Unleashing Vision Foundation Models for Coronary Artery Segmentation: Parallel ViT-CNN Encoding and Variational Fusion

Caixia Dong[1,2], Duwei Dai[2], Xinyi Han[3], Fan Liu[2], Xu Yang[2], Zongfang Li[1,2(✉)], and Songhua Xu[1,2(✉)]

[1] National-Local Joint Engineering Research Center of Biodiagnosis & Biotherapy, the Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710004, China
[2] Institute of Medical Artificial Intelligence, the Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710004, China
lzf2568@xjtu.edu.cn, songhuaxu@mail.xjtu.edu.cn
[3] Viadrina European University, Frankfurt (Oder), 15230, Germany

Abstract. Accurate coronary artery segmentation is critical for computer-aided diagnosis of coronary artery disease (CAD), yet it remains challenging due to the small size, complex morphology, and low contrast with surrounding tissues. To address these challenges, we propose a novel segmentation framework that leverages the power of vision foundation models (VFMs) through a parallel encoding architecture. Specifically, a vision transformer (ViT) encoder within the VFM captures global structural features, enhanced by the activation of the final two ViT blocks and the integration of an attention-guided enhancement (AGE) module, while a convolutional neural network (CNN) encoder extracts local details. These complementary features are adaptively fused using a cross-branch variational fusion (CVF) module, which models latent distributions and applies variational attention to assign modality-specific weights. Additionally, we introduce an evidential-learning uncertainty refinement (EUR) module, which quantifies uncertainty using evidence theory and refines uncertain regions by incorporating multi-scale feature aggregation and attention mechanisms, further enhancing segmentation accuracy. Extensive evaluations on one in-house and two public datasets demonstrate that the proposed framework significantly outperforms state-of-the-art methods, achieving superior performance in accurate coronary artery segmentation and showcasing strong generalization across multiple datasets. The code is available at https://github.com/d1c2x3/CAseg.

Keywords: Coronary artery segmentation· Vison foundation model· Parallel encoding · Variational fusion.

1 Introduction

Coronary artery disease (CAD) is the most common type of heart disease and a leading cause of global mortality [20]. Given the significant clinical challenges

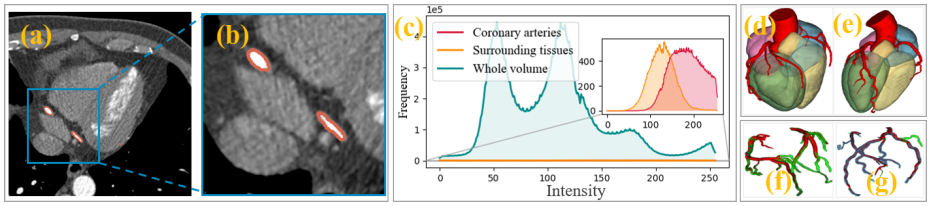

Fig. 1. The major challenges of accurate coronary artery segmentation. In (f) and (g), red, green, and cyan indicate correct, under, and over-segmentation, respectively.

associated with CAD, accurate imaging techniques are critical for early diagnosis and effective treatment planning. Coronary computed tomography angiography (CCTA) has become as the standard non-invasive modality for evaluating coronary artery anatomy and pathologies [15]. Precise segmentation of coronary arteries in CCTA images is essential for assessing stenosis severity, plaque morphology, and guiding clinical decision-making in CAD management.

Despite advancements in imaging technology, accurate coronary artery segmentation in CCTA images remains challenging due to several inherent factors: small vessel size (Fig. 1(a)-(b)), low contrast with surrounding tissues (Fig. 1(c)), and complex vascular morphology (Fig. 1(d)-(e)), all of which complicate the task of delineating vascular structures.

