps compared to the U-Net baseline. This indicates that the autoencoder-based methods suffer from under-segmentation in both applications.

In the ablation study, the label refinement method with synthetic errors (LR+Syn) achieves better Dice, leakage, and number of gaps scores than the baseline refinement network (LR), for both applications. For airway segmentation, the (LR+Syn) method has slightly lower completeness, while this is similar for vessel segmentation. Moreover, adding synthetic errors to the initial segmentations (LR+Syn(init)), in contrast to doing so to the ground truth segmentations (LR+Syn), achieves similar results in all metrics when compared to the baseline U-Net, for both applications. This suggests that the initial segmentations are too incomplete to add sufficient useful synthetic errors to train the refinement network. The proposed method, combining the synthetic errors and the label appearance simulation network (LR+Syn+LASN), achieves a much higher completeness, with similar Dice, leakage and number of gaps scores when compared to the method with only synthetic errors (LR+Syn), for both applications.

5.2. Semi-supervised results

We conducted experiments using semi-supervised learning to train the proposed label refinement method, to investigate the benefit of using additional unlabeled data for training. As segmentations in which to synthesize errors for the unlabeled data, we used the results on the same data obtained by the proposed method (LR+Syn+LASN) trained on the labeled data. We denote these results as "pseudo labels". The error generation in these pseudo labels follows the same strategy and hyperparameters as in the previous experiments (Sections 3.2 and 4.2). The pseudo labels are also used as ground truth segmentations for the unlabeled images. These unlabeled data together with the labeled data in the previous experiments are then used to train a new label refinement network.

The results of our semi-supervised learning experiments for airway and brain vessel segmentation are shown in Tables 3 and 4, respectively. Adding unlabeled data for training significantly improves the Dice score while the leakage remains similar, for both applications. The completeness is also improved for airway segmentation, while for vessels, the number of gaps is improved.

5.3. Influence of the synthetic error rate

The results of our experiments varying the maximum synthetic error rates (Section 4.2) are shown in Fig. 5. For airway segmentation, with a smaller amount of "discontinuity" errors (0.1) the completeness is increased. Between 0.1 and 0.5, changing the amount of "discontinuity" errors in the segmentations with synthetic errors does not affect much the method performance. In contrast, increasing the amount of "missing terminal branches" errors improves both Dice and completeness scores, reaching a peak when the maximum error rate is ≈0.75. This supports our hypothesis that missing terminal branches are relevant errors to be corrected in the initial airway segmentations. For vessel segmentation, a moderate amount (0.6) of "discontinuity" errors has a positive effect in the method performance.

When compared to the LR+Syn and LR+Syn(init) methods, the proposed label refinement method is able to learn from higher amounts of synthetic errors, thereby improving the label refinement performance.

