VasoMIM: Vascular Anatomy-Aware Masked Image Modeling for Vessel Segmentation

De-Xing Huang[1,2], Xiao-Hu Zhou[1,2*], Mei-Jiang Gui[1,2], Xiao-Liang Xie[1,2], Shi-Qi Liu[1], Shuang-Yi Wang[1,2], Tian-Yu Xiang[1,2], Rui-Ze Ma[1], Nu-Fang Xiao[1], Zeng-Guang Hou[1,2*]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
{huangdexing2022, xiaohu.zhou, zengguang.hou}@ia.ac.cn

Abstract

Accurate vessel segmentation in X-ray angiograms is crucial for numerous clinical applications. However, the scarcity of annotated data presents a significant challenge, which has driven the adoption of self-supervised learning (SSL) methods such as masked image modeling (MIM) to leverage large-scale unlabeled data for learning transferable representations. Unfortunately, conventional MIM often fails to capture vascular anatomy because of the severe class imbalance between vessel and background pixels, leading to weak vascular representations. To address this, we introduce **Vas**cular anat**o**my-aware **M**asked **I**mage **M**odeling (**VasoMIM**), a novel MIM framework tailored for X-ray angiograms that explicitly integrates anatomical knowledge into the pre-training process. Specifically, it comprises two complementary components: *anatomy-guided masking strategy* and *anatomical consistency loss*. The former preferentially masks vessel-containing patches to focus the model on reconstructing vessel-relevant regions. The latter enforces consistency in vascular semantics between the original and reconstructed images, thereby improving the discriminability of vascular representations. Empirically, VasoMIM achieves state-of-the-art performance across three datasets. These findings highlight its potential to facilitate X-ray angiogram analysis.

Project Page — https://dxhuang-casia.github.io/VasoMIM
Extended Version — https://arxiv.org/abs/2508.10794

Introduction

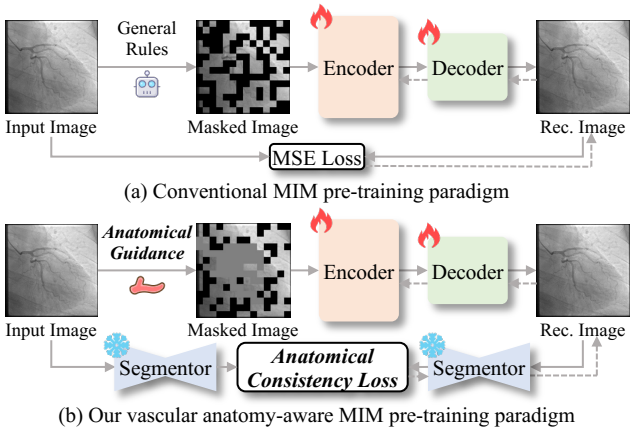
Cardiovascular diseases (CVDs) constitute a global health crisis and remain the leading cause of mortality worldwide (Vaduganathan et al. 2022). X-ray angiography is considered the gold standard for diagnosing CVDs (Kheiri et al. 2022), planning treatment (Members et al. 2022), and guiding intraoperative procedures (Huang et al. 2025). However, radiologists often struggle to accurately delineate vessels in X-ray angiograms because of low contrast, motion artifacts, and overlapping anatomical structures (Huang et al. 2024). Consequently, there is an urgent need for automated vessel segmentation methods.

Over the past decade, numerous vessel segmentation algorithms have been proposed (Ronneberger et al. 2015; Huang et al. 2024; Wu et al. 2025), helping to alleviate radiologists' workload. However, training high-performance models requires large-scale datasets of X-ray angiograms with pixel-level labels, and producing such annotations remains labor-intensive, time-consuming, and dependent on specialized domain knowledge (Esteva et al. 2019). Self-supervised learning (SSL) provides an attractive alternative by learning generalizable representations from vast unlabeled data, thereby boosting downstream performance (Gui et al. 2024). In particular, masked image modeling (MIM) (He et al. 2022; Fu et al. 2025; Zhuang et al. 2025b; Tang et al. 2025) has achieved remarkable success in natural and medical image analysis by training models to reconstruct masked image patches, as shown in Fig. 1 (a).