Deep learning has shown significant potential in coronary artery segmentation, offering improved scalability and accuracy. UNet and its variants remain foundational to many state-of-the-art models [19,5,7]. For example, 3D-FFR-UNet [19] improves feature fusion with dense convolutional blocks, while Dong et al. [7] leverage multi-scale attention to capture finer vessel details. However, while these convolutional neural network (CNN)-based methods are effective at extracting local features, they often struggle to preserve the anatomical continuity of vessels, resulting in fragmented and anatomically inconsistent segmentations, particularly in complex vascular regions (Fig. 1(f)). Vision transformer (ViT)-based approaches [26,10], in contrast, excel at modeling global structural features but often lack the spatial resolution needed to preserve fine-grained vascular details essential for delineating thin and tortuous vessels (Fig. 1(g)). Hybrid approaches that combine CNNs and ViTs offer promising solutions. For instance, Pan et al. [18] propose a cross-transformer network that integrates UNet for local features and Transformers for long-range dependencies. Similarly, Ensembled-SAMs [1] integrates nnU-Net [11] with SAMs [13] but rely on 2D slice processing and result merging, neglecting feature-level fusion and 3D inter-slice continuity.

In this work, we propose a novel segmentation framework that leverages the power of vision foundation models (VFMs) through a parallel encoding architecture (Fig. 2). **First**, the ViT encoder within the VFM [21] captures global structural features, enhanced by the activation of the final two ViT blocks and the integration of an attention-guided enhancement (AGE) module, which improves the model's ability to capture vascular continuity and topology; meanwhile, the CNN encoder extracts local details, ensuring a comprehensive representation.

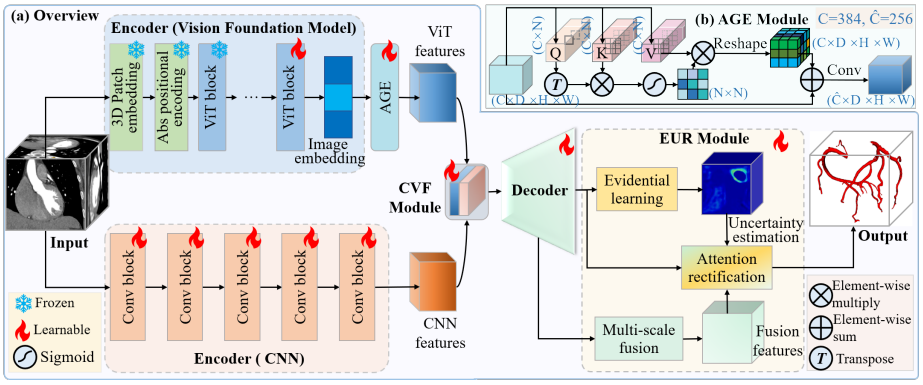

Fig. 2. Illustrates the pipeline of the proposed framework. It consists of three main components. (1) A parallel encoding architecture integrates a 3D foundation model-driven encoder and a 3D UNet-based shape encoder to extract 3D volumetric representations from global and local perspectives. (2) A CVF module adaptively fuses these features by modeling latent distributions and applying variational attention. (3) An EUR module refines predictions in uncertain regions to enhance segmentation accuracy.

Second, to effectively fuse global and local information, we introduce a cross-branch variational fusion (CVF) module, which models latent distributions and applies a variational attention mechanism to adaptively assign modality-specific weights. **Additionally**, we design the evidential-learning uncertainty refinement (EUR) module to quantify segmentation uncertainty using evidence theory and refine predictions by aggregating multi-scale features and attention mechanisms. These components collectively enhance segmentation accuracy and robustness, particularly in complex vascular structures.

2 Methodology

An overview of the proposed framework is shown in Fig. 2(a). It employs a parallel encoding architecture, combining a 3D foundation model-driven encoder (e.g., based on a pre-trained like SAM-Med3D [21]) with a 3D UNet-based shape encoder. These two encoders work in parallel to extract complementary 3D volumetric representations, capturing global contextual information and local structural details simultaneously. To enhance global feature perception, the final two ViT blocks are activated, and an AGE module is integrated (Fig. 2(b)), which leverages attention mechanisms [7] and a fusion layer to emphasize vascular continuity and morphology. Next, we introduce a CVF module to fuse these features obtained from the two encoders. Finally, an EUR module is introduced to refine predictions in uncertain regions.

The details of our method are elaborated in the following sections.

2.1 Cross-branch Variational Fusion Module

The CVF module is designed to integrate global and local features extracted from the ViT and CNN branches. The module comprises two core components: latent distribution learning and variational attention fusion.

