TUBULARITY-AWARE SELF-SUPERVISED PRETRAINING FOR 3D VESSEL SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

3D vessel segmentation requires delineating sparse, slender, and intricately branching tubular structures, yet acquiring annotations for learning these structures is costly. Self-supervised pretraining offers an opportunity to learn transferable representations from unlabeled images, but a key question remains: how can pretraining encourage vascular structure learning without manual annotations? Our premise is that, even when vessel locations are unknown, their elongated, curved morphology and fine-scale characteristics can guide the construction of self-supervised tasks. We propose Tubularity-Aware Pretraining (TAP), a self-supervised framework for 3D vessel segmentation that uses these priors to construct explicit restoration, localization, and detail prediction targets. Specifically, TAP combines cubic and curvilinear masking to learn contextual restoration of missing content; localizes intensity-matched donor content implanted into synthetic tubular regions to learn dense region discrimination; and predicts removed wavelet detail from coarse views to provide dedicated supervision for fine-scale information. Together, these tasks introduce explicit learning objectives for structural information relevant to vessel segmentation without requiring manual vessel annotations. Experiments demonstrate the effectiveness of TAP. Across four CT vessel datasets, TAP achieves state-of-the-art average DSC and clDice scores of 85.00% and 80.72%, respectively, outperforming the strongest competing method by 2.40 and 4.04 percentage points. Further experiments demonstrate its transferability to ASOCA, which is excluded from pretraining, and its effectiveness under limited-label fine-tuning. Code and models will be made publicly available.

1 INTRODUCTION

3D vessel segmentation aims to delineate vascular structures from volumetric medical images and plays an important role in vascular analysis, disease diagnosis, and intervention planning (Lesage et al., 2009). However, as illustrated in Fig. 1(a), vessels occupy only a small fraction of image voxels while forming complex networks with small calibers, hierarchical branches, and fragile connections. Missing a small branch or disrupting a local connection, even when only a few voxels are affected, can compromise the integrity of the vascular network.

To address these challenges, existing methods for vessel and tubular structure segmentation improve feature extraction, contextual modeling, and structural constraints (Qi et al., 2023; Pan et al., 2022; Zhou et al., 2025a; Xu et al., 2025; Shi et al., 2026; Shit et al., 2021). Advances in general medical image segmentation also yield strong performance on vascular structures through improved network architectures, downsampling, and feature fusion (Zhou et al., 2025b; Huang et al., 2026b). Despite these advances, supervised models still learn to recognize vascular structures primarily from annotations for the target task. When annotations are limited, learning representations suited to vessel segmentation before downstream training offers an opportunity for further improvement.

Pretraining provides effective initialization for downstream segmentation by learning transferable representations from source data. Previous studies have demonstrated that 3D medical image representations learned with segmentation supervision can transfer across tasks (Liu et al., 2023; Li et al., 2024; Zhang et al., 2021). For vessels, vesselFM combines annotated real data, domain-randomized data, and generated data to learn a vessel segmentation model that generalizes across imaging domains (Wittmann et al., 2025). However, such approaches still rely on real or synthetic segmentation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

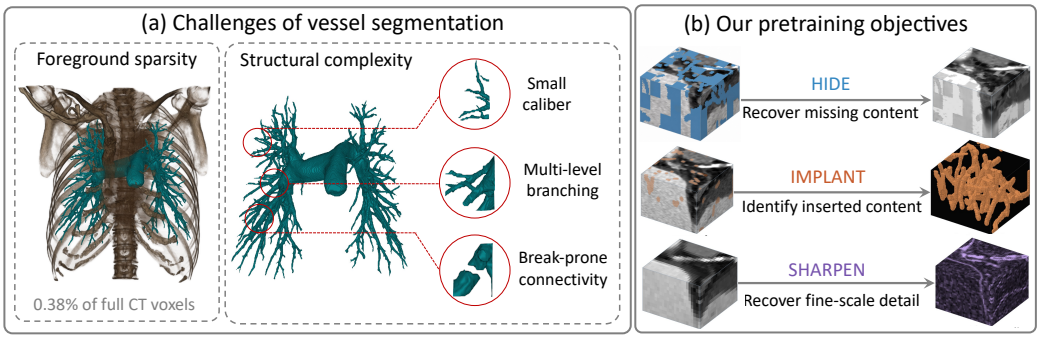

Figure 1: **Vessel segmentation challenges and the motivation for TAP.** (a) Vessels occupy only 0.38% of voxels in the illustrated CT volume, yet form a complex network with small calibers, multi-level branching, and fragile connectivity. (b) TAP addresses this mismatch through three complementary pretext tasks: Hide reconstructs missing content within cubic and synthetic curvilinear masks; Implant localizes intensity-matched tubular content inserted from another volume; and Sharpen predicts removed wavelet detail from a coarse view.

supervision, while the high cost of dense manual annotation for 3D vessels limits the amount of real data available for supervised pretraining. In contrast, self-supervised pretraining constructs learning targets directly from images, enabling the use of large-scale unlabeled volumes. Existing methods learn transferable 3D medical image representations through image restoration, local perturbation, multi-view learning, and volumetric contrastive learning (Zhou et al., 2021; Valanarasu et al., 2023; Tang et al., 2022; Wang et al., 2023; Wu et al., 2024). Although these approaches provide effective initialization for downstream segmentation, their objectives are primarily designed to model general image content or large anatomical structures and may therefore be insufficiently sensitive to the sparse tubular regions, local connectivity, and fine-scale morphology that are important for vessel segmentation. This raises a central question: *How can we guide models to learn representations relevant to vascular structures without relying on manual vessel annotations?*

Our starting point is that the absence of vessel location annotations does not preclude the use of vessel morphology. The elongated, curvilinear, and fine-scale characteristics of vessels are not only major sources of difficulty in vessel segmentation, but also useful priors for designing self-supervised tasks. Based on these morphological priors, we construct restoration, localization, and detail prediction tasks, encouraging the model to learn representations that are more sensitive to elongated structures and local details.

Building on this idea, we propose TAP, a tubularity-aware self-supervised pretraining approach for 3D vessel segmentation. As illustrated in Fig. 1(b), TAP jointly constructs three tasks: restoration, localization, and detail prediction. First, it combines conventional cubic masking with synthetic curvilinear tubular masking, training the model to recover missing content across different masking geometries using contextual information. Second, it implants content from another volume into synthetic tubular regions and predicts the implanted regions, introducing a dense region discrimination task into pretraining. Intensity matching before implantation reduces simple differences in intensity statistics between the implanted content and its surroundings. Finally, TAP uses a 3D Haar wavelet decomposition (Mallat, 1989) to construct a coarse-scale view and the removed detail subbands. Predicting these subbands with an energy-normalized loss provides dedicated supervision for fine-scale information.

We evaluate TAP on four public 3D vessel segmentation benchmarks. TAP achieves the best average DSC and clDice of 85.00% and 80.72%, outperforming vesselFM by 2.40 and 4.04 percentage points, respectively. On ASOCA, which is not used during pretraining, TAP achieves 86.13% DSC and 84.54% clDice after fine-tuning with the available training annotations. It also consistently improves performance under limited-annotation settings, demonstrating strong transferability and label efficiency. Our main contributions are as follows:

- We propose TAP, a tubularity-aware self-supervised pretraining framework for 3D vessel segmentation that leverages morphological priors to learn representations from unlabeled images.

- We introduce complementary restoration, localization, and detail prediction tasks to learn representations sensitive to tubular structures and fine-scale morphology.
- We demonstrate strong performance in segmentation, cross-dataset transfer, and label-efficient fine-tuning across five CT vessel datasets.

2 Related Work

2.1 3D Vessel Segmentation

3D vessel segmentation is fundamental to vascular analysis. Existing studies address the elongated morphology and connectivity of vessels through structural priors, training objectives, and network design. Vesselness filters analytically characterize tubular structures (Frangi et al., 1998), while clDice encourages connectivity through overlap between segmentation masks and their skeletons, and topology-aware losses constrain segmentation topology (Shit et al., 2021; Hu et al., 2019). Segmentation networks improve vessel delineation through feature extraction tailored to elongated structures, contextual modeling, and structural constraints (Qi et al., 2023; Pan et al., 2022; Zhou et al., 2025a; Xu et al., 2025; Shi et al., 2026). General medical image segmentation networks also achieve strong performance through multi-scale representations and feature fusion (Zhou et al., 2025b; Huang et al., 2026b). These networks primarily learn structural discrimination through downstream segmentation supervision. TAP incorporates tubular morphological priors into pretraining tasks, learning relevant representations before accessing downstream annotations without introducing additional modules into the downstream network.

2.2 Pretraining for 3D Medical Images

3D medical image pretraining mainly follows supervised and self-supervised approaches. Supervised pretraining learns representations from multi-organ segmentation annotations (Liu et al., 2023; Li et al., 2024) or jointly learns from multiple partially labeled datasets (Zhang et al., 2021). For vascular structures, vesselFM combines annotated real data, domain-randomized data, and generated data to improve cross-domain generalization (Wittmann et al., 2025). These methods demonstrate supervised pretraining's transferability but still rely on real or synthetic segmentation labels.

Self-supervised pretraining derives learning targets from images themselves. Existing tasks include image restoration (Zhou et al., 2021), recovery from local masking and low-level perturbations (Valanarasu et al., 2023), multi-task learning with inpainting and rotation prediction (Tang et al., 2022), and multi-view learning and spatially guided volumetric contrast (Wang et al., 2023; Wu et al., 2024). These tasks learn general or local image representations but are not specifically designed around vessel morphology. Other studies exploit flow-based consistency to learn vessel-related representations (Jena et al., 2021) or use anatomical information to guide masking and reconstruction in 2D angiography (Huang et al., 2026a). TAP targets 3D vessel segmentation by using tubular morphology to construct synthetic perturbations and image decomposition to generate detail prediction targets. This provides self-supervision for content restoration, region localization, and detail prediction without first localizing actual vessels or estimating flow fields.

