3D Vascular Segmentation Supervised by 2D Annotation of Maximum Intensity Projection

Zhanqiang Guo, Zimeng Tan, Jianjiang Feng, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*— Vascular structure segmentation plays a crucial role in medical analysis and clinical applications. The practical adoption of fully supervised segmentation models is impeded by the intricacy and time-consuming nature of annotating vessels in the 3D space. This has spurred the exploration of weakly-supervised approaches that reduce reliance on expensive segmentation annotations. Despite this, existing weakly supervised methods employed in organ segmentation, which encompass points, bounding boxes, or graffiti, have exhibited suboptimal performance when handling sparse vascular structure. To alleviate this issue, we employ maximum intensity projection (MIP) to decrease the dimensionality of 3D volume to 2D image for efficient annotation, and the 2D labels are utilized to provide guidance and oversight for training 3D vessel segmentation model. Initially, we generate pseudo-labels for 3D blood vessels using the annotations of 2D projections. Subsequently, taking into account the acquisition method of the 2D labels, we introduce a weakly-supervised network that fuses 2D-3D deep features via MIP to further improve segmentation performance. Furthermore, we integrate confidence learning and uncertainty estimation to refine the generated pseudo-labels, followed by fine-tuning the segmentation network. Our method is validated on five datasets (including cerebral vessel, aorta and coronary artery), demonstrating highly competitive performance in segmenting vessels and the potential to significantly reduce the time and effort required for vessel annotation. Our code is available at: https://github.com/gzq17/Weakly-Supervised-by-MIP.**

***Index Terms*— Vessel Segmentation, Weakly-Supervised, Maximum Intensity Projection, Pseudo-Label Refinement**

I. Introduction

Tree-like vascular structures are ubiquitously present within the human body, often characterized by intricate complexities observed at a microscale. Prominent instances of such intricate networks encompass cerebral vessels, aorta, and coronary arteries. Computed Tomography Angiography (CTA) and Magnetic Resonance Angiography (MRA) have emerged as invaluable imaging modalities, facilitating the acquisition of extensive vascular image datasets that have advanced vascular structures research. CTA imaging techniques usually require the injection of contrast agent to highlight blood flow, and it is a contrast-based, minimally invasive, and cost-efficient imaging modality [1]. And MRA techniques rely on blood flow or inflow angiography, augmenting flowing blood's radiance in comparison to stationary tissue through the employment of a short echo time and flow compensation [2].

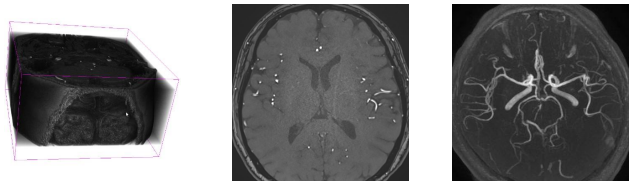
The automatic and accurate segmentation of vessels from CTA and MRA is an essential prerequisite in clinical diagnosis and intervention for vascular diseases. Convolutional Neural Networks (CNNs)-based algorithms have demonstrated impressive performance across diverse computer vision tasks, including the segmentation of vascular structures [3]–[7]. However, the performance of CNNs is contingent upon large annotated datasets, which are tedious and expensive to obtain, especially for vascular images. Consequently, it is meaningful to develop weakly-supervised methods that leverage weak annotations instead of voxel-wise annotations.

Various weakly supervised annotations have been used for different types of segmentation tasks, including image-level category labels [8], bounding boxes [9], [10], scribbles [11], [12], and key points [13], [14]. While these weak annotations demonstrate favorable performance in natural images and large organ segmentation, their utility for sparse blood vessel segmentation remains limited. Image-level annotation is not suitable for segmentation tasks where object classes in images are usually fixed, such as blood vessels and background in our task. The vascular structure typically occupies a small portion of the overall image in the number of voxels, yet exhibits extensive spatial extension, rendering bounding box annotations insufficient in providing substantial informative cues. Scribbles annotation is primarily feasible for 2D images, but given the small size and large number of blood vessels in 2D slices, it is difficult and time-consuming to label, as illustrated in Fig. 1(b). Key points annotation, such as hundreds of bifurcation points and endpoints of blood vessels, is also laborious to locate and label. Consequently, these weak labels have been scarcely employed in vessel segmentation investigations. Another weakly supervised method for 3D segmentation is to fully annotate a subset of slices within the training volume [15], [16]. While this approach does alleviate the segmentation burden, the process of annotating these specific slices remains time-intensive, particularly when dealing with vessel slices (Fig. 1(b)). Furthermore, the annotation of 2D slices lacks the essential information pertaining to vascular connectivity, which is crucial for segmentation of 3D vessels.

Reducing the dimensionality of 3D space to 2D image for annotation and supervision is another intuitive approach in

Manuscript received June 11, 2023. This study got ethical approval of Wuhan Union Hospital of China and Xuanwu Hospital of Capital Medical University (2020009) for using the clinically collected dataset. *(Corresponding author: Jianjiang Feng.)*

Zhanqiang Guo, Zimeng Tan, Jianjiang Feng, and Jie Zhou are with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: guozq21@mails.tsinghua.edu.cn; tzm19@mails.tsinghua.edu.cn; jfeng@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).



(a) Original image (b) Cross-section (c) MIP

Fig. 1. (a) is an original 3D image. The characteristics of vessels on 2D sections are ambiguous and scattered, as shown in (b). (c) is the MIP image of (a). Compared with (a), annotating vessels in MIP image is obviously much easier.

weakly supervised learning. And this has been widely used in 3D point cloud human pose estimation tasks, where the pose information from 2D image is used to aid in the supervision of 3D point cloud pose estimation [17]–[19]. MRA and CTA techniques harness the principle of blood flow or inflow angiography to impart a brighter appearance to blood vessels relative to surrounding tissue during imaging. Leveraging this property, the maximum intensity projection (MIP) technique projects the maximum voxel value along a specified direction onto the resulting plane [20]. By compressing 3D data into 2D projected image, MIP achieves data dimensionality reduction and it is a widely-used scientific method for visualization of vessel structures, vascular analysis, diagnosis, and surgical planning [21]. In clinical practice, radiologists routinely conduct a swift examination of MIP images to rapidly identify the location, shape, and blood flow characteristics of vessels [22]. MIP images offer abundant information about blood vessels for 3D images, with annotating vessels in MIP images proving significantly easier than directly annotating in 3D space, as depicted in Fig. 1(c). Regrettably, limited efforts have been made towards utilizing MIP images directly in the study of 3D vessel segmentation.

In this paper, to alleviate the reliance on 3D vascular annotation and tackle the inapplicability of current weakly supervised labels in 3D vascular segmentation, we introduce a novel approach for weakly supervised learning incorporating the concept of data dimensionality reduction supervision. Leveraging the features unveiled by MIP images, we propose to guide the segmentation of blood vessels in 3D space via 2D MIP annotations. Specifically, we undertake annotation on the MIP image obtained from 3D volume. To utilize effectively of this weak annotation, we back-project the 2D label to generate a sparse 3D foreground image. Subsequently, a region growing algorithm is applied to obtain a more complete labeling of the image. Furthermore, we propose a weakly-supervised segmentation network based on 2D-3D feature fusion to enhance the accuracy of vessel segmentation, taking into account the acquisition method of the 2D labels. Moreover, we integrate the confidence learning (CL) [23] and uncertainty estimation (UE) via Monte Carlo dropout to further improve the reliability of the generated pseudo-labels, followed by a fine-tuning procedure on the segmentation network to refine its performance. Our contributions can be summarized as follows:

- Considering the sparsity of 3D blood vessels, we propose a weakly supervised segmentation framework based on MIP annotations. To the best of our knowledge, this is the first work to utilize MIP image annotations as weakly supervised labels for 3D vessel segmentation.
- A segmentation network that fuses 2D-3D features is developed to make full use of designed weak label. And we integrate confident learning and uncertainty estimation to further improve the network's performance.
- We validate the effectiveness and generalization of our method on five datasets. Additionally, through carefully designed experiments, we demonstrate that our method achieves superior performance compared to fully supervised segmentation methods while requiring less annotation time when utilizing larger quantities of weakly annotated data.

II. Related Work

A. Weakly-Supervised Segmentation

Benefits from the development of CNNs, remarkable progress has been accomplished in the field of weakly supervised segmentation. Li et al. [8] implemented an Online Easy Example Mining method for weakly-supervised segmentation of glands using patch-level category labels. Dorent et al. [10] combined the features of extreme points and bounding boxes to supervise the segmentation of vestibular schwannoma and achieved good result. Scribble is obtainable for most segmentation tasks and Zhang et al. [12] adopted the mixup strategy with a dedicated design of random occlusion to perform increments and decrements of scribbles. Meanwhile, as a label for weakly supervised segmentation, key points are often used for object segmentation with regular shape. Guo et al. [14] proposed a weakly supervised learning method for nuclei segmentation that required annotation of the nuclear centroid. Nevertheless, due to the sparsity characteristic displayed by 3D vessels, these weak labels (image-level labels, bounding boxes, graffiti and points) are unsuitable for our specific task, as discussed in Sec. I.

In practice, most existing weakly supervised approaches of vascular segmentation rely on traditional vessel enhancement techniques to obtain initial segmentation results, which are then iteratively refined manually [24], [25]. Fu et al. [26] introduced to supervise the segmentation of LSCI images by choosing best binary labels acquired through various combinations of thresholds. However, the effectiveness and annotation workload of these methods depend on the quality of initial labels. Moreover, since it is time-consuming to correct the labels in 3D images, these methods are mostly used in 2D vessel segmentation. Aiming to 3D hepatic vessel segmentation, Xu et al. [27] used both high-quality labeled data and noisy labeled data to train their proposed Mean-Teacher-Assisted network. Nevertheless, different from weakly supervised training, this method still required high-quality annotations, and the outcomes were reliant on the quality of noisy labels. In this work, we utilize the annotation of 2D MIP image to supervise the vessel segmentation in 3D space, which greatly reduces the annotation time.