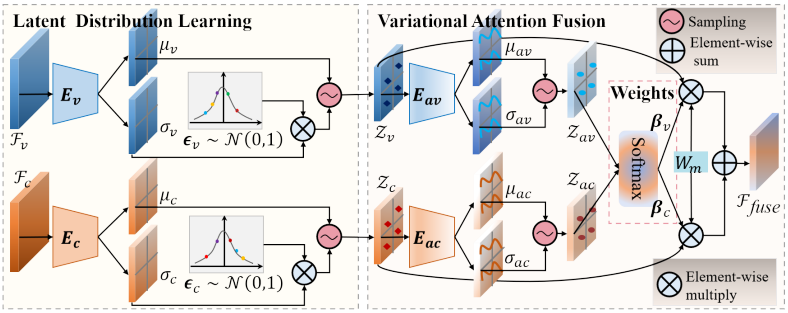

Fig. 3. Structure of the cross-branch variational fusion module. It integrates global and local features through latent distribution learning and variational attention fusion.Two encoders E_v and E_c parameterize the mean and variance of latent distributions, while E_{av} and E_{ac} compute adaptive weights β_v and β_c for feature fusion.

Latent Distribution Learning. The CVF module employs independent encoders, E_v and E_c, for the ViT and CNN branches to capture the inherent variability and complementarity of global and local features. These encoders utilize multi-layer perceptrons (MLPs) to parameterize the latent distributions of global ($\mathbf{F}_v$) and local ($\mathbf{F}_c$) features, modeling them as Gaussian distributions with learnable means and standard deviations:

$$\boldsymbol{\mu}_v = \mathrm{MLP}(\mathbf{F}_v), \quad \boldsymbol{\sigma}_v = \mathrm{MLP}(\mathbf{F}_v), \quad \boldsymbol{\mu}_c = \mathrm{MLP}(\mathbf{F}_c), \quad \boldsymbol{\sigma}_c = \mathrm{MLP}(\mathbf{F}_c). \quad (1)$$

The latent variables $\mathbf{Z}_v$ and $\mathbf{Z}_c$ are then sampled as $\mathbf{Z}_v \sim \mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2)$ and $\mathbf{Z}_c \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$. To ensure differentiability during training, the reparameterization trick is applied: $\mathbf{Z}_v = \boldsymbol{\mu}_v + \boldsymbol{\sigma}_v \cdot \boldsymbol{\epsilon}_v$ and $\mathbf{Z}_c = \boldsymbol{\mu}_c + \boldsymbol{\sigma}_c \cdot \boldsymbol{\epsilon}_c$, where $\boldsymbol{\epsilon}_v, \boldsymbol{\epsilon}_c \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

This mechanism enables the CVF module to learn robust feature representations that account for both deterministic and stochastic variations. Consequently, the latent variables encapsulate richer contextual information, which is critical for downstream tasks.

Variational Attention Fusion. The global and local latent features are integrated using a variational attention mechanism. First, the latent variables $\mathbf{Z}_v$ and $\mathbf{Z}_c$ are processed through MLP-based encoders E_{av} and E_{ac}, generating intermediate latent distributions: $\mathbf{Z}_{av} \sim \mathcal{N}(\boldsymbol{\mu}_{av}, \boldsymbol{\sigma}_{av}^2)$ and $\mathbf{Z}_{ac} \sim \mathcal{N}(\boldsymbol{\mu}_{ac}, \boldsymbol{\sigma}_{ac}^2)$. Similar to E_v and E_c, these encoders ensure consistent and robust latent feature representation. Fusion weights $(\boldsymbol{\beta}_v, \boldsymbol{\beta}_c)$ are then computed via the softmax function: $(\boldsymbol{\beta}_v, \boldsymbol{\beta}_c) = \mathrm{Softmax}(\mathbf{Z}_{av}, \mathbf{Z}_{ac})$. The final fused feature representation $\mathbf{F}_{\text{fuse}}$ is computed as a weighted combination of the latent variables from both