3 Tubularity-Aware Pretraining

TAP trains a single volumetric encoder-decoder f_θ on unlabeled CT crops with three pretext tasks, illustrated in Fig. 2. Let $x \in \mathbb{R}^{D\times H\times W}$ denote an intensity-normalized crop. The network has three lightweight output heads: a reconstruction head that predicts a volume $\hat{x}$, an implant head that predicts a probability map $\hat{T}$, and a detail head that predicts wavelet subbands. Each task is defined by what it removes from x and what it requires the network to bring back.

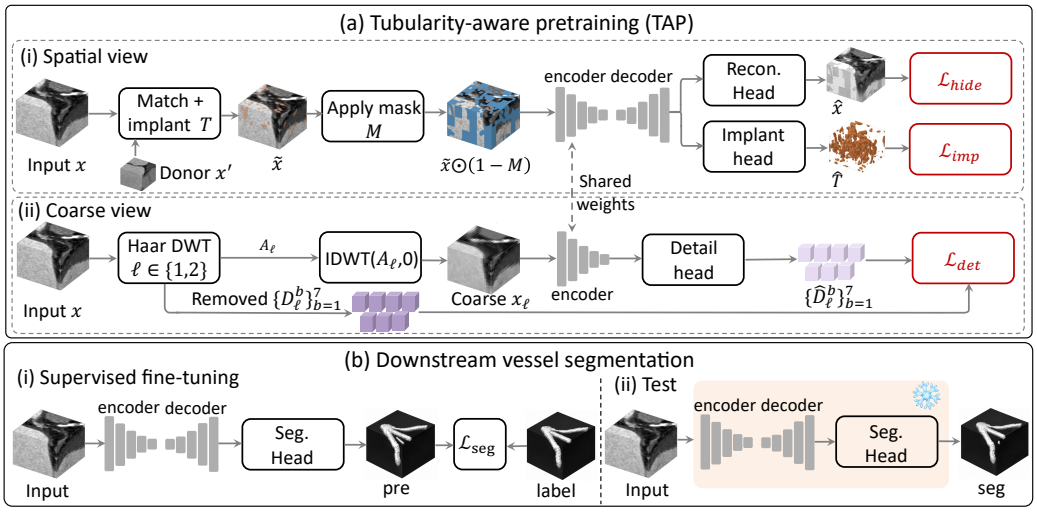

Figure 2: **Overview of TAP.** (a) Tubularity-aware pretraining uses spatial and coarse views of each CT crop. In the spatial view, intensity-matched donor content is implanted into a synthetic tubular region, followed by mixed cubic and curvilinear masking. A shared encoder-decoder reconstructs the masked content and localizes the implants under $\mathcal{L}\mathrm{hide}$ and $\mathcal{L}\mathrm{imp}$. In the coarse view, detail coefficients at all levels from 1 to the sampled level $\ell \in \{1, 2\}$ are set to zero during inverse reconstruction. The detail head predicts only the seven subbands at level ℓ under $\mathcal{L}_{\mathrm{det}}$. (b) After pretraining, the heads are discarded, and the pretrained encoder-decoder is fine-tuned for vessel segmentation. Inference uses only the segmentation network.

3.1 Hiding with Curvilinear Masks

Curvilinear mask generation. We generate tube-shaped regions with a random walk. Starting from a random voxel p_0 and a random unit direction d_0, we iterate

$$p_{k+1} = p_k + \delta\, d_k, \qquad d_{k+1} = \frac{d_k + \sigma\,\epsilon_k}{\|d_k + \sigma\,\epsilon_k\|_2}, \qquad \epsilon_k \sim \mathcal{N}(0, I), \tag{1}$$

for K steps, where δ is the step length and σ controls the curvature. A tube is formed by the union of balls centered along the path, with radii r_k that can be sampled from a predefined range. Several tubes are drawn until they cover the required volume. The walk produces regions that bend and cross tissue boundaries, so no fixed template can be memorized, while their thin and elongated shape matches the structure that segmentation must resolve.

Mixed mask and reconstruction loss. Let $M \in \{0, 1\}^{D\times H\times W}$ be the binary mask with overall masking ratio ρ, of which a fraction η of the masked volume is drawn from the curvilinear generator and the remainder from randomly selected cubic blocks of side s. The network receives the masked volume $x \odot (1 - M)$ and is trained to reconstruct the hidden voxels:

$$\ell_{\mathrm{hide}} = \frac{1}{\|M\|_1}\,\|M \odot (\hat{x} - x)\|_2^2. \tag{2}$$

Here, x denotes the intensity-normalized crop before implantation or masking, and the reconstruction loss is evaluated only at masked voxels. The cubic blocks preserve the context learning that makes masked reconstruction effective in the first place; the curvilinear tubes add the requirement that the network complete paths.

3.2 Implanting and Localizing Tubular Fragments

Reconstruction asks what should be present in a hidden region; segmentation also asks the converse, whether the content that is present belongs to the structure. We give the network a discriminative task of exactly this form by implanting tube-shaped content from a different volume and training the network to find it.

Implant construction. Let x' be a donor crop from another volume in the minibatch and $T \in \{0,1\}^{D\times H\times W}$ a tube-shaped region produced by the generator of Section 3.1. Copying x' into T directly would make the implant detectable from its intensity statistics alone. We therefore match the donor content to the host region before pasting,

$$\phi(x') = \frac{x' - \mu'_T}{\max(\sigma'_T, \varepsilon)}\sigma_T + \mu_T, \qquad \tilde{x} = (1-T)\odot x + T \odot \phi(x'), \quad \varepsilon = 10^{-6}. \tag{3}$$

where (μ_T, σ_T) and (μ'_T, σ'_T) are the mean and standard deviation of x and x' inside T. After matching, the implant has the intensity of its surroundings but the texture and structure of another patient. What remains inconsistent is precisely what a segmentation network must be sensitive to: a tube-shaped region whose content does not continue the anatomy around it.

Localization loss. The hide and implant tasks share one forward pass on $\tilde{x} \odot (1 - M)$, while reconstruction targets the original crop x only at masked voxels. The implant head predicts a voxel-wise probability map $\hat{T}$, supervised by the binary implant mask T:

$$\ell_{\text{imp}} = \text{BCE}(\hat{T}, T) + \text{DiceLoss}(\hat{T}, T). \tag{4}$$

Both terms are evaluated over the full crop, including masked regions and non-implant background. BCE provides voxel-wise supervision, while Dice promotes overlap with sparse implant regions.

3.3 Sharpening with Wavelet Detail Prediction

To provide explicit supervision for these details, we use a 3D Haar wavelet decomposition to separate coarse image content from detail components. We then train the model to predict the removed detail components from the coarse-scale view.

Coarse view and detail target. We apply the 3D Haar wavelet transform to x for ℓ levels, yielding a low-pass approximation A_ℓ and seven detail subbands $\{D_j^b\}_{b=1}^7$ at each level $j \in \{1, \ldots, \ell\}$. We construct the coarse view by setting all detail subbands at levels $1, \ldots, \ell$ to zero and inverting the transform,

$$x_\ell = \text{IDWT}\left(A_\ell, \{\mathbf{0}_j\}_{j=1}^{\ell}\right), \tag{5}$$

where $\mathbf{0}_j$ denotes the seven zero-valued detail subbands at level j. This produces a full-resolution volume containing only the coarse information retained after ℓ levels of decomposition. The network receives x_ℓ and predicts the seven detail subbands at level ℓ, denoted by $\{\hat{D}_\ell^b\}_{b=1}^7$. We sample ℓ uniformly from $\{1, 2\}$ for each crop, exposing the model to detail prediction at two spatial scales.

Energy-normalized detail loss. Detail subbands differ in energy by orders of magnitude: the axial subbands of a CT volume carry far more energy than the diagonal ones, and a plain regression loss would be dominated by the former. We normalize each subband by its own root-mean-square energy before applying the smooth-L_1 loss,

$$\ell_{\text{det}} = \frac{1}{7}\sum_{b=1}^{7} \text{SmoothL}_1\left(\frac{\hat{D}_\ell^b}{e_\ell^b + \varepsilon}, \frac{D_\ell^b}{e_\ell^b + \varepsilon}\right), \qquad e_\ell^b = \sqrt{\tfrac{1}{n_\ell}\left\|D_\ell^b\right\|_2^2}, \tag{6}$$

where n_ℓ is the number of coefficients in a subband and ε is a small constant. RMS normalization reduces scale disparities across orientations and prevents high-energy subbands from dominating the optimization. It thereby encourages learning from both strong and weak detail components while reducing the sensitivity of the loss to variations in crop contrast.

3.4 Joint Objective and Transfer to Segmentation

Each crop produces two views per iteration. The first view is implanted and masked, and supplies the hide and implant losses; the second view is the coarse view of Equation 5 and supplies the detail loss. The joint objective is

$$\ell = \ell_{\text{hide}} + \lambda_{\text{imp}}\,\ell_{\text{imp}} + \lambda_{\text{det}}\,\ell_{\text{det}}. \tag{7}$$

All three tasks share the encoder, while Hide and Implant also share the decoder but use different heads. The shared representation must therefore complete tubular gaps, detect contextual mismatch, and recover missing fine-scale details.

Table 1: **Comparison on four CT vessel segmentation datasets.** The upper block lists supervised architectures trained from scratch, the lower block lists pretraining methods followed by fine-tuning. The last two columns average DSC and clDice over the four datasets. Bold and underlined entries mark the best and second-best results in each column.

