Deep-learning approach to automate the segmentation of aorta in non-contrast CTs

Qixiang Ma, Antoine Lucas, Houda Hammami, Huazhong Shu, Adrien Kaladji, Pascal Haigron

Deep-learning approach to automate the segmentation of aorta in non-contrast CTs

Qixiang Ma[a,b,*] **Antoine Lucas**[a,b] **Houda Hammami**[a,b] **Huazhong Shu**[b,c,d] **Adrien Kaladji**[a,b] **and Pascal Haigron**[a,b]

[a]University of Rennes, Inserm, CHU Rennes, LTSI - UMR 1099, Rennes, France
[b]Centre de Recherche en Information Biomédicale sino-français (CRIBs), Université de Rennes, Inserm, Rennes, France, and Southeast University, Nanjing, China
[c]Southeast University, Laboratory of Image Science and Technology, Nanjing, China
[d]Southeast University, Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Nanjing, China

Abstract

Purpose: Segmentation of vascular structures in preoperative computed tomography (CT) is a preliminary step for computer-assisted endovascular navigation. It is a challenging issue when contrast medium enhancement is reduced or impossible, as in the case of endovascular abdominal aneurysm repair for patients with severe renal impairment. In non-contrast–enhanced CTs, the segmentation tasks are currently hampered by the problems of low contrast, similar topological form, and size imbalance. To tackle these problems, we propose a novel fully automatic approach based on convolutional neural network.

Approach: The proposed method is implemented by fusing the features from different dimensions by three kinds of mechanisms, i.e., channel concatenation, dense connection, and spatial interpolation. The fusion mechanisms are regarded as the enhancement of features in non-contrast CTs where the boundary of aorta is ambiguous.

Results: All of the networks are validated by three-fold cross-validation on our dataset of non-contrast CTs, which contains 5749 slices in total from 30 individual patients. Our methods achieve a Dice score of 88.7% as the overall performance, which is better than the results reported in the related works.

Conclusions: The analysis indicates that our methods yield a competitive performance by overcoming the above-mentioned problems in most general cases. Further, experiments on our non-contrast CTs demonstrate the superiority of the proposed methods, especially in low-contrast, similar-shaped, and extreme-sized cases.

 [DOI: 10.1117/1.JMI.10.2.024001]

Keywords: non-contrast enhanced computed tomographies; abdominal aortic aneurysm; convolutional neural network; segmentation; feature fusing.

Paper 22211GR received Aug. 11, 2022; accepted for publication Feb. 13, 2023; published online Mar. 2, 2023.

1 Introduction

Vascular disease is one of the most common causes of death in industrialized countries.[1] Abdominal aortic aneurysm (AAA), regarded as a severe vascular disease, involves an excessive enlargement of the abdominal aorta. The rupture of the AAA is a surgical emergency that leads to a high death rate in many countries, with an overall mortality of 80 to 90 percent.[2]

To treat the AAA diseases, endovascular aneurysm repair (EVAR) is an interventional procedure that is widely performed to avoid the risks of open surgery.[3] The guidance of AAA

*Address all correspondence to Qixiang Ma, qixiang.ma@univ-rennes1.fr

2329-4302/2023/$28.00 © 2023 SPIE

treatment via EVAR is currently based on the observation of computerized tomography angiography (CTA) images, which are acquired following a contrast agent injection. The contrast agent enhances the radio-density within the vascular lumen, which consequently yields a salient region on reconstructed CTA images in which the lumen of the aorta is distinguished from the surrounding tissues. Therefore, CTA images are crucial for the intervention.[4,5] Further, the segmentation of the aorta on such CTA images is a prerequisite for the planning and guidance of intervention.[6,7]

However, although CTA imaging is generally used to guide AAA diseases, the contrast agent yields potential risk in some special cases. A typical example is the condition of kidney disease, e.g., renal impairment, renal failure, or end-stage kidney failure. When following the EVAR with contrast injection, the patients who suffer from such pre-existing kidney diseases are prone to serious risk of contrast-induced nephropathy,[8,9] an inducement of further degeneration of the kidney function. To avoid the adverse effects of contrast agents, only non-contrast CTs are supposed to be used in these special cases. In a previous work,[10] the authors showed that the EVAR without the use of contrast agent appears to be safe and accurate for the patients with severe chronic kidney disease, with the guidance of EVAR being achieved by manual segmentation of aortas in non-contrast CTs. In this paper, our objective is to automate the segmentation process.

The aorta segmentation in non-contrast CTs is more challenging compared with CTAs. The main difficulties are as follows:

1. Low contrast of the adjacent regions: without the enhancement of contrast agent, the aorta (both lumen and outer wall) exhibits homogeneity with its surrounding tissues. Therefore, it is tough to distinguish the boundary of the target object from the background according to the voxel intensity.
2. Topological form similar to the surrounding tissues: along the axial dimension, the aorta often appears in the form of a circle or ellipse, resembling surrounding tissues. That makes determining the aorta more challenging in non-contrast CTs.
3. Size imbalance between foreground and background areas: compared with the background, the region of an aorta only occupies a tiny limited area, so the extreme inequality between background and foreground pixels impedes segmentation methods from achieving satisfactory results.

For example, Fig. 1(a) shows a 2D non-contrast axial slice of a patient who suffers from the AAA. Figure 1(b) shows the ground truth, including the targeted object, i.e., aorta (in red), and the background areas (in blue). Comparing Figs. 1(a) and 1(b), it is observed that the region of the aorta has low contrast with its surrounding tissues. In addition, Fig. 1(c) shows the similar topological form (in blue boxes) with the aorta (in red box), which have proved to potentially yield false positives (FPs) in the subsequent experiments. Furthermore, in Fig. 1(d), the histogram of the aorta and background confirms the issue of the imbalance of the pixels. The number of pixels of the aorta is much less than that of the background. In addition, the pixel space of the aorta is fully covered by the background, which means that it is difficult to distinguish the object according to pixel intensity.

To deal with the segmentation task of AAA, it was common practice for experts to perform manual measurements, which was time-consuming and subjective, being operator-dependent. During past decades, several classical segmentation methods have been proposed to automate the process. However, most of them are based on hand-crafted features, such as active contours[11–13] and graph cut algorithms,[14,15] restraining the representation capability. However, to the best of our knowledge, no conventional method is related to non-contrast CTs.

Recently, the development of deep-learning (DL) techniques has promoted various computer vision tasks,[16] including medical image processing.[17] DL-based approaches surmount the obstacles that the conventional methods are consistently tied to the particular man-made features. convolutional neural network (CNN),[18] a typical implementation of DL techniques, is extensively used in the tasks of medical image segmentation. Among the proposed techniques, U-Net[19] is regarded as one of the most critical baseline models for segmentation tasks. It works following the mechanism of a combination of an encoder and a decoder, with skip connections to concatenate the feature map of the encoder and its related one from the decoder. Based on U-net,

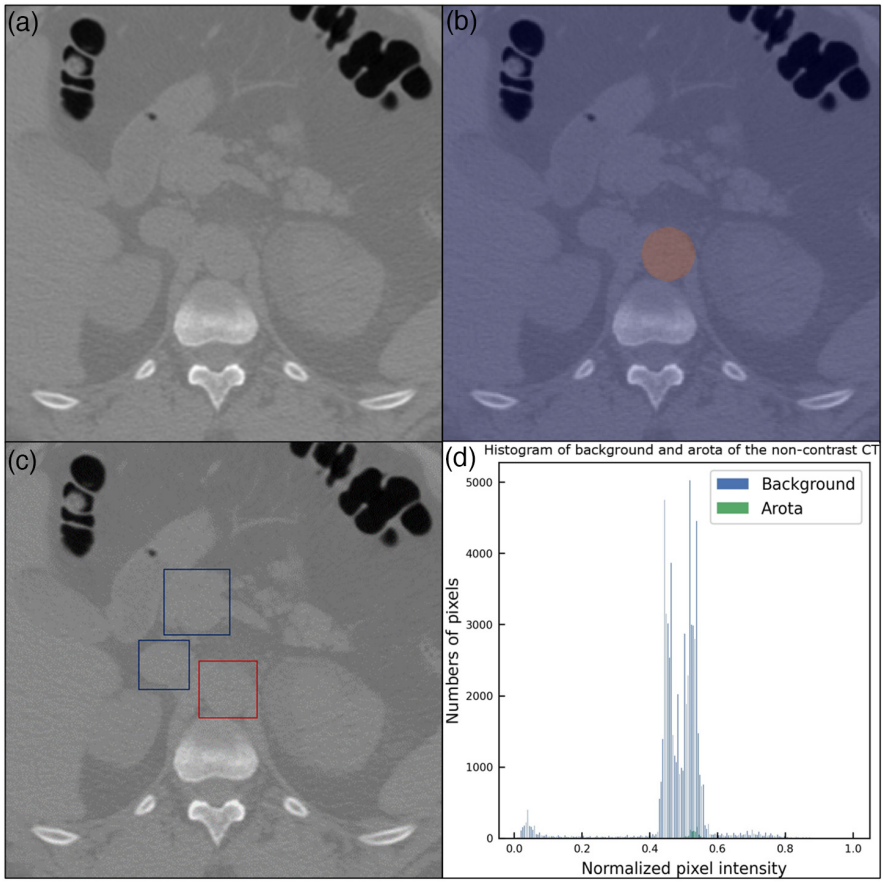

Fig. 1 Axial non-contrast CT slice (cropped to 256 × 256) and its related information obtained from a patient suffering from the AAA. (a) The original axial non-contrast slice; (b) the ground truth including the aorta (in red) and background (in blue); (c) the surrounding tissues with similar topological form (in blue boxes). (d) Histogram of the aorta and background.

a large number of variations, such as 3D U-Net,[20] V-Net,[21] and Res U-Net,[22] have been proposed to further promote the performance.

In terms of aorta segmentation, contemporary research has shown the efforts on DL-based methods. For 2D axial slice segmentation, deep belief and U-Net-based models have been proposed to segment the AAA region, respectively.[23,24] Cheng et al.[25] introduced the squeeze-and-excitation block to the U-Net-based model for segmentation and detection of aortic dissection. The main weakness of such 2D methods is that the contextual information cannot be captured because the CT slices are treated as individual samples. For 3D volume segmentation, Habijan et al.[26] proposed a modified 3D U-Net for aorta segmentation on volumetric CTs; however, more computing resources were required. Furthermore, López-Linares et al.[27] leveraged both the 2D and 3D strategies in a cascaded way to segment the aorta as well as aortic thrombus. Even though the mechanism of cascaded 2D and 3D extracted both the intra-/inter-slice contextual information, it was implicitly restricted because these two kinds of features can only be obtained in their related depths.

In addition to such limitations, all of these CNN-based methods solely focus on contrast-enhanced CTs, without verifying their performance on non-contrast CTs. To the best of our knowledge, there are only two CNN-based studies that partially involve the segmentation of AAA for CT images without a contrast agent.[28,29] One of them, DeepAAA,[29] leveraged a variant of 3D U-Net to train and test the mixture of CTAs and non-contrast CTs. The other study[28] developed a cascaded attention-based 3D U-Net, performing individually on the paired contrast-enhanced and non-contrast CTs, respectively. Neither of the studies specifically focus on analyzing the non-contrast CT segmentation. In addition, they developed the 3D U-Net-based

(or cascaded 3D U-Net-based) model, which requires 3D volume data and has a high demand for GPU memory as well as computational cost.