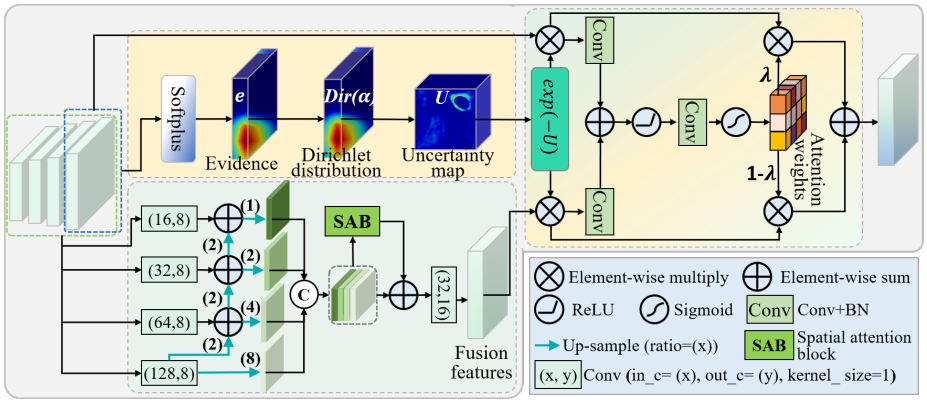

Fig. 4. Structure of the evidential-learning uncertainty refinement module. The module refines segmentation by combining evidential uncertainty modeling, multi-scale feature fusion, and uncertainty-guided refinement.

branches:

$$\mathbf{F}_{\text{fuse}} = \mathbf{W}_m \cdot (\boldsymbol{\beta}_v \cdot \mathbf{Z}_v + \boldsymbol{\beta}_c \cdot \mathbf{Z}_c), \tag{2}$$

where $\mathbf{W}_m$ is a learnable weight matrix for optimal feature transformation.

This mechanism adaptively balances global and local contributions, enhancing the model's ability to capture both macro- and micro-level vessel structures.

2.2 Evidential-learning Uncertainty Refinement Module

The EUR module enhances segmentation robustness in ambiguous and low-contrast regions through evidential uncertainty estimation, multi-scale feature fusion, and uncertainty-guided refinement.

Uncertainty Quantification. Deep models often exhibit overconfidence, reducing reliability in medical segmentation. To address this, the EUR module employs evidential learning paradigm based on Subjective Logic theory [12], modeling uncertainty through evidence rather than direct probabilities.

A Dirichlet distribution [23] is adopted to capture voxel-wise uncertainty, with the evidence map $\mathbf{e}$ computed via a non-negative activation function, Softplus: $\mathbf{e} = \text{Softplus}(\mathbf{F})$. Here $\mathbf{F}$ represents the input feature map. The Dirichlet parameters are then given by $\boldsymbol{\alpha} = \mathbf{e} + 1$, where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$, and K is the number of classes. Uncertainty is estimated as: $U = \frac{K}{S}$, where $S = \sum_{k=1}^{K} \alpha_k$ denotes the Dirichlet strength. This formulation highlights high-uncertainty regions such as boundaries and low-contrast areas, guiding more informed segmentation.

Multi-scale Feature Fusion. To enhance the network's ability to capture contextual information, the EUR module integrates multi-scale features from different decoder stages. Lower-resolution features are upsampled to align with higher-resolution ones, followed by progressive fusion:

$$\mathbf{F}_i^t = \begin{cases} \text{Conv}(\mathbf{F}_i), & i = 4 \\ \text{Conv}(\mathbf{F}_i) + \text{Up}_{\times 2}(\mathbf{F}_{i+1}^t), & i = 3, 2, 1 \end{cases}, \quad \mathbf{F}_i' = \text{Up}_{\times 2^{i-1}}(\mathbf{F}_i^t), \tag{3}$$

where $\mathbf{F}_i$ denotes features from the i-th scale, $\mathrm{Conv}(\cdot)$ denotes a $1\times1\times1$ convolution for channel alignment, and $\mathrm{Up}_{\times k}(\cdot)$ denotes upsampling with a ratio of k. The fused features are concatenated as $\mathbf{F}_c = \mathrm{Cat}(\mathbf{F}'_1, \mathbf{F}'_2, \mathbf{F}'_3, \mathbf{F}'_4)$, where $\mathrm{Cat}(\cdot)$ denotes the concatenation operation. A spatial attention block (SAB) [14] further enhances spatial localization: $\mathbf{F}_{\text{fusion}} = \mathbf{F}_c + \mathrm{SAB}(\mathbf{F}_c)$. This fusion strategy enhances cross-scale interaction and spatial sensitivity, ensuring robust feature representations.