Method	ImageCAS			CAS2023			Parse2022			SEG.A			Average	
	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑
Supervised training from scratch														
Swin UNETR	69.67	70.62	34.79	86.53	81.85	1.73	75.73	60.04	6.27	63.01	45.18	9.87	73.73	64.42
CTNet	68.83	67.93	43.47	85.83	81.18	2.97	80.23	64.46	5.72	73.42	47.06	19.60	77.08	65.16
DSCNet	73.49	78.93	22.78	83.14	77.45	2.96	75.04	58.39	6.49	76.21	50.02	25.28	76.97	66.20
GLCP	67.09	69.01	42.10	85.81	81.56	2.31	82.06	69.39	7.93	72.98	38.07	37.73	76.98	64.51
VSNet	50.79	16.42	66.11	74.67	63.87	16.35	71.42	36.30	35.30	71.69	20.46	66.17	67.14	34.26
COMMA	72.26	76.80	15.90	85.18	79.36	2.84	81.81	66.56	2.63	79.78	61.44	6.61	79.76	71.04
nnWNet	78.33	80.35	14.23	82.44	77.92	3.15	84.63	74.13	3.82	82.55	58.55	13.54	81.99	72.74
CROWn	67.87	70.46	31.49	85.95	81.53	3.25	78.03	63.75	4.87	67.96	46.06	18.48	74.95	65.45
Pretraining followed by supervised fine-tuning														
VoCo	71.72	73.95	25.89	86.61	81.56	1.84	77.67	61.18	5.19	66.25	46.16	11.57	75.56	65.71
SwinMM	73.40	79.36	21.99	86.44	81.60	1.81	77.34	61.44	5.24	67.73	48.84	8.85	76.23	67.81
DAE	72.25	73.27	25.99	86.16	81.79	1.92	75.87	60.16	7.61	66.28	48.67	8.62	75.14	65.97
Universal Model	72.87	75.75	23.06	86.15	81.67	1.52	77.80	60.83	4.84	72.70	57.58	5.54	77.38	68.96
SuPreM	73.21	77.10	19.09	86.82	82.45	**1.50**	80.04	64.22	4.47	75.42	59.77	5.17	78.87	70.89
vesselFM	75.25	78.47	20.47	87.09	82.82	1.92	85.79	78.47	1.85	82.25	66.94	3.68	82.60	76.68
Models Genesis	73.80	79.54	18.53	86.37	82.61	2.44	81.20	68.35	3.96	82.19	65.20	3.45	80.89	73.92
DODNet	71.78	73.13	23.32	84.48	81.36	2.16	77.21	60.78	5.38	73.24	54.05	4.87	76.68	67.33
TAP (Ours)	**80.43**	**86.46**	**6.77**	**87.33**	**83.47**	1.74	**87.39**	**81.63**	**1.46**	**84.83**	**71.33**	**2.27**	**85.00**	**80.72**

4 Experiments

4.1 Experimental Setup

Datasets and evaluation metrics. We use five publicly available 3D vessel segmentation benchmarks: ImageCAS (Zeng et al., 2023), CAS2023 (Sunhaozhong, 2023), Parse2022 (Luo et al., 2023), SEG.A (Radl et al., 2022), and ASOCA (Gharleghi et al., 2022). We split ImageCAS, CAS2023, Parse2022, SEG.A, and ASOCA into 850/50/100, 85/5/10, 85/5/10, 46/4/6, and 28/4/8 training/validation/test cases, respectively. For pre-training, we use only the training images from ImageCAS, CAS2023, Parse2022, and SEG.A, resulting in a total of 1,066 images. All annotations are discarded during pre-training, and the images are therefore used in an unlabeled manner. Notably, ASOCA is excluded from pretraining and is used to evaluate cross-dataset transfer after supervised fine-tuning on its training split. We report the Dice similarity coefficient (DSC) (Dice, 1945), the Jaccard coefficient (JC) (Jaccard, 1901), and the normalized surface dice (NSD) (Nikolov et al., 2021) in percent, the average surface distance (ASD, lower is better) (Heimann et al., 2009), and clDice (Shit et al., 2021) in percent for centerline connectivity. The main-paper tables report DSC, clDice, and ASD; additional JC and NSD results are provided in the appendix.

Pretraining protocol. All pretraining experiments are implemented in MONAI and trained on four NVIDIA A6000 GPUs with a total batch size of 16. We instantiate f_θ as the Swin UNETR encoder-decoder (Hatamizadeh et al., 2021) and pretrain it on $96 \times 96 \times 96$ crops, sampled uniformly at random. Each crop is independently standardized to zero mean and unit variance before corruption, with the standard deviation clamped to 10^{-6}. The masking ratio is set to $\rho = 0.5$, with cubic blocks of side length $s = 16$ and a curvilinear fraction of $\eta = 0.3$. For curvilinear masks, the random-walk generator uses a step length of $\delta = 8$ voxels, directional perturbation magnitude $\sigma = 0.35$, $K = 8$ steps per walk, and a fixed radius of $r_k = 3$ voxels, with at most 256 walks per mask. Implant regions are generated using the same random-walk process, with donor samples obtained by cyclically shifting the minibatch on each GPU and intensity-matched to the recipient mean and standard deviation. The detail level ℓ is sampled uniformly from $1, 2$. Unless otherwise specified in dedicated ablation experiments, we use $\lambda_{\text{imp}} = 0.1$ and $\lambda_{\text{det}} = 0.05$ throughout all experiments. We optimize the model for 40,000 iterations using AdamW with an initial learning rate of 10^{-4} and polynomial decay $\alpha_t = 10^{-4}(1 - t/40{,}000)^{0.9}$.

Table 2: **Transfer to ASOCA and fine-tuning with limited labels.** DSC, clDice, and ASD on CAS2023 and on ASOCA when fine-tuning uses 100%, 50%, 20%, or 10% of the training labels. ASOCA is excluded from the pretraining corpus of TAP; Swin UNETR is trained from scratch. Bold and underlined entries mark the best and second-best results in each column of each block.

Method	100%			50%			20%			10%		
	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓
Dataset CAS2023												
Swin UNETR	86.53	81.85	1.73	85.86	81.68	2.34	84.26	79.64	3.12	79.86	76.93	2.84
VoCo	86.61	81.56	1.84	86.08	81.88	2.39	85.27	80.62	1.93	82.00	76.65	4.14
SwinMM	86.44	81.60	1.81	85.73	81.29	2.42	84.86	79.22	2.10	81.42	76.82	4.30
DAE	86.16	81.79	1.92	85.11	81.51	3.23	84.79	79.52	2.27	81.25	**77.70**	2.45
Universal Model	86.15	81.67	1.52	85.32	81.00	2.44	84.57	80.13	1.90	76.53	74.90	4.53
SuPreM	86.82	82.45	**1.50**	85.90	82.11	2.34	85.19	80.62	1.90	80.50	76.84	3.83
vesselFM	87.09	82.82	1.92	86.48	82.22	2.70	85.44	81.31	1.94	81.50	75.13	**1.58**
Models Genesis	86.37	82.61	2.44	86.45	**82.96**	2.80	85.32	81.07	2.19	81.09	76.80	1.94
DODNet	84.48	81.36	2.16	84.20	81.19	2.39	83.49	79.93	**1.36**	66.31	64.41	1.80
TAP (Ours)	**87.33**	**83.47**	1.74	**86.86**	82.93	**2.23**	**85.79**	**81.58**	1.82	**83.21**	77.16	1.66
Dataset ASOCA												
Swin UNETR	80.71	79.09	13.25	79.51	79.05	10.23	73.16	74.10	14.76	69.34	64.19	25.10
VoCo	82.94	80.93	8.98	82.46	79.46	8.82	76.35	78.00	6.61	72.25	71.80	8.96
SwinMM	84.15	83.65	6.53	81.82	82.07	4.60	77.54	78.60	7.04	72.60	70.81	7.48
DAE	82.01	80.33	10.20	81.06	80.07	7.64	76.79	77.02	8.89	71.22	69.37	11.94
Universal Model	84.19	81.75	8.50	82.12	82.13	5.64	77.90	78.47	5.90	71.39	72.51	7.45
SuPreM	84.99	83.87	4.69	83.66	82.38	3.75	77.78	77.31	10.99	73.11	72.10	8.46
vesselFM	83.00	82.25	6.06	81.95	80.49	3.14	76.67	78.61	5.11	70.40	71.29	3.15
Models Genesis	82.47	82.92	**4.56**	81.07	80.83	3.26	76.31	76.52	5.27	70.85	71.93	**2.21**
DODNet	82.40	81.80	7.03	80.33	78.92	9.04	75.68	77.17	6.48	71.10	**73.02**	5.20
TAP (Ours)	**86.13**	**84.54**	4.59	**84.17**	**82.44**	**1.74**	**78.71**	**78.72**	**5.02**	**74.40**	72.60	3.00

Downstream protocol. Unless otherwise specified, all compared pretrained models are fine-tuned under the same downstream protocol. TAP is initialized with our pretrained backbone, while the baselines use their publicly released checkpoints with the corresponding backbone architectures; segmentation heads are randomly initialized. We fine-tune TAP on each downstream segmentation task using random 96^3 crops and an equally weighted sum of Dice and cross-entropy losses, with no topology-specific loss. Optimization uses AdamW for $40{,}000$ iterations with a batch size of 4 on a single NVIDIA A6000 GPU. The initial learning rate is 10^{-4} and follows polynomial decay, $\alpha_t = 10^{-4}(1 - t/40{,}000)^{0.9}$. During inference, we apply sliding-window prediction with a patch size of $96 \times 96 \times 96$ and a stride of 16 voxels along each spatial axis. For voxels covered by multiple windows, we average the predicted softmax probabilities across all overlapping windows. The final segmentation is obtained by assigning each voxel to the class with the highest averaged probability.