To tackle these problems, in this paper, we propose a CNN-based approach fusing 2D and 3D features while utilizing them simultaneously during each stage of forward propagation. It is implemented by three kinds of mechanisms, i.e., channel concatenation, dense connection, and spatial interpolation. Compared with the related and state-of-the-art techniques, the three mechanisms of fusion yield superior performance while overcoming the barriers of segmentation tasks on our non-contrast CT dataset. Our main contributions are as follows: (1) a novel insight of the integration of CNN features and three mechanisms for flexible implementation and (2) results of aorta segmentation in non-contrast-enhanced CTs. To the best of our knowledge, it is the first CNN-based methodological attempt entirely concentrating on aorta segmentation tasks for non-contrast CTs.

The remainder of this paper is organized as follows. In Sec. 2, we describe the related work. The proposed methods are then introduced in Sec. 3. The experimental approaches and analysis of the results are reported in Secs. 4 and 5, respectively. In Sec. 7, we draw the conclusion and provide perspectives.

2 Related Work

2.1 *Conventional Methods of Segmentation of Aorta*

The conventional approaches are mainly based on the active contours and graph cut algorithms. For active-contours-based methods, de Bruijne et al.[11] applied the active shape model (ASM) in sequential slices to decide the boundary of AAA; and Dhibi et al.[12] and Chen et al.[13] proposed two active-contour-based methods for estimating the volume of the thrombus and aneurysm, respectively. As for graph-cut-based methods, Freiman et al.[14] presented an iterative model-constrained graph-cut algorithm, and Lee et al.[15] proposed a 3D graph search approach based on a triangular mesh. Both of these aim at extracting thrombus from surrounding tissues. Duquette et al.[30] made an attempt on segmentation of both lumen and thrombus for CTA as well as MRI based on graph cut theory.

All of the aforementioned classical methods process the contrast-enhanced CTs, aiming at distinguishing the inner lumen as well as its surrounding tissues such as the thrombus. However, for non-contrast-enhanced CTs, there is no salient region of lumen appearing on the images. Therefore, for the segmentation on non-contrast CTs, the aim is to segment the whole region of the aorta from the surrounding tissues [as shown in Fig. 1(b)]. Even though the inner lumen and the thrombus are not identified individually, it still makes sense for the guidance of EVAR. The existing classical method related to non-contrast CTs focus on the research of aortic calcification,[31] thoracic aorta territory,[32] and pulmonary artery.[33] But few studies involve the AAA. Our approach specifically concentrates on the segmentation task for abdominal aorta as well as aneurysm based on non-contrast-enhanced CTs.

2.2 *State-of-the-art CNN-based Architectures*

Recently, CNN-based models facilitate the segmentation of medical images without resorting to man-made features. According to the strategy of data processing flow, the state-of-the-art models can be roughly divided into the 2D-based, 3D-based, and cascade of 2D and 3D-based models. The milestone of 2D-based models is U-Net,[19] which effectively fuses low-level and high semantic information by combining the related feature maps from both the analysis and synthesis paths through skip connections. This mechanism has made a breakthrough in segmentation tasks for medical images. Thus, U-Net is regarded as a benchmark for the subsequent derivatives.[34,35] However, in practice, most of the medical data are constructed as 3D volumes, which requires the model to further extract the contextual information among the adjacent slices instead of treating the data as individual 2D images. Motivated by this idea, çiçek et al. proposed 3D U-Net[20] to deal with the structural medical data in a 3D manner. The V-Net[21] introduces a residual connection[36] and extends the depth of the 3D U-based model to further improve the performance.

The most notorious limitation of 3D models is the over-dependency on expensive computational resources as well as hardware configurations. Some of the 3D U-Net-based models divide the 3D volume into local patches to match the limited available GPU memory.[37,38] Fabian et al. supposes in Ref. 39 that these patch-based training strategies may restrict the field of view of the models, so the contextual information is insufficiently collected. For aorta segmentation, the patch-based approach is not practical because the aorta only appears in a small region of the CTs.

Recently, the hybrid 2D and 3D networks have been implemented to learn the representation of 2D and 3D features in a cascading way. Li et al.[40] proposed H-DenseUNet to initially extract 2D features by 2D DenseUNet and further extract the 3D contextual information by a subsequent counterpart 3D model. Zhang et al.[41] integrated this mechanism into a pair-ware encoder-decoder to make it more lightweight. The cascading mechanism has proved to be more competitive in segmentation tasks. However, it is intuitively limited because the 2D and 3D features are only learned in their related depth, e.g., the 3D features are only extracted in the high-level layers. Compared with existing cascading models, our goal is to fuse the 2D and 3D features in each single layer, so the two kinds of features are integrated more flexibly.

2.3 *CNN-based Methods for Aorta Segmentation with Non-contrast Enhanced CTs*

We have sketched the literature related to CNN-based methods for aorta segmentation with CTAs in Sec. 1. But there are only a few research studies based on CNN involving segmentation of abdominal aorta and aneurysm on non-contrast-enhanced CTs.[28,29] Lu et al. proposed DeepAAA[29] and conducted experiments with both CTAs and non-contrast CTs from 223 patients. The model is derived from 3D U-Net,[20] containing four encoder/decoder modules and an additional bottleneck layer with dropout regularization. They mainly investigated the Dice score as well as the maximum diameter error. However, they made a mixture of CTAs (52%) and non-contrast CTs (48%) in both the training and testing stages, which did not specifically emphasize the effect of non-contrast CTs. Chandrashekar et al.[28] proposed a 3D cascaded attention-based model that integrates additive attention gates into the original 3D U-Net. Furthermore, it combines two models in a cascaded way to implement coarse-fine segmentation. The non-contrast CTs that they used were collected from 26 patients, 13 of them for training and the rest for validation and testing. An offline data augmentation of divergence transformations was applied to the training set with an extensive ratio of 10:1. We suppose that this mechanism may have over-dependence on the augmented data instead of the proposed model itself. In this paper, we compare our model with these two methods by conducting experiments on our non-contrast-enhanced CTs without any data augmentation except for horizontal flip. In addition, Suzuki et al.,[42] in their recent study, used a voting mechanism to merge several basic CNN models to segment the aorta and main pulmonary artery in non-contrast CTs. Although it is a different focus from our work because there are no AAA cases in their study, we choose the representative approach in their study (VGG19-U-Net) for comparison.

3 Method

3.1 *Overview*

In this paper, we propose an encoder-decoder-based CNN architecture. As shown in Fig. 2(a), the architecture contains four encoder modules and five decoder modules in total, following a 1×1 convolutional layer to reduce channels. Each encoder module (an EnBlock plus a pooling layer) is used for the contraction of the features while a decoder module (DeBlock) makes the expansion. Each pair of En/DeBlock models are connected by a skip connection which fuses the low- and high-level features in a concatenated way. This architecture accepts any scale of ordered CT slices (the lengths of the sides are even numbers) as input. As Fig. 2(a) shows, the input gray image has a spatial size of $w \times h$ and a mini-batch size of N, and the spatial size of the input feature maps of the n'th ($n \leq 5$) En/DeBlock is

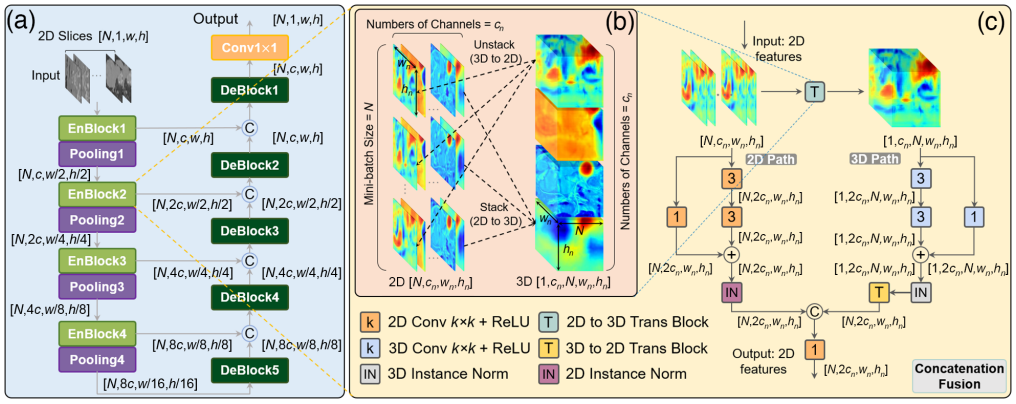

Fig. 2 Overall architecture and the CTF pattern. (a) Overall architecture; (b) the mechanism of 2D and 3D features transformation; and (c) the CTF pattern.

$$w_n = 2^{1-n}w, \tag{1}$$

$$h_n = 2^{1-n}h, \tag{2}$$

With the input channel size of each EnBlock being

$$c_n = \begin{cases} 1, & n = 1 \\ 2^{n-2}c, & \text{otherwise}, \end{cases} \tag{3}$$

where c is the initial channel size. In each EN/DeBlock, we fuse the 2D and 3D features in three different ways to extract the intra-/inter-slice contextual features simultaneously.

We pre-define a series of notations for the three following subsections. In the following parts, x_{in}^{2D} and x_{in}^{3D} represent the 2D and 3D input of each En/DeBlock, and x_{out} denotes the output of the En/DeBlock. $T^{2\text{ to }3}$ and $T^{3\text{ to }2}$ are the transformation to transfer the features from 2D/3D to its counterpart 3D/2D features, respectively, where $x_{\text{in}}^{3\text{D}} = T^{2\text{ to }3}(x_{\text{in}}^{2\text{D}})$. F_k^{nD} represents a n-dimensional convolution with a kernel size of k in each dimension, followed by an activation function rectified linear unit. IN^{nD} is a n-dimensional instance normalization (IN)[43] operation. P^{2D} and P^{3D} are the features generated from the 2D and 3D paths, respectively.

3.2 *Concatenation Fusion Pattern*

The first approach integrates the 2D and 3D features in a simple but efficient way, i.e., concatenating the feature maps along channels.

Because 3D and 2D features have different dimensions, the fusion can be processed only after the dimensional transformation. Figure 2(b) shows the mechanism of 2D/3D dimensional transformation in the n'th EN/DeBlcok. The left part of Fig. 2(b) is the 2D features with mini-batch N and channel size c_n. To transform the 2D feature to 3D features, we stack the features along the direction of the mini-batch to generate a 3D feature that contains the spatial size of $[N, w_n, h_n]$ with a channel size c_n. That yields a counterpart 3D feature map shown on the right part of Fig. 2(b). The transformation makes an extension of the dimension, so we regard the whole 3D feature as a volume with a new mini-batch $N_{3D} = 1$. In a word, from left to right the 2D features ($[N, c_n, w_n, h_n]$) are stacked to generate a counterpart 3D feature ($[1, c_n, N, w_n, h_n]$), and vice versa from right to left.

Figure 2(c) shows the flowchart of concatenation fusion (CTF) in the n'th EnBlock. The 2D and 3D features are first processed in the two paralleled paths individually and then fused by concatenation. The transformation of 2D/3D features is achieved by the 2D/3D Trans Block shown in (b). The input feature map x_{in}^{2D} is 2D with a shape of $[N, c_n, w_n, h_n]$. On the left path, it is first fed into two cascaded 3×3 convolutional layers with 1 padding and 1 stride. Additionally, a sideway 1×1 layer is used to provide the residual signal.[36] On the right side, the 3D features x_{in}^{3D} are processed by the counterpart layer with 3D convolutional kernels.

Because the 3D features contain a single mini-batch with $N_{3D} = 1$, we use IN[43] instead of batch normalization (BN)[44] to normalize the features in each path, which enables the normalization in the 3D path to be performed for the whole volume, whereas in the 2D path, it processes every single slice. This mechanism implicitly considers the intra- and inter- contextual information in 2D and 3D features, respectively. Furthermore, to reduce the redundant weights, we modify the conventional pattern convolution-normalization-activation to convolution-activation in each convolutional module. In each En/DeBlock, the normalization parts are only applied before the last convolutional layer. It is worth mentioning that the 3D features are supposed to be transformed back to 2D features before fusion, which is implemented by the unstack part in Fig. 2(b). The fused features are then fed into a 2D 1×1 convolutional layer to reduce the channel size.