Uncertainty-guided Refinement. The EUR module refines the segmentation results by integrating the initial prediction $\mathbf{P}$, uncertainty map $\mathbf{U}$, and fused features $\mathbf{F}_{\text{fusion}}$. First, a reliable mask $\mathbf{M}_r$ is constructed to suppress uncertain regions: $\mathbf{M}_r = (\mathbf{P}+\mathbf{F}_{\text{fusion}})\cdot\exp(-\mathbf{U})$. Here, $\exp(-\mathbf{U})$ suppresses high-uncertainty areas, focusing on more reliable regions. Next, an attention mechanism [17] adaptively highlights important spatial regions by generating a dynamic weight map $\boldsymbol{\lambda} = \mathrm{Sigmoid}(\mathrm{Conv}(\mathrm{ReLU}(\mathbf{M}_r)))$, where $\boldsymbol{\lambda} \in [0, 1]$. The final refined representation is obtained as: $\mathbf{F}_{\text{refined}} = \boldsymbol{\lambda} \cdot \mathbf{P} + (1 - \boldsymbol{\lambda}) \cdot \mathbf{F}_{\text{fusion}}$, where $\boldsymbol{\lambda}$ balances the contributions of the initial prediction and fused features to improve accuracy.

2.3 Loss Function

Our training objective integrates a combined segmentation loss ($\mathcal{L}_{\text{seg}}$ [6]) and an evidential regularization loss ($\mathcal{L}_{\text{KL}}$ [28]), which uses a Dirichlet-based term to guide uncertainty estimation. The total loss is formulated as $L = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{KL}}$. The segmentation loss is a weighted sum of Dice and weighted cross-entropy (WCE) losses, $\mathcal{L}_{\text{seg}} = \gamma\mathcal{L}_{\text{Dice}} + (1 - \gamma)\mathcal{L}_{\text{WCE}}$, with γ empirically set to 0.6.

3 Experiments and Results

3.1 Datasets and Implementation

Dataset. We evaluated our method on three datasets. CCTA119 is our in-house dataset, which includes 119 CCTA volumes from a Grade III Level A medical institution, with a resolution of $a \times a \times b\,\mathrm{mm}^3$ ($a \in [0.28, 0.41]$, $b \in [0.5, 1.0]$) and a matrix size of $512 \times 512 \times N$ ($N \in [155, 353]$), annotated by three radiologists with at least 5 years of experience. The second is the MICCAI 2020 ASOCA challenge dataset [9], containing 40 CCTA scans. The third, ICAS-100, is a subset of the ImageCAS dataset [24], consisting of 100 CCTA scans.

Evaluation Metrics. We evaluate our proposed method using two metrics: dice similarity coefficient (DSC) and average symmetric surface distance (ASSD).

Implementation Details. We evaluated the proposed framework using five-fold cross-validation on the CCTA119 dataset (95 training, 24 testing subjects), ASOCA (32 training, 8 testing subjects) and ICAS-100 (80 training, 20 testing subjects). All experiments were conducted using the PyTorch framework on NVIDIA 3090 GPUs. The networks were trained with the Adam optimizer, an initial learning rate of 1×10^{-4}, 600 epochs, and a batch size of 2. During training, sub-volumes of size $160\times160\times128$ were randomly cropped from the full volumes. In the testing phase, a sliding window approach with the same sub-volume size was used, moving in steps of half the window size to cover the entire volume.

Table 1. Comparison of different methods for coronary artery segmentation.