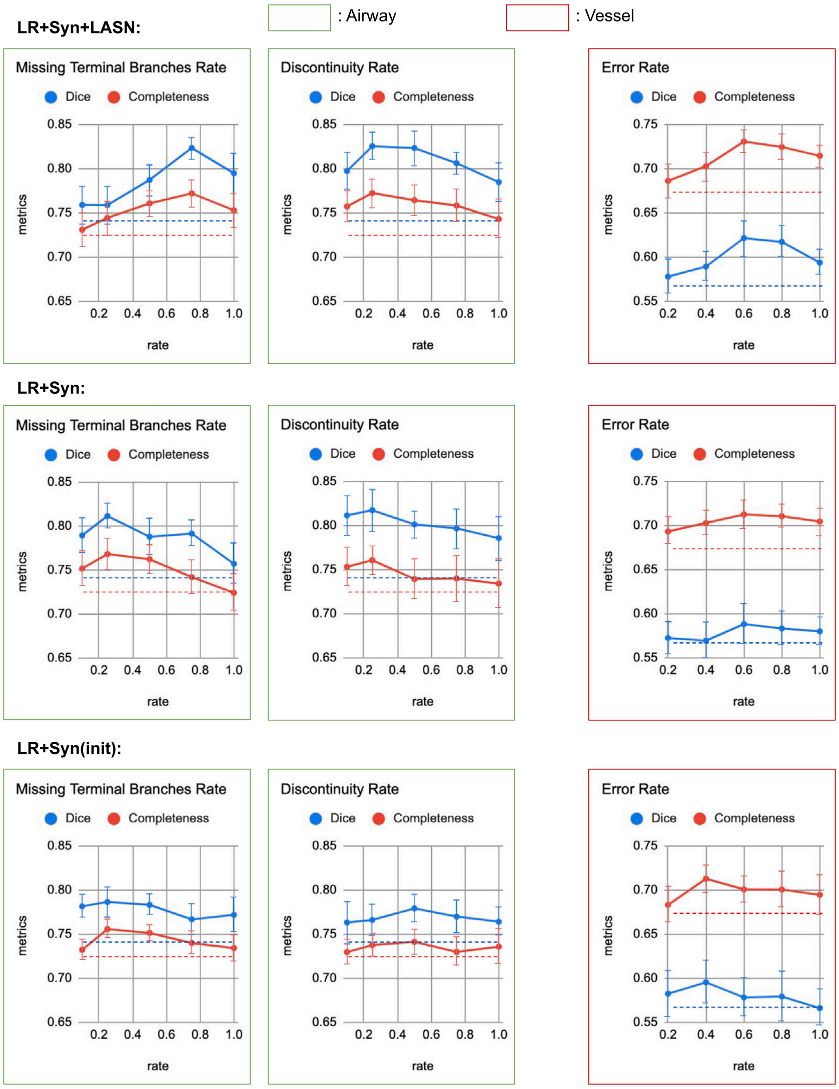

Fig. 5. Influence of the hyperparameters of the proposed method, the maximum synthetic error rates, in the method performance, for airway and brain vessel segmentation. Results are shown as average performance with standard deviation (error bars), for Dice and completeness metrics, over three random data splits. The results for the baseline (LR) are displayed as dashed line.

6. Discussion

In this paper, we propose a novel label refinement method that can correct errors in the initial segmentations from a standard deep segmentation network such as the U-Net. The novelty of our method is that it uses segmentations augmented with realistic synthetic errors as training samples, from where the label refinement network can learn to correct the errors. The synthetic errors are automatically generated to simulate common errors observed in the initial segmentations and are then refined by a label appearance simulation network to resemble the appearance of real errors in the initial segmentations.

We evaluated our method on the segmentation of airways from chest CT scans and brain vessels from CTA images of the brain. In both applications, our method achieved significantly higher Dice overlap and completeness scores, with a lower number of gaps and a comparable leakage, when compared to the baseline U-Net and other previous label refinement methods. When segmenting branching structures, a higher completeness means that more and/or longer branches are detected, especially the smaller ones that are challenging to segment.

The ability of our method to segment highly complete tree-like structures with more branches is clinically important, as this could lead to more sensitive biomarkers. For example, in airway analysis, the airway-artery ratio (Kuo et al., 2017) and airway tapering (Kuo et al., 2020) measures can be used to assess cystic fibrosis lung disease, and including more measurements from the smaller peripheral branches can allow earlier detection of the disease (Tiddens et al., 2010). Moreover, the ability of our method to correct continuity errors and thereby connect the segmentation is beneficial, as most methods to measure branches assume a fully connected segmentation and discard branches after a discontinuity.

The proposed method outperformed the four other label refinement methods: DoubleU-Net (Jha et al., 2020), SCAN (Dai et al., 2018), Post-DAE (Larrazabal et al., 2020) and DVAE (Araújo et al., 2019); and the U-Net trained with the clDice loss (Shit et al., 2021). Moreover, using semi-supervised learning techniques to train our method with additional unlabeled data we can further improve the method performance when compared to the fully supervised setting.