4.2 Main Results

Overall comparison. Tab. 1 compares TAP with a broad range of supervised and pretrained baselines across four CT vessel segmentation datasets. TAP achieves the best DSC and clDice on all four datasets and the lowest ASD on three, reaching average scores of 85.00 DSC and 80.72 clDice. Compared with the strongest pretraining baseline, vesselFM, TAP improves the average DSC and clDice by 2.40 and 4.04 points, respectively. Notably, the gain in clDice is substantially larger than that in DSC, suggesting that TAP improves not only voxel-wise overlap but also the preservation of vessel connectivity and topology. These improvements are particularly pronounced on ImageCAS and Parse2022, where TAP improves clDice over the best competing method by 6.11 and 3.16 percentage points, respectively, while also reducing ASD. Similar gains are observed on SEG.A. Overall, these results indicate that the learned representation transfers consistently across distinct vascular anatomies and segmentation settings. Qualitative comparisons are presented in Fig. 3.

Pretraining versus architectural design. TAP outperforms the supervised architectures in Tab. 1 on DSC and clDice across all four datasets using a standard Swin UNETR. Compared with the same backbone trained from scratch, TAP improves average DSC and clDice by 11.27 and 16.30

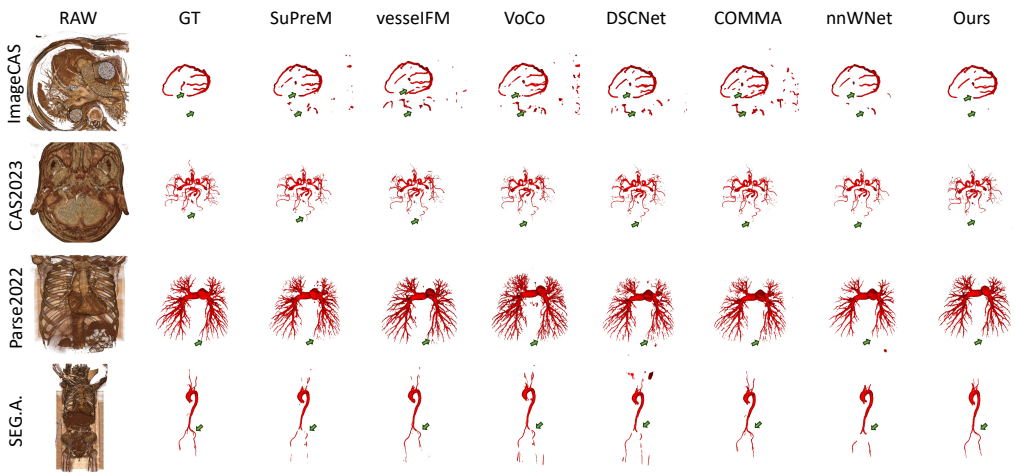

Figure 3: Qualitative comparison on four 3D vessel segmentation datasets. For each dataset, the raw image and ground truth (GT) are shown for reference, followed by predictions from representative competing methods and our TAP model. Red structures denote vessel masks, and green arrows highlight challenging regions such as thin branches, discontinuities, and false-positive responses.

percentage points, respectively. These gains demonstrate the benefit of structure-aware pretraining without introducing specialized modules into the downstream network.

Cross-dataset transfer. The ASOCA results in Tab. 2 provide a transfer test, as no ASOCA images are seen during TAP pretraining. Under full supervision, TAP achieves the best DSC (86.13) and clDice (84.54), outperforming the strongest competing method by 1.14 and 0.67 points, respectively, while its ASD (4.59) is within 0.03 of the best result. Compared with Swin UNETR trained from scratch, TAP improves DSC and clDice by 5.42 and 5.45 points and reduces ASD from 13.25 to 4.59. These gains on an unseen dataset suggest that TAP learns transferable vascular representations rather than relying on dataset-specific exposure during pretraining.

Label-efficient fine-tuning. As shown in Tab. 2, we further evaluate TAP using 100%, 50%, 20%, and 10% of the training labels on CAS2023 and ASOCA. Across the eight settings, TAP ranks first in DSC in every case and achieves the best clDice in five settings, ranking second in the remaining three. Notably, with only 50% of the labels, TAP already surpasses a fully supervised Swin UNETR trained from scratch in DSC and clDice. The advantage persists in the most label-scarce regime: with 10% labels, TAP exceeds the strongest competing method by 1.21 DSC points on CAS2023 and 1.29 points on ASOCA, and outperforms the corresponding from-scratch models by 3.35 and 5.06 points. Together, these results show that the pretrained representation remains effective as supervision becomes increasingly limited, including when transferred to an unseen dataset.

4.3 Ablation Studies

Component analysis. Tab. 3 isolates the contribution of each pretraining component under an identical backbone and training protocol. Cubic masked reconstruction alone substantially improves over training from scratch, increasing the average DSC and clDice by 7.11 and 8.62 points and reducing ASD from 13.16 to 6.79. Replacing part of the cubic mask with curvilinear regions provides a further improvement, with the largest effect on ImageCAS, where clDice increases by 5.21 points and ASD decreases from 15.46 to 7.67, supporting the benefit of structure-aware corruption for vascular representation learning. Implant localization and detail prediction both provide additional gains over mixed masking, with particularly pronounced improvements on Parse2022 and SEG.A. The full model consistently achieves the best DSC and clDice across all four datasets and the lowest ASD on three, reaching 85.00 DSC, 80.72 clDice, and 3.06 ASD on average. Its consistent improvement over either auxiliary task alone indicates that the three components provide complementary pretraining signals rather than redundant supervision.

Table 3: **Ablation study of components.** All variants use the same Swin UNETR backbone, pretraining corpus, masking ratio, and schedule, and are fine-tuned identically. Row 2 is standard masked reconstruction with cubic blocks; rows 4 and 5 add implant localization and detail prediction, respectively, to the mixed mask of row 3; row 6 is the full model. DSC and clDice are in percent, ASD is a distance (lower is better), and the last three columns average over the four datasets.

Variant	ImageCAS			CAS2023			Parse2022			SEG.A			Average		
	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓
1 No pretraining	69.67	70.62	34.79	86.53	81.85	1.73	75.73	60.04	6.27	63.01	45.18	9.87	73.73	64.42	13.16
2 Cubic masking	76.70	80.12	15.46	86.70	82.62	2.02	82.71	69.19	3.03	77.26	60.24	6.64	80.84	73.04	6.79
3 Mixed masking	79.58	85.33	7.67	86.76	82.73	1.25	82.27	68.57	3.29	77.22	60.56	5.50	81.46	74.30	4.43
4 Mixed + implant	79.35	85.02	8.20	87.20	83.15	1.32	85.75	76.41	1.95	83.68	68.40	3.19	84.00	78.25	3.67
5 Mixed + detail	80.09	86.04	6.95	87.01	83.02	**1.11**	86.02	76.80	1.93	84.05	66.56	4.97	84.29	78.11	3.74
6 Full model	**80.43**	**86.46**	**6.77**	**87.33**	**83.47**	1.74	**87.39**	**81.63**	**1.46**	**84.83**	**71.33**	**2.27**	**85.00**	**80.72**	**3.06**

Table 4: **Ablation of masking strategies and wavelet detail prediction.** Metrics are averaged over the four datasets. (a) Overall masking ratio ρ with cubic masked reconstruction only. (b) Curvilinear fraction η in the full model at $\rho = 0.5$, where the cubic-to-curvilinear ratio is $(1-\eta) : \eta$. (c) Wavelet level selection and loss normalization in the Mixed + detail variant, without implant localization. "Levels 1+2" denotes uniform sampling of $\ell \in \{1, 2\}$; "no normalization" uses the same sampling with an unnormalized Smooth-L_1 loss.

(a) Masking ratio ρ

ratio ρ	DSC↑	clDice↑	ASD↓
0.3	78.89	72.13	7.56
0.5	80.84	73.04	6.79
0.7	80.52	72.91	7.08
0.9	78.66	71.13	6.50

(b) Curvilinear fraction η

fraction η	DSC↑	clDice↑	ASD↓
0	82.01	75.83	4.25
0.3	85.00	80.72	3.06
0.7	83.98	77.55	3.62
1.0	83.28	77.45	3.74

(c) Wavelet detail prediction

detail target	DSC↑	clDice↑	ASD↓
level 1	83.52	77.50	4.78
level 2	83.67	76.94	4.30
levels 1+2	84.29	78.11	3.74
no normalization	83.85	76.85	4.26

Masking ratio and mask composition. Tab. 4(a) evaluates the overall masking ratio ρ using cubic masked reconstruction alone. Among the tested ratios, $\rho = 0.5$ yields the highest average DSC and clDice. Tab. 4(b) evaluates mask composition in the full TAP model by fixing $\rho = 0.5$ and varying the curvilinear fraction η, with both implant localization and detail prediction retained. Here, $\eta = 0$ uses only cubic masks for reconstruction. Unlike the baseline in (a), this configuration retains both auxiliary tasks, isolating the effect of mask composition within the full model. Increasing η from 0 to 0.3 improves average DSC and clDice by 2.99 and 4.89 percentage points, respectively, while reducing ASD from 4.25 to 3.06. Further increasing η degrades all three metrics, favoring a cubic-to-curvilinear ratio of $7 : 3$ among the tested settings.

Wavelet detail prediction. Tab. 4(c) examines wavelet level selection and loss normalization in the Mixed + detail variant, without implant localization. Uniformly sampling $\ell \in \{1, 2\}$ outperforms either fixed level across all three metrics, suggesting that detail targets at different scales provide complementary supervision. With this sampling strategy retained, removing energy normalization reduces average DSC from 84.29 to 83.85 and clDice from 78.11 to 76.85. This comparison supports normalizing the loss by subband energy to balance contributions from detail subbands. Together, these results favor two-level sampling with energy-normalized detail prediction.