B. Noisy Pseudo-Label Refinement

The primary step in weakly- and semi-supervised learning is generating pseudo-labels of training data, leveraging weakly labeled or pre-existing fully labeled data. To improve the robustness of trained models, recent studies have focused on refining the noisy pseudo-labels. Particularly, uncertainty estimation has also emerged as a common approach for optimizing noisy labels [29], [30]. For instance, Cao et al. [31] estimated uncertainty to discern potential noise in the generated pseudo-labels, consequently mitigating the detrimental impact on network performance. Additionally, Northcutt et al. [23] proposed a confident learning method capable of identifying potentially incorrect samples in noisy labels through uncertainty estimation, subsequently removing them during training. This approach is gradually being adopted for optimizing generated noisy pseudo-labels [14], [27]. However, these techniques exclusively address already labeled data. And, the generated pseudo-labels via MIP labels used in our work only cover a part of the voxels. Consequently, we employ the confidence learning method to refine the already-labeled voxels during the pseudo-label refinement, while simultaneously integrating the uncertainty estimation to assign labels to unlabeled voxels.

C. Dimensionality Reduction Supervision and MIP

The utilization of annotation information derived from low-dimensional data to enhance analysis in high-dimensional spaces has found widespread application across various domains, especially in the field of 3D point cloud pose estimation [17]–[19]. Zhang et al. [19] employed adversarial learning to leverage weakly supervised data comprising solely annotations of 2D human joints, enabling the recovery of human pose. Similarly, Wu et al. [18] introduced a refined point set network structure to transfer annotation information obtained from 2D human pose estimation within existing large-scale RGB datasets to the 3D task.

For 3D CTA and MRA images, MIP is an intuitive method for dimensionality reduction, projecting 3D voxels onto a projection plane based on their maximum intensity. MIP images prove valuable in facilitating rapid observation of vascular structures and blood flow characteristics by medical professionals. Salvi et al. [22] trained a vision transformer using MIP images for the diagnosis of peripheral arterial disease. And recent studies have emphasized the combination of MIP image features to enhance algorithmic performance when analyzing 3D images. Chen et al. [32] leveraged prior knowledge demonstrating the similarity in tree structures between 2D and 3D blood vessels, employing an adversarial learning method to utilize existing 2D blood vessel annotations to supervise the fidelity of the MIP image of the 3D segmentation result. However, the use of projection information is considerably constrained in these studies. For instance, in the study by Dima et al. [33], the reliance on preprocessing steps and the limitations imposed by the type of vascular tissue and imaging method influenced the utilization of projection images. Furthermore, some researchers have utilized projection information to enhance the learning of image features. For instance, Zheng et al. [34] utilized MIP images with varying plate thicknesses as input to augment the spatial information of CT images and aid in discriminating between nodules and blood vessels. Wang et al. [35] integrated MIP image embedding into 3D MRA to extract vessel structures. However, these studies primarily employed MIP to extract feature and required complete 3D vascular annotations. In our work, we employ MIP technique to achieve dimensionality reduction of 3D volume to 2D image, which serves the purpose of facilitating annotation and supervision. The proposed method brings about a remarkable reduction in the required annotation time, while simultaneously guaranteeing the quality of vessel segmentation.

III. Method

The overall pipeline of the proposed weakly-supervised vascular segmentation framework is illustrated in Fig. 2. Due to the limited number of pixels in the 2D MIP labels, we first generate 3D pseudo-labels based on them. Subsequently, the proposed 2D-3D feature fusion network is trained with 2D weak annotations and the newly generated 3D pseudo-labels. To enhance the credibility of the pseudo-labels, confidence learning in combination with uncertainty estimation is employed to optimize the labels, followed by fine-tuning of the network. Each step of the proposed methodology will be described in detail in this section.

A. MIP and 3D Pseudo-labels Generation

Let $X \in R^{H\times W\times D}$ denote a 3D image and $\Omega = \{(x,y,z)\}^{H\times W\times D}$ denote the set of all points in 3D space. In the following description, for convenience, we project the image in the transverse plane. We perform MIP of X to obtain the 2D projection image X_{MIP} and index map X_{index}, mathematically expressed as $X_{\mathrm{MIP}}(x,y) = \max\limits_{z=1,2..,D} X(x,y,z)$ and $X_{\mathrm{index}}(x,y) = \arg\max\limits_{z=1,2..,D} X(x,y,z)$. Y_{MIP} is the vessel annotation of projection image, a weak label of 3D image X. However, Y_{MIP} is extremely sparse for 3D volume, so it is necessary to generate credible 3D pseudo-label to train the segmentation model. Using the index map, we back-project the labeled 2D image to get a series of discrete points S_{p} in 3D space, which are foreground voxels. The back-project operation is expressed as:

$$S_{\mathrm{p}} = \{(x,y,z);\ Y_{\mathrm{MIP}}(x,y) = 1, X_{\mathrm{index}}(x,y) = z\}. \quad (1)$$

To increase the amount of foreground voxels for supervision, we treat S_{p} as seed points and employ a region growing algorithm to obtain the set S_1. The algorithm starts from the seed points S_{p} and gradually adds adjacent voxels to the foreground until the predefined stopping criterion is reached. The criterion we choose is that the difference between the voxel value of the candidate point and the average voxel value of seed points is lower than a preset threshold α.

A straightforward approach of generating the background voxels set is to utilize all columns in the 3D image that correspond to the background points in the 2D projection label, denoting as $T_{\mathrm{b1}} = \{(x,y,z); Y_{\mathrm{MIP}}(x,y) = 0, z = 1,2,...,D\}$. However, two facts are ignored: (i) Most voxels

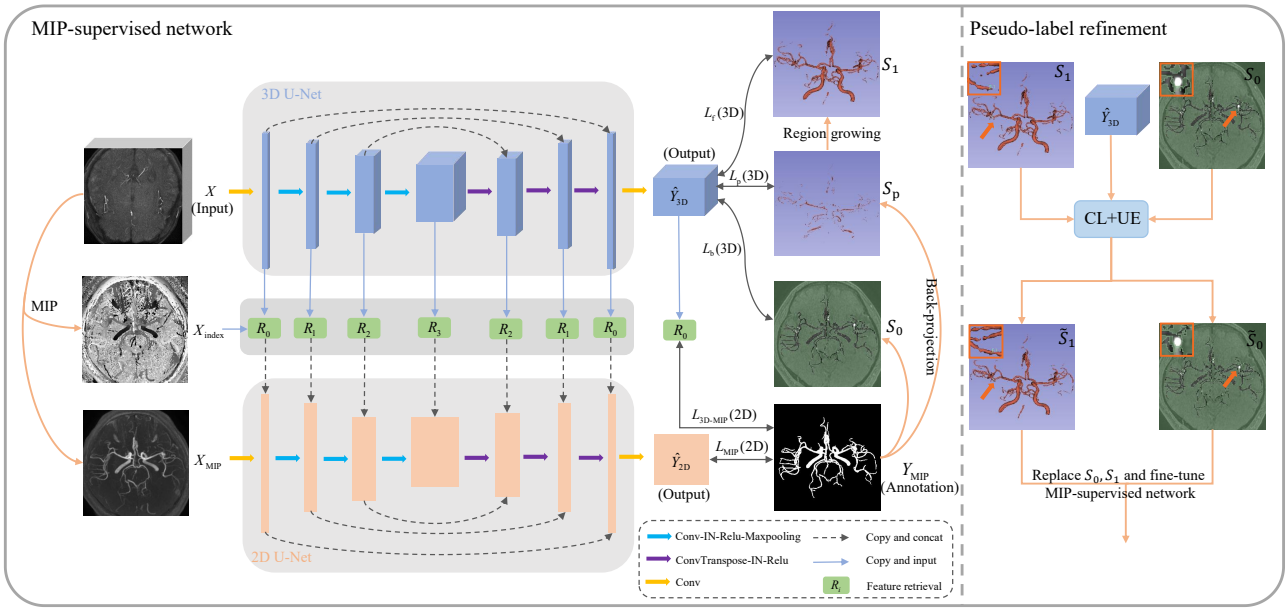


Fig. 2. The proposed weakly-supervised vessel segmentation framework. Initially, we employ MIP to reduce dimensionality of 3D volume for annotating. Subsequently, we introduce a novel 2D-3D feature fusion network, which is trained with pseudo labels generated from 2D annotations. To enhance the efficacy of the network, we integrate confidence learning and uncertainty estimation methods to refine the pseudo labels, followed by fine-tuning of the network.

on the columns correspond to foreground pixels in the 2D label are also background, but are not included in T_{b1}. These voxels are denoted as $T_{\mathrm{b2}} = \{(x,y,z); X(x,y,z) < \beta_{\mathrm{th}}$ and $Y_{\mathrm{MIP}}(x,y) = 1\}$; (ii) During projection, certain blood vessels may be obstructed by brighter noise, resulting in being displayed as background in 2D projection image. $T_{\mathrm{b3}} = \{(x,y,z); \gamma_{\mathrm{th}} < X(x,y,z) < \eta_{\mathrm{th}}$ and $Y_{\mathrm{MIP}}(x,y) = 0\}$ represents these voxels. So the corrected background voxels set S_0 is computed as $S_0 = T_{\mathrm{b1}} \cup T_{\mathrm{b2}} / T_{\mathrm{b3}}$.

The parameters β_{th}, γ_{th} and η_{th} are related to the average gray value v_{ave} of the known foreground, computed as $v_{\mathrm{ave}} = \frac{1}{|S_{\mathrm{p}}|}\sum_{p\in S_{\mathrm{p}}} X(p)$. In our experiment, $\beta_{\mathrm{th}} = 0.2v_{\mathrm{ave}}$, $\gamma_{\mathrm{th}} = 1.2v_{\mathrm{ave}}$, $\eta_{\mathrm{th}} = 1.6v_{\mathrm{ave}}$.