6.1. Comparison with other methods

The main difference between the proposed method and other label refinement methods is that ours provides a more general and powerful way of using a training dataset that includes segmentations augmented with synthetic errors. Instead, DoubleU-Net (Jha et al., 2020) uses the original images masked by the initial segmentations to train the second network. Although the increased model capacity of DoubleU-Net may improve the segmentations, its ability to correct the errors may be limited by the fact that no new errors are introduced to the input of the refinement network. This makes it less efficient to implicitly exploit the label structural information similar to a standard U-Net. Post-DAE (Larrazabal et al., 2020) and DVAE (Araújo et al., 2019) aim to refine the initial segmentation by learning and recovering the overall label structure. In both applications, the autoencoder-based methods show worse segmentation performance than our method. This may reflect the limitations of autoencoders in reconstructing the complicated label structure of airways and brain vessels. SCAN (Dai et al., 2018) refines the segmentation by making it indistinguishable from the ground truth segmentation through an adversarial loss, where the distribution of the learned features may also provide the general structural information of the objects to be segmented. SCAN mainly focuses on simulating the appearance of the ground truth segmentations. However, SCAN is not designed to learn to correct structural errors explicitly, thus it may not capture the local continuity information as efficiently as our method. This is reflected by the significantly worse completeness reported by SCAN in Tables 1 and 2, for both applications. Our method provides an implicit way to enhance the network awareness of the structural information in the ground truth segmentations. For example, after seeing many continuity errors, the refinement network is expected to understand the local continuity within elongated structures, and consequently to be able to correct these errors in the initial segmentations. Finally, the clDice loss (Shit et al., 2021) focuses the training on the centerline structures and penalizes errors (discontinuities) in them. Although using the clDice loss is a straightforward approach, it only corrects discontinuity errors and not missing terminal branches.

6.2. Synthetic errors for semi-supervised learning

With the proposed method, synthetic errors can be added to any pseudo labels obtained on unlabeled data, to be used in semi-supervised learning. In Section 5.2 we have shown that our method performance was significantly improved when using additional unlabeled data for training. Our approach to generating synthetic errors could be used together with other common semi-supervised methods using pseudo labels, e.g., to optimize the prediction consistency of the same image from different models (Tarvainen and Valpola, 2017), or the prediction consistency of the same image with different transformations (Bortsova et al., 2019). Using synthetic errors in these methods may improve the segmentation quality of pseudo labels from the unlabeled data, which could provide more informative features from these data and thereby improve the segmentation performance.

6.3. Importance of realistic synthetic errors

The proposed label refinement network may underperform if the segmentations with synthetic errors used for training are too different from the initial segmentations. In our method, the synthetic errors are added to the ground truth segmentations, which have a fine and smooth appearance. In contrast, the initial segmentations are more irregular. Our proposed solution is to use a label appearance simulation network trained with an adversarial loss to make the appearance of the segmentations with synthetic errors resemble that of the real initial segmentations. The results in Tables 1 and 2 clearly show the benefit of using the LASN network in our method. In both applications, without the LASN network could our method (LR+Syn) only slightly improve the segmentation performance with a reduced leakage, when compared to the baseline (LR). This may be due to the positive regularization effect of increasing the variety in the training data by including the segmentations with synthetic errors. Only after introducing the LASN network was our method able to improve the completeness while retaining an adequate leakage.

6.4. Possible application to other segmentation tasks

While we studied only two applications in this work, the proposed method can be applied to other segmentation tasks of tree-like structures, such as neurons or vessels in retina, liver, or lungs. Indeed, the types of errors studied (i.e. missing terminal branches and discontinuities) are relevant for any tree-like structures, and not limited to airways and brain vessels.

The proposed label refinement method via error synthesis can be applied to other non-tree-like segmentation tasks. The core step is to identify common types of errors in the initial segmentations. For example, a common error we observed in prior work using the U-Net for the segmentation of the aorta and pulmonary arteries from chest CT scans (Chen et al., 2021) was that the segmentation of one of the structures often leaked into the other one, while being both independent anatomical structures. This is mostly due to the obscured boundaries of both arteries on the CT scan. This type of error can be simulated by locally removing the boundaries between the aorta and pulmonary artery classes. Applying our method to correct such errors may improve the overall segmentation performance for this application.

In some segmentation tasks, a discontinuity in the segmentation could be due to a pathological abnormality (e.g. mucus plugs in airways or occlusions in vessels), rather than an error to be corrected by our method. In this scenario, since the label refinement model f_2 uses as inputs both the original image and the segmentation with synthetic errors, from the image it may be able to differentiate between cases where there is a label discontinuity but some evidence of an organ (a segmentation error) and cases of a pathological abnormality (not an error). Whether our method can handle pathological abnormalities correctly will depend on whether the training dataset contains sufficiently many examples of these, and should be tested on data containing such abnormalities.