Nevertheless, adapting MIM to X-ray angiograms remains challenging due to the *extreme class imbalance* between vessel and background pixels. We attribute this difficulty to a lack of explicit anatomical awareness in two key aspects of the MIM framework. **First, vessel-containing patches are more likely to be ignored than background-only patches during the masking process.** Current masking strategies are typically based on general rules, which can be divided into data-independent and data-adaptive (Hinojosa, Liu, and Ghanem 2024). The former includes random masking (He et al. 2022), block-wise masking (Bao et al. 2022), uniform masking (Li et al. 2022b), *etc*. The latter involves designing masking strategies based on specific feedback, *e.g.,* attention maps (Kakogeorgiou et al. 2022; Liu, Gui, and Luo 2023), loss predictions (Wang et al. 2025), reward functions (Xu et al. 2025), *etc*. However, none of these methods adequately focus on vascular anatomy, resulting in few vessel-containing patches being masked and hindering the effective learning of vascular representations. **Second, the pixel-level reconstruction loss fails to preserve semantic consistency during reconstruction.** Most MIM methods minimize the mean squared error (MSE) between original and reconstructed patches, but this pixel-level objective ignores vascular anatomy and thus fails to encourage learning of discriminative vascular representations.

To address these challenges, we introduce **Vas**cular anat**o**my-aware **M**asked **I**mage **M**odeling (**VasoMIM**), a

*Corresponding Authors.



(a) Conventional MIM pre-training paradigm

(b) Our vascular anatomy-aware MIM pre-training paradigm

Figure 1: Comparison of conventional MIM and VasoMIM. (a) Conventional MIM masks patches based on *general rules* and learns to reconstruct patches via minimizing *pixel-level loss*. (b) VasoMIM guides patch masking with *vascular anatomy* and enforces *anatomical consistency* during reconstruction, enabling the model to learn richer vascular representations. Dark gray patches are vessel-relevant regions.

novel MIM paradigm tailored for X-ray angiograms, as presented in Fig. 1 (b). The core insight of VasoMIM is to inject vascular anatomical knowledge into MIM. **To address the first challenge, we introduce an anatomy-guided masking strategy** that biases the masking process toward vessel-containing patches, guiding the model to reconstruct anatomically relevant regions. **For the second challenge, we propose an anatomical consistency loss** that enforces consistency between the vascular anatomy from the original and the reconstructed angiograms, thereby encouraging the model to learn more discriminative vascular representations. This raises the question of how to extract vascular anatomy from X-ray angiograms. To resolve this, we apply Frangi filter (Frangi et al. 1998) to obtain vessel segmentation masks in an unsupervised manner. To integrate this filter into end-to-end SSL, we train a segmentor in advance using pseudo-labels produced by it.

In summary, our main contributions are as follows:

- A novel vascular anatomy-aware MIM framework, VasoMIM, is proposed to enhance the model's ability to understand vascular content in X-ray angiograms.
- Two complementary components are proposed: (1) an anatomy-guided masking strategy that preferentially masks vessel-containing patches, guiding the model to focus on vascular regions, and (2) an anatomical consistency loss that ensures semantic consistency between the original and reconstructed images, thereby boosting the discriminability of vascular representations.
- Extensive experiments demonstrate the benefits of integrating anatomical knowledge into the MIM framework. Meanwhile, VasoMIM consistently outperforms state-of-the-art SSL alternatives on vessel segmentation tasks.