B. Weakly-Supervised Network

To effectively leverage the information provided by the weak annotation, a 2D-3D deep feature fusion network is designed based on U-Net [36] and 3D U-Net [16], denoted as $g_{2\mathrm{D}}$ and $g_{3\mathrm{D}}$ respectively, as shown in Fig. 2. In fact, most fully supervised methods can also serve as backbone for our scheme. The 3D image X and the corresponding 2D projection image X_{MIP} are fed into two networks to obtain the prediction probability maps, $\hat{Y}_{3\mathrm{D}} = g_{3\mathrm{D}}(X)$, $\hat{Y}_{2\mathrm{D}} = g_{2\mathrm{D}}(X_{\mathrm{MIP}})$.

The use of MIP image effectively captures the spatial information, geometric attribute, and interconnectivity of 3D blood vessels. Moreover, the availability of ground truth supervision for MIP image enhances the reliability of $g_{2\mathrm{D}}$ prediction, so it is important to flow the information from $g_{2\mathrm{D}}$ to $g_{3\mathrm{D}}$ during training. Within our proposed network framework, we establish a linkage between the feature map of two networks based on the inherent relationship exhibited by their respective inputs. This facilitates utilization of the segmentation information generated by the 2D network within the 3D network. Specifically, the features extracted from two networks are connected using the index map X_{index} obtained during MIP. For the extracted 3D feature of the i-th layer $f_i^{3\mathrm{D}} \in R^{C\times\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}}$, the corresponding 2D feature $f_i^{3\mathrm{D}\mapsto 2\mathrm{D}} \in R^{C\times\frac{H}{2^i}\times\frac{W}{2^i}}$ is calculated by the feature retrieval module R_i as follows:

$$\begin{aligned} f_i^{3\mathrm{D}\mapsto 2\mathrm{D}}(c,x,y) &= R_i(f_i^{3\mathrm{D}}) \\ &= f_i^{3\mathrm{D}}(c,x,y,X_{\mathrm{index}}^i(x,y)), \\ & i \in \{0,1,2,3\}, \end{aligned} \tag{2}$$

where $X_{\mathrm{index}}^i \in R^{\frac{H}{2^i}\times\frac{W}{2^i}}$ is obtained by interpolating the index map $\frac{X_{\mathrm{index}}}{2^i}$. And the feature $f_i^{3\mathrm{D}\mapsto 2\mathrm{D}}$ is concatenated with the corresponding feature layer of the 2D U-Net.

C. Confident Learning and Uncertainty Estimation

Two challenges arise in the pseudo-labels (S_0, S_1) generated in Sec. III-A: (i) the pseudo-label generation process, which relies on seed points and grayscale information, unavoidably introduces noise into the labels; (ii) the pseudo-labels offer coverage only for a subset of voxels, whereas many voxels remain unlabeled ($|S_0 \cup S_1| < |\Omega|$). To tackle these issues, we propose the incorporation of confidence learning (CL) and uncertainty estimation (UE) to further refine the pseudo-labels.

1) Noisy Labeled Voxels Refinement with CL: To identify and address the presence of noisy labels within the pre-labeled voxels ($\Omega_{\mathrm{L}} = S_0 \cup S_1$), the true (latent) foreground and background sets are estimated using the network output ($\hat{Y}_{3\mathrm{D}}$):

$$S_i^* = \{p; p \in \Omega_{\mathrm{L}}, i = \arg\max_j \hat{y}_{3\mathrm{D}}^p(j), \hat{y}_{3\mathrm{D}}^p(i) > t_i\}, \tag{3}$$

where t_i is average self-confidence of the labeled set S_i, that is $t_i = \frac{1}{|S_i|}\sum_{q\in S_i} \hat{y}_{3\mathrm{D}}^q(i)$. And $\hat{y}_{3\mathrm{D}}^p(i)$ is the predicted

probability ($\hat{Y}_{3D}$) belonging to the i-th category at point p. And then we calculate the normalized count matrix $\tilde{C}_{S,S^*}$ as:

$$\tilde{C}_{S,S^*}[i][j] = \frac{|S_i \cap S_j^*|}{\sum_{j\in\{0,1\}} |S_i \cap S_j^*|} \cdot |S_i|, \tag{4}$$

where the reason for normalization is $|S_0^* \cup S_1^*| \leq |\Omega_L|$ affected by the threshold t_i. Subsequently, we estimate the joint distribution based on $\tilde{C}_{S,S^*}$:

$$\hat{Q}_{S,S^*}[i][j] = \frac{\tilde{C}_{S,S^*}[i][j]}{\sum_{i\in\{0,1\}} \sum_{j\in\{0,1\}} \tilde{C}_{S,S^*}[i][j]}. \tag{5}$$

The mislabeled voxels set (the set to remove) is selected by $\hat{Q}_{S,S^*}[i][j]$ with the Prune by Noise Rate (PBNR) strategy [23], expressed as:

$$\begin{aligned} S_i^{(\mathrm{re})} =& \{p;\, y_{3D}^p(1-i) - y_{3D}^p(i) > i_{\mathrm{th}}, p \in S_i \cap S_{1-i}^*\} \\ & \cup (S_0 \cap S_1), \end{aligned} \tag{6}$$

where i_{th} is the minimum value of the top $|\Omega_L| \cdot \hat{Q}_{S,S^*}[i][1-i]$ in the set $\{y_{3D}^q(1-i) - y_{3D}^q(i); q \in S_i \cap S_{1-i}^*\}$. And, the voxels that exist simultaneously in S_0 and S_1 are also removed. We select some of wrongly labeled background voxels to add to the foreground set based on prior knowledge, and vice versa:

$$S_i^{(\mathrm{add1})} = \{p;\ p \in S_{1-i}^{(\mathrm{re})},\ \text{and prior condition}\}, \tag{7}$$

where the *prior condition* is $X(p) < \varepsilon_1 v_{\mathrm{ave}}$ when $i = 0$, while $D(p, S_1) < d_{\mathrm{th1}}$ when $i = 1$. And $D(p, S_1)$ is the minimum distance from point p to set S_1.

2) Unlabeled Voxels Refinement with UE: To address the issue of unlabeled voxels ($\Omega_U = \Omega / \Omega_L$), we adopt uncertainty estimation method to assign labels to reliable voxels. This process begins with the measurement of uncertainty for each voxel, utilizing the Monte Carlo dropout method. For each training data X, we execute multiple forward passes (K times, $K = 6$ in our experiments) using g_{3D} with dropout to obtain prediction probabilities $\{\hat{Y}_k\}_{k=1}^K$:

$$\hat{Y}_k = g_{3D}(X + \mathcal{N}_k(\mu, \sigma^2)), \tag{8}$$

where $\mathcal{N}_k(\mu, \sigma^2)$ is a stochastic Gaussian distribution with mean μ and variance σ^2, with the dimensions matching those of X ($\mu = 0$, $\sigma = 0.1$ in our experiments). Meanwhile, we compute prediction probability $\hat{Y} = g_{3D}(X)$ without dropout and the probability result with dropout $\hat{Y}_{\mathrm{dp}} = \frac{1}{K}\sum_k \hat{Y}_k$. Subsequently, the uncertainty of each voxel is computed as:

$$u_p = -\sum_{i\in\{0,1\}} \left(\frac{1}{K}\sum_k \hat{y}_k^p(i)\right) \cdot \log_2\left(\frac{1}{K}\sum_k \hat{y}_k^p(i)\right). \tag{9}$$

Finally, based on the uncertainty, we determine the additional set of foreground and background points:

$$u_{\mathrm{ave}}^i = \frac{1}{\sum_{p\in\Omega_U} [y^p = y_{\mathrm{dp}}^p = i]} \cdot \sum_{p\in\Omega_U} [y^p = y_{\mathrm{dp}}^p = i] \cdot u_p, \tag{10}$$

$$\begin{aligned} S_i^{(\mathrm{add2})} = \{&p; p \in \Omega_U, y^p = y_{\mathrm{dp}}^p = i,\ u_p < u_{\mathrm{ave}}^i, \\ &\text{and prior condition}\}, \end{aligned} \tag{11}$$

where $y^p = \arg\max_j \hat{y}^p(j)$, $y_{\mathrm{dp}}^p = \arg\max_j \hat{y}_{\mathrm{dp}}^p(j)$, $[\cdot]$ is indicator function. And the *prior condition* is $X(p) < \varepsilon_2 v_{\mathrm{ave}}$ when $i = 0$, while $D(p, S_1) < d_{\mathrm{th2}}$ when $i = 1$.

The foreground and background sets after refinement are calculated as:

$$\tilde{S}_i = (S_i \cup S_i^{(\mathrm{add1})} / S_i^{(\mathrm{re})}) \cup S_i^{(\mathrm{add2})}. \tag{12}$$

D. Loss Function

As shown in Fig. 2, the loss function consists of two components, L_{2D} and L_{3D}. The output of 2D U-Net is under supervision via MIP annotation, whereas the output of the 3D network is under the guidance of 3D pseudo-labels. As the 3D pseudo-labels do not cover all the voxels, we employ a weighted cross-entropy loss expressed as:

$$\begin{aligned} L_{3D} &= L_f(3D) + L_p(3D) + L_b(3D) \\ &= -\frac{1}{|S_f|}\sum_{p\in S_f} \log(\hat{y}_{3D}^p(1)) - \frac{1}{|S_p|}\sum_{p\in S_p} \log(\hat{y}_{3D}^p(1)) \\ &\quad - \frac{1}{|S_b|}\sum_{p\in S_b} \log(\hat{y}_{3D}^p(0)), \end{aligned} \tag{13}$$

where $S_f = S_1$ and $S_b = S_0$ in the initial training phase, while $S_f = \tilde{S}_1$ and $S_b = \tilde{S}_0$ during fine-tuning the network. As Y_{MIP} serves as the ground truth for X_{MIP}, this component is supervised by the Dice loss, a commonly used loss function for segmentation task:

$$\begin{aligned} L_{2D} &= L_{3D\text{-}\mathrm{MIP}}(2D) + L_{\mathrm{MIP}}(2D) \\ &= \mathrm{Dice}(R_0(\hat{Y}_{3D}), Y_{\mathrm{MIP}}) + \mathrm{Dice}(\hat{Y}_{2D}, Y_{\mathrm{MIP}}), \end{aligned} \tag{14}$$

where $R_0(\cdot)$ is defined in the Eq. 2. The final loss L_{all} is expressed as:

$$L_{\mathrm{all}} = L_{3D} + \lambda L_{2D}. \tag{15}$$

IV. Experiments

A. Datasets and Preprocess

We evaluate our method on five datasets, including three cerebrovascular datasets, a coronary CTA dataset, and an aortic CTA dataset.