The process of the CTF in each En/DeBlock is formulated as

$$x_{\text{out}} = F_1^{2D}([P^{2D}, T^{3\,\text{to}\,2}(P^{3D})]), \tag{4}$$

$$P^{2D} = IN^{2D}(F_3^{2D}(F_3^{2D}(x_{\text{in}}^{2D})) + F_1^{2D}(x_{\text{in}}^{2D})), \tag{5}$$

$$P^{3D} = IN^{3D}(F_3^{3D}(F_3^{3D}(x_{\text{in}}^{3D})) + F_1^{3D}(x_{\text{in}}^{3D})), \tag{6}$$

where $[\ldots]$ denotes a concatenation operation along the direction of the channel.

3.3 *Dense Fusion Pattern*

The second approach that we propose is fusing the 2D and 3D features in a dense-based mechanism. The concept of "dense" for a CNN model was first proposed by Huang et al.[45] to address the issues of the vanishing/exploding gradients in the deeper layer. It contributes the insight that the direct connection can be added to all subsequent layers to improve the performance and the dense connection implicitly enhances the deep supervision. Based on a dense connection, Dolz et al.[38] proposed HyperDense-Net to process multi-modal images. It leverages two paralleled dense models to extract the features from $T1$ and $T2$-weighted magnetic resonance (MR) images, respectively. There are interactive concatenations of features across the two paths.

Inspired by HyperDense-Net,[38] our second fusion pattern is designed as dense fusion (DSF). As Fig. 3(a) shows, in the n'th En/DeBlock, the input 2D features x_{in}^{2D} are initially transformed to generate counterpart 3D features x_{in}^{3D}, and then they are fed into two paralleled convolutional paths, respectively. During each path, the features are processed by their convolutional models with 2D/3D kernels while interacting with another convolutional path. For example, on the 2D side, the feature maps are first fed into a 3×3 convolutional layer that does not alter the channel size. The output, therefore, has the same size as the input ($[N, c_n, w_n, h_n]$). Then the output is concatenated with the features transformed from its related 3D counterpart. At the same time, the 2D feature is transformed to a 3D counterpart, concatenated in the 3D path. The concatenated features double the channel size, with shapes of $[N, 2c_n, w_n, h_n]$ and $[1, 2c_n, N, w_n, h_n]$ in the 2D and 3D paths, respectively. They are not only fed into subsequent convolutional layers, which reduces the channel size to 1, but also concatenated with the subsequent outputs. The last output of each path is normalized respectively, fused by element-wise summing, and then fed into a 1×1 convolutional layer. The final output of DSF has the same shape as the input.

Our DSF and HyperDense-Net possess two main differences. First, HyperDense-Net treats two different modalities of 3D MRI in two paralleled paths, respectively, whereas DSF processes 2D features as well as its counterpart 3D representations simultaneously. The interaction of the two paths in HyperDense-Net is direct because they have the exact same size, whereas in DSF, before the interaction, it is necessary to perform 2D/3D transformation to unify the size. Second, even though both of them utilize dense connections inside or across the paths, HyperDense-Net is regarded as a global dense connection network because it connects the whole architecture from the beginning layer to the end. However, DSF limits the dense connection to a single En/DeBlock, which is packaged as a modularized element, e.g., the second EnBlock does not interact with the third EnBlock with a dense connection. This modularization facilitates the design of architecture, especially when it is necessary to add or remove some layers.

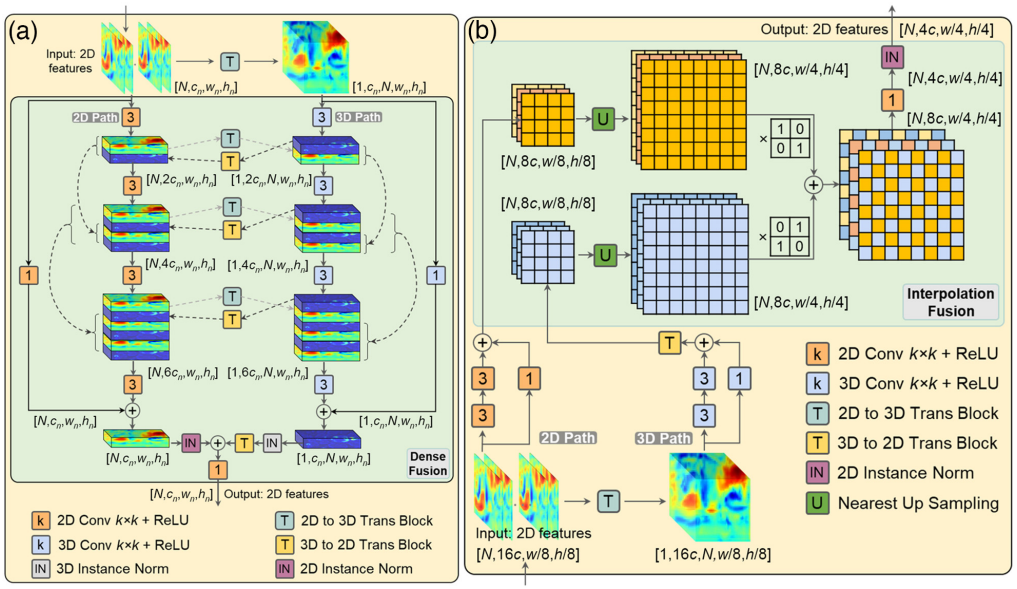

Fig. 3 Mechanism of (a) DSF and (b) IPF pattern, respectively.

The process of the DSF in each En/DeBlock is formulated as

$$x_{\text{out}} = F_1^{2D}(P^{2D} + T^{3\,\text{to}\,2}(P^{3D})), \tag{7}$$

$$P^{2D} = IN^{2D}(F_3^{2D}(\text{Dense}_j^{2D}) + F_1^{2D}(x_{\text{in}}^{2D})), \tag{8}$$

$$P^{3D} = IN^{3D}(F_3^{3D}(\text{Dense}_j^{3D}) + F_1^{3D}(x_{\text{in}}^{3D})), \tag{9}$$

$$\text{Dense}_j^{2D} = \begin{cases} [F_3^{2D}(x_{\text{in}}^{2D}), T^{3\,\text{to}\,2}(f_0^{3D})], & j = 0 \\ [F_3^{2D}(\text{Dense}_{j-1}^{2D}), T^{3\,\text{to}\,2}(f_j^{3D}), \text{Dense}_{j-1}^{2D}], & \text{otherwise} \end{cases}, \tag{10}$$

$$\text{Dense}_j^{3D} = \begin{cases} [F_3^{3D}(x_{\text{in}}^{3D}), T^{2\,\text{to}\,3}(f_0^{2D})], & j = 0 \\ [F_3^{3D}(\text{Dense}_{j-1}^{3D}), T^{2\,\text{to}\,3}(f_j^{2D}), \text{Dense}_{j-1}^{3D}], & \text{otherwise} \end{cases}, \tag{11}$$

$$f_j^{2D} = \begin{cases} F_3^{2D}(x_{\text{in}}^{2D}), & j = 0 \\ F_3^{2D}(\text{Dense}_{j-1}^{2D}), & \text{otherwise} \end{cases}, \tag{12}$$

$$f_j^{3D} = \begin{cases} F_3^{3D}(x_{\text{in}}^{3D}), & j = 0 \\ F_3^{3D}(\text{Dense}_{j-1}^{3D}), & \text{otherwise} \end{cases}, \tag{13}$$

where $[\ldots]$ denotes a concatenation operation along the direction of the channel. $j \in [0,1,2]$ represents the index of the sub-layer of the dense fusion.

3.4 *Interpolation Fusion Pattern*

The third pattern of fusion that we propose is interpolation fusion (IPF). This fusing method is only applied in the path of the decoder. In general, the contraction phase (encoder) generates the feature maps containing high semantic information but low spatial resolution. Therefore, during the expansion phase (decoder), it is necessary to recover the resolution. The conventional approaches to expand the spatial resolution in U-Net-based models are up-convolution[19,20] and transposed convolution.[21] Up-convolution is an up-sampling operation with a successive convolution layer, whereas transposed convolution inherently contains trainable weights. In terms of our model, because we leverage both 2D and 3D features in each DeBlock, a relevant strategy is to recover the spatial resolution by fusing the two features by interpolation in the spatial domain.

Figure 3(b) shows the pattern of fusion specifically in the fourth DeBlock. The 2D and 3D features (x_{in}^{2D} and x_{in}^{3D}) are initially processed by 2D and 3D convolutional kernels, respectively, to obtain both the intra- and inter- contextual information of CT slices. Then, instead of combining along the channels as CTF and DSF, the two paralleled features are fused within the 2D spatial domain. The features are first 2× upsampled by nearest neighbor resampling and then multiplied by a binary mask to keep half of the spatial information, respectively. The subsequent element-wise sum operation enables the interpolation between pixels. Finally, a 1×1 convolutional layer is utilized to reduce the channel size. IPF implicitly leverages spatial information of 3D features as a complementary component to enhance the 2D features.

The process of IPF in each En/DeBlock is formulated as

$$x_{\text{out}} = IN^{2\text{D}}(F_1^{2\text{D}}(Mask_{10} \times U(P^{2\text{D}}) + Mask_{01} \times U(T^{3\,\text{to}\,2}(P^{3\text{D}})))), \tag{14}$$

$$P^{2\text{D}} = F_3^{2\text{D}}(F_3^{2\text{D}}(x_{\text{in}}^{2\text{D}})) + F_1^{2\text{D}}(x_{\text{in}}^{2\text{D}}), \tag{15}$$

$$P^{3\text{D}} = F_3^{3\text{D}}(F_3^{3\text{D}}(x_{\text{in}}^{3\text{D}})) + F_1^{3\text{D}}(x_{\text{in}}^{3\text{D}}), \tag{16}$$

where $Mask_{10}$ and $Mask_{01}$ represent the two binary masks in Fig. 3(b), respectively. U denotes the 2× up-sampling operation with the nearest-neighbor interpolation.

3.5 *Loss Function*

We optimize the weights of the model by minimizing the binary cross entropy between the labels and the predicted results. The loss function is formulated as

$$\text{BCE}(p, g) = -\sum_{i \in \Omega} g_i \log(p_i) + (1 - g_i) \log(1 - p_i), \tag{17}$$

where g_i and p_i represent the ground truth and predicted result of voxel i, respectively. Ω denotes the space of predicted result p and ground truth g. The two classes of our study are the background and the aorta.

4 Experiments

4.1 *Data Set*

The imaging data of this study are obtained from 30 patients who suffer from infrarenal AAAs. This is an observational study for which the data were collected retrospectively at Rennes University Hospital from patients who underwent the EVAR procedure. Patient informed consent was obtained for being registered anonymously in the research database. A pre-operative non-contrast-enhanced CT scan was performed on each patient. The original images were given in Digital Imaging and Communications in Medicine format with a spatial size of 512×512 and a thickness of 0.625 to 5 mm for each axial slice.

Two vascular surgeons (A and B) manually delineated the ground truth masks for supervising segmentation tasks. Surgeon A delineated all of the CTs, regarded as ground truth g. To assess intra- and inter-observer variability of the manual segmentation, we follow the strategies of the related work,[28] randomly selecting a subset s of the scans ($|s| = 10$). Surgeon B delineated s independently, generating g_B. Surgeon A delineated s for a second time after a gap of 10 days to yield g_A. g_A and g_B were compared against g in terms of Dice score to assess intra- and inter-observer variability of the manual segmentation, respectively.