5 Conclusion

This work introduces TAP, a tubularity-aware self-supervised pretraining framework for 3D vessel segmentation. TAP translates vessel morphology into pretraining supervision through complementary restoration, localization, and detail prediction tasks, learning representations sensitive to elongated structures and local details without requiring vessel location annotations. Experiments on four public benchmarks demonstrate that TAP consistently achieves state-of-the-art performance, with further results supporting its transferability to ASOCA, which is excluded from pretraining, and its effectiveness under limited-label fine-tuning. We hope this work inspires self-supervised learning approaches that turn knowledge of target structures into learning objectives, enabling unlabeled data to better support structurally demanding medical image analysis.

AI USE STATEMENT

The scientific direction of this work was established by the authors, who identified the research problem, developed the central idea and methodological design, determined the experimental protocol, made the research decisions, analyzed the experimental evidence, and drew the conclusions presented in this paper. Generative AI tools were used only in supporting roles, including as a discussion aid for ideas and methodological choices, with any suggestions evaluated and selected by the authors. AI-assisted coding tools were also used to accelerate implementation after the intended algorithms, components, and code structure had been specified by the authors, and generative AI was occasionally used to translate passages of relevant literature.

Generative AI tools were used for auxiliary tasks such as organizing and summarizing experimental outputs and improving the clarity and readability of manuscript text written by the authors. AI was not used to determine the interpretation, claims, or conclusions of the work, and experimental-output assistance was limited to presentation and consolidation. We did not use generative AI to generate synthetic datasets, formulate mathematical claims or proofs, clean or reformat datasets, conduct qualitative or thematic data analysis, or independently interpret experimental results.

All AI-assisted material was verified by the authors: implementations were inspected and tested against the intended algorithms and experimental design, experimental summaries were checked against the original outputs, and language edits were reviewed to preserve the intended technical meaning. The authors retained control over the research process and all scientific decisions throughout the project. The authors take responsibility for the final manuscript, its claims, results, code, and other associated artifacts.

ETHICS STATEMENT

TAP investigates self-supervised representation learning for 3D vessel segmentation, with the aim of improving segmentation performance using unlabeled medical images. This study uses only publicly available datasets: ImageCAS, CAS2023, Parse2022, SEG.A, and ASOCA. It involves no private patient datasets, new participant recruitment, interventions involving human participants, or collection of personally identifiable information. Use of these datasets is subject to their respective licenses and terms of use.

REPRODUCIBILITY STATEMENT

Section 3 describes TAP in detail, including curvilinear tubular masking and reconstruction, intensity-matched tubular content implantation and localization, and detail prediction using the 3D Haar wavelet transform. It also specifies the individual loss functions and the joint training objective. Section 4.1 documents the dataset splits, pretraining data, model architecture, key hyperparameters, and protocols for pretraining, downstream fine-tuning, inference, and evaluation. Controlled ablations in Section 4.3 isolate the contributions of the main design choices. To support independent reproduction and further research, we will make our code and trained models publicly available.

REFERENCES

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.

Alejandro F Frangi, Wiro J Niessen, Koen L Vincken, and Max A Viergever. Multiscale vessel enhancement filtering. In *International conference on medical image computing and computer-assisted intervention*, pp. 130–137. Springer, 1998.

Ramtin Gharleghi, Dona Adikari, Katy Ellenberger, Sze-Yuan Ooi, Chris Ellis, Chung-Ming Chen, Ruochen Gao, Yuting He, Raabid Hussain, Chia-Yen Lee, et al. Automated segmentation of normal and diseased coronary arteries–the asoca challenge. *Computerized Medical Imaging and Graphics*, 97:102049, 2022.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE transactions on medical imaging*, 28(8):1251–1265, 2009.

Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. *Advances in neural information processing systems*, 32, 2019.

De-Xing Huang, Xiao-Hu Zhou, Mei-Jiang Gui, Xiao-Liang Xie, Shi-Qi Liu, Shuang-Yi Wang, Tian-Yu Xiang, Rui-Ze Ma, Nu-Fang Xiao, and Zeng-Guang Hou. VasoMIM: Vascular anatomy-aware masked image modeling for vessel segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4985–4993, 2026a. doi: 10.1609/aaai.v40i6.42503.

Xingru Huang, Shuanghua Ye, Zhao Huang, Wenwen Tang, Huiyu Zhou, Zhiwen Zheng, Jin Liu, and Xiaoshuai Zhang. Crown: A unified framework for anti-aliased downsampling and phase-calibrated fusion in 3d medical segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8514–8524, 2026b.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

Rohit Jena, Sumedha Singla, and Kayhan Batmanghelich. Self-supervised vessel enhancement using flow-based consistencies. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12902 of *Lecture Notes in Computer Science*, pp. 242–251. Springer, 2021. doi: 10.1007/978-3-030-87196-3_23.

David Lesage, Elsa D Angelini, Isabelle Bloch, and Gareth Funka-Lea. A review of 3d vessel lumen segmentation techniques: Models, features and extraction schemes. *Medical image analysis*, 13 (6):819–845, 2009.

Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised 3d models transfer to medical imaging tasks? In *International Conference on Learning Representations*, volume 2024, pp. 16389–16404, 2024.

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21152–21164, 2023.

Gongning Luo, Kuanquan Wang, Jun Liu, Shuo Li, Xinjie Liang, Xiangyu Li, Shaowei Gan, Wei Wang, Suyu Dong, Wenyi Wang, et al. Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge. *arXiv preprint arXiv:2304.03708*, 2023.

Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. *Journal of medical Internet research*, 23(7):e26151, 2021.

Chengwei Pan, Baolian Qi, Gangming Zhao, Jiaheng Liu, Chaowei Fang, Dingwen Zhang, and Jinpeng Li. Deep 3d vessel segmentation based on cross transformer network. In *2022 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 1115–1120. IEEE, 2022.

Yaolei Qi, Yuting He, Xiaoming Qi, Yuan Zhang, and Guanyu Yang. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6047–6056. IEEE, 2023.

Lukas Radl, Yuan Jin, Antonio Pepe, Jianning Li, Christina Gsaxner, Fen-hua Zhao, and Jan Egger. Avt: Multicenter aortic vessel tree cta dataset collection with ground truth segmentation masks. *Data in brief*, 40:107801, 2022.

Gen Shi, Hui Zhang, and Jie Tian. Comma: coordinate-aware modulated mamba network for 3d dispersed vessel segmentation. *IEEE Transactions on Image Processing*, 2026.

Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. cldice-a novel topology-preserving loss function for tubular structure segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16555–16564. IEEE, 2021.

Sunhaozhong. Cerebral artery segmentation challenge (cas) 2023. CodaLab Competitions, 2023. https://codalab.lisn.upsaclay.fr/competitions/9804.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20698–20708. IEEE, 2022.

Jeya Maria Jose Valanarasu, Yucheng Tang, Dong Yang, Ziyue Xu, Can Zhao, Wenqi Li, Vishal M Patel, Bennett Landman, Daguang Xu, Yufan He, et al. Disruptive autoencoders: Leveraging low-level features for 3d medical image pre-training. *arXiv preprint arXiv:2307.16896*, 2023.

Yiqing Wang, Zihan Li, Jieru Mei, Zihao Wei, Li Liu, Chen Wang, Shengtian Sang, Alan L Yuille, Cihang Xie, and Yuyin Zhou. Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 486–496. Springer, 2023.

Bastian Wittmann, Yannick Wattenberg, Tamaz Amiranashvili, Suprosanna Shit, and Bjoern Menze. vesselfm: A foundation model for universal 3d blood vessel segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20874–20884. IEEE, 2025.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22873–22882. IEEE, 2024.

Jichen Xu, Anqi Dong, Yang Yang, Shuo Jin, Jianping Zeng, Zhengqing Xu, Wei Jiang, Liang Zhang, Jiahong Dong, and Bo Wang. Vsnet: Vessel structure-aware network for hepatic and portal vein segmentation. *Medical Image Analysis*, 101:103458, 2025.

An Zeng, Chunbiao Wu, Guisen Lin, Wen Xie, Jin Hong, Meiping Huang, Jian Zhuang, Shanshan Bi, Dan Pan, Najeeb Ullah, 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.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1195–1204. Ieee, 2021.

Feixiang Zhou, Zhuangzhi Gao, He Zhao, Jianyang Xie, Yanda Meng, Yitian Zhao, Gregory YH Lip, and Yalin Zheng. Glcp: Global-to-local connectivity preservation for tubular structure segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 237–247. Springer, 2025a.

Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20852–20862. IEEE, 2025b.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

Table 5: **Additional metrics on four vessel segmentation datasets.** JC and NSD complement the comparison in main-paper Tab. 1 under the same experimental settings. Average denotes the unweighted mean across datasets. Bold and underlined values indicate the highest and second-highest displayed scores, respectively.