1) TubeTK: The publicly available dataset TubeTK[1] comprises 42 3D time-of-flight MRA volumes with labeled vessels (centerline + radius). To facilitate further analysis, we convert the annotation into voxel data using the MetaIO[2].

2) Cerebral MRA: This dataset consists of 96 MRA cerebrovascular volumes acquired from various imaging systems. The images are retrospectively collected from Xuanwu Hospital of Capital Medical University, China. Each sample has 3D vessel annotation. During annotating, we employ Frangi filtering [37] to generate the initial segmentation result for the vessels. Then, fine corrections are made by two radiologists to obtain the final vascular label.

[1] https://public.kitware.com/Wiki/TubeTK/Data
[2] https://itk.org/Wiki/MetaIO

3) Cerebral CTA: It comprises 47 3D CTA cerebrovascular volumes with blood vessel annotations. The source and labeling process of this dataset are consistent with those of the Cerebral MRA dataset. However, for CTA data, the skull region is highlighted, affecting the segmentation of blood vessels and causing obstruction during MIP. To mitigate this issue, the CTA data has been performed skull-stripping to remove the bright skull regions [38].

4) Coronary CTA: The coronary dataset contains 52 3D CTA volumes, which are retrospectively collected from Wuhan Union Hospital of China. And the annotation of coronary arteries is completed by a radiologist. The grayscale value of the ascending aorta, left atrium, and other some parts is found to be higher than that of the coronary artery, resulting in occlusion during MIP. To address this, we apply a region growing method to remove the ascending aorta, and subsequently employ a combination of the threshold method and region growing method for each slice to remove the remaining high-intensity areas, as shown in Fig. 3. Notably, this step is solely performed during the MIP process of training samples, and the original image served as the input during training. So no processing steps is needed during inference.

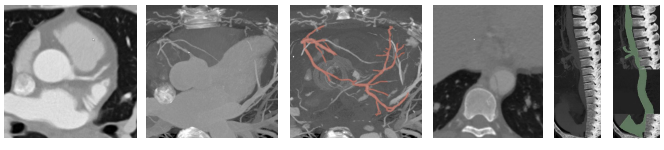


Fig. 3. Preprocessing of Coronary CTA and Aorta CTA datasets: (a) and (d) depict a slice of the coronary and aorta volume, respectively. (b) and (e) show the MIP images obtained through direct projection, revealing that a significant portion of vessels is obscured by other anatomical structures. The MIP images of processed volumes are displayed in (c) and (f), with a clear display of the majority of the blood vessel.

5) Aorta CTA: This dataset comprises 50 aorta volumes from Wuhan Union Hospital of China. Annotation of the aorta is done by a radiologist. In contrast to other datasets mentioned above, the unique structure of aorta limits the utility of MIP image projected in transverse plane. Therefore, we perform MIP in the sagittal plane on aorta dataset. Similar to Coronary CTA dataset, the aorta is also affected by other brighter parts during projection, primarily concentrated in the middle layers. And we employ a traditional algorithm based on shape prior [39] to process the middle part of layers and remove possible occluded regions (in experiments, we process the middle 230-380 layers, as shown in Fig. 3). This step is, again, exclusive to the MIP process.

TABLE I
THE RESOLUTION (SPACING) AND SIZE (CROPPING) OF ADJUSTED DATA. THE NUMBER OF TRAIN/VALIDATION/TEST SETS (NUMBER) AND WHETHER SOME PARTS WERE REMOVED (REMOVING).

Dataset	Spacing(mm^3)	Cropping	Number	Removing
TubeTK	0.5×0.5×0.8	384× 384×128	30/4/8	✘
Cerebral MRA	0.5×0.5×0.75	320× 320×128	30/4/62	✘
Cerebral CTA	0.5×0.5×0.75	320× 320×128	30/5/12	✓
Coronary CTA	0.4×0.4×0.4	320× 320×256	30/4/18	✓
Aorta CTA	1.0×1.0×1.0	128× 160×480	30/4/16	✓

The resolution of the data in each dataset is standardized and subsequently the volumes are cropped to ensure consistent size, leaving the middle vessel area. In cases where the image size is insufficient, zero padding is employed to achieve uniform data size. Additionally, a gray value normalization step is applied, mapping the intensity range to 0-1. The images of each dataset are randomly partitioned into training, validation, and test sets. Table I provides an overview of the adjusted data, encompassing information on resolution, data size, the specific allocation of images, as well as any pre-processing steps executed to eliminate potential occlusions within the data that might impede blood vessel visibility during MIP.

B. Metrics and Implementation Details

1) Metrics: We utilize the following metrics to evaluate our method: Dice Similarity Coefficient (DSC), ClDice [40], which is tailored to evaluate tubular structures while accounting for vascular connectivity, and Average Hausdorff Distance (AHD) [41], which incorporates voxel localization considerations [42]. Furthermore, to indicate the statistical significance of improvements of the proposed method, we also present the p-values for DSC using a paired t-test with each comparison method.

2) Implementation Details: In our proposed weakly supervised network, we employ 3D U-net [16] and 2D U-net [36] architectures as the backbone. The down-sampling path of two models features convolution layers with filter numbers of [8, 16, 32, 64]. The implementation of the network is conducted using the PyTorch framework. Training process is performed on a NVIDIA GeForce GTX 3090 GPU with 24G memory. During the training process, we utilize the adaptive moment estimation (Adam) optimizer, initialized with a learning rate of 0.001. A decay factor of 0.9 is applied to the learning rate after each iteration. The maximum number of training iterations is set to 1000. We employ the preset parameter $\alpha = 0.1$ in pseudo-labels generation by region growing. And the prior parameters d_{th1}, d_{th2}, ε_1 and ε_2 are respectively set to 1.5, 4.0, 0.7 and 0.2 in Eq. 7 and Eq. 11. Importantly, it should be noted that when working with the public dataset (TubeTK dataset), the priori information is not set during pseudo-labels refinement. In other words, the parameters d_{th1}, d_{th2}, ε_1 and ε_2 are considered ∞. This adjustment is made due to the noisy labels in TubeTK dataset, where the annotated vessels are thin and incorporate some venous structures [43]. Consequently, during training with pseudo-labels, the TubeTK dataset has a higher tolerance for noise in the labels. Additionally, the balance parameter λ in the loss function is set to 1.0. We will shortly make our code publicly available.

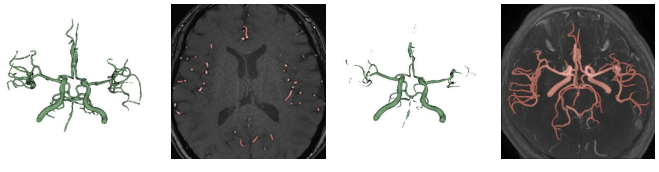


Fig. 4. Annotations of different categories.

C. Compared Methods and Annotation Time

1) Compared Methods: As shown in Table I, our proposed method utilize a training set size of $N = 30$ for all five datasets, under supervision via the annotation of 2D MIP image (as shown in Fig. 4(d)). To illustrate the effectiveness of our proposed method, we implement the following algorithms for comparison.

- **Full-sup:** We report the results of fully supervised model training with N images as the upper bound. The model structure of Full-sup is consistent with our proposed method.
- **Baseline3D [16]:** We use m_1 ($m_1 < N$) fully labeled samples (Fig. 4(a)) to train a fully supervised model, whose structure is consistent with 3D Unet used in our proposed method.
- **Baseline2D:** For each sample, $s_1\%$ layers are randomly chosen and labeled vessel in 2D slice (Fig. 4(b)). The N volumes with these annotations are then used to train the 3D Unet.
- **MTCL [27]:** This method represents a semi-supervised approach for blood vessel segmentation that relies on a small number of fully labeled data and a larger amount of noisy labeled data. Specifically, we apply the Frangi vessel enhancement method [37] and get a noisy label (Fig. 4(c)) by only manually removing obvious noise due to the arduous task of adding blood vessels. And we employ m_2 fully labeled data and $N - m_2$ noisy labeled data for MTCL.
- **SLD [32]:** To ensure fair comparison, we employ the supervision of the MIP annotation as an alternative to the supervision of the adversarial learning component within SLD. And the number of training images is consistent with our approach (N).
- **SPDS [33]:** Similar to SLD [32], to ensure methodological consistency, we align the supervision labels employed in SPDS with our approach, and the 3D Unet structure is applied as the backbone in SPDS.