Method	CCTA119		ASOCA		ICAS-100	
	DSC %↑	ASSD mm↓	DSC %↑	ASSD mm↓	DSC %↑	ASSD mm↓
3D-UNet[3]	81.65±0.73	1.62±0.12	84.78±0.85	1.41±0.13	76.17±1.04	1.57±0.16
S^2CA-Net[27]	84.01±0.68	1.12±0.10	87.11±0.79	1.16±0.13	78.88±0.87	1.28±0.14
I^2U-Net [4]	84.56±0.64	0.96±0.09	87.47±0.72	0.85±0.12	79.25±0.85	1.12±0.12
UNETR[10]	83.22±0.57	1.56±0.11	86.54±0.82	1.21±0.13	78.64±0.83	1.37±0.14
TransUNet[2]	83.53±0.62	1.52±0.09	87.38±0.69	1.18±0.11	78.87±0.82	1.32±0.13
nnFormer[26]	84.61±0.56	1.38±0.11	87.46±0.67	1.03±0.10	79.22±0.79	1.15±0.14
CS^2Net[16]	84.34±0.61	1.04±0.10	87.01±0.62	0.88±0.12	79.07±0.71	1.09±0.12
3D-FFR -UNet[19]	84.58±0.55	0.91±0.09	87.12±0.64	0.82±0.10	78.94±0.83	1.05±0.12
VSNet [22]	85.07±0.51	0.87±0.09	88.04±0.63	0.79±0.11	79.19±0.86	1.02±0.13
Ours	87.31±0.42	0.71±0.08	90.15±0.57	0.66±0.09	82.15±0.73	0.86±0.12

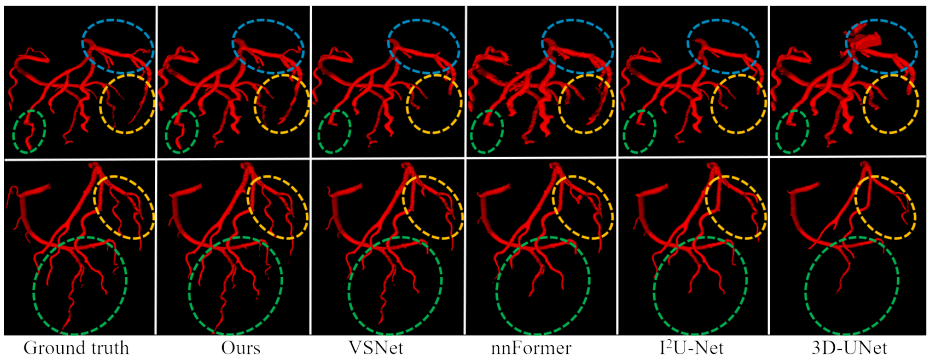

Fig. 5. Visual results. The cyan, yellow and green dashed circles highlight the regions for better visual comparison.

3.2 Comparison with State-of-the-Art Methods

We compared our method against nine state-of-the-art approaches, including CNN-based methods (3D-UNet [3], S^2CA-Net [27], I^2U-Net [4]), transformer-based methods (UNETR [10], TransUNet [2], nnFormer [26]), and vessel segmentation methods (CS^2Net [16], 3D-FFR-UNet [8], VSNet [22]). All comparisons used publicly available codes for fairness. Quantitative and qualitative results are shown in Table 1 and Fig. 5, respectively, while Table 2 presents cross-validation results, demonstrating the generalization capability of the proposed method.

Quantitative Results. As shown in Table 1, we conducted extensive comparative experiments on CCTA119, ASOCA, and ICAS-100. On *CCTA119*, our method consistently outperforms all comparison methods, achieving a 5.66% higher DSC and a 0.91mm lower ASSD than 3D-UNet, as well as a 2.75% improvement in DSC over I^2U-Net, the best-performing CNN model. Furthermore, it surpasses the strongest transformer-based and vessel segmentation methods, with DSC gains of 2.70% over nnFormer and 2.24% over VSNet. On *ASOCA*, our method also achieves the best performance among all compared methods,

Table 2. Cross-validation results: Model trained on CCTA119 and tested on ASOCA and ICAS-100.