6.5. Limitations

The main limitation of the proposed method lies in the two-step design and implementation: (1) analyze the errors in the initial segmentations to identify the relevant types of errors, and (2) design and generate the synthetic errors based on these results. The first step requires observation and interpretation by human experts. The synthetic errors we used in this paper are suitable for the segmentation of tree-like structures. However, the relevant types of errors generally differ across different applications and datasets, and therefore the synthetic errors we used are not directly applicable to other segmentation tasks. The second step is typically a complex image processing task. Nevertheless, once the synthetic errors are successfully designed for a given application, the training of our label refinement method can be done fully automatically.

A limitation of our validation of the proposed method is that we considered only two types of false negative errors (i.e., missing terminal branches and errors in continuity). We did not consider false positive errors because these were much less frequent in the initial segmentations and often appeared as disconnected blobs that could be easily removed without the need for more complex label refinement. Nevertheless, from the results obtained in this paper we expect that our method can successfully correct other types of errors as well.

7. Conclusion

We presented a novel label refinement method that can learn from synthetic errors to refine the initial segmentations from a base segmentation network. A label appearance simulation network was applied to reduce the appearance difference between the segmentations with synthetic errors and the real initial segmentations, thereby improving the generalizability of our method. On two segmentation tasks for branching structures, the proposed method achieved significantly better segmentation results when compared to four previous label refinement methods, and a U-Net trained with a loss tailored for tubular structures. The segmentation performance of our method was further improved by using additional unlabeled data for training with semi-supervised learning techniques.

CRediT authorship contribution statement

Shuai Chen: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Antonio Garcia-Uceda:** Writing – review & editing, Visualization, Software, Resources, Methodology, Investigation, Data curation. **Jiahang Su:** Writing – review & editing, Visualization, Software, Methodology, Investigation, Data curation. **Gijs van Tulder:** Writing – review & editing, Supervision, Methodology. **Lennard Wolff:** Writing – review & editing, Resources, Investigation, Data curation. **Theo van Walsum:** Writing – review & editing, Supervision, Resources, Data curation. **Marleen de Bruijne:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Shuai Chen reports financial support was provided by Chinese Scholarship Council. Shuai Chen reports a relationship with Erasmus Medical Center that includes: employment. Antonio Garcia-Uceda reports a relationship with Erasmus Medical Center that includes: employment. Jiahang Su reports a relationship with Erasmus Medical Center that includes: employment. Gijs van Tulder reports a relationship with Radboud University that includes: employment. Lennard Wolf reports a relationship with Erasmus Medical Center that includes: employment. Theo van Walsum reports a relationship with Erasmus Medical Center that includes: employment. Marleen de Bruijne reports a relationship with Erasmus Medical Center that includes: employment.

Data availability

The authors do not have permission to share data.

Acknowledgments

We want to thank Thirona and H. Tiddens, M. van de Corput and M. Bonte (Erasmus MC - Lung Analysis group) for providing the airway segmentations and anonymous chest CT data used in this study. This work was partially funded by Chinese Scholarship Council (File No. 201706170040), Netherlands Organisation for Scientific Research (NWO) project VI.C.182.042, the MARBLE project, Netherlands (EFRO/OP-Oost: PROJ-00887), the Contrast project (Dutch Heart Foundation (CVON2015-01: CONTRAST), the Brain Foundation Netherlands (HA2015.01.06) and additional funding by the Ministry of Economic Affairs, Netherlands by means of the PPP Allowance made available by the Top Sector Life Sciences & Health to stimulate public–private partnerships (LSHM17016)).

Ethics approval

This work involved medical images from human subjects acquired for clinical studies. The CT scans used for airway and brain vessel experiments are from the study protocols Update CF and MR CLEAN Registry, respectively. Approval of the study protocols was granted by the central medical ethics committee of the hospital Erasmus MC, Rotterdam, the Netherlands (No. MEC-2013-338 and MEC-2014-235, respectively).