To improve the training and testing efficiency, we extracted a 256×256 region of interest (RoI) from each slice by cropping the slice with a uniform spatial position. The contextual information outside the RoIs or the samples in which the abdominal aorta is absent are excluded. Finally, the dataset contains 5749 axial slices belonging to 30 patients. It is divided into three sub-datasets (non-overlapping for patients), marked as D_0, D_1, and D_2 for 3-fold

cross-validation. We use dynamic programming to ensure that the distribution in terms of size and shape of the aorta is balanced among the sub-datasets.

4.2 *Implementation Details and Optimization*

We train each model 100 epochs with the early stopping strategy. RMSprop[46] is applied to optimize the training stage with the following hyper-parameters: $\alpha = 0.99$, $\epsilon = 1 \times 10^{-8}$. The initial learning rate is 0.001 and decreases by 0.9 every 40 epochs. All of the weights of the models are initialized by the Kaiming initialization.[47] The experiments are implemented on the PyTorch[48] platform, deployed on Ubuntu 18.04 with a GPU of Nvidia GeForce GTX 1080 (12 GB memory). The input size for our model is 256×256 with an ordered mini-batch $N = 16$. For the 3D models compared in this study, the input size is a volume of $16 \times 256 \times 256$. The intensities of input are normalized to zero mean and unit deviation. We do not leverage any other data augmentation except for an on-the-fly horizontal flipping.

To achieve high efficiency, we set our models with a smaller initial channel size c to limit the size of intermediate features as well as the number of the convolutional kernels. Instead of setting c with a large number of 32 or 64 as in the current CNN-based models, we set $c = 10$ for CTF and IPF and $c = 20$ for DSF.

4.3 *Evaluation Metrics*

The performance for the segmentation tasks are evaluated following different metrics, i.e., Dice similarity coefficient (DSC), sensitivity (SEN), volumetric similarity (VS), and Hausdorff distance (HD), where DSC and SEN are overlap-based, VS is volume-based, and HD is surface-based metric. In addition, because the maximum transverse diameter (MTD) is a crucial indicator for verifying an aortic aneurysm, we also evaluate the error of MTD, which is a specific measurement for aorta segmentation. To obtain the MTD of an aorta, we first apply the ellipse fitting[49] to each segmented contour of the aorta to get a simulated ellipse. Then the MTD is regarded as the long axis of the ellipse.

In addition, to comprehend the performance of the methods on multi-sized cross sections of the aorta, we introduce a metric called vascular cross-section (VCS) ratio. For a 2D slice s belonging to patient p, we define a VCS ratio $r = a(s)/a(s_{\max})$, where $a(s)$ is the number of pixels of the aorta in slice s and $s_{\max}$ denotes the slice that belongs to p and contains the most pixels of aorta compared with other slices of p. This VSC ratio is independent of the pixel spacing; therefore, it can be universally used in CTs with different pixel spacings.

5 Results

5.1 *Comparison with Aorta-specific CNN-based Methods*

We first compare our approaches with three pre-existing methods, i.e., VGG19-U-net,[42] DeepAAA,[29] and 3D cascaded attention-based U-Net (3D-CA),[28] which are related to aorta segmentation on non-contrast-enhanced CTs; VGG19-U-Net[42] is a 2D-based method, and the other two are 3D-based. For comparability, all of the training strategies and the data pre-processing approaches for the related methods are identical to ours. Because 3D-CA[28] has a mechanism of two-stage segmentation (a cascaded coarse-fine segmentation pipeline), we set the input size of the two stages at $16 \times 256 \times 256$ (S1) and $16 \times 224 \times 224$ (S2), respectively.

Table 1 shows the quantitative results according to four evaluation metrics. It manifests that all of the best results are achieved from our approaches, especially for the CTF-CTF and DSF-DSF models. The former yields the best performance in D_0, whereas the latter generates most of the competitive results among the three subsets as well as the best performance in the overall evaluation. The gap is specifically obvious in D_2; ours obtains three DSCs exceeding 87.0% whereas the others are below 85%. For each model, we also record the requirement of memory and the capability of processing the samples during a unit interval [frame per second (fps)]. It is obvious that the 2D-based method has a significantly higher calculation rate (106.1 fps) than the

Table 1 Performance comparison among our methods and three related works that involve aorta segmentation on non-contrast-enhanced CTs. For ours, the form "*X*-*X*" means the mechanism of encoder and decoder, respectively. For example, "CTF-IPF" represents a model that contains CTF in all of the EnBlocks and IPF in all of the DeBlocks. The bold values represent the optimal results of their columns.

			D_0				D_1				D_2				Average			
Model	Memory	fps	DSC	SEN	HD	VS	DSC	SEN	HD	VS	DSC	SEN	HD	VS	DSC	SEN	HD	VS
VGG19-U-Net[42]	24.7M	106.1	89.0	89.5	7.75	96.7	84.0	82.0	10.12	95.6	83.9	81.1	8.80	93.1	85.6	83.0	8.89	95.1
DeepAAA[29]	75.2M	44.4	87.6	91.7	7.75	95.1	87.1	89.1	8.15	94.7	84.7	85.2	13.42	94.1	86.5	88.6	9.78	95.3
3D-CA[28] (S1)	25.8M	30.3	85.9	88.2	19.98	93.7	83.9	85.8	18.3	95.3	82.9	85.9	14.82	94.1	84.2	86.6	17.7	94.4
D-CA[28] (S2)	25.8M	39.2	87.3	88.8	14.91	96.6	84.8	86.8	14.96	93.2	82.8	82.5	20.27	96.1	85.0	86.3	16.71	95.3
CTF-CTF (ours)	15.9M	64.3	**90.9**	**92.8**	**5.69**	**97.4**	87.1	88.2	8.02	96.0	87.4	**88.4**	9.12	96.8	88.5	**89.8**	7.61	96.7
CTF-IPF (ours)	10.5M	59.5	90.1	88.6	6.91	97.3	87.6	88.2	8.60	**96.2**	87.1	87.3	8.83	96.2	88.3	88.0	8.11	96.5
DSF-DSF (ours)	20.2M	42.3	90.2	91.2	6.84	97.2	**88.1**	**89.1**	**6.83**	96.1	**87.8**	87.9	**8.53**	**97.4**	**88.7**	89.4	**7.40**	**96.9**

other methods. Because we set a smaller initial channel size c for each of our models, they do not occupy extreme memory space while yielding a relatively high segmentation speed, except for DSF. We hypothesize that, in DSF, the large number of tasks on dimension transformation prolongs the processing time.

We randomly visualize three types of segmentation results in Fig. 4, i.e., the regular/irregular-shaped and large-sized cross sections of the aorta. The first row shows the regular-shaped ones, which means the shape of the aorta approximates a circle. It is observed that both false negatives

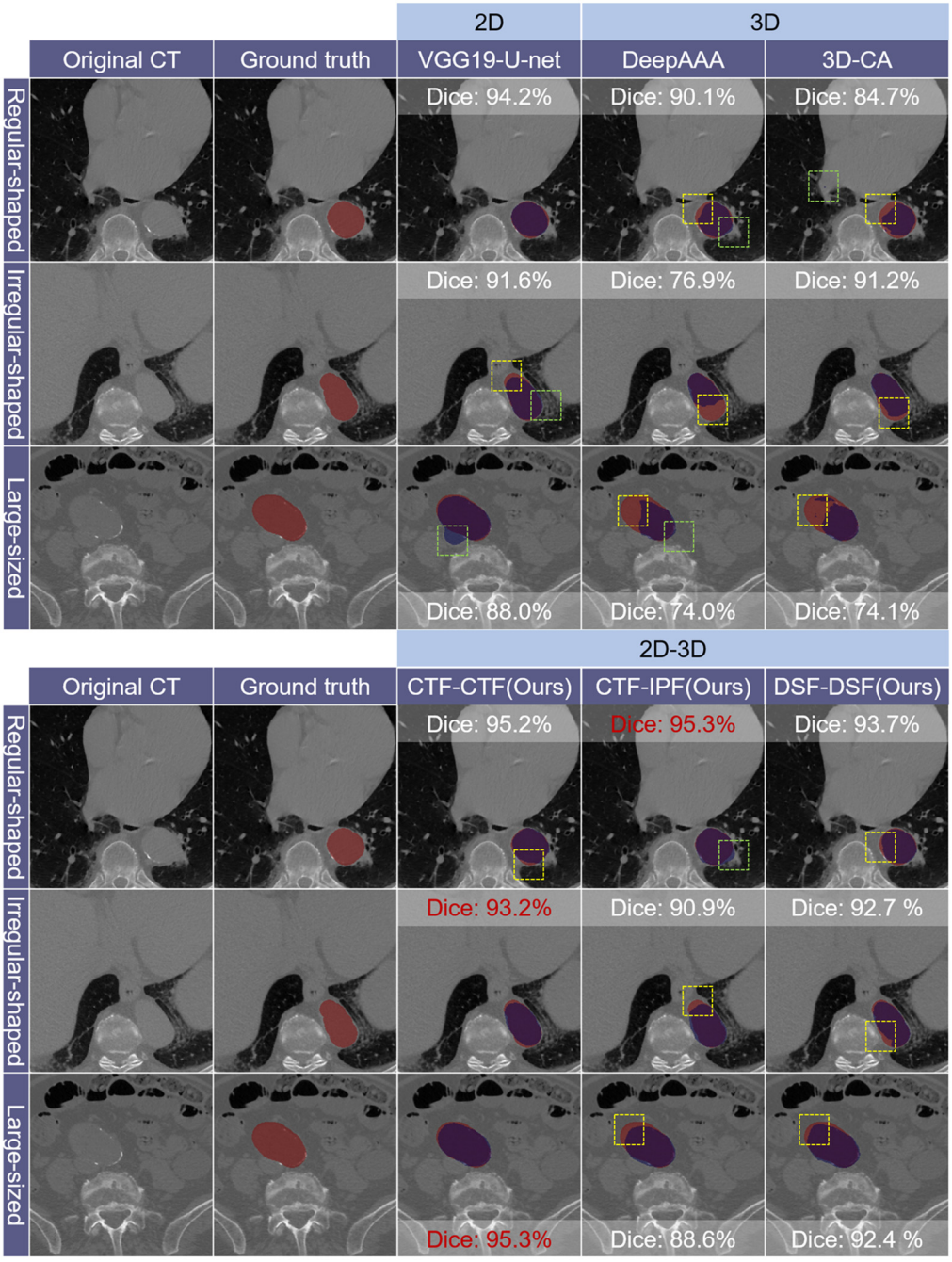

Fig. 4 Visualization of the segmentation results sampled from patient-1. They are generated from our methods and the related methods in terms of 2D, 3D, and 2D-3D mechanisms. The red and blue regions represent the ground truth and predicted results, respectively. The purple regions represent the overlay areas. From top to bottom, the ordered rows show the regular, irregular-shaped, and large-sized cross sections of the aorta. The Dice coefficient similarity (DSC) related to each slice is marked. The green- and yellow-dotted boxes denote the FPs and FNs, respectively.

(FNs) and FPs are present in DeepAAA and 3D-CA, respectively, whereas ours eliminate most of the faults. The second row consists of the samples with irregular-shaped aortas, i.e., the shape of the aorta is far from a circle, more similar to an ellipse. All segmentation results present various degrees of FNs except for CTF. The DeepAAA fails to segment out the majority of the aorta, whereas CTF obtains the comparatively best result. The last row illustrates the large-sized ones in which the aorta involves enlargement and has the potential to be an aneurysm. Both of the related methods yield severe FPs, which has an adverse effect on the guidance of the intervention. In contrast, the segmentation results of CTF and DSF occupy most of the enlarged aorta. Overall, compared with the two related methods, our approaches yield superior performance on aortas with various shapes and sizes.