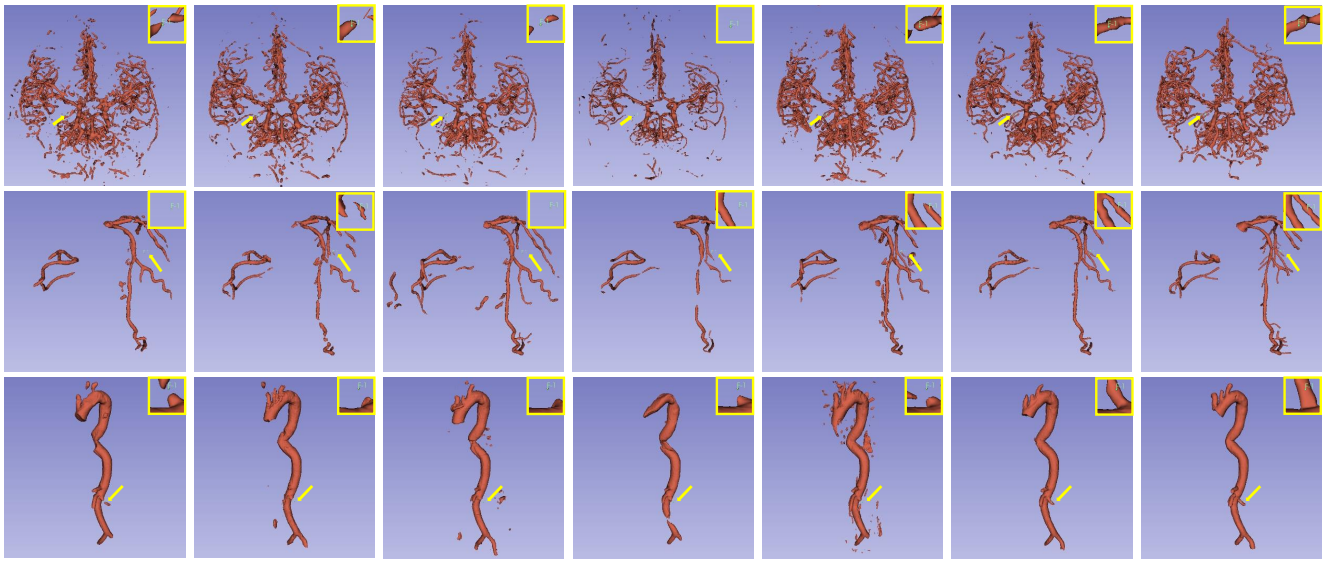
2) Annotation Time of Each Method: For a fair comparison, we endeavor to maintain consistency between the annotation time of the training data employed by the compared methods and the annotation time of N MIP images used in our proposed method. We present the average time of labeling data and the total annotation time (taken by a radiologist to manually annotate the images) of training samples for each method on Cerebral MRA, as shown in Table II.

TABLE II
THE TYPE OF ANNOTATED DATA (ANNOTATION), THE NUMBER OF ANNOTATIONS (NUMBER), THE AVERAGE ANNOTATION TIME (AVE), AND THE TOTAL ANNOTATION TIME OF THE TRAINING SAMPLES (ALL).

Method	Ave (min)	All (min)	Number	Annotation
Full-sup	73.41	2202.30	$N = 30$	Fig. 4(a)
Baseline3D [16]	73.41	220.23	$m_1 = 3$	Fig. 4(a)
Baseline2D	10.07	302.10	$s_1 = 10$	Fig. 4(b)
MTCL [27]	-	237.82	$m_2 = 3$	Fig. 4(a) and (c)
SLD [32]	6.78	203.40	$N = 30$	Fig. 4(d)
SPDS [33]	6.78	203.40	$N = 30$	Fig. 4(d)
Ours	6.78	203.40	$N = 30$	Fig. 4(d)

D. Comparative Results

1) Quantitative results: We present the quantitative results of the five datasets in Table III and Table IV. Under approximately consistent annotation time, our proposed method leverages the annotation of MIP image to provide the model with enhanced information regarding blood vessel direction and connectivity, thereby yielding superior outcomes. The weakly supervised labels proposed in our method offer effective information for blood vessel segmentation while significantly reducing the annotation workload. And the proposed approach effectively harnesses the information, resulting in compelling results that closely approach the performance achieved through



(a) Baseline3D (b) Baseline2D (c) MTCL (d) SPDS (e) SLD (f) Ours (g) Label

Fig. 5. Segmentation results on three testing images from three datasets (TubeTK, Coronary CTA, Aorta CTA dataset in order). The red boxes highlight close-ups of some vessels.

TABLE III

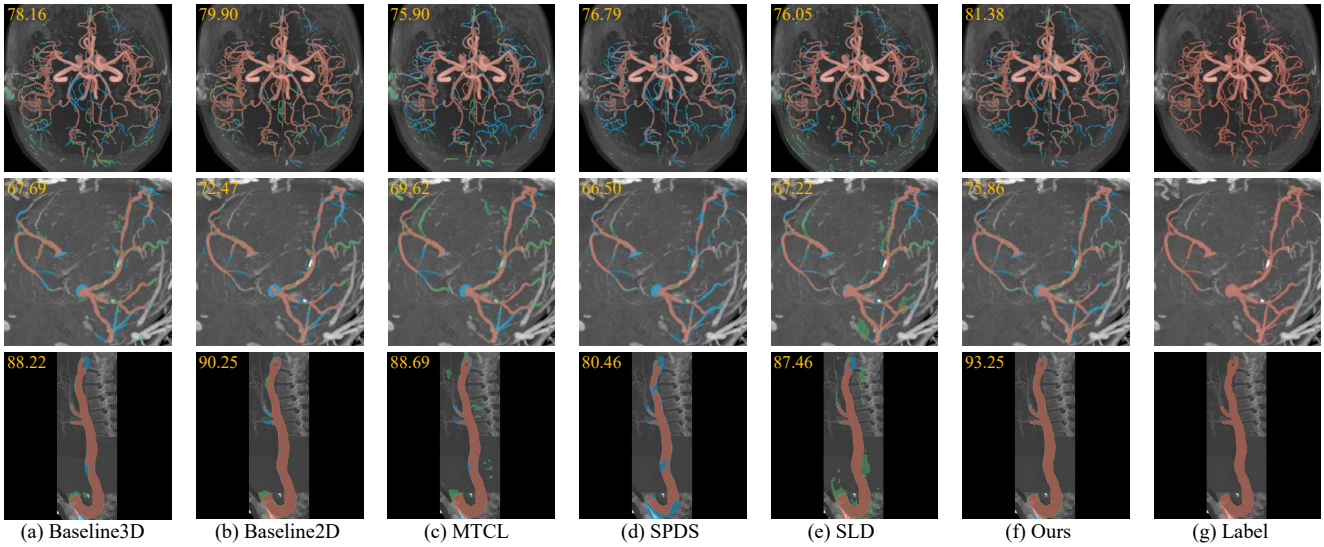
COMPARISON WITH OTHER METHODS ON THREE CEREBROVASCULAR DATASETS, WITH THE BEST PERFORMANCE HIGHLIGHTED IN BOLD. THE p OF $DSC(p)$ REPRESENTS THE P-VALUE CALCULATED BY THE T-TEST, AND $*$ INDICATES THE STATISTICAL DIFFERENCE BETWEEN OURS AND OTHER METHODS. ($* : p < 0.05$, $** : p < 0.01$, $*** : p < 0.001$)

Method	TubeTK			Cerebral MRA			Cerebral CTA		
	DSC(%) (p)	ClDice(%)	AHD(mm)	DSC(%)	ClDice(%)	AHD(mm)	DSC(%)	ClDice(%)	AHD(mm)
Full-sup	64.52 (***)	77.60	0.917	85.07 (***)	89.00	0.303	85.34 (***)	88.92	0.288
Baseline3D (2016) [16]	55.39 (***)	62.04	1.602	79.81 (***)	71.96	1.011	78.18 (***)	79.96	0.486
Baseline2D	59.09 (***)	74.11	1.029	82.28 (***)	85.40	0.436	81.19 (***)	81.80	0.406
MTCL (2022) [27]	60.95 (0.127)	74.89	1.017	76.01 (***)	73.04	1.502	72.34 (***)	66.40	1.487
SLD (2023) [32]	58.40 (***)	71.08	1.237	81.09 (***)	83.06	0.453	79.77 (***)	77.68	0.641
SPDS (2023) [33]	55.12 (***)	69.14	1.794	79.93 (***)	82.31	0.426	78.75 (***)	77.35	0.476
Ours	**61.10** (-)	**75.93**	**0.843**	**84.35** (-)	**87.40**	**0.336**	**83.84** (-)	**83.10**	**0.255**

TABLE IV

COMPARISON WITH OTHER METHODS ON CORONARY CTA AND AORTA CTA DATASETS, WITH THE BEST PERFORMANCE HIGHLIGHTED IN BOLD. THE p OF $DSC(p)$ IS CONSISTENT WITH TABLE III.

Method	Coronary CTA			Aorta CTA		
	DSC(%)	ClDice(%)	AHD(mm)	DSC(%)	ClDice(%)	AHD(mm)
Full-sup	77.40 (**)	75.88	0.745	91.75 (0.052)	89.20	0.339
Baseline3D (2016) [16]	67.42 (***)	67.23	1.706	87.85 (**)	86.21	0.698
Baseline2D	73.59 (**)	**73.95**	0.686	90.47 (*)	86.69	0.367
MTCL (2022) [27]	67.77 (**)	64.36	2.093	87.05 (**)	82.18	1.185
SLD (2023) [32]	70.45 (***)	68.17	1.792	86.81 (***)	83.50	0.919
SPDS (2023) [33]	69.46 (***)	68.93	1.053	83.38 (***)	79.26	1.185
Ours	**75.43** (-)	73.17	**0.660**	**90.84** (-)	**91.25**	**0.348**



(a) Baseline3D (b) Baseline2D (c) MTCL (d) SPDS (e) SLD (f) Ours (g) Label

Fig. 6. The MIPs of images and segmentation results. The first to third rows represent the MIP images corresponding to the segmentation results of one testing sample from the Cerebral MRA, Coronary CTA, and Aorta CTA datasets, respectively. The red pixels, blue pixels and green pixels denote true positives, false negatives and false positives respectively. The Dice score (%) of corresponding 3D segmentation result is shown in the upper left corner of MIP image.

full supervision. MTCL [27] achieves performance comparable to our proposed method on the TubeTK dataset, potentially due to the resemblance between the noisy labels employed in MTCL and the provided annotations, which contain noise itself [43]. Notably, the methods SLD [32] and SPDS [33] employ the same weak labels as our method to promote fair comparison; however, their utilization of these labels is severely restricted, leading to suboptimal performance. Utilizing identical labels, our proposed framework integrates the projection label acquisition method with the design of the 2D-3D feature fusion network, while also optimizing the resultant pseudo labels, resulting in superior performance.