Method	CCTA119→ASOCA		CCTA119→ICAS-100	
	DSC %↑	ASSD mm↓	DSC %↑	ASSD mm↓
3D-UNet	79.14±0.81	1.64±0.17	73.55±1.06	1.62±0.18
I^2U-Net	82.36±0.75	1.10±0.11	75.76±0.87	1.33±0.14
nnFormer	82.39±0.69	1.09±0.12	75.62±0.91	1.35±0.15
VSNet	82.67±0.74	1.02±0.13	75.87±0.94	1.28±0.14
Ours	85.26±0.61	0.83±0.11	78.74±0.82	1.05±0.13

Table 3. Ablation studies of our method on the CCTA119 dataset.

Net	DSC %↑	ASSD mm↓
Net1	82.92±0.53	1.39±0.09
Net2	84.17±0.46	1.15±0.08
Net3	85.38±0.44	0.95±0.10
Net4	84.11±0.47	1.12±0.09
Ours	87.31±0.42	0.71±0.08

outperforming VSNet by 2.11% in DSC and 0.13 mm in ASSD. On *ICAS-100*, our method performs consistently best among all comparisons, despite potential labeling errors in the dataset leading to relatively low overall metrics.

Qualitative Results. Fig. 5 provides a qualitative comparison, highlighting that comparison methods exhibit over-segmentation, under-segmentation, or both, leading to suboptimal performance. In contrast, our method closely aligns with ground truth, particularly in complex vascular structures, further demonstrating its effectiveness.

Cross-validation Results. As shown in Table 2, our method demonstrates superior performance in cross-validation. When trained on the CCTA119 dataset and tested on the ASOCA dataset, it achieves a DSC of 85.26%, surpassing 3D-UNet by 6.12% (79.14%) and VSNet by 2.59% (82.67%). Moreover, our method achieves an ASSD of 0.83 mm, outperforming 3D-UNet by 0.81 mm (1.64 mm) and VSNet by 0.19 mm (1.02 mm). Consistent results on the *CCTA119→ICAS-100* setup further validate the method's strong generalization capability.

3.3 Ablation Study

We conduct an ablation study to evaluate the contributions of the key components in our proposed method: the Enhanced-ViT (ViT encoder enhanced by activating the final two ViT blocks and integrating the AGE module), the CVF module, and the EUR module. Starting from the baseline encoder-decoder network (Net1) [25], we incrementally integrate the following components: Net2 adds Enhanced-ViT with sum-based fusion, Net3 replaces the sum fusion with CVF, Net4 extends Net1 by adding EUR, and Ours combines all components.

As shown in Table 3, Net2 achieves a 1.25% improvement in DSC over Net1. Net3 further improves DSC by 1.21% over Net2, while Net4 shows a 1.19% gain over Net1. By integrating Enhanced-ViT, CVF, and EUR, Ours achieves a significant DSC improvement of 4.39% over Net1, demonstrating the effectiveness of each component and their synergistic effects in enhancing the framework.

4 Conclusion

In this study, we propose a novel segmentation framework that leverages the power of the VFM through a parallel encoding architecture for accurate coronary

artery segmentation. The framework incorporates: 1) a ViT encoder to capture global high-level features and a CNN encoder to extract local low-level details, 2) a CVF module for adaptive feature integration via latent distribution modeling and variational attention, and 3) an EUR module to quantify uncertainty and refine segmentation by incorporating multi-scale feature information and attention mechanisms. Extensive experiments on in-house and public datasets demonstrate that our method outperforms state-of-the-art approaches, showcasing its effectiveness, robustness, and strong generalization capability. These results highlight its potential for advancing CAD diagnosis and clinical decision-making.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (Grant Nos. 62301413, 82302309, 62371270, 12326617, and 12026609) and Natural Science Basic Research Program of Shaanxi (Grant No. 2025JC-YBQN-1134).

Disclosure of Interests

This study and its authors have no competing interests.