Figure 5 shows the visualization of 3D reconstruction of the aorta from a random patient. Without any post-processing, the part of the aneurysm (the region of the enlargement of the bottom) is relatively well segmented by each method. However, all of the methods yield various degrees of FPs in the top region, which is especially obvious in 3D-CA, with several parts of the background being predicted as the aorta. In addition, the result of DeepAAA shows severe FNs in which no aorta is segmented out and it is presented as a separation of the aorta. Both of the FPs/FNs adversely affect the guidance of intervention. Meanwhile, CTF-CTF also yields a result of the separation of the aorta in which the aorta is incorrectly segmented out in the irrelevant region. In this case, CTF-IPF and DSF-DSF show superior overall VS as well as segmentation results. It is worth noting that the VS does not represent the overlap between the segments; therefore, a segment result may contain a lower VS with a higher Dice score (CTF-CTF).

We also calculate the error of the MTD. Table 2 shows the error of the MTD between the predicted results and the ground truth. It shows that CTF-IPF achieves the best results on D_2 (4.80 ± 7.12 mm), whereas CTF-CTF yields the best overall performance (3.95 ± 0.95 mm). Figure 6 illustrates the fitted ellipse and the MTD of the samples from Fig. 4, it is observed that both FPs and FNs introduce nuisance variability on the evaluation of the MTD. For example, the tiny FP in 3D-CA (P1-17) consequently yields an unexpected larger transverse diameter; in

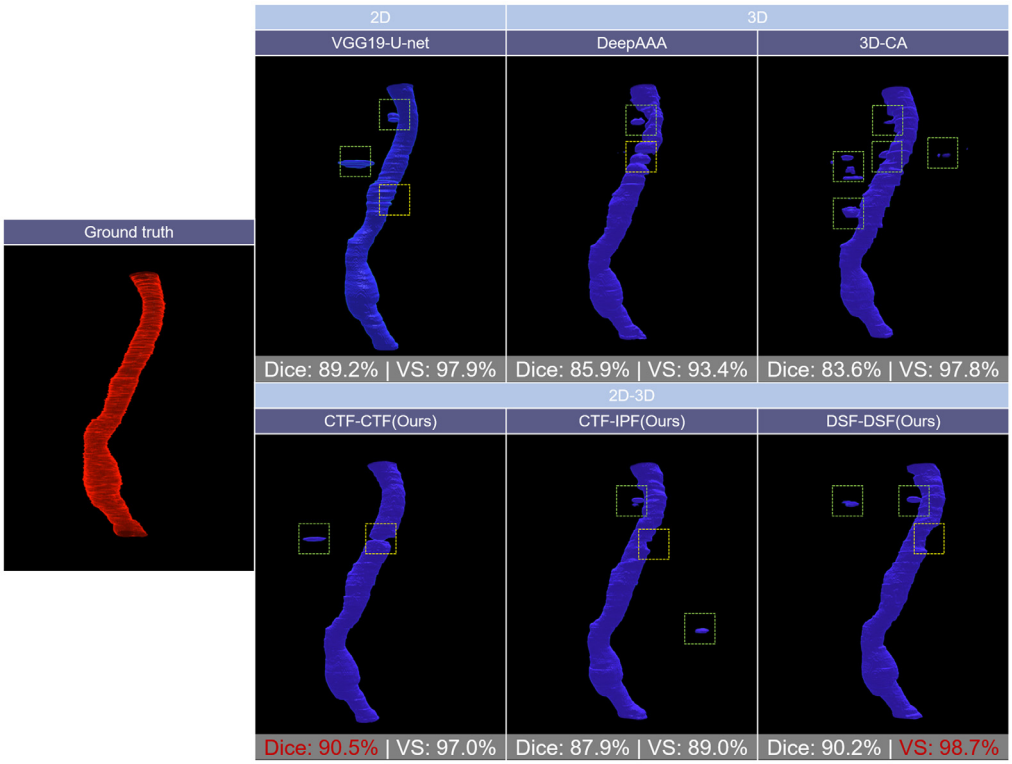

Fig. 5 Visualization of 3D reconstruction of the aorta (from patient-16). The Dice score and VS related to each volume are marked. The green- and yellow-dotted boxes denote the FPs and FNs, respectively.

Table 2 The error of the MTD between the predicted results and the ground truth. The bold values represent the optimal results of their columns.

	MTD error of aortic (mm)			
Model	D_0	D_1	D_2	Mean ± *Std*
VGG19-U-Net[42]	4.00±1.65	7.02±4.27	5.78±4.56	5.60±1.24
DeepAAA[29]	4.72±2.46	5.09±3.67	7.98±12.05	5.93±1.79
D-CA[28]	9.56±7.17	8.24±6.37	10.38±16.98	9.54±1.30
CTF-CTF (ours)	**2.95±1.06**	**4.07±2.30**	4.84±5.52	**3.95±0.95**
CTF-IPF (ours)	3.46±1.31	4.93±2.96	**4.80±7.12**	4.39±0.81
DSF-DSF (ours)	3.51±1.91	4.27±3.14	4.98±7.58	4.25±0.73

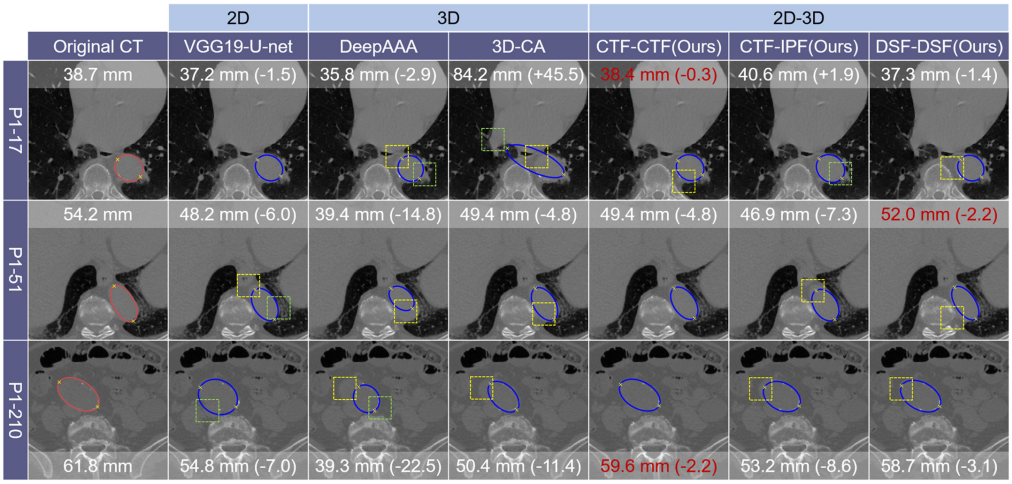

Fig. 6 Visualization of the evaluation of ellipse fitting and the related MTD of the aorta. The red and dark blue contours represent the ellipse fitting of the ground truth and predicted results, respectively. The two yellow crosses of each contour mark the long axis of the ellipse, where the length and error are also shown. The green and yellow dotted boxes denote the FPs and FNs, respectively.

addition, the false negative (FN) observed in P1-210 leads to the shortened evaluated transverse diameter. All of the evidence reveals that a tiny FP/FN may not largely affect the overlap-based metrics, but it is prone to yielding large negative effects on the evaluation of the MTD. The overall low degree of error on the MTD of our approaches shows the capability of eliminating the FPs/FNs.

5.2 *Comparison with Generic CNN-based Methods*

We also compare our method with several state-of-the-art CNN-based methods not specifically designed for aorta segmentation; they are based on 2D (Res-U-Net[50]), 3D (V-net,[21] 3D Res-U-Net[22]) and 2D-3D cascaded (LW-HCN[41]) mechanisms. To make the results comparable, we assign the compared methods identical training strategies and optimization as ours. Table 3 gives the average performance of three-fold cross-validation obtained from the state-of-the-art methods and ours. With moderate memory utilization and fps, our methods yield the best overall performance in comparison with all three kinds of methods in various metrics. In particular, for the overlap-based ones, DSF exceeds Res-U-Net by 1.6% and 1.8% on DSC and SEN, respectively. Surprisingly, the 2D method shows superior performance compared with the

Table 3 Performance comparison among our and the state-of-the-art methods based on five evaluation metrics. For ours, the form *X-X* means the mechanism of encoder and decoder, respectively. For example, "CTF-IPF" represents a model that contains CTF in all of the EnBlocks and IPF in all of the DeBlocks. The bold values represent the optimal results of their columns.

	Model	Memory	fps	DSC (%)	SEN (%)	HD	VS (%)	MTD error (mm)
2D	Res-U-Net[50]	8.9M	147.3	87.1±1.9	87.6±2.8	8.6±1.4	**96.9±0.7**	5.2±1.4
3D	V-Net[21]	45.6M	48.2	76.4±2.2	78.0±3.2	10.8±4.0	78.6±6.6	9.4±2.5
	3D Res-U-Net[22]	76.6M	41.5	86.0±0.9	86.4±1.7	10.6±2.2	96.2±1.3	6.8±1.2
2D-3D	LW-HCN[41]	2.5M	128.9	84.6±1.1	82.1±2.7	25.0±7.6	86.8±1.7	17.7±5.6
Ours	CTF-CTF	15.9M	64.3	88.5±2.1	**89.8±2.6**	7.6±1.7	96.7±0.7	**4.0±1.0**
	CTF-IPF	10.5M	59.5	88.3±1.6	88.0±0.7	8.1±1.1	96.5±0.6	4.4±0.8
	DSF-DSF	20.2M	42.3	**88.7±1.3**	89.4±1.7	**7.4±1.0**	**96.9±0.7**	4.3±0.3

3D and 2D-3D cascaded ones; we hypothesize that the intra-slice contextual information contributes more to the decision making for the segmentation.

5.3 *Ablation Studies*

5.3.1 *Performance of 2D/3D mechanisms in encoder/decoder modules*

To reveal the effectiveness of the mechanisms of 2D/3D feature fusion, ablation studies of various combinations of mechanisms was conducted. Table 4 summarizes the performance of different combinations. According to Table 4, the following points are suggested.

1. Naive 2D/3D in encoder/decoder combinations have a negative impact on the performance. In baseline models, both the "2D-3D" and "3D-2D" mechanisms yield inferior performance to the "2D-2D"[50] and "3D-3D" models.[22] It illustrates that a 2D/3D combination using encoder/decoder as the base unit degrades the performance. We hypothesize that the different mechanisms prevent the decoder from making effective use of the semantic context information provided by the encoder.
2. Leveraging any fusion mechanisms of CTF, IPF, or DSF in an encoder/decoder outperforms pure 2D/3D mechanisms. It shows the effectiveness of the fusion mechanisms in improving the performance by the gap between our approaches with the baseline.
3. Using the fusion mechanisms in both the encoder and decoder generates a better performance than using them separately in an encoder or decoder. It shows that stacking these mechanisms positively affects the model.