2) Qualitative results: Fig. 5 exhibits the 3D results of testing data from different datasets. Our proposed method yields better connectivity and more accurate boundary detection of small vessels when compared to other algorithms. These findings are consistent with the results in Table III and Table IV. Additionally, the outcome of Baseline2D on Aorta CTA dataset is comparable to that of our proposed approach on metrics of DSC and AHD. This can be attributed to the relatively simple structure of the aorta in comparison to coronary arteries and cerebral vessels, allowing 2D slices

TABLE V

THE RESULTS OF ABATION STUDY. WE DEMONSTRATE THE EFFECTIVENESS OF EACH COMPONENT. THE p OF $\boldsymbol{DSC(p)}$ IS CONSISTENT WITH TABLE III.

Method	TubeTK		Cerebral MRA		Cerebral CTA		Coronary CTA		Aorta CTA	
	DSC(%)	AHD(mm)	DSC(%)	AHD(mm)	DSC(%)	AHD(mm)	DSC(%)	AHD(mm)	DSC(%)	AHD(mm)
L_{2D}	55.39 (∗∗∗)	1.602	72.73 (∗∗∗)	1.680	44.35 (∗∗∗)	5.471	72.24 (∗∗∗)	1.044	88.71 (∗∗∗)	0.650
L_{3D}	58.01 (∗∗∗)	0.870	81.66 (∗∗∗)	0.601	80.52 (∗∗∗)	0.349	73.63 (∗∗)	0.783	87.34 (∗∗∗)	0.615
L_{all}	60.08 (∗∗)	0.971	83.74 (∗∗∗)	0.375	83.53 (∗)	0.289	74.17 (∗)	0.699	89.90 (∗)	0.460
L_{all}+RF	**61.10** (-)	**0.843**	**84.35** (-)	**0.336**	**83.84** (-)	**0.255**	**75.43** (-)	**0.660**	**90.84** (-)	**0.348**

to provide sufficient and effective information for segmentation. However, the performance of Baseline2D in vascular connectivity is even poorer (shown in the final row of Fig. 5(b) and (e)) due to the difficulty of focusing on the 3D structural features from slice annotation. Fig. 6 shows the MIP images of testing volumes and their corresponding segmentation results from three datasets. The efficacy of our method can also be seen from the distribution of false negatives and false positives.

E. Ablation Study

1) The Effectiveness of Each Component: To verify the efficacy of each component in the proposed framework, we conduct experiments deploying solely 2D features or 3D features, recorded as L_{2D} and L_{3D}, respectively. Additionally, we present the outcomes without the pseudo-label refinement module (L_{all}). The results (as shown in Table V) demonstrate that relying solely on either 2D or 3D features leads to a degradation in performance, especially when only 2D features are utilized. Our proposed method, which effectively integrates two types of features, achieves better performance. Furthermore, comparison with L_{all}, the framework with enhancing the accuracy of pseudo-labels by confidence learning and uncertainty estimation (L_{all}+RF) achieves superior outcomes across all five datasets.

2) Confident Learning and Uncertainty Estimation: We present an analysis of the performance achieved by incorporating confident learning for the refinement of existing noisy labels (CL), as well as incorporating uncertainty estimation to refine unlabeled voxels (UE), on two datasets. We conduct an assessment of the quality of generated pseudo labels and the segmentation performance using two optimization strategies in comparison to absence of pseudo-label refinement. A higher quantity (Num) and accuracy (Acc) of the generated foreground and background labels relative to the real labels is indicative of higher pseudo-label quality. The results, presented in Table VI, indicate that both the CL and UE methods effectively enhance label quality. The UE method primarily concentrates on refining unlabeled voxels, leading to a significant impact on the number of voxels in pseudo-label (Num). Furthermore, comparison of the results from the two datasets reveals that the addition of background voxels mainly occurs in the TubeTK dataset, while foreground voxels are primarily added in the Coronary CTA dataset. This disparity may stem from the differences in annotation quality between the two datasets, with the TubeTK dataset exhibiting inherent noise (some venous) in its annotations, resulting in lower prediction uncertainty for the background. Moreover, we identify a positive correlation between the quality of generated pseudo labels and the final segmentation results, in line with our expectations. And comparative analysis against pseudo-labels generated solely through traditional methods (No) reveals an improvement in segmentation performance of network upon both these two strategies. And the combined utilization of confident learning and uncertainty estimation yields best outcomes in the network's performance.

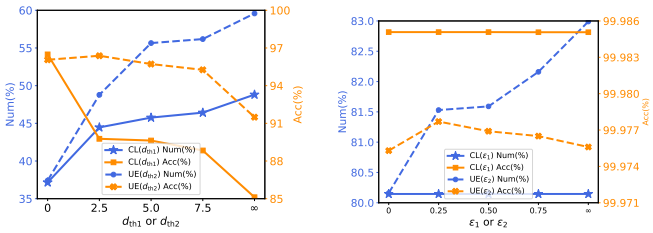
3) The Parameter of Pseudo-Label Refinement: The control over foreground generation primarily rests with parameters d_{th1} and d_{th2}, while background generation is primarily regulated by ε_1 and ε_2 (Eq. 7 and Eq. 11). To explore the impact of parameter variations on the generation of foreground and background, we conduct experiments on the Coronary CTA dataset, as depicted in Fig. 7. The *Num* metric indicates the proportion of the number of generated voxels to the actual number, while *Acc* represents the accuracy rate. Within the confidence learning module, increasing the value of d_{th1} results in the inclusion of more foreground voxels from $S_0^{(re)}$. However, this increment is accompanied by a decline in Acc. Similarly, within the uncertainty estimation module, increasing d_{th2} leads to an increased allocation of foreground voxels to unlabeled voxels and a decrease in accuracy. Regarding

TABLE VI

THE EFFECTIVENESS OF CONFIDENCE LEARNING AND UNCERTAINTY ESTIMATION. THE p OF $\boldsymbol{DSC(p)}$ IS CONSISTENT WITH TABLE III.

Dataset	Method	Foreground		Background		Segmentation Result		
		Num(%)	Acc(%)	Num(%)	Acc(%)	DSC(%)	ClDice(%)	AHD(mm)
TubeTK	No	38.31	82.47	92.94	99.86	60.08 (∗∗)	75.25	0.971
	CL	39.19	82.59	92.92	99.87	60.18 (∗∗∗)	74.99	0.866
	UE	40.42	82.35	96.17	99.85	60.64 (∗∗∗)	75.30	0.858
	CL+UE	41.30	82.49	96.15	99.86	61.10 (-)	75.93	0.843
Coronary CTA	No	37.14	96.05	80.16	99.97	74.17 (∗∗)	71.10	0.699
	CL	37.52	96.39	80.14	99.98	75.30 (0.115)	72.00	0.666
	UE	50.93	95.77	80.50	99.97	74.56 (∗∗∗)	71.62	0.663
	CL+UE	51.04	96.03	80.48	99.98	75.43 (-)	73.17	0.660

background voxels, the variation of ε_1 has minimal observable impact on Acc and Num of generated background due to $|S_1^{(\mathrm{re})}| \ll |S_0|$. And similar to d_{th2}, as ε_2 increases, the number of allocated background voxels increases while the accuracy declines. The observed variations align with our initial expectations. The numerical selection of d_{th1} and d_{th2} aims to strike a balance between the quantity and accuracy of the generated foreground set. And these adjustments additionally impact the subsequent fine-tuning process of the segmentation network. Similar considerations apply to ε_1 and ε_2. During the experimental analysis, we maintain consistent parameter settings across the four datasets, with the exception of TubeTK dataset. This consistent performance demonstrates the robust generalization capabilities of our proposed method.



(a) Foreground (b) Background

Fig. 7. (a) shows the impact of d_{th1} and d_{th2} on the generated foreground, and (b) is the impact of ε_1 and ε_2 on the background.

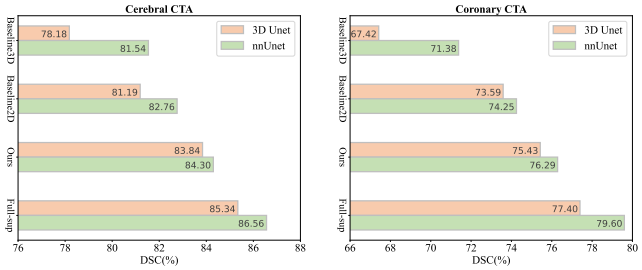
F. Generalization of Method

TABLE VII
CROSS-VALIDATION ON TUBETK AND CEREBRAL CTA DATASETS. THE p OF $DSC(p)$ IS CONSISTENT WITH TABLE III.

Method	TubeTK		Cerebral CTA	
	DSC(%)	AHD(mm)	DSC(%)	AHD(mm)
Full-sup	63.90 (∗ ∗ ∗)	0.998	84.96 (∗ ∗ ∗)	0.387
Baseline3D	56.85 (∗ ∗ ∗)	1.387	78.72 (∗ ∗ ∗)	0.543
Baseline2D	59.40 (∗ ∗ ∗)	1.035	81.86 (∗ ∗ ∗)	0.488
MTCL	60.05 (∗)	1.008	74.75 (∗ ∗ ∗)	1.347
SLD	58.18 (∗ ∗ ∗)	1.307	81.49 (∗ ∗ ∗)	0.645
SPDS	52.53 (∗ ∗ ∗)	1.602	78.15 (∗ ∗ ∗)	0.459
Ours	**60.38** (-)	**0.936**	**84.96** (-)	**0.297**

1) Cross-Validation Experiments: To further demonstrate the generalization of our method, we conduct cross-validation experiments on two relatively small datasets, as presented in Table VII. By comparing the results, we can draw similar conclusions to those observed from Table III and Table IV: our method effectively utilizes the carefully designed weak labels and achieves superior performance compared to other methods within similar annotation timeframes. Moreover, in the cross-validation experiments, the entire dataset is used for testing, resulting in more robust and stable outcomes. Consequently, compared to the results in Table III, our method exhibits higher statistical significance when evaluated against MTCL on the TubeTK dataset. This observation further supports that our approach can deliver superior results compared to MTCL on datasets with noisy labels.