References

1. Chen, F., Ge, J., Zheng, Y., Guo, K., Cao, F., Liang, J.: Ensembled-sams for enhanced small coronary artery segmentation in ccta images. IEEE Journal of Biomedical and Health Informatics (2024)
2. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
4. Dai, D., Dong, C., Yan, Q., Sun, Y., Zhang, C., Li, Z., Xu, S.: I2u-net: A dual-path u-net with rich information interaction for medical image segmentation. Medical Image Analysis p. 103241 (2024)
5. Dong, C., Dai, D., Li, Y., Xu, S.: High-quality coronary artery segmentation via fuzzy logic modeling coupled with dynamic graph convolutional network. Pattern Recognition p. 111891 (2025)
6. Dong, C., Dai, D., Li, Z., Xu, S.: A novel deep network with triangular-star spatial–spectral fusion encoding and entropy-aware double decoding for coronary artery segmentation. Information Fusion **112**, 102561 (2024)
7. Dong, C., Xu, S., Dai, D., Zhang, Y., Zhang, C., Li, Z.: A novel multi-attention, multi-scale 3d deep network for coronary artery segmentation. Medical Image Analysis **85**, 102745 (2023)

8. Dong, C., Xu, S., Li, Z.: A novel end-to-end deep learning solution for coronary artery segmentation from ccta. Medical Physics (2022)
9. Gharleghi, R., Adikari, D., Ellenberger, K., Webster, M., Ellis, C., Sowmya, A., Ooi, S., Beier, S.: Annotated computed tomography coronary angiogram images and associated data of normal and diseased arteries. Scientific Data **10**(1), 128 (2023)
10. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
12. Jsang, A.: Subjective Logic: A formalism for reasoning under uncertainty. Springer Publishing Company, Incorporated (2018)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
14. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE transactions on pattern analysis and machine intelligence **45**(1), 919–931 (2022)
15. Marano, R., Rovere, G., Savino, G., Flammia, F.C., Carafa, M.R.P., Steri, L., Merlino, B., Natale, L.: Ccta in the diagnosis of coronary artery disease. La radiologia medica **125**, 1102–1113 (2020)
16. Mou, L., Zhao, Y., Fu, H., Liu, Y., Cheng, J., Zheng, Y., Su, P., Yang, J., Chen, L., Frangi, A.F., et al.: Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging. Medical image analysis **67**, 101874 (2021)
17. Oktay, O.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
18. Pan, C., Qi, B., Zhao, G., Liu, J., Fang, C., Zhang, D., Li, J.: Deep 3d vessel segmentation based on cross transformer network. In: 2022 IEEE international conference on bioinformatics and biomedicine (BIBM). pp. 1115–1120. IEEE (2022)
19. Song, A., Xu, L., Wang, L., Wang, B., Yang, X., Xu, B., Yang, B., Greenwald, S.E.: Automatic coronary artery segmentation of ccta images with an efficient feature-fusion-and-rectification 3d-unet. IEEE Journal of Biomedical and Health Informatics **26**(8), 4044–4055 (2022)
20. Virani, S.S., Alonso, A., Aparicio, H.J., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Cheng, S., Delling, F.N., et al.: Heart disease and stroke statistics—2021 update: a report from the american heart association. Circulation **143**(8), e254–e743 (2021)
21. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. arXiv preprint (2023)
22. Xu, J., Dong, A., Yang, Y., Jin, S., Zeng, J., Xu, Z., Jiang, W., Zhang, L., Dong, J., Wang, B.: Vsnet: Vessel structure-aware network for hepatic and portal vein segmentation. Medical Image Analysis **101**, 103458 (2025)
23. Xu, Y., Tang, J., Men, A., Chen, Q.: Eviprompt: A training-free evidential prompt generation method for adapting segment anything model in medical images. IEEE Transactions on Image Processing (2024)

24. Zeng, A., Wu, C., Lin, G., Xie, W., Hong, J., Huang, M., Zhuang, J., Bi, S., Pan, D., Ullah, N., et al.: Imagecas: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. Computerized Medical Imaging and Graphics **109**, 102287 (2023)
25. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749–753 (2018)
26. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE Transactions on Image Processing (2023)
27. Zhou, L., Jiang, Y., Li, W., Hu, J., Zheng, S.: Shape-scale co-awareness network for 3d brain tumor segmentation. IEEE Transactions on Medical Imaging (2024)
28. Zou, K., Yuan, X., Shen, X., Wang, M., Fu, H.: Tbrats: Trusted brain tumor segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 503–513. Springer (2022)