5.3.2 *Performance in relation to the number of training data*

We conducted experiments with different numbers of training data to assess the performance of 2D, 3D, and 2D-3D fusion approaches. Figure 7 shows the overall performance of three mechanisms related to the numbers of training data n. It can be observed that, when the number of training data is small ($n = 5$ or $n = 10$), the 2D mechanism generates a superior performance over the 3D and 2D-3D fusion approaches. However, as the number of training data accumulates, the performances of the 3D and 2D-3D fusion methods increase rapidly. The 2D-3D fusion method shows the best performance when the number of training data $n \geq 15$. We hypothesize that, when the amount of training data is small, the 3D and 2D-3D fusion approaches are prone to over-fitting because their constructions are more complex with more trainable weights than 2D models (note that the number of trainable weights of the 2D method is 8.9M whereas there are 20.2M for CTF-CTF and 76.6M for the 3D method according to Table 4). Therefore, the

Table 4 Performance comparison for ablation studies of various combinations of mechanisms. For each approach, the form "*X*-*Y*" means the model contains a mechanism "*X*" in the encoder and "*Y*" in the decoder. The "baseline" represents the pure 2D/3D mechanisms. For example, the "2D-3D" means the training model contains pure 2D modules in the encoder and pure 3D modules in the decoder (with residual path). For our approaches, the "base" means the pure 2D mechanism. For example, "CTF-base" represents a training model containing CTF in the encoder and pure 2D modules in the decoder. The bold values represent the optimal results of their columns.

	Model	Memory	fps	DSC (%)	SEN (%)	HD
Baseline	2D-2D[50]	8.9M	147.3	87.1±1.9	87.6±2.8	8.6±1.4
	2D-3D	16.8M	67.8	84.5±2.7	80.3±4.4	22.3±9.9
	3D-2D	19.8M	43.2	85.5±1.9	85.8±2.4	16.5±2.4
	3D-3D[22]	76.6M	41.5	86.0±0.9	86.4±1.7	10.6±2.4
Ours	CTF-Base	4.9M	209.9	88.0±1.1	88.8±1.4	8.3±0.8
	Base-CTF	14.5M	77.9	87.9±0.9	88.9±1.2	9.9±0.3
	CTF-CTF	15.9M	64.3	88.5±2.1	**89.8±2.6**	7.6±1.7
	Base-IPF	9.0M	78.2	87.5±1.4	89.1±2.7	8.6±1.2
	CTF-IPF	10.5M	59.5	88.3±1.6	88.0±0.7	8.1±1.1
	DSF-Base	4.4M	136.5	88.0±0.4	88.0±2.7	8.6±1.1
	Base-DSF	19.3M	86.3	88.6±1.2	88.9±1.4	7.6±0.6
	DSF-DSF	20.2M	42.3	**88.7±1.3**	89.4±1.7	**7.4±1.0**

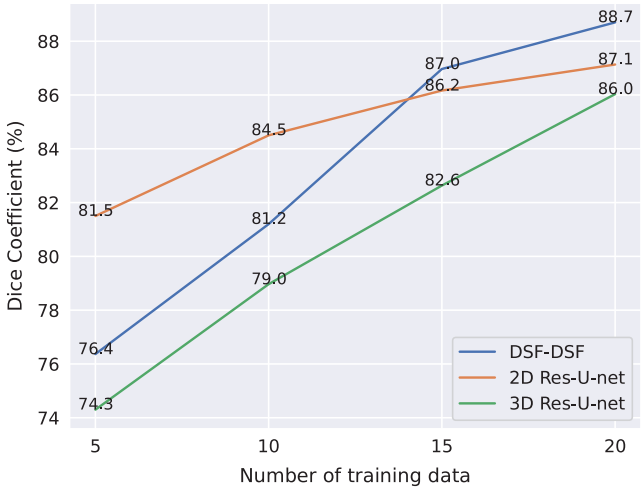

Fig. 7 Overall performances of the 2D, 3D, and 2D-3D fusion approaches in relation to different numbers of training data *n*.

mechanism of 2D-3D fusion generates superior performance when the number of training data is adequate.

5.4 *Analysis of the Results of Individual Patients*

To have a sufficient understanding of the performance of our methods, we further analyze the results shown on the ordered slices of individual patients. Figure 8 illustrates the results of a

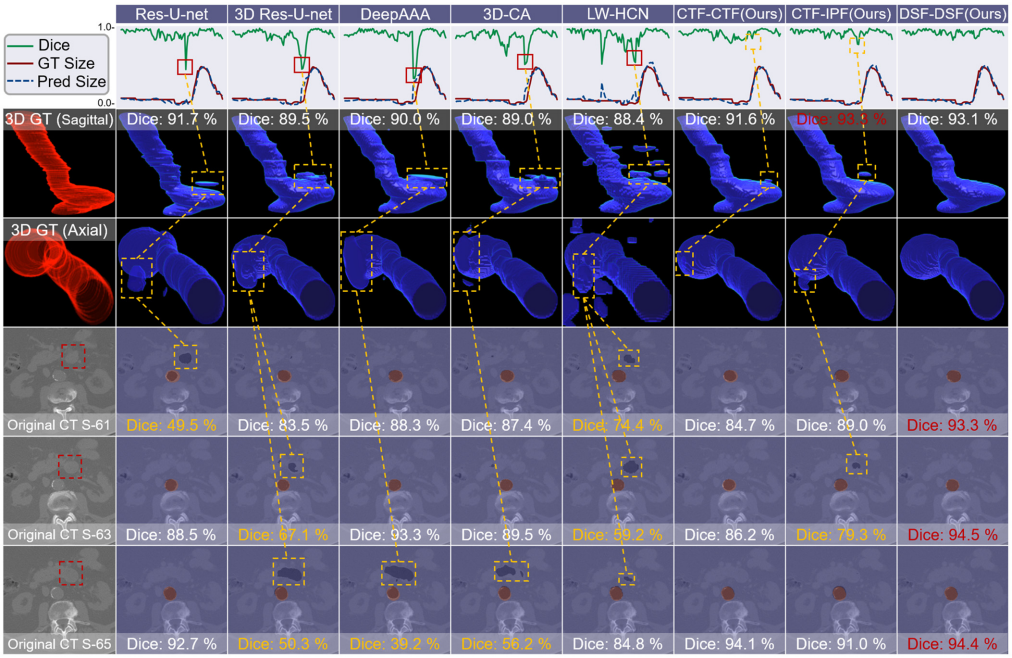

Fig. 8 Results of a single patient obtained from various methods (sampled from patient-27). The first row: the variation of the Dice score and the size of the aorta. For each line chart, the *x*-axis represents the ordered indexes of the slices, the *y*-axis is a unit range where the green line is the variation of Dice, and the red and blue dotted lines are the normalized sizes of the aorta, which are obtained from the ground truth and prediction, respectively. The red boxes mark the sudden decline of the Dice score where the region is close to the enlargement of the aorta. The second and third rows: the 3D visualization of the results from the perspectives of sagittal and axial, respectively. The Dice score of the whole volume is denoted. The last three rows: the visualization of 2D original CTs and the segmentation results related to the Dice-declined parts. For original CTs, the red dotted boxes mark the areas that are peculiarly prone to being falsely predicted as aortas because of their shapes and pixel intensities. For the results, the red and dark blue overlays represent the ground truth and predicted results, respectively. The highest Dice score (red) and the ones lower than 80% (orange) are marked. The yellow dotted boxes represent the FPs of 2D and 3D forms related to the decline of Dice.

single patient obtained from various methods. The first row indicates that, except for ours, each method yields a sudden decline of the Dice score (red-solid box) in which the size of the aorta starts to turn into an enlargement. The segmentation results related to the decline are visualized on its 3D and 2D form. By observing a single line chart from left to right and its related 3D visualization from top to bottom, we realize that the decline is caused by the FPs that appear at the region where the size of the aorta varies rapidly along the axial direction (yellow dotted boxes), i.e., the aorta with a regular size initially turns into an aneurysm. We hypothesize that it is induced by the inability to capture the inter-slice contextual information in this region. By contrast, ours, especially for DSF, eliminates most of the FPs.

Furthermore, in Fig. 8, each FP of 3D visualization is accompanied by its related 2D shape. Comparing the original CT slices and the results of segmentation, we observe that all of the FPs appear at the region where the surrounding tissues resemble the aorta in the aspect of size, shape, and pixel intensity (red dotted boxes), which we regarded as the main difficulties for aorta segmentation on non-contrast-enhanced CTs in Sec. 1. The compared state-of-the-art methods generate different degrees of FPs among these regions (yellow-dotted boxes); thus it consequently leads to a decline of the Dice score. However, for our methods, except for one case of CTF-IPF, all of the results indicate that the three mechanisms of feature fusing are able to overcome the mentioned difficulties within a single slice.

To summarize, according to the analysis of the results of the ordered slices of an individual patient, our methods have more ability to deal with the difficulties that occur in both the inner

and intra-slices. We attribute it to the fusion of 2D and 3D features, which facilitates utilizing the intra-/inter-slice contextual information.

5.5 *Analysis of the Variation of Dice Coefficient on Multi-sized Cross Sections of Aorta*

We also analyze the performance of various methods in terms of VCS ratios r, which is introduced in Sec. 4.3. Generally, for a CT slice S, a larger r means that the location of the aortic cross section is closer to the aneurysm or that the cross section area $a(S)$ is larger; otherwise the opposite is the case. Figure 9 shows an abdominal aorta structure with various cases of VCS ratios r and a heat map for a visual explanation. Each cross section S of the aorta possesses a VCSr r, where the VCSr of the largest cross section is $r_M = 1$. Figure 10 shows the overall performance on all validation sets, which are divided into four parts according to the VCS ratio r. It indicates that, for all of the methods, segmenting the cross-section of an aorta with a smaller r is more challenging. This phenomenon can be partially observed from the 3D segmentation results of Fig. 5, in which a large number of FPs/FNs appear at locations away from the enlargement of the aorta, and in these locations, the cross-sections of the aorta are relatively small. We suppose it is related to the size imbalance mentioned in Sec. 1. In this experiment, our methods show relative superiority within the parts of $r \in (0.0, 0.25]$ (85.1% versus 82.1%) and $r \in (0.75, 1.0]$ (91.5% versus 90.5%), which manifests that they generate a more competitive performance in the parts of the aorta where the r is extremely small or large.

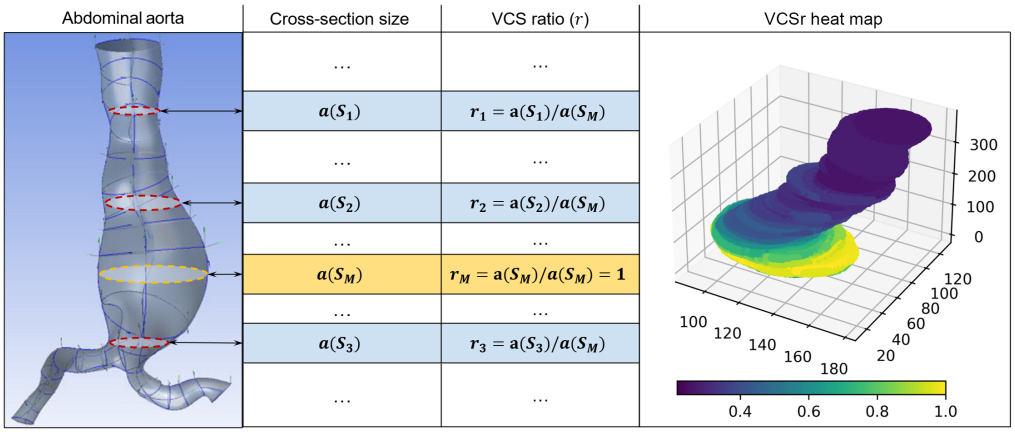

Fig. 9 Illustration of an abdominal aorta structure with various cases of VCS ratios r of different cross sections S, where $a(S)$ represents the cross section area and S_M is the cross section possessing the largest area. A structure of an aorta with its VSC ratios in a heat-map pattern is shown on the right. For each cross section of the aorta, the VCS ratio $r \in (0,1]$. The darker cross section means it has a smaller VSC ratio r, whereas the lighter cross-section represents a larger VSC ratio r.