Related Work

SSL in Medical Imaging

Self-supervised learning (SSL) (Chen et al. 2020; Chen, Xie, and He 2021; Caron et al. 2021; He et al. 2022; Wang et al. 2025) offers a practical solution to mitigate annotation scarcity in medical imaging. Existing approaches can be categorized into two main paradigms: contrastive learning (CL)-based and masked image modeling (MIM)-based. CL-based methods aim to pull positive pairs together while pushing negative pairs apart in feature space, such as C2L (Zhou et al. 2020), MICLe (Azizi et al. 2021), VoCo (Wu, Zhuang, and Chen 2024), RAD-DINO (Pérez-García et al. 2025), *etc*. However, because these methods focus on image-level representations, they may perform poorly on dense prediction tasks like segmentation (Chaitanya et al. 2020). MIM-based approaches train models to reconstruct masked patches, yielding finer-grained representations (Yuan et al. 2023). The choice of training objective is critical. Most methods use a pixel-level reconstruction loss (Kang et al. 2024; Li et al. 2024; Xu et al. 2025), while some incorporate a contrastive loss to improve multi-view alignment in 3D medical images (Zhuang et al. 2025a,b). However, these objectives do not ensure semantic consistency in the reconstructed images, often resulting in weak vascular representations. Our work addresses this limitation by introducing a semantic-level anatomical consistency loss.

Masking Strategies in MIM

The design of the masking strategy is crucial in MIM given the high redundancy of images (He et al. 2022). Most methods (Xie et al. 2024; Kang et al. 2024) adopt random masking at high ratios (*e.g.,* 75%). To create more challenging pretext tasks, many carefully designed masking strategies have been proposed. AMT (Liu, Gui, and Luo 2023) masked patches based on attention maps. SemMAE (Li et al. 2022a) learned semantic parts of images first and used part segmentation results to guide patch masking. HAP (Yuan et al. 2023) exploited human-structure priors to guide the masking process for human-centric perception pre-training. Methods such as HPM (Wang et al. 2025), AnatoMask (Li et al. 2024) and, AHM (Xu et al. 2025) identified patches that are difficult to reconstruct through a loss predictor, a self-distillation framework and a policy network, respectively, and preferentially masked these hard patches. However, none of these approaches can explicitly incorporate vascular anatomy into the masking process, which is essential for guiding models to focus on vascular regions.

Vessel Segmentation

Traditional vessel segmentation methods, such as Frangi filter (Frangi et al. 1998), active contours (Taghizadeh Dehkordi et al. 2014), and graph cuts (Wang et al. 2020), rely on handcrafted features that generalize poorly across the wide variability of X-ray angiograms. The emergence of deep learning has significantly advanced the field. Early CNN-based architectures like U-Net (Ronneberger et al. 2015) and its variants (Zhou et al. 2019; Li et al. 2020) capture hierarchical features but remain constrained by their local re-