2) Robustness: In this subsection, we primarily focus on the robustness of our proposed weakly-supervised segmentation framework. Our scheme can leverage most fully supervised methods as backbones, allowing us to analyze their impact on the segmentation performance. Due to the superior performance, nnUnet [44] is widely used in various medical image segmentation tasks. Fig. 8 illustrates the results achieved by various methods using 3D Unet and nnUnet as backbones, respectively. Notably, we utilize initially generated pseudo labels (S_0, S_1) to obtain the 'data fingerprint' and 'pipeline fingerprint' (design parameters of nnUnet) of the backbone for proposed framework when employing nnUnet. Based on the findings presented in Fig. 8, it is evident that the methods exhibit similar characteristics under both backbones. And with the same backbone (3D Unet or nnUnet), our approach effectively harnesses the labels of MIP images, yielding superior results compared to Baseline2D and Baseline3D, while approaching the performance of fully supervised methods. Furthermore, a horizontal comparison between the two backbones reveals that nnUnet outperforms 3D Unet, aligning with expectations due to nnUnet's ability to fully exploit the dataset characteristics.



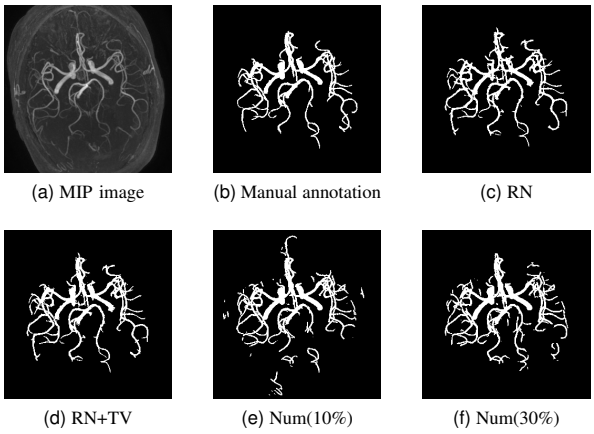
(a) Cerebral CTA (b) Coronary CTA

Fig. 8. DSC metric results under different backbones on two datasets.

G. Trade-off Between Annotation Time and Performance

1) Further Optimization of Annotation Method: In our method, the annotation of 2D projected blood vessels is an essential aspect. Despite being simpler than annotating in 3D space, it necessitates manual intervention by radiologists, with an average annotation time per image of 6-7 minutes (shown in Table II). To segment 2D vascular images, conventional methods or deep learning methods can be applied for automatic or semi-automatic segmentation. Although these approaches may impact annotation accuracy, it offers significant reduction of annotation workload and potential for unsupervised and semi-supervised segmentation. In this section, we investigate the impact of reducing annotation time on results by employing both a conventional method and a learning-based approach on Cerebral MRA dataset (consistent with Table II). We combine Frangi filtering [37] and homomorphic filtering techniques for the segmentation of MIP images, followed by the manual removing of obvious noise (RN) and labeling of thick vessels (TV) to acquire MIP annotations of varying qualities. These results are then integrated into our framework as 3D weak labels to derive the final 3D segmentation result. Furthermore,

varying numbers (10%, 30%) of MIP images from training set are annotated, allowing us to train a 2D Unet [36] for obtaining segmentation results of the remaining MIP images. Subsequently, the manually annotated labels and the segmentation results serve as weakly supervised labels for training our proposed network. Fig. 9 presents the annotated images obtained through different methods.

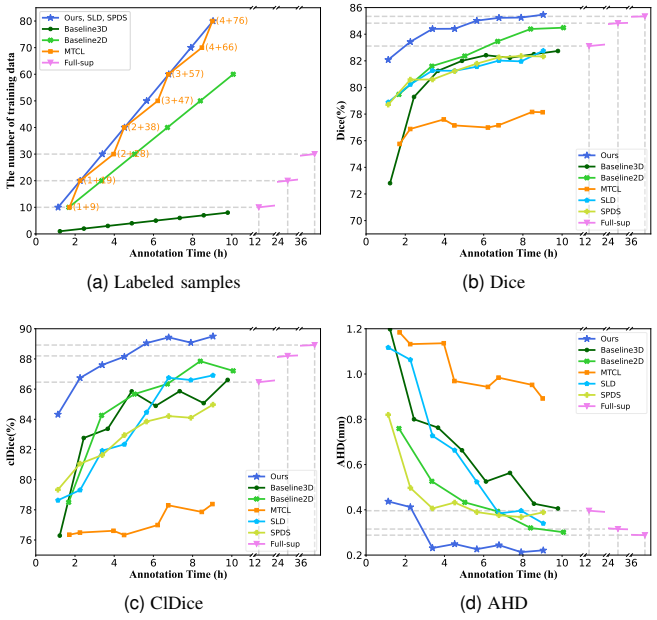


(a) MIP image (b) Manual annotation (c) RN

(d) RN+TV (e) Num(10%) (f) Num(30%)

Fig. 9. Different annotations of one MIP image, where (b) is the label with completely manual annotation. And (e) and (f) are the predicted results of 2D Unet trained with the annotated MIP images (Num(10%), Num(30%)).

Table VIII presents the average labeling time required for weakly supervised labels obtained through various methods, in addition to the MIP labeling quality (Dice coefficient and accuracy) and their implications on the final segmentation results. Initially, as expected, it is observed that modifying all or part of the manual annotations by incorporating algorithm-generated pseudo labels leads to a reduction in annotation time, resulting in a decrease in the quality of MIP annotation and a negative impact on segmentation results. Furthermore, the comparison of the results obtained through two different methods reveals that, when utilizing 2D Unet, the quality of MIP annotations and the performance of the final segmentation results outperform those of traditional method under similar annotation time. This can be attributed to the inclusion of manually obtained correct labels in the 2D Unet method, enabling the network to assimilate more valuable information during the learning process. Additionally, the comparison between the results of $Num(30\%)$ and all manual annotations ($Manually$) indicates that employing 2D Unet can significantly reduce annotation time with only a slight decrease in segmentation performance. This highlights the potential of our framework to integrate with other 2D segmentation methods for further reduction of the annotation workload without significant performance compromise, a topic to be further discussed in Sec. IV-H.



(a) Labeled samples (b) Dice

(c) ClDice (d) AHD

Fig. 10. (a) shows the number of training images and annotation time of each method, where the numbers of full annotation and noisy label are shown for MTCL. Notably, the lines of SLD and SPDS in (a) align with our method, as identical training labels are utilized to ensure fair comparison (mentioned in Sec. IV-C). (b)-(d) show the segmentation performance vs. annotation time. Notably, the annotation time of fully supervised method (Full-sup) trained with the fewest number of samples (12.2h, 10 training samples) is longer than that of the proposed method (Ours) trained with the greatest number of samples (9.0h, 80 training samples), which is why there is no overlap between the two methods on the axis of Annotation Time.

2) The Number of Weakly Annotations: The performance of weakly-supervised method theoretically does not exceed that of fully supervised learning under the same amount of training data. However, it is meaningful to study whether weakly-supervised method can outperform fully supervised method in less annotation time by adding weakly labeled training samples. This area of research has received limited attention in prior studies on weakly supervised learning. In our study, we randomly select 12 samples as testing set and 4 samples

TABLE VIII

THE IMPACT OF DIFFERENT ANNOTATION METHODS ON THE SEGMENTATION RESULTS. THE p OF $DSC(p)$ IS CONSISTENT WITH TABLE III. WHEN ASSESSING THE QUALITY OF MIP ANNOTATION IN 2D UNET METHOD, WE CONCURRENTLY EVALUATE THE MANUALLY ANNOTATED TRAINING DATA (DICE=100%, ACC=100%) AND THE GENERATED SEGMENTATION RESULTS, AS THEY COLLECTIVELY SERVE AS WEAK LABELS FOR OUR METHOD, FACILITATING EASIER COMPARISON WITH FRANGI+HOMOMORPHIC.

Mehod	Scheme	MIP Annotation				Segmentation Result		
		Ave(min)	Dice(%)	Acc(%)	Example	Dice(%)	ClDice(%)	AHD(mm)
Frangi+Homomorphic	RN	0.86	75.83	95.75	Fig. 9(c)	73.84 (∗∗∗)	64.48	1.282
	RN+TV	2.39	86.13	97.41	Fig. 9(d)	79.79 (∗∗∗)	77.56	0.830
2D Unet	Num(10%)	0.68	81.35	96.36	Fig. 9(e) and (b)	78.85 (∗∗∗)	79.51	1.191
	Num(30%)	2.26	87.73	97.66	Fig. 9(f) and (b)	82.42 (∗∗∗)	84.36	0.509
Mannually	-	6.78	100	100	Fig. 9(b)	84.35 (-)	87.40	0.336

as the validation set from the 96 volumes of Cerebral MRA dataset. In Fig. 10, we present the results of monitoring the performance trends of each method under varying labeling times (different numbers of training samples) on Cerebral MRA dataset, with the same testing volumes. Additionally, we evaluate the performance of full supervision with 10, 20, and 30 training samples which require labeling times of 12.2h, 24.5h, and 36.7h respectively.