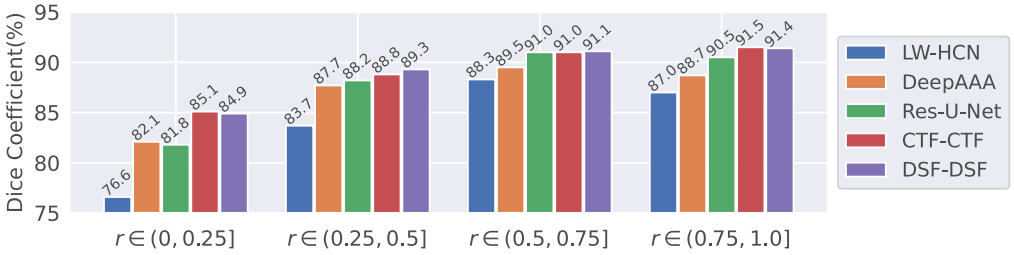

Fig. 10 Overall performance on aorta parts with different VCS ratios r.

6 Discussion

Segmentation of the abdominal aorta in non-contrast CTs is challenging, especially in the cases of AAAs. The problems of low contrast, similar topological form, and size imbalance hamper it. In previous research, Suzuki et al.[42] conducted the segmentation of the aorta and main pulmonary artery by a voting mechanism to merge several conventional 2D and 3D CNN models. Although they also focused on the non-contrast CTs, there was no case of AAA. Lu et al. proposed DeepAAA[29] to treat the task in a pure 3D pattern. Chandrashekar et al. proposed a 3D cascaded attention-based[28] model that introduced additive attention gates with coarse-fine segmentation. These two studies partially focus on the segmentation of the abdominal aorta in non-contrast CTs. Our study is entirely devoted to the segmentation of the abdominal aorta in non-contrast CTs, in which AAA is present in each sample. Different from the related methods, we proposed a 2D-3D feature fusing method implemented by three kinds of fusion mechanisms. These fusion mechanisms leverage the intra- and inter-slice context information.

Compared with the pure 2D and 3D methods, ours generates superior performances in the aortas of various cross-section shapes, e.g., the aortas of regular, irregular-shaped, and large-sized patterns (Figs. 4 and 6). According to Table 3, DSF-DSF generates the best results in terms of DSC, HD, and VS, whereas CTF-CTF presents the best SEN that represents the highest true positive (TP) rate. Therefore, CTF-CTF also yields the best result of the MTD associated closely with TP. The ablation studies proved the flexibility and effectiveness of the fusion mechanisms (Table 4). In addition, it is suggested that the fusion mechanisms can eliminate the risks of FPs caused by the diversity between adjacent CT slices (Fig. 8), which is supposed to be one of the reasons for the improvement of the performance. Compared with the related methods, ours mainly raised the performance in the slices in which the cross-section size of the aorta is relatively small (containing a smaller VCSr r). We hypothesize that the small-sized cross sections are more challenging to segment well because of our proposed size imbalance. However, with various mechanisms of feature fusing by leveraging the intra- and inter-slice context, the performance of segmentation in this part is improved significantly (Fig. 10).

There are also certain limitations of our method. According to Fig. 7, it is prone to over-fitting when the training data is extremely inadequate, which is caused by the more complex model structures compared with the conventional 2D models. In addition, even for our method, there is still a considerable gap between the performance on the small and large cross sections of aortas. Resorting to multi-scale learning or weighted loss functions could be further solutions.

Furthermore, we analyze the intra- and inter-observer variability, the significant difference, and the visualized feature maps in the following subsections.

6.1 *Intra- and Inter-observer Variability Assessment*

We assess intra- and inter-observer variability in terms of Dice score as we mentioned in the Sec. 4.1. Table 5 summarizes the Dice scores for intra- and inter-observer variability according to different VCS ratios (r). It shows that the overall inter-observer variability is greater than the intra-observer variability (as seen by the Dice score of $96.1\% \pm 1.4$ being lower than $96.6\% \pm 1.1$). Generally, the Dice scores of both variabilities increase with VCSr, which is consistent with our analysis in the Sec. 5.5. These results suggest that aortas with small cross-

Table 5 Dice score (%) for intra-/inter-observer in terms of different VCS ratios (r).

VCSr [(r)]	Intra	Inter
$r \in (0, 0.25]$	95.8±1.1	96.0±1.2
$r \in (0.25, 0.50]$	96.6±1.6	95.5±1.2
$r \in (0.50, 0.75]$	97.2±0.6	96.6±0.8
$r \in (0.75, 1.00]$	97.4±0.4	96.9±1.0
Overall	96.6±1.1	96.1±1.4

sections are more difficult to segment, either manually or fully automatically. For manual segmentation, in our practice, aortas with smaller cross-sections are more challenging to be delineated because the boundary between the aorta and the surrounding tissue is visually more ambiguous, which makes it more difficult for the observers to determine the location of the boundaries. Even so, the results of Table 5 strongly support the reliability of the manual segmentation used for training and testing.

6.2 *Significant Difference between Methods*

We leverage the pairwise Wilcoxon Rank Sum Test to provide p-values to reflect the significant differences. Figure 11 shows the heat map of p-values of the results of different methods in terms of Dice score and sensitivity, where each 2D slice is regarded as an individual sample. According to Fig. 11, the p-values between the Dice scores of the proposed methods are larger than 0.05, which represents less significant differences in terms of the Dice score. However, the p-values between the sensitivities of proposed methods are much <0.05. This means that the significant differences of the proposed methods mainly present in sensitivity, which is related to the ability of positive test. In addition, the proposed method shows significant differences from related method[50] in terms of both the Dice score and sensitivity.

6.3 *Visualization of Features Maps*

We exhibit the intermediate results to analyze the details of the fusion of features. Figs. 12 and 13 illustrate the comparison of feature maps of 2D, 3D, and 2D-3D fusion approaches; Fig. 12 shows the feature maps of Res-U-Net (2D) and DSF-DSF, and Fig. 13 shows the feature maps of DeepAAA (3D) and DSF-DSF. For each approach, observing from encoders to decoders, the spatial information first decreased due to down-sampling and then recovered by up-sampling. Meanwhile, semantic information is increasing continuously, with the active regions of the aorta gradually becoming remarkable.

Comparing the features from the three methods, the feature space of the 2D-3D method is visually richer than that of the 2D and 3D methods (e.g., the features of DeBlock1); we assume that it sufficiently explores both the 2D and 3D domains rather than the separate 2D and 3D paths in the other two methods. Furthermore, focusing on the feature maps of the 2D-3D method, we

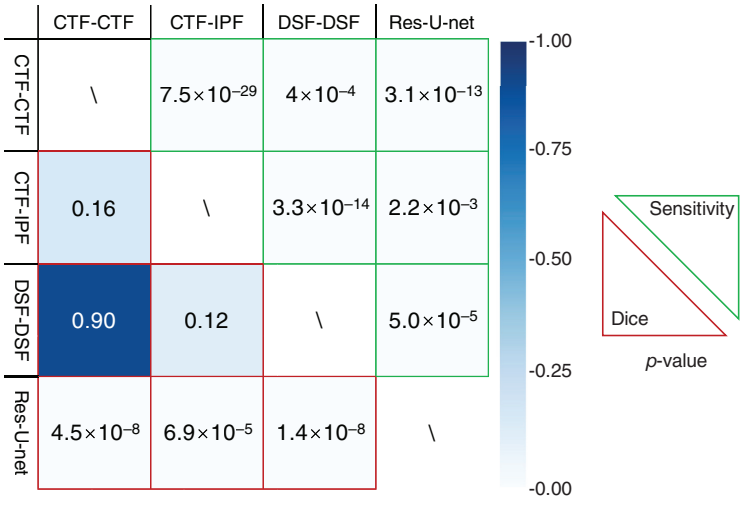

Fig. 11 Heat map of p-values of the results of different methods in terms of the Dice score and sensitivity, where each 2D slice is regarded as an individual sample. The red cells show p-values of Dice scores between a paired method, and the green cells show the p-values of sensitivities. For example, the third row of the first column shows that the p-value between the Dice scores of CTF-CTF and DSF-DSF is 0.90, and the first row of the third column means that the p-value of their sensitivity is 4×10^{-4}.

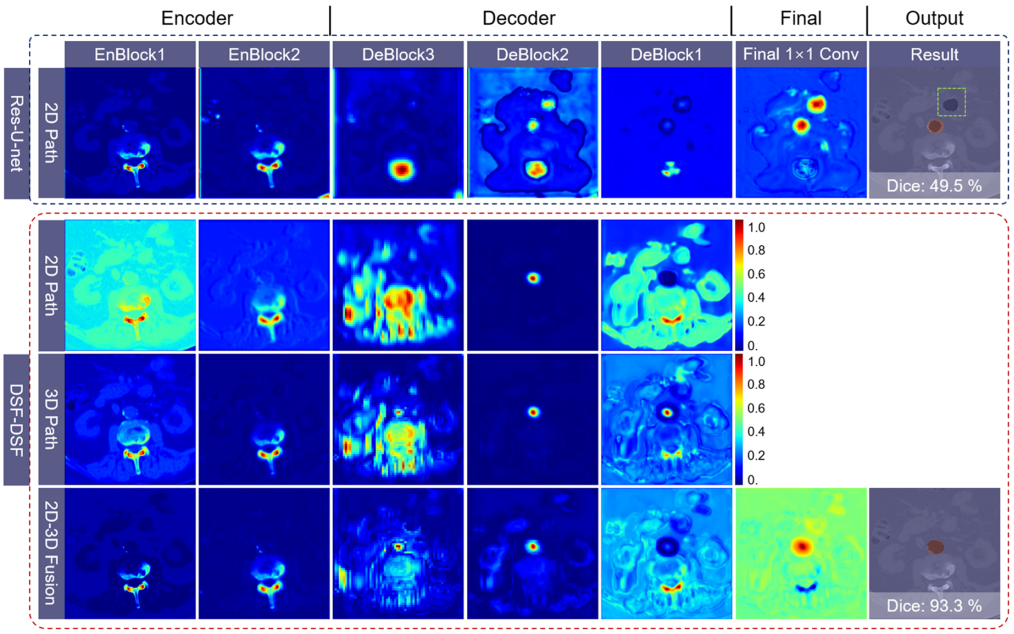

Fig. 12 Feature maps of Res-U-Net (blue-dotted box) and DSF-DSF (red-dotted box), sampled from Patient-27, slice-61. From left to right, we sample the output feature maps from six stages during the workflow, i.e., encoder (EnBlock1-2), decoder (DeBlcok1-3), and final convolutional layer. All of the multi-channel feature maps are applied channel-wise mean and 0-1 normalization for visualization. For the visualization of 3D feature maps, we extract the related 2D feature from the axial direction. The feature maps with different spatial sizes are resized to 256×256 for visualization. The final results and the Dice score are attached. The green-dotted boxes denote the FPs.