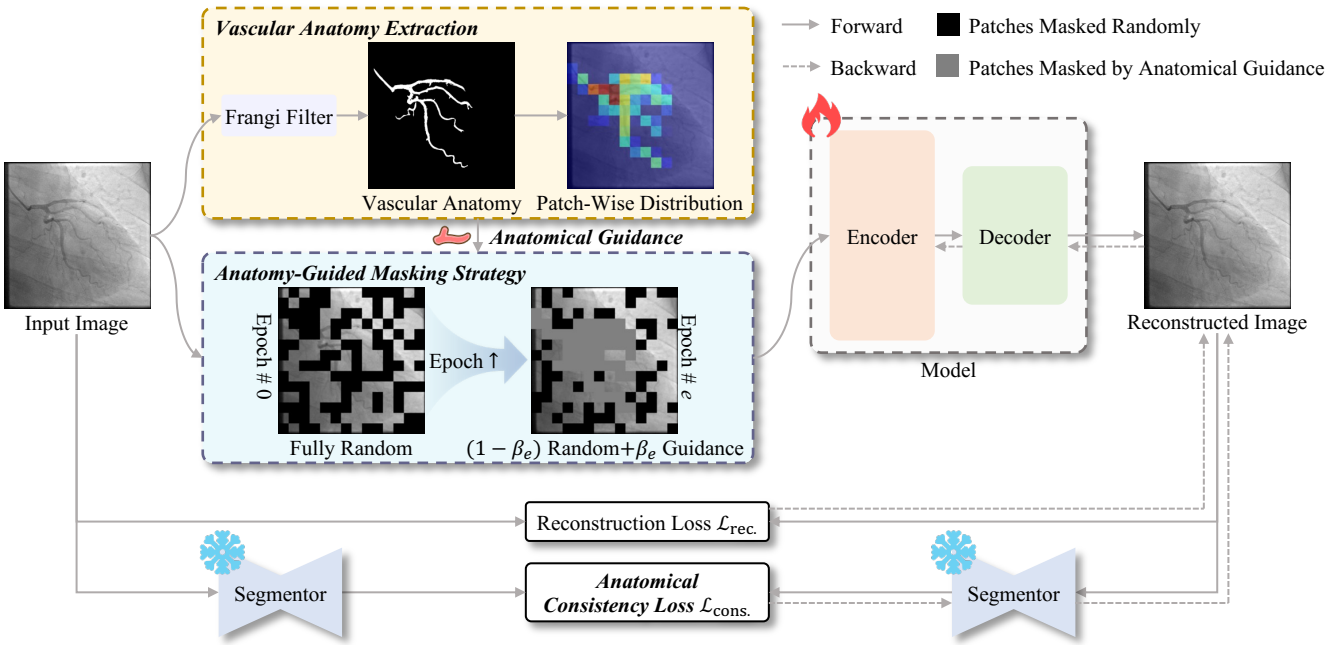


Figure 2: Overall framework of VasoMIM. During pre-training, each X-ray angiogram is first processed by Frangi filter to extract its vascular anatomy. From this anatomy, we derive a patch-wise vascular anatomical distribution f to guide the masking process. Finally, the model is optimized by minimizing $\mathcal{L}_{\text{train}}$, which is a combination of standard pixel-wise reconstruction loss $\mathcal{L}_{\text{rec.}}$ and the designed anatomical consistency loss $\mathcal{L}_{\text{cons.}}$.

ceptive fields. More recent transformer- and Mamba-based models (Chen et al. 2024; Hatamizadeh et al. 2021; Ruan, Li, and Xiang 2024; Wang et al. 2024) address this limitation by integrating global context, thereby improving the delineation of complex vascular structures. However, these deep models are highly data-hungry, and the scarcity of labeled angiograms continues to limit their gains (Zhou et al. 2021). This work leverages MIM to learn generalizable representations from large-scale unlabeled data, significantly boosting the performance of segmentors. We use U-Net as a representative case study to demonstrate the broad applicability of our method.

Method

The overall framework is illustrated in Fig. 2. First, we extract vascular anatomy from X-ray angiograms using Frangi filter (Frangi et al. 1998). Next, this anatomical knowledge is used to guide the masking process. Finally, an anatomical consistency loss is incorporated into the training objective to capture discriminative vascular representations.

Vascular Anatomy Extraction

Frangi filter effectively enhances vascular anatomy in an unsupervised manner by highlighting vessel-like features. Following the implementation in (Wu et al. 2025), our approach proceeds in three stages.

Multi-Scale Hessian Vesselness. Given an X-ray angiogram $I \in \mathbb{R}^{C\times H\times W}$, we smooth it with a Gaussian kernel $G(\sigma)$ at scales $\sigma \in \{1, 2, 3, 4\}$ and calculate the multi-scale Hessian $H_\sigma(i) = \nabla^2 [I * G(\sigma)] (i)$, where C, H, and W denote the number of channels, height, and width of the image. At each pixel i, we obtain eigenvalues $|\lambda_1| \leq |\lambda_2|$ of $H_\sigma(i)$ and define the vesselness response

$$V_\sigma(i) = \begin{cases} |\lambda_1(i)|, & \lambda_2(i) < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The final vesselness map $V(i)$ is the maximum response over all scales, capturing vessels of varying diameters.

Adaptive Percentile Thresholding. Subsequently, we apply adaptive thresholding to the vesselness map V at its α-th percentile value $T = \text{Percentile}_\alpha [V(i)]$, creating a binary map $\hat{B}(i) = \mathbb{I}[V(i) \geq T]$. This approach effectively filters out low-intensity noise while retaining prominent vascular anatomy.