Method	ImageCAS		CAS2023		Parse2022		SEG.A		Average	
	JC ↑	NSD ↑	JC ↑	NSD ↑	JC ↑	NSD ↑	JC ↑	NSD ↑	JC ↑	NSD ↑
Swin UNETR	53.97	67.48	76.33	89.78	61.08	67.33	47.89	45.48	59.82	67.52
CTNet	52.95	65.93	75.45	90.13	67.08	73.59	58.96	58.93	63.61	72.15
DSCNet	58.09	73.81	71.14	85.72	60.06	64.28	62.14	61.68	62.86	71.37
GLCP	51.02	66.08	75.22	89.16	69.69	78.35	58.66	54.23	63.65	71.96
VSNet	35.50	42.80	60.10	74.20	56.40	55.60	56.70	44.80	52.18	54.35
COMMA	56.94	71.69	74.27	88.12	69.36	75.95	67.02	65.55	66.90	75.33
nnWNet	<u>64.60</u>	<u>80.54</u>	70.37	88.00	73.52	82.24	<u>72.77</u>	<u>73.91</u>	70.32	<u>81.17</u>
CROWn	51.86	66.40	75.42	89.16	64.12	71.00	51.49	49.98	60.72	69.14
VoCo	56.43	70.59	76.44	89.91	63.64	69.42	51.12	48.22	61.91	69.54
SwinMM	58.42	73.25	76.18	89.71	63.20	69.65	52.66	50.85	62.62	70.87
DAE	57.03	70.89	75.75	89.37	61.27	67.51	51.47	47.95	61.38	68.93
Universal Model	57.77	71.65	75.73	89.71	63.80	69.86	58.15	56.72	63.86	71.99
SuPreM	58.13	72.59	76.81	90.16	66.85	72.84	61.42	60.64	65.80	74.06
vesselFM	60.70	75.77	<u>77.20</u>	<u>90.60</u>	<u>75.20</u>	<u>82.38</u>	70.50	72.57	<u>70.90</u>	80.33
Models Genesis	58.92	74.12	76.13	90.26	68.55	76.21	70.43	71.35	68.51	77.99
DODNet	56.44	70.38	73.22	89.55	63.04	69.95	58.43	54.84	62.78	71.18
TAP (Ours)	**67.54**	**82.59**	**77.60**	**90.85**	**77.75**	**87.41**	**74.08**	**76.88**	**74.24**	**84.43**

APPENDIX

A ADDITIONAL METRICS

We evaluate segmentation performance using five metrics: the Dice similarity coefficient (DSC), Jaccard coefficient (JC), centerline Dice (clDice), average surface distance (ASD), and normalized surface Dice (NSD). The main paper presents DSC, clDice, and ASD, while this appendix reports the complementary JC and NSD results. DSC, JC, clDice, and NSD are expressed as percentages, with higher values indicating better performance. Surface distances are computed on the voxel grid with unit spacing along all three axes, without incorporating physical voxel spacing. NSD uses a surface tolerance of one voxel, and ASD measures the mean distance from the predicted surface to the ground-truth surface in voxel units, with lower values indicating better performance. All experiments follow the corresponding main-paper protocols.

A.1 OVERALL COMPARISON

Tab. 5 reports JC and NSD for the methods compared in main-paper Tab. 1. TAP achieves the highest scores on all four datasets, supporting its effectiveness in foreground overlap and surface agreement.

A.2 TRANSFER AND LABEL EFFICIENCY

Tab. 6 supplements main-paper Tab. 2 with JC and NSD under different annotation fractions. TAP achieves the best result in 14 of the 16 comparisons. On both CAS2023 and ASOCA, TAP fine-tuned with 50% of the training annotations outperforms Swin UNETR trained from scratch with all annotations on both metrics.

Table 6: **Additional metrics for transfer and label-efficient fine-tuning.** JC and NSD on CAS2023 and ASOCA using different fractions of the training annotations, following main-paper Tab. 2. ASOCA is excluded from TAP pretraining. Bold and underlined values indicate the highest and second-highest displayed scores within each dataset block.

Method	100%		50%		20%		10%	
	JC↑	NSD↑	JC↑	NSD↑	JC↑	NSD↑	JC↑	NSD↑
Dataset CAS2023								
Swin UNETR	76.33	89.78	75.37	89.51	72.92	87.78	66.53	84.67
VoCo	76.44	89.91	75.68	89.76	74.46	88.76	<u>69.54</u>	85.59
SwinMM	76.18	89.71	75.16	89.15	73.84	88.01	68.68	85.29
DAE	75.75	89.37	74.18	88.67	73.73	88.08	68.49	<u>85.69</u>
Universal Model	75.73	89.71	74.50	89.09	73.35	88.23	62.02	81.58
SuPreM	76.81	90.16	75.40	89.70	74.33	88.75	67.39	84.12
vesselFM	<u>77.20</u>	<u>90.60</u>	<u>76.31</u>	90.16	74.81	89.04	68.84	84.27
Models Genesis	76.13	90.26	76.24	<u>90.19</u>	**75.43**	<u>89.21</u>	68.26	85.63
DODNet	73.22	89.55	72.85	89.25	71.84	88.16	52.19	68.70
TAP (Ours)	**77.60**	**90.85**	**77.00**	**90.37**	<u>75.30</u>	**89.40**	**71.50**	**86.70**
Dataset ASOCA								
Swin UNETR	68.04	84.92	66.17	85.15	57.96	80.61	53.40	73.16
VoCo	71.08	86.93	70.37	86.21	62.18	**85.03**	<u>58.23</u>	79.45
SwinMM	72.85	88.70	69.48	87.93	63.60	<u>84.85</u>	57.35	78.36
DAE	69.83	86.26	68.31	86.13	62.61	83.23	55.68	76.78
Universal Model	72.85	88.11	69.98	88.07	<u>64.13</u>	84.23	56.01	79.23
SuPreM	<u>74.07</u>	<u>89.19</u>	<u>72.10</u>	<u>88.25</u>	63.99	83.84	57.98	79.06
vesselFM	71.23	87.89	69.97	86.41	62.77	84.11	55.46	77.36
Models Genesis	70.50	87.96	68.74	86.27	62.31	82.50	55.65	78.29
DODNet	70.30	88.24	67.33	86.02	61.28	83.61	55.57	<u>80.31</u>
TAP (Ours)	**75.73**	**89.84**	**72.79**	**88.52**	**65.15**	84.66	**60.00**	**81.00**

Table 7: **Additional metrics for the component ablation.** JC and NSD for the six variants in main-paper Tab. 3, using the same training and evaluation settings. Average denotes the unweighted mean across datasets. Bold and underlined values indicate the highest and second-highest displayed scores, respectively.

Method	ImageCAS		CAS2023		Parse2022		SEG.A		Average	
	JC ↑	NSD ↑	JC ↑	NSD ↑	JC ↑	NSD ↑	JC ↑	NSD ↑	JC ↑	NSD ↑
1 No pretraining	53.97	67.48	76.33	89.78	61.08	67.33	47.89	45.48	59.82	67.52
2 Cubic masking	62.58	77.42	76.62	90.12	70.66	77.70	63.58	62.42	68.36	76.92
3 Mixed masking	66.50	81.60	76.74	90.37	70.01	76.62	63.46	62.60	69.18	77.80
4 Mixed + implant	66.10	81.08	<u>77.43</u>	90.51	75.21	83.57	72.40	74.14	72.79	82.33
5 Mixed + detail	<u>67.12</u>	<u>82.11</u>	77.13	<u>90.71</u>	<u>75.61</u>	<u>84.20</u>	<u>72.95</u>	<u>74.15</u>	<u>73.20</u>	<u>82.79</u>
6 Full model	**67.54**	**82.59**	**77.60**	**90.85**	**77.75**	**87.41**	**74.08**	**76.88**	**74.24**	**84.43**

A.3 Component Ablation

Tab. 7 reports JC and NSD for the component ablations in main-paper Tab. 3. Adding either Implant or Sharpen to mixed masking improves both metrics on average. TAP achieves the highest JC and NSD on every dataset, suggesting that the auxiliary objectives provide complementary supervision.

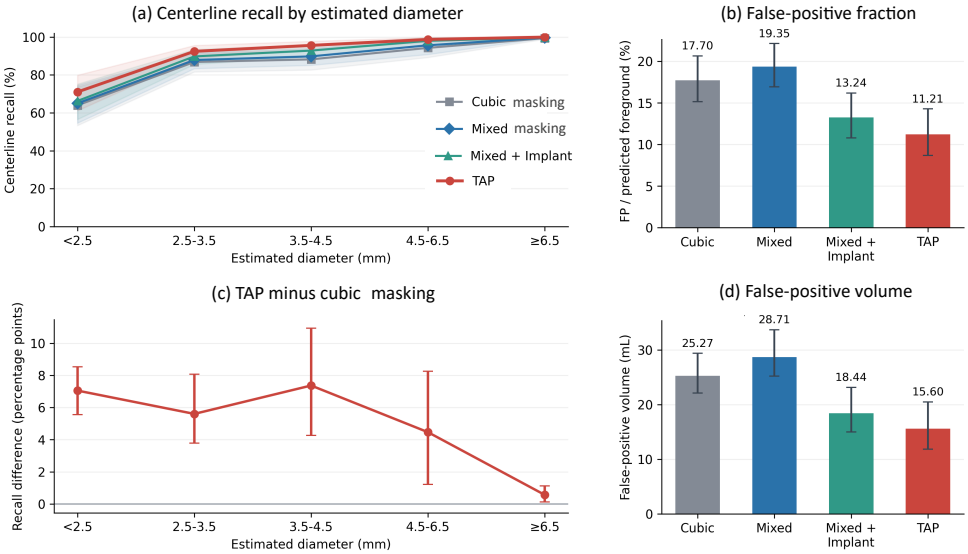

Figure 4: **Vessel-caliber and false-positive analysis on Parse2022.** (a) Centerline recall grouped by ground-truth vessel diameter. (b) False-positive fraction of predicted foreground. (c) Patient-level paired recall differences between TAP and Cubic masking. (d) False-positive volume. Curves and bars show patient-level means; shaded regions and error bars denote 95% bootstrap percentile confidence intervals.

B Centerline Coverage across Vessel Calibers

Global overlap metrics can obscure errors in small vessels, as larger vessels contribute more foreground voxels. We therefore examine centerline coverage across vessel diameters on Parse2022 to identify where pretraining improves vascular recovery. Because increased coverage may also result from overly broad foreground predictions, we jointly evaluate false-positive fraction and volume to assess whether coverage gains are accompanied by improved foreground localization.