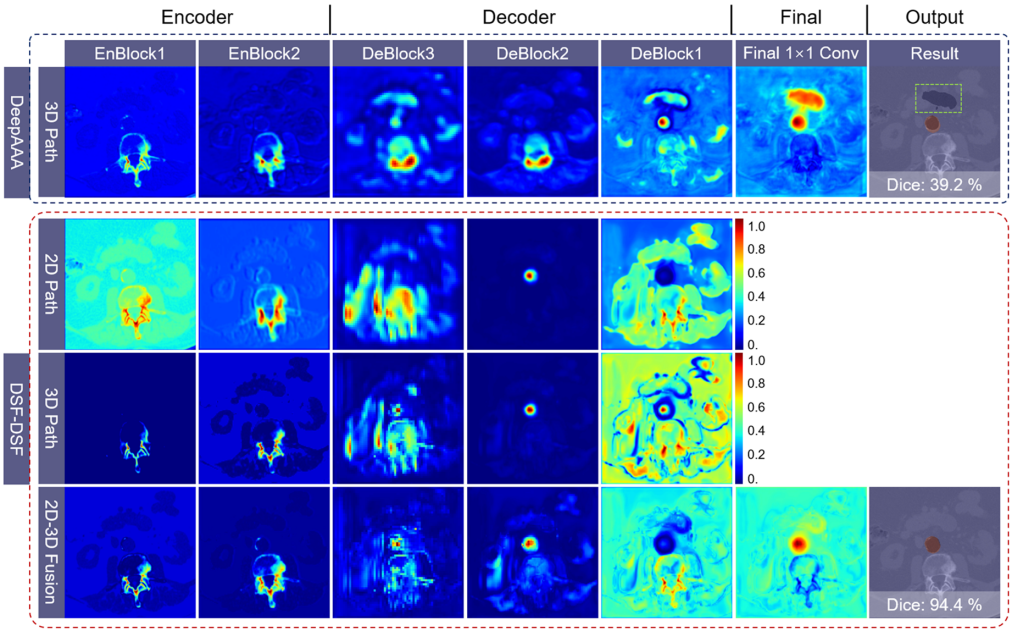

Fig. 13 Feature maps of DeepAAA (blue-dotted box) and DSF-DSF (red-dotted box), sampled from patient-27, slice-65. From left to right, we sample the output feature maps from six stages during the workflow, i.e., encoder (EnBlock1-2), decoder (DeBlcok1-3), and final convolutional layer. All of the multi-channel feature maps are applied channel-wise mean and 0-1 normalization for visualization. For the visualization of 3D feature maps, we extract the related 2D feature from the axial direction. The feature maps with different spatial sizes are resized to 256×256 for visualization. The final results and the Dice score are attached. The green dotted boxes denote the FPs.

observe that the fusion of the features enriches the semantic information. For example, in DeBlock3, the feature maps of the 2D path and 3D path display large and non-targeted active regions. However, the fused feature highlights the active area of the aorta. The 2D and 3D paths are supposed to yield different representations for the same spatial area. The fusion takes full advantage of these representations according to particular mechanisms rather than simple superposition. Furthermore, the segmentation results of the 2D-3D method show fewer FPs than the separate 2D and 3D approaches. We hypothesize that one of the reasons is that the fusion mechanism eliminates FPs from the 2D or 3D paths.

7 Conclusion

The segmentation of the aorta in non-contrast CTs is impeded by the difficulties of low contrast, similar shape, and size imbalance. To address these issues, this study proposes a mechanism to integrate the 2D and 3D features of the convolutional neural network, with application to non-contrast CTs in the context of EVAR. In particular, the mechanism of integration is featured as the fusion of features based on concatenation, dense connection, and interpolation. These simple but efficient mechanisms provide novel insights into the construction of CNN models. In addition, extensive experiments on our dataset of non-contrast CTs demonstrate the superiority of our methods, especially in low-contrast, similar-shaped, and extreme-sized cases. The best results were obtained from the DSF approach. Our future work will deal with the application of the proposed methods to various types of vascular structures.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgment

This study was partially supported by the French National Research Agency (ANR) in the framework of the Investissement d'Avenir Program through Labex CAMI (ANR-11- LABX-0004). This work was supported in part by the National Key Research and Development Program of China (Grant No. 2022YFE0116700). The first author is grateful for the support of China Scholarship Council (Grant No. 201906090389).

References

1. G. H. Gibbons and V. J. Dzau, "Molecular therapies for vascular diseases," *Science* **272**(5262), 689–693 (1996).
2. W. P. Robinson et al., "Derivation and validation of a practical risk score for prediction of mortality after open repair of ruptured abdominal aortic aneurysms in a US regional cohort and comparison to existing scoring systems," *J. Vasc. Surg.* **57**(2), 354–361 (2013).
3. S. R. Walsh et al., "Renal consequences of endovascular abdominal aortic aneurysm repair," *J. Endovasc. Ther.* **14**(1), 92–100 (2008).
4. M. V. Parker et al., "What imaging studies are necessary for abdominal aortic endograft sizing? A prospective blinded study using conventional computed tomography, aortography, and three-dimensional computed tomography," *J. Vasc. Surg.* **41**(2), 199–205 (2005).
5. C. Caradu et al., "Fully automatic volume segmentation of infrarenal abdominal aortic aneurysm computed tomography images with deep learning approaches versus physician controlled manual segmentation," *J. Vasc. Surg.* **74**(1), 246–256 (2021).
6. A. Kaladji et al., "Interest of fusion imaging and modern navigation tools with hybrid rooms in endovascular aortic procedures," *J. Vasc. Surg.* **58**(3), 458–466 (2017).
7. A. Kaladji et al., "Sizing for endovascular aneurysm repair: clinical evaluation of a new automated three-dimensional software," *Ann. Vasc. Surg.* **24**(7), 912–920 (2010).

8. R. Mehran and E. Nikolsky, "Contrast-induced nephropathy: definition, epidemiology, and patients at risk," *Kidney Int. Suppl.* **100**, S11–S15 (2006).
9. R. Solomon, "Contrast media nephropathy—how to diagnose and how to prevent?" *Nephrol. Dial. Transplant.* **22**(7), 1812–1815 (2007).
10. A. Kaladji et al., "Safety and accuracy of endovascular aneurysm repair without preoperative and intra-operative contrast agent," *Eur. J. Vasc. Endovasc. Surg.* **49**(3), 255–261 (2015).
11. M. de Bruijne et al., "Active-shape-model-based segmentation of abdominal aortic aneurysms in CTA images," *Proc. SPIE* **4684**, 463–474 (2002).
12. M. Dhibi et al., "Détection des Contours des Thrombus Veineux dans les Images Echographiques," SETIT: Sciences Electroniques, Technologies de l'Information et des Télécommunications (2005).
13. Y. Chen et al., "Segmentation of the thrombus of giant intracranial aneurysms from CT angiography scans with lattice Boltzmann method," *Med. Image Anal.* **18**(1), 1–8 (2014).
14. M. Freiman et al., "An iterative model-constrained graph-cut algorithm for abdominal aortic aneurysm thrombus segmentation," in *Proc. IEEE Int. Symp. Biomed. Imaging: From Nano to Macro*, IEEE, pp. 672–675 (2010).
15. K. Lee et al., "Three-dimensional thrombus segmentation in abdominal aortic aneurysms using graph search based on a triangular mesh," *Comput. Biol. Med.* **40**(3), 271–278 (2010).
16. Y. LeCun et al., "Deep learning," *Nature* **521**(7553), 436–444 (2015).
17. G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.* **42**, 60–88 (2017).
18. Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, Vol. **3361**, M.A. Arbib, Ed., MIT Press (1995).
19. O. Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
20. ö. çiçek et al., "3D U-net: learning dense volumetric segmentation from sparse annotation," *Lect. Notes Comput. Sci.* **9901**, 424–432 (2015).
21. F. Milletari et al., "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, pp. 565–571 (2016).
22. L. Yu et al., "Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images," in *Proc. 31st AAAI Conf. Artif. Intell.*, pp. 240–246 (2017).
23. H. A. Hong and U. U. Sheikh. "Automatic detection, segmentation and classification of abdominal aortic aneurysm using deep learning," in *Proc. IEEE 12th Int. Colloquium on Signal Process. & its Appl. (CSPA)*, March, IEEE, pp. 242–246 (2016).
24. J. Q. Zheng et al., "Abdominal aortic aneurysm segmentation with a small number of training subjects," arXiv:1804.02943 (2018).
25. J. Chen et al., "A deep learning algorithm using contrast-enhanced computed tomography (CT) images for segmentation and rapid automatic detection of aortic dissection," *Biomed. Signal Process. Control* **62**, 102145 (2020).
26. M. Habijan et al., "Abdominal aortic aneurysm segmentation from CT images using modified 3D U-net with deep supervision," in *Proc. Int. Symp. ELMAR*, IEEE, pp. 123–128 (2020).
27. K. López-Linares et al., "Fully automatic detection and segmentation of abdominal aortic thrombus in post-operative CTA images using deep convolutional neural networks," *Med. Image Anal.* **46**, 202–214 (2018).
28. A. Chandrashekar et al., "A deep learning approach to automate high-resolution blood vessel reconstruction on computerised tomography images with or without the use of contrast agents," *Eur. Heart J.* **41**(Supplement2), ehaa946-0154 (2020).
29. J.T. Lu et al., "DeepAAA: clinically applicable and generalizable detection of abdominal aortic aneurysm using deep learning," *Lect. Notes Comput. Sci.* **11765**, 723–731 (2019).
30. A. A. Duquette et al., "3D segmentation of abdominal aorta from CT-scan and MR images," *Comput. Med. Imaging Graph.* **36**(4), 294–303 (2010).

31. S. Kurugol et al., "Aorta segmentation with a 3D level set approach and quantification of aortic calcifications in non-contrast chest CT," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, IEEE, pp. 2343–2346 (2012).
32. K. Tokunaga and H. Hanaizumi, "Extraction of the thoracic aorta territory and aneurysm from CT images," in *SICE Annu. Conf.*, IEEE, Vol. 3 (2004).
33. Z. S. Gamechi et al., "Aorta and pulmonary artery segmentation using optimal surface graph cuts in non-contrast CT," *Proc. SPIE* **10574**, 105742D (2018).
34. O. Oktay et al., "Attention U-net: learning where to look for the pancreas," arXiv:1804.03999 (2018).
35. Z. Zhou et al., "UNet++: A nested U-Net architecture for medical image segmentation," *Lect. Notes Comput. Sci.* **11045**, 3–11 (2018).
36. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June, pp. 770–778 (2016).
37. H. Chen et al., "VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage* **170**, 446–455 (2018).
38. J. Dolz et al., "HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imaging* **38**(5), 1116–1126 (2019).
39. F. Isensee et al., "NnU-Net: self-adapting framework for U-net-based medical image segmentation," arXiv:1809.10486 (2018).
40. X. Li et al., "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imaging* **37**(12), 2663–2674 (2018).
41. J. Zhang et al., "Light-weight hybrid convolutional network for liver tumor segmentation," *IJCAI Int. Jt. Conf. Artif. Intell.* **19**, 4271–4277 (2019).
42. H. Suzuki et al., "Segmentation of aorta and main pulmonary artery of non-contrast CT images using U-Net for chronic thromboembolic pulmonary hypertension: evaluation of robustness to contacts with blood vessels," *Proc. SPIE* **12033**, 560–565 (2022).
43. D. Ulyanov et al., "Instance normalization: the missing ingredient for fast stylization," arXiv:1607.08022 (2016).
44. S. Ioffe et al., "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. ICML*, pp. 448–456 (2015).
45. G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June, pp. 4700–4708 (2017).
46. G. Hinton et al., "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent," Cited on 14(8), 2 (2012).
47. K. He et al., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June, pp. 1026–1034 (2015).
48. A. Paszke et al., "Pytorch: an imperative style, high-performance deep learning library," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, December, Vol. 32, pp. 8026–8037 (2019).
49. A. Fitzgibbon et al., "Direct least squares fitting of ellipses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, May, IEEE, pp. 476–480 (1999).
50. Z. Zhang et al., "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.* **15**(5), 749–753 (2018).

Qixiang Ma is currently a PhD student at Université de Rennes 1, Signal and Image Processing Laboratory. He received his MS and his BS degrees in computer science and technology from Southeast University, Nanjing, China in 2020 and 2017, respectively. His research interests are in the domain of deep-learning-based medical image processing and computer-aided diagnosis.

Biographies of the other authors are not available.