Seed Region Growing. Finally, an optimal seed pixel $s = \arg\max_i V(i)$ is automatically identified based on the highest vesselness intensity value. A region-growing algorithm (Adams and Bischof 1994) is then applied to the binary threshold map $\hat{B}$, iteratively expanding the region to include neighboring pixels with high vesselness values. This process yields a binary mask $B \in \mathbb{R}^{1\times H\times W}$ that accurately delineates the vascular anatomy.

Anatomy-Guided Masking Strategy

Compared to natural images, X-ray angiograms exhibit far greater spatial redundancy because vessels occupy only a small fraction of each image. Existing general rule-based masking strategies lack any anatomical awareness and therefore mask background-only patches far more often. As a re-

sult, pre-training is dominated by background content rather than learning vascular representations.

To address this issue, we introduce an anatomy-guided masking strategy. Our key idea is that patches containing vessels are more informative and should be masked with higher probability. Formally, a patch-wise vascular anatomical distribution f is defined. In the pre-training process, X-ray angiograms I and the corresponding masks B are split into non-overlapping patches $x \in \mathbb{R}^{N \times \left(P^2 C\right)}$ and $m \in \mathbb{R}^{N \times P^2}$, respectively, where P is the patch size and $N = HW/P^2$ is the sequence length. Let $m_i \in \{0,1\}^{P^2}$ be the i-th patch of m. Then $f(m_i)$ is defined as follows:

$$f\left(m_i\right) = \frac{\sum_{j=1}^{P^2} \mathbb{I}\left(m_{ij}=1\right)}{\sum_{i,j=1}^{N,P^2} \mathbb{I}\left(m_{ij}=1\right)}, \quad i = 1, 2, \cdots, N \tag{2}$$

where $\mathbb{I}\left(\cdot\right)$ is the indicator function.

Weak-to-Strong Anatomical Guidance. By leveraging the patch-wise vascular anatomical distribution, we perform anatomy-aware patch masking. However, during the early training stages, masking too many vessel-containing patches may impair the model's ability to reconstruct those patches rich in vascular content. To this end, we adopt a weak-to-strong anatomy-guided strategy. As illustrated in Fig. 2, for a specific training epoch e, β_e of the masked patches are sampled according to f, and the remaining $1-\beta_e$ are randomly selected. β_e is increased linearly during pre-training.

$$\beta_e = \beta_0 + \frac{e}{E}\left(\beta_E - \beta_0\right) \tag{3}$$

where E is the maximum pre-training epoch, and $\beta_0, \beta_E \in [0,1]$ are hyper-parameters. Under this masking strategy, $\beta_e \gamma N$ patches are masked by anatomical guidance and the remaining $(1-\beta_e)\gamma N$ patches are randomly masked, where γ represents the masking ratio.

Anatomical Consistency Loss

In conventional MIM, the reconstruction objective typically minimizes the mean squared error (MSE) between masked and reconstructed patches. However, this loss fails to preserve semantic consistency and does not encourage the model to learn the discriminative vascular representations essential for downstream tasks.

To overcome this limitation, we introduce an anatomical consistency loss that explicitly directs the model to preserve vascular anatomy during reconstruction:

$$\mathcal{L}_{\text{cons.}} = \mathcal{L}\left[\mathcal{S}\left(I\right), \mathcal{S}\left(I'\right)\right] \tag{4}$$

where I' is the reconstructed X-ray angiogram. $\mathcal{L}\left(\cdot,\cdot\right)$ is an abstract metric function, and we use cross-entropy loss by default. Here, $\mathcal{S}\left(\cdot\right)$ represents the segmentor used to extract vascular anatomy. Although Frangi filter provides high-quality vascular masks, its non-differentiability prevents direct integration into the end-to-end pre-training process. To solve this problem, we train a lightweight UNeXt-S network ($\sim$ 0.3 M) (Valanarasu and Patel 2022) on pseudo-labels generated by Frangi filter, then freeze its weights during subsequent pre-training.