For each ground-truth skeleton voxel v, the local vessel diameter is approximated as $2\,\mathrm{EDT}_G(v)$ in millimeters, where EDT_G is computed using the image voxel spacing. The false-positive fraction is defined as $FP/(TP + FP)$ and reported as a percentage. False-positive volume is computed as $FP\, s_x s_y s_z/1000$ mL, where FP denotes the number of false-positive voxels and s_x, s_y, s_z are the case-specific voxel spacings in millimeters. Both measures are computed per patient and then averaged across patients. Paired recall differences between TAP and Cubic masking are computed at the patient level, with uncertainty summarized by 95% percentile bootstrap confidence intervals using 10,000 resamples and a fixed random seed of 1337. Fig. 4(a,c) shows that TAP improves centerline coverage over cubic masking across the displayed diameter groups, with larger gains below 4.5 mm. The improvement reaches approximately 7 percentage points in both the < 2.5 mm and 3.5–4.5 mm groups, but falls below 1 point for vessels ≥ 6.5 mm, where all methods approach complete coverage.

These gains are accompanied by fewer false positives. Compared with cubic masking, TAP reduces the mean false-positive fraction from 17.70% to 11.21% and the false-positive volume from 25.27 to 15.60 mL. Adding implant localization already reduces these measures to 13.24% and 18.44 mL, respectively, and full TAP further improves both while maintaining higher centerline coverage.

Together, these results indicate improved recovery of small- and intermediate-caliber vessels without simply increasing foreground extent. Centerline coverage does not, however, directly establish correct branch connectivity and therefore complements the overlap, surface, and clDice evaluations.

Table 8: Comparison of self-supervised objectives under a common training protocol. Cubic masking refers to the cubic masked-reconstruction baseline in Tab. 3, retrained under the common protocol. Bold and underlined entries indicate the best and second-best results, respectively.

Method	ImageCAS			CAS2023			Parse2022			SEG.A			Average	
	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑	ASD↓	DSC↑	clDice↑
Controlled self-supervised pretraining														
Cubic masking	76.70	80.12	15.46	86.70	82.62	2.02	82.71	69.19	3.03	77.26	60.24	6.64	80.84	73.04
VoCo	77.47	82.85	12.20	86.87	82.78	1.96	83.74	71.27	3.13	78.85	63.26	5.51	81.74	75.04
DAE	77.29	81.72	<u>10.43</u>	86.92	<u>82.92</u>	1.87	<u>84.17</u>	<u>73.32</u>	2.77	80.63	62.38	6.12	82.25	75.09
SwinMM	<u>78.31</u>	<u>83.17</u>	10.59	<u>87.07</u>	82.40	**1.56**	83.76	71.49	<u>2.55</u>	<u>81.04</u>	<u>63.77</u>	<u>4.65</u>	<u>82.55</u>	<u>75.21</u>
TAP (Ours)	**80.43**	**86.46**	**6.77**	**87.33**	**83.47**	<u>1.74</u>	**87.39**	**81.63**	**1.46**	**84.83**	**71.33**	**2.27**	**85.00**	**80.72**

C Additional Ablation Studies

C.1 Controlled Comparison of Self-supervised Objectives

This experiment tests whether TAP's advantage comes from its pretraining objectives rather than differences in pretraining data or backbone architecture. While main-paper Tab. 1 compares complete pretraining solutions using publicly released checkpoints, Tab. 8 retrains cubic masking, VoCo, DAE, SwinMM, and TAP under a common protocol. Method-specific objectives and hyperparameters follow their original implementations, while the pretraining data, backbone, training budget, optimizer, and downstream protocol are kept fixed.

All methods are pretrained from scratch on the same 1,066 unlabeled CT volumes from ImageCAS, CAS2023, Parse2022, and SEG.A using a Swin UNETR backbone, 96^3 crops, and 40,000 optimization steps. The optimizer, learning-rate schedule, effective batch size, downstream splits, fine-tuning, inference, and checkpoint selection are also shared, so the comparison mainly varies the self-supervised objective.

Under these controlled conditions, TAP achieves an average DSC/clDice of 85.00/80.72, outperforming the strongest baseline, SwinMM, by 2.45/5.51 points. It also achieves the highest DSC and clDice on all four datasets. The improvements are particularly pronounced on ImageCAS, Parse2022, and SEG.A, while remaining positive on the more saturated CAS2023 benchmark. These results show that TAP retains its advantage after aligning the pretraining corpus, backbone, and training protocol, supporting the benefit of incorporating tubular morphology into restoration, localization, and detail-prediction objectives rather than relying on additional data or a different architecture.

C.2 Implant Construction and Shortcut Sensitivity

Implant localization should provide transferable structural supervision rather than reward simple intensity differences or boundary artifacts. We therefore compare intensity treatment, boundary treatment, and implant geometry while retaining Hide and Sharpen.

For random-blob controls, we smooth an i.i.d. standard Gaussian noise field using an isotropic Gaussian filter with $\sigma = 3$ voxels, kernel radius 9 voxels, and reflective padding, and then select the K highest-valued voxels, where K equals the foreground voxel count of the corresponding tubular mask. This exactly matches foreground volume for each crop, while connectivity is not enforced. For feathered implantation, the binary implant mask T is smoothed using a $3\times3\times3$ average-pooling filter with stride 1 and zero padding of 1 voxel to obtain a soft mask. Its blending strength is randomly scaled by a factor sampled from $\mathcal{U}(0.5, 1)$ for each crop, while the localization target remains the original binary mask T.

Tab. 9 summarizes the results. Direct pasting yields the highest localization Dice but weaker downstream performance than intensity-matched implantation, indicating that easier implant detection does not necessarily produce more transferable representations. The benefit of implantation is largely preserved after boundary feathering, suggesting that the gain is not primarily driven by sharp paste boundaries. With foreground volume matched exactly, tubular implants consistently outperform random-blob implants under both hard and feathered blending. These controls support the contribution of tubular geometry, rather than implant size or simple intensity and boundary cues, to downstream transfer.

Table 9: Implant construction controls with Hide and Sharpen retained. Random-blob masks are volume-matched to the corresponding tubular masks. Downstream results are averaged over the four datasets; localization Dice measures implant prediction. Bold indicates the best value.

Implant geometry	Intensity treatment	Boundary treatment	Localization Dice ↑	DSC ↑	clDice ↑	ASD ↓
No implant	–	–	–	84.29	78.11	3.74
Tubular	Direct paste	Hard	**98.22**	84.40	78.65	3.62
Tubular (default)	Mean/std matching	Hard	91.57	**85.00**	**80.72**	**3.06**
Tubular	Mean/std matching	Feathered	88.75	84.93	80.45	3.12
Random blob	Mean/std matching	Hard	93.17	84.55	79.10	3.48
Random blob	Mean/std matching	Feathered	89.84	84.48	78.95	3.53

Table 10: Detail-prediction formulations with Hide and Implant retained. Results are four-dataset means. Bold indicates the best value in each metric column.

Detail prediction formulation	DSC ↑	clDice ↑	ASD ↓
No detail prediction	84.00	78.25	3.67
Gaussian high-pass residual	84.65	79.70	3.30
Laplacian	84.48	79.25	3.46
Gradient / 3D Sobel	84.72	79.96	3.24
Haar DWT, levels 1+2 (default)	**85.00**	**80.72**	**3.06**

C.3 Alternative Detail-prediction Formulations

To examine the value of explicit high-frequency supervision and the choice of its formulation, Tab. 10 compares Gaussian high-pass residuals, Laplacian responses, 3D Sobel gradients, and Haar detail prediction while retaining Hide and Implant. This complements main-paper Tab. 4(c), which studies wavelet levels and loss normalization. Gaussian high-pass targets use an isotropic Gaussian filter with $\sigma = 1$ voxel and a $7 \times 7 \times 7$ kernel; Laplacian targets use a six-neighbor 3D stencil, and Sobel targets retain three signed directional responses. All alternatives use unit-ℓ_2-normalized filter kernels, the same coarse input views and scale sampling as Haar, and the same channel-wise normalized Smooth-L1 loss with $\lambda_{\text{det}} = 0.05$.

All three alternative targets improve over the no-detail baseline. Gaussian high-pass, Laplacian, and Sobel prediction increase average DSC by 0.65, 0.48, and 0.72 points, and clDice by 1.45, 1.00, and 1.71 points, respectively, while also reducing ASD. Explicit high-frequency supervision thus adds useful information beyond reconstruction and implant localization. Haar prediction achieves the highest DSC of 85.00 and clDice of 80.72, and the lowest ASD of 3.06. Relative to the no-detail baseline, it improves DSC and clDice by 1.00 and 2.47 points. It also outperforms the strongest alternative, Sobel prediction, by 0.28 DSC points and 0.76 clDice points.

These results show that explicit detail supervision complements reconstruction and localization, with multiscale directional targets from Haar decomposition providing the best overall performance among the tested formulations. Together with the benefits of two-level sampling and energy normalization in main-paper Tab. 4(c), the results support Sharpen's use of dedicated supervision across scales and directions to improve vessel foreground recovery.

C.4 Sensitivity to Auxiliary Loss Weights

The auxiliary loss weights control the contributions of localization and detail prediction relative to reconstruction. Tab. 11 fixes $\lambda_{\text{hide}} = 1$ and varies one auxiliary coefficient at a time to examine performance around the default $(1, 0.1, 0.05)$ and identify the effects of larger changes.

Performance remains similar under moderate changes around the default, whereas larger deviations generally reduce DSC and clDice. TAP is reasonably stable to nearby auxiliary-loss weights rather than depending on a narrowly tuned coefficient choice.

Table 11: Sensitivity to the Implant and Sharpen loss weights. The Hide weight is fixed to 1.