Based on the results presented in Fig. 10, it is evident that the performance of each method shows a gradual increase with the expansion of training data, consistent with anticipated outcomes. And our method outperforms other methods across all indicators in the case where the annotation time of training data is similar. Furthermore, when the number of weakly-labeled samples is sufficient, our proposed method can outperform full supervision while still requiring far less labeling time. For example, our approach achieves better results than fully supervised segmentation (trained on 30 images) with only about 7.9h of data annotation time, which is significantly less than the 36.7h required for the latter. Similarly, in much less annotation time (about 2.3h, 5.7h), our method outperforms the performance under full supervision with the labeling time of 12.2h and 24.5h.

H. Limitation and Future Works

One limitation of our study pertains to the impact of blood vessel occlusion on the accurate labeling of the MIP image. Our proposed methodology necessitates the blood vessel annotation of the MIP image. But the presence of diverse types of blood vessels introduces varying degrees of occlusion challenges due to dissimilar surrounding tissues. Consequently, some preprocessing procedures may be required for certain datasets (e.g., the Aorta CTA dataset and the Coronary CTA dataset) in the training phase. Furthermore, occlusion is present in the images of the patients who undergo surgery and implant metal materials, which is not considered in our work. Hence, it is intriguing to investigate how to attain superior performance in the presence of occlusion, even when it is severe.

Additionally, our approach mandates the annotation of 2D MIP blood vessel images, which is still time-consuming. In Sec. IV-G.1, we attempt to replace manual full annotation of MIP images with automatic methods (with minor manual annotation); however, these methods yield unsatisfactory results, introducing noise that adversely affects the final segmentation outcome. Exploring the utilization of existing 2D blood vessel weakly-supervised and semi-supervised methods to minimize the annotation workload while upholding accurate blood vessel segmentation represents a promising avenue for future research.

V. Conclusion

In this study, we present a framework for weakly supervised segmentation of vessels in 3D volumes with dimensionality reduction annotation, leveraging the sparse structural characteristics of vascular structure. To this end, MIP image is employed as a means of annotation and supervision. Initially, we obtain pseudo-label of 3D vessels through MIP annotation. Subsequently, we design a 2D-3D feature fusion network to make best use of the weak label, taking into account the acquisition method of the 2D labels. During the pseudo-label generation, it is inevitable that some noise is introduced and certain voxels may be overlooked. To mitigate these issues and enhance network performance, we integrate confidence learning and uncertainty estimation methods to refine the pseudo-labels. We conduct comprehensive experiments across five vascular datasets. And the results demonstrate that our proposed method achieves high-quality vascular segmentation, approaching the performance of fully-supervised segmentation under the same number of training samples. Furthermore, we design experiments to validate that our proposed weakly supervised segmentation framework achieves superior performance to fully supervised segmentation with much less annotation time by increasing training samples, showing the immense potential of our method in the field of vessel segmentation.

References

[1] F. Fu *et al.*, "Rapid vessel segmentation and reconstruction of head and neck angiograms using 3D convolutional neural network," *Nature Commun.*, vol. 11, no. 1, pp. 1–12, Sep. 2020.

[2] B. Zhang *et al.*, "Cerebrovascular segmentation from TOF-MRA using model-and data-driven method via sparse labels," *Neurocomputing*, vol. 380, pp. 162–179, Mar. 2020.

[3] G. Tetteh *et al.*, "Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-D angiographic volumes," *Frontiers Neurosci.*, vol. 14, pp. 1285, Dec. 2020.

[4] L. Yao *et al.*, "Tag-net: Topology-aware graph network for centerline-based vessel labeling," *IEEE Trans. Med. Imag.*, early access, Jan. 2023, doi: 10.1109/TMI.2023.3240825.

[5] Y. Wang *et al.*, "Deep distance transform for tubular structure segmentation in CT scans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3833–3842.

[6] W. Jiang *et al.*, "Ori-net: Orientation-guided neural network for automated coronary arteries segmentation," *Expert Syst. Appl.*, vol. 238, pp. 121905, 2024.

[7] Y. Tan, K. Yang, S. Zhao, and Y. Li, "Retinal vessel segmentation with skeletal prior and contrastive loss," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2238–2251, Sep. 2022.

[8] Y. Li, Y. Yu, Y. Zou, T. Xiang, and X. Li, "Online easy example mining for weakly-supervised gland segmentation from histology images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2022, pp. 578–587.

[9] H. Du, Q. Dong, Y. Xu, and J. Liao, "Weakly-supervised 3D medical image segmentation using geometric prior and contrastive similarity," *IEEE Trans. Med. Imag.*, early access, Apr. 2023, doi: 10.1109/TMI.2023.3269523.

[10] R. Dorent *et al.*, "Inter extreme points geodesics for end-to-end weakly supervised image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 615–624.

[11] Q. Chen and Y. Hong, "Scribble2d5: Weakly-supervised volumetric image segmentation via scribble annotations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2022, pp. 234–243.

[12] K. Zhang and X. Zhuang, "Cyclemix: A holistic strategy for medical image segmentation from scribble supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11656–11665.

[13] H. Qu *et al.*, "Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3655–3666, Nov. 2020.

[14] R. Guo, M. Pagnucco, and Y. Song, "Learning with noise: Mask-guided attention model for weakly supervised nuclei segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 461–470.

[15] U. Wickramasinghe, P. Jensen, M. Shah, J. Yang, and P. Fua, "Weakly supervised volumetric image segmentation with deformed templates," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2022, pp. 422–432.

[16] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 424–432.

[17] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3D human pose using multi-view geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1077–1086.

[18] C. Wu, Y. Xiao, B. Zhang, M. Zhang, Z. Cao, and J. Zhou, "C3P: Cross-domain pose prior propagation for weakly supervised 3D human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 554–571.

[19] Z. Zhang, L. Hu, X. Deng, and S. Xia, "Weakly supervised adversarial learning for 3D human pose estimation from point clouds," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 5, pp. 1851–1859, May 2020.

[20] S. Napel *et al.*, "CT angiography with spiral CT and maximum intensity projection.," *Radiology*, vol. 185, no. 2, pp. 607–610, Nov. 1992.

[21] G. Kiefer, H. Lehmann, and J. Weese, "Fast maximum intensity projections of large medical data sets by exploiting hierarchical memory architectures," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 2, pp. 385–394, Apr. 2006.

[22] A. Salvi, R. Shah, L. Higgins, and P. G. Menon, "Vision transformers for AI-driven classification of peripheral artery disease from maximum intensity projections of runoff CT angiograms," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2022, pp. 3870–3872.

[23] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *J. Artif. Intell. Res.*, vol. 70, pp. 1373–1411, Apr. 2021.

[24] A. Vepa *et al.*, "Weakly-supervised convolutional neural networks for vessel segmentation in cerebral angiography," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 585–594.

[25] Y. Zhao, L. Rada, K. Chen, S. P. Harding, and Y. Zheng, "Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retinal images," *IEEE Trans. Med. Imag.*, vol. 34, no. 9, pp. 1797–1807, Sep. 2015.

[26] S. Fu *et al.*, "Robust vascular segmentation for raw complex images of laser speckle contrast based on weakly supervised learning," *IEEE Trans. Med. Imag.*, early access, Jun. 2023, doi: 10.1109/TMI.2023.3287200.

[27] Z. Xu *et al.*, "Anti-interference from noisy labels: Mean-teacher-assisted confident learning for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3062–3073, Nov. 2022.

[28] A. J. Bekker and J. Goldberger, "Training deep neural-networks based on unreliable labels," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 2682–2686.

[29] Y. Wang, J. Peng, and Z. Zhang, "Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 9092–9101.

[30] W. Yang, T. Zhang, X. Yu, T. Qi, Y. Zhang, and F. Wu, "Uncertainty guided collaborative training for weakly supervised temporal action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 53–63.

[31] X. Cao, H. Chen, Y. Li, Y. Peng, S. Wang, and L. Cheng, "Uncertainty aware temporal-ensembling model for semi-supervised ABUS mass segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 431–443, Jan. 2021.

[32] H. Chen, X. Wang, and L. Wang, "3D vessel segmentation with limited guidance of 2D structure-agnostic vessel annotations," *arXiv preprint arXiv:2302.03299*, 2023.

[33] A.F. Dima *et al.*, "3D arterial segmentation via single 2D projections and depth supervision in contrast-enhanced CT images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2023, pp. 141–151.

[34] S. Zheng, J. Guo, X. Cui, R. N. J. Veldhuis, M. Oudkerk, and P. M. A. van Ooijen, "Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection," *IEEE Trans. Med. Imag.*, vol. 39, no. 3, pp. 797–805, Mar. 2020.

[35] Y. Wang *et al.*, "JointVesselNet: Joint volume-projection convolutional embedding networks for 3D cerebrovascular segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2020, pp. 106–116.

[36] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[37] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 1998, pp. 130–137.

[38] J. Muschelli, N. L. Ullman, W. A. Mould, P. Vespa, D. F. Hanley, and C. M. Crainiceanu, "Validated automatic brain extraction of head CT images," *NeuroImage*, vol. 114, no. 1, pp. 379–385, Jul. 2015.

[39] A. Biesdorf, S. Worz, H. von Tengg-Kobligk, and K. Rohr, "Automatic detection of supraaortic branches and model-based segmentation of the aortic arch from 3D CTA images," in *Proc. IEEE 6th Int. Symp. Biomed. Imag.*, 2009, pp. 486–489.

[40] S. Shit *et al.*, "ClDice-a novel topology-preserving loss function for tubular structure segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 16560–16569.

[41] M. Beauchemin, K.P.B. Thomson, and G. Edwards, "On the hausdorff distance used for the evaluation of segmentation results," *Can. J. Remote Sens.*, vol. 24, no. 1, pp. 3–8, Mar. 1998.

[42] A.A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, pp. 1–28, Dec. 2015.

[43] A. Hilbert *et al.*, "BRAVE-NET: fully automated arterial brain vessel segmentation in patients with cerebrovascular disease," *Frontiers Artif. Intell.*, vol. 3, pp. 78, Sep. 2020.

[44] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen and K.H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2020.