Usage	Dataset	# Train	# Test
Pre-Training	ARCADE	$2,000$	–
	CADICA	$6,594$	–
	Stenosis	$7,492$	–
	SYNTAX	$2,943$	–
	XCAD	$1,621$	–
	Total	$20,650$	–
Vessel Segmentation	ARCADE	200	300
	XCAV	175	46
	CAXF	337	201
	Total	712	547

Table 1: Details of datasets used for pre-training and vessel segmentation.

Training Objective. In addition to the anatomical consistency loss, we include the pixel-level reconstruction loss (*i.e.,* MSE loss) following conventional MIM approaches. The overall training loss is given by:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{rec.}} + \mathcal{L}_{\text{cons.}} \tag{5}$$

Results

Datasets

Table 1 summarizes the datasets used for both pre-training and vessel segmentation.

Pre-Training. We assemble a corpus of $20,650$ coronary X-ray angiograms from five public datasets: ARCADE (Popov et al. 2024), CADICA (Jiménez-Partinen et al. 2024), Stenosis (Danilov et al. 2021), SYNTAX (Mahmoudi et al. 2025), and XCAD (Ma et al. 2021).

Vessel Segmentation. VasoMIM is evaluated on three benchmarks: one in-domain dataset, ARCADE, and two out-of-domain datasets, CAXF (Li et al. 2020) and XCAV (Wu et al. 2025). Note that the images of ARCADE used for vessel segmentation are not included in the pre-training set.

Implementation Details

Pre-Training. Our implementation is based on MAE (He et al. 2022). By default, ViT-B/16 (Dosovitskiy et al. 2021) is used as the backbone following previous works. VasoMIM is pre-trained for 800 epochs on an NVIDIA A6000 GPU, employing the AdamW optimizer (Loshchilov and Hutter 2019) with a batch size of 256 and an input resolution of 224×224.

Vessel Segmentation. We adopt U-Net (Ronneberger et al. 2015) as the segmentation decoder and fine-tune it end-to-end on ARCADE, CAXF and XCAV for 500 epochs each, using input images resized to 224×224. Optimization is performed using AdamW with an initial learning rate of $1e^{-4}$ and a weight decay of 0.05. A cosine-annealing schedule with $T\_\text{max} = 500$ is employed. All experiments are conducted on an NVIDIA A6000 GPU.