$(\lambda_{\text{hide}}, \lambda_{\text{imp}}, \lambda_{\text{det}})$	ImageCAS					CAS2023				
	DSC↑	JC↑	ASD↓	NSD↑	clDice↑	DSC↑	JC↑	ASD↓	NSD↑	clDice↑
$(1, 0.1, 0.05)$	**80.43**	67.54	6.77	82.59	**86.46**	87.33	77.60	1.74	**90.85**	83.47
$(1, 0.05, 0.05)$	80.25	67.95	6.91	82.32	86.12	87.20	77.60	1.68	90.04	83.22
$(1, 0.2, 0.05)$	80.36	**68.10**	**6.70**	**82.65**	86.38	**87.36**	**77.85**	1.65	90.20	**83.51**
$(1, 0.3, 0.05)$	79.98	67.57	7.08	81.96	85.83	87.10	77.44	1.79	89.90	83.01
$(1, 0.1, 0.01)$	79.82	67.34	7.31	81.70	85.58	87.22	77.63	**1.61**	90.08	83.28
$(1, 0.1, 0.1)$	80.31	68.02	6.85	82.42	86.25	87.28	77.73	1.70	90.13	83.40
$(1, 0.1, 0.15)$	80.06	67.68	7.02	82.05	85.91	87.16	77.54	1.82	89.95	83.13

$(\lambda_{\text{hide}}, \lambda_{\text{imp}}, \lambda_{\text{det}})$	Parse2022					SEG.A				
	DSC↑	JC↑	ASD↓	NSD↑	clDice↑	DSC↑	JC↑	ASD↓	NSD↑	clDice↑
$(1, 0.1, 0.05)$	**87.39**	**77.75**	**1.46**	87.41	**81.63**	**84.83**	**74.08**	**2.27**	**76.88**	**71.33**
$(1, 0.05, 0.05)$	87.06	77.43	1.58	87.37	80.71	84.46	73.65	2.49	75.96	70.35
$(1, 0.2, 0.05)$	87.25	77.73	1.50	87.41	81.22	84.68	73.98	2.34	76.20	71.02
$(1, 0.3, 0.05)$	86.80	77.03	1.67	87.10	80.20	84.15	73.19	2.63	75.52	69.84
$(1, 0.1, 0.01)$	86.55	76.64	1.78	86.82	79.65	84.02	73.00	2.81	75.31	69.15
$(1, 0.1, 0.1)$	87.21	77.67	1.52	**87.58**	81.10	84.61	73.87	2.40	76.11	70.88
$(1, 0.1, 0.15)$	86.94	77.24	1.62	87.24	80.47	84.29	73.40	2.59	75.73	70.06

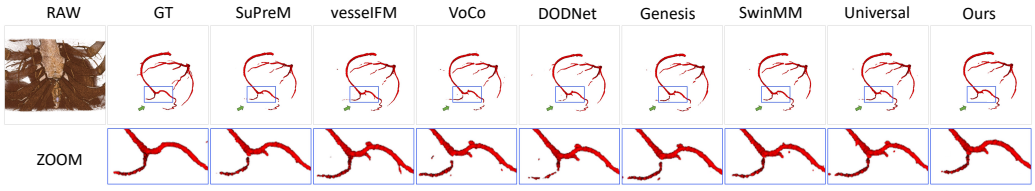

Figure 5: Qualitative comparison on ASOCA, which is excluded from TAP pretraining. The first column shows the input volume rendering, followed by the ground truth and predictions from representative competing methods and TAP. The second row enlarges the regions marked by blue boxes. Red structures denote vessel masks, and green arrows highlight local differences in thin-branch recovery and continuity.

D Qualitative Results

D.1 Transfer to ASOCA

ASOCA is excluded from TAP pretraining, providing a setting to examine whether the learned representation transfers to an unseen vessel dataset after supervised fine-tuning. Fig. 5 complements the quantitative results in Tab. 2 with whole-vessel renderings and matched local enlargements across representative competing methods.

At the whole-vessel level, most methods recover the major coronary trajectories, making their differences difficult to distinguish from the global renderings alone. The enlarged region reveals clearer differences in local structure: several competing predictions contain missing or interrupted thin segments, incomplete branches near the bifurcation, or locally inconsistent vessel extent, whereas TAP more closely preserves the corresponding continuous structure in the ground truth. These observations are consistent with the quantitative gains in DSC and clDice on ASOCA, and illustrate that the transfer improvement is reflected particularly in local branch recovery and continuity rather than only in the major vessel trunks.

D.2 Slice-level Comparisons on Four Datasets

Three-dimensional renderings reveal overall vessel trajectories, whereas slice-level views provide a closer inspection of small vessel cross-sections and local boundary differences. Fig. 6 therefore compares predictions with ground truth at matched slice locations and enlarged regions.

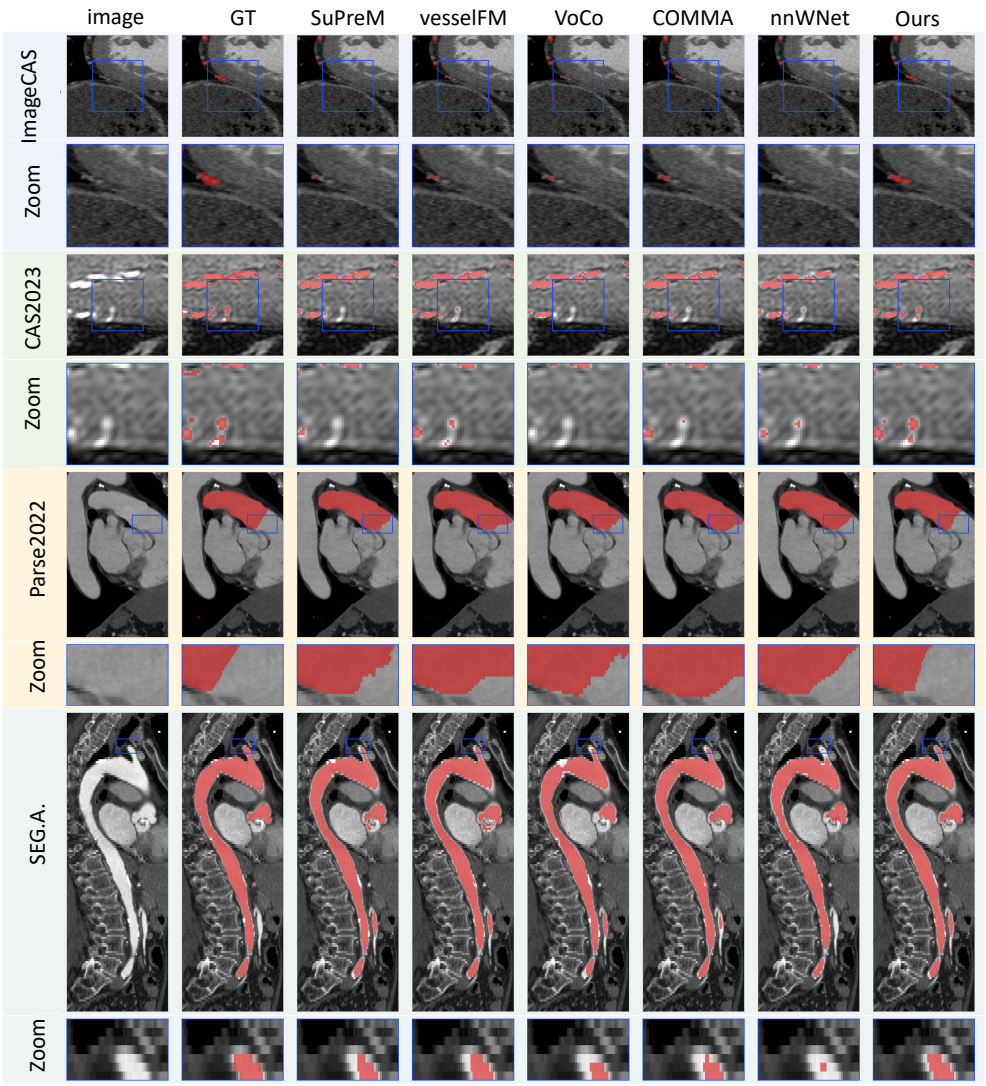

Figure 6: Slice-level comparisons on ImageCAS, CAS2023, Parse2022, and SEG.A. Columns show the image, ground truth, SuPreM, vesselFM, VoCo, COMMA, nnWNet, and TAP. Red overlays indicate vessel masks; blue boxes define the enlarged regions below each example.

On ImageCAS, TAP recovers the highlighted small vessel cross-section that is missed by some competing methods. On CAS2023, it preserves several small, separate vascular regions that are absent in some predictions, including VoCo. For Parse2022, the enlarged view highlights differences in local contour placement, with TAP more closely following the ground-truth boundary. Similar differences are observed on SEG.A, where TAP maintains a foreground extent and boundary shape closer to the annotation. Overall, the slice-level comparisons further illustrate TAP's ability to recover small vascular structures while preserving local vessel boundaries.

E Limitations and Future Work

While the present study demonstrates the effectiveness of tubularity-aware self-supervised pretraining for 3D CT vessel segmentation, it also leaves several natural directions for further investigation. ***I) Cross-modality and temporal extension.*** The present study focuses on 3D CT vessel imaging. Extending TAP to other vascular imaging modalities, such as magnetic resonance angiography, as

well as time-resolved vascular imaging, would broaden the scope of morphology-aware pretraining and provide an opportunity to incorporate modality-specific appearance and temporal vascular information. ***II) Data and model scaling.*** Our experiments prioritize controlled comparisons of pretraining objectives under a fixed training setting. An important next step is to scale TAP to substantially larger unlabeled corpora and broader backbone families. Such studies could further characterize how data scale and model capacity interact with tubularity-aware supervision and influence representation quality and label efficiency. ***III) Beyond vessel segmentation.*** The downstream evaluation in this work centers on vessel segmentation. Since TAP is designed to learn representations sensitive to elongated structures, local continuity, and fine-scale vascular morphology, a natural extension is to investigate whether these representations also benefit related vascular analysis tasks, including anatomical branch identification, centerline extraction, and vessel caliber estimation. Exploring these directions could broaden TAP from a segmentation-oriented pretraining framework toward more general representation learning for vascular image analysis.