References

Araújo, R.J., Cardoso, J.S., Oliveira, H.P., 2019. A deep learning design for improving topology coherence in blood vessel segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 93–101.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., de Bruijne, M., 2019. Semi-supervised medical image segmentation via learning consistency under transformations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 810–818.

Bouma, N., Janssens, H., Andrinopoulou, E., Tiddens, H., 2020. Airway disease on chest computed tomography of preschool children with cystic fibrosis is associated with school-age bronchiectasis. Pediatr. Pulmonol. 55 (1), 141–148. http://dx.doi.org/10.1002/ppul.24498.

Chen, S., Gamechi, Z.S., Dubost, F., van Tulder, G., de Bruijne, M., 2021. An end-to-end approach to segmentation in medical images with CNN and posterior-CRF. Med. Image Anal. 102311.

Cheng, G., Wu, X., Xiang, W., Guo, C., Ji, H., He, L., 2021. Segmentation of the airway tree from chest CT using tiny atrous convolutional network. IEEE Access 9, 33583–33594. http://dx.doi.org/10.1109/ACCESS.2021.3059680.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Dai, W., Dong, N., Wang, Z., Liang, X., Zhang, H., Xing, E.P., 2018. SCAN: Structure correcting adversarial network for organ segmentation in chest x-rays. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 263–273.

Fiorio, C., Gustedt, J., 1996. Two linear time union-find strategies for image processing. Theoret. Comput. Sci. 154 (2), 165–181.

Garcia-Uceda, A., Selvan, R., Saghir, Z., Tiddens, H., de Bruijne, M., 2021. Automatic airway segmentation from computed tomography using robust and efficient 3-D convolutional neural networks. Sci. Rep. 11 (1), 16001.

Graham, M., Gibbs, J., Cornish, D., Higgins, W., 2010. Robust 3-D airway tree segmentation for image-guided peripheral bronchoscopy. IEEE Trans. Med. Imaging 29 (4), 982–997.

Hilbert, A., Madai, V.I., Akay, E.M., Aydin, O.U., Behland, J., Sobesky, J., Galinovic, I., Khalil, A.A., Taha, A.A., Wuerfel, J., et al., 2020. BRAVE-NET: fully automated arterial brain vessel segmentation in patients with cerebrovascular disease. Front. Artif. Intell. 3, 78.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Jansen, I.G., Mulder, M.J., Goldhoorn, R.-J.B., 2018. Endovascular treatment for acute ischaemic stroke in routine clinical practice: prospective, observational cohort study (MR CLEAN registry). BMJ 360.

Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D., 2020. DoubleU-Net: A deep convolutional neural network for medical image segmentation. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems. CBMS, IEEE, pp. 558–564.

Kingma, D., Ba, J., 2017. Adam: A method for stochastic optimization. arXiv e-prints arXiv:arXiv:1412.6980.

Kuo, W., de Bruijne, M., Petersen, J., Nasserinejad, K., Ozturk, H., Chen, Y., Perez-Rovira, A., Tiddens, H., 2017. Diagnosis of bronchiectasis and airway wall thickening in children with cystic fibrosis: Objective airway-artery quantification. Eur. Radiol. 27 (11), 4680–4689.

Kuo, W., Perez-Rovira, A., Tiddens, H., de Bruijne, M., study group, N.C.C., 2020. Airway tapering: an objective image biomarker for bronchiectasis. Eur. Radiol. 30 (5), 2703–2711.

Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E., 2020. Post-DAE: anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE Trans. Med. Imaging 39 (12), 3813–3820.

Lee, T., Kashyap, R., Chu, C., 1994. Building skeleton models via 3-D medial surface axis thinning algorithms. CVGIP, Graph. Models Image Process. 56 (6), 462–478.

Livne, M., Rieger, J., Aydin, O.U., Taha, A.A., Akay, E.M., Kossen, T., Sobesky, J., Kelleher, J.D., Hildebrand, K., Frey, D., et al., 2019. A U-Net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease. Front. Neurosci. 13, 97.