Method	ARCADE		CAXF		XCAV	
	DSC (%)	clDice (%)	DSC (%)	clDice (%)	DSC (%)	clDice (%)
Traditional						
Frangi Filter (Frangi et al. 1998)	41.30	40.91	64.01	65.73	58.46	57.15
From Scratch						
U-Net (Ronneberger et al. 2015)	$58.27_{\pm 1.33}$	$59.70_{\pm 1.40}$	$78.72_{\pm 0.74}$	$82.68_{\pm 0.87}$	$68.63_{\pm 2.80}$	$63.47_{\pm 3.33}$
Contrastive Learning						
MoCo v3 (Chen, Xie, and He 2021)	$60.99_{\pm 0.30}$	$62.68_{\pm 0.18}$	$77.76_{\pm 0.51}$	$80.91_{\pm 0.31}$	$70.85_{\pm 0.34}$	$63.97_{\pm 0.71}$
DINO (Caron et al. 2021)	$65.86_{\pm 0.49}$	$67.84_{\pm 0.52}$	$80.13_{\pm 0.53}$	$82.90_{\pm 0.51}$	$72.28_{\pm 0.96}$	$66.36_{\pm 1.17}$
Masked Image Modeling						
MAE (He et al. 2022)	$68.17_{\pm 0.29}$	$69.89_{\pm 0.22}$	$83.53_{\pm 0.14}$	$87.37_{\pm 0.21}$	$76.43_{\pm 0.17}$	$72.58_{\pm 0.49}$
SimMIM (Xie et al. 2022)	$66.92_{\pm 0.43}$	$68.93_{\pm 0.71}$	$82.24_{\pm 0.34}$	$85.77_{\pm 0.17}$	$75.10_{\pm 0.36}$	$69.98_{\pm 0.42}$
AMT (Liu, Gui, and Luo 2023)	$68.15_{\pm 0.23}$	$69.77_{\pm 0.38}$	$83.47_{\pm 0.09}$	$87.40_{\pm 0.04}$	$76.51_{\pm 0.20}$	$72.60_{\pm 0.44}$
DeblurringMIM† (Kang et al. 2024)	$\underline{68.60}_{\pm 0.44}$	$70.21_{\pm 0.37}$	$\underline{83.85}_{\pm 0.09}$	$\underline{87.78}_{\pm 0.20}$	$\underline{77.02}_{\pm 0.08}$	$\underline{73.58}_{\pm 0.19}$
CrossMAE (Fu et al. 2025)	$62.40_{\pm 0.33}$	$64.23_{\pm 0.27}$	$80.07_{\pm 0.13}$	$83.45_{\pm 0.19}$	$72.25_{\pm 0.24}$	$65.94_{\pm 0.15}$
HPM (Wang et al. 2025)	$66.82_{\pm 0.28}$	$68.49_{\pm 0.41}$	$82.61_{\pm 0.21}$	$86.18_{\pm 0.10}$	$75.48_{\pm 0.19}$	$70.79_{\pm 0.26}$
CheXWorld† (Yue et al. 2025)	$67.95_{\pm 0.26}$	$\underline{70.31}_{\pm 0.48}$	$80.64_{\pm 0.31}$	$82.65_{\pm 0.31}$	$73.74_{\pm 0.24}$	$67.13_{\pm 0.32}$
VasoMIM	$\mathbf{68.85}_{\pm 0.47}$	$\mathbf{70.56}_{\pm 0.36}$	$\mathbf{84.49}_{\pm 0.17}$	$\mathbf{88.33}_{\pm 0.09}$	$\mathbf{77.52}_{\pm 0.26}$	$\mathbf{74.18}_{\pm 0.34}$

Table 2: Comparison of state-of-the-art methods on ARCADE, CAXF and XCAV. All methods are reimplemented using their official codebases. The best results are highlighted in **bold** and the second-best results are underlined. Results are reported as "$\mathrm{mean} \pm \mathrm{std}$" over three random seeds, except for Frangi filter. † indicates that the model is specialized in medical imaging.

Evaluation Metrics. Dice similarity coefficient (DSC) and centerlineDice (clDice) (Shit et al. 2021) are adopted. Compared to DSC, clDice better captures topological correctness by measuring overlap between predicted and ground-truth vascular centerlines.

Main Results on Vessel Segmentation

All baselines are pre-trained on the angiogram dataset in Table 1 using their official implementations. Each model is fine-tuned with three random seeds, and we report all metrics as "$\mathrm{mean} \pm \mathrm{std}$".

In-Domain Dataset (ARCADE). When trained from scratch, U-Net achieves 58.27% DSC and 59.70% clDice on ARCADE. By pre-training on large-scale unlabeled data, our VasoMIM improves performance by 10.58% in DSC and 10.86% in clDice, reaching 68.85% DSC and 70.56% clDice, respectively, surpassing leading SSL baselines by a clear margin. Compared to Frangi filter, VasoMIM yields an absolute gain of 27.55% in DSC and 29.65% in clDice, underscoring the crucial role of our anatomy-aware pre-training in boosting segmentation performance.

Out-of-Domain Datasets (CAXF and XCAV). We further conduct experiments on out-of-domain datasets. Our VasoMIM consistently achieves the best performance, with 84.49% and 77.52% DSC on CAXF and XCAV, respectively. Notably, it outperforms the best baseline by $+0.64\%$ DSC on CAXF and $+0.50\%$ DSC on XCAV. This performance gap is even larger than that on the in-domain dataset (*i.e.*, $+0.25\%$ DSC), highlighting VasoMIM's strong generalizability and robustness.

Qualitative Results. Several cases from the three datasets are presented in Fig. 4. VasoMIM produces precise vessel segmentations, even for very thin or blurred vessels.