Lo, P., van Ginneken, B., Reinhardt, J., Tarunashree, Y., de Jong, P., Irving, B., Fetita, C., Ortner, M., Pinho, R., Sijbers, J., Feuerstein, M., Fabijanska, A., Bauer, C., Beichel, R., Mendoza, C., Wiemker, R., Lee, J., Reeves, A., Born, S., Weinheimer, O., van Rikxoort, E., Tschirren, J., Mori, K., Odry, B., Naidich, D., Hart-mann, I., Hoffman, E., Prokop, M., Pedersen, J., de Bruijne, M., 2012. Extraction of airways from CT (EXACT'09). IEEE Trans. Med. Imaging 31 (11), 2093–2107.

Lo, P., Sporring, J., Ashraf, H., Pedersen, J., de Bruijne, M., 2010. Vessel-guided airway tree segmentation: A voxel classification approach. Med. Image Anal. 14 (4), 527–538.

Lo, P., Sporring, J., Pedersen, J., de Bruijne, M., 2009. Airway tree extraction with locally optimal paths. Med. Image Comput. Comput. Assist. Interv. MICCAI 51–58.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Meijs, M., Patel, A., van de Leemput, S.C., Prokop, M., van Dijk, E.J., de Leeuw, F.-E., Meijer, F.J., van Ginneken, B., Manniesing, R., 2017. Robust segmentation of the full cerebral vasculature in 4D CT of suspected stroke patients. Sci. Rep. 7 (1), 1–12.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Peter, R., Emmer, B.J., van Es, A.C., van Walsum, T., 2017. Cortical and vascular probability maps for analysis of human brain in computed tomography images. In: 2017 IEEE 14th International Symposium on Biomedical Imaging. ISBI 2017, IEEE, pp. 1141–1145.

Qin, Y., Zheng, H., Gu, Y., Huang, X., Yang, J., Wang, L., Yao, F., Zhu, Y., Yang, G., 2021. Learning tubule-sensitive CNNs for pulmonary airway and artery-vein segmentation in CT. IEEE Trans. Med. Imaging 40 (6), 1603–1617.

Rodriguez-Luna, D., Dowlatshahi, D., Aviv, R.I., Molina, C.A., Silva, Y., Dzialowski, I., Lum, C., Czlonkowska, A., Boulanger, J.-M., Kase, C.S., et al., 2014. Venous phase of computed tomography angiography increases spot sign detection, but intracerebral hemorrhage expansion is greater in spot signs detected in arterial phase. Stroke 45 (3), 734–739.

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

Sanchesa, P., Meyer, C., Vigon, V., Naegel, B., 2019. Cerebrovascular network segmentation of MRA images with deep learning. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, IEEE, pp. 768–771.

Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P., Bauer, U., Menze, B.H., 2021. clDice-a novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16560–16569.

Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V., 2021. U-Net and its variants for medical image segmentation: A review of theory and applications. IEEE Access 9, 82031–82057. http://dx.doi.org/10.1109/ACCESS.2021.3086020.

Su, J., Wolff, L., van Es, A.C.M., van Zwam, W., Majoie, C., Dippel, D.W., van der Lugt, A., Niessen, W.J., Van Walsum, T., 2020. Automatic collateral scoring from 3D CTA images. IEEE Trans. Med. Imaging 39 (6), 2190–2200.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Adv. Neural Inf. Process. Syst. 30.

Tiddens, H., Donaldson, S., Rosenfeld, M., Pare, P., 2010. Cystic fibrosis lung disease starts in the small airways: Can we treat it more effectively? Pediatr. Pulmonol. 45 (2), 107–117. http://dx.doi.org/10.1002/ppul.21154.

Yang, Y., Wang, Z., Liu, J., Cheng, K.-T., Yang, X., 2019. Label refinement with an iterative generative adversarial network for boosting retinal vessel segmentation. arXiv preprint arXiv:1912.02589.

Zheng, H., Qin, Y., Gu, Y., Xie, F., Sun, J., Yang, J., Yang, G., 2021. Refined local-imbalance-based weight for airway segmentation in CT. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2021. pp. 410–419. http://dx.doi.org/10.1007/978-3-030-87193-2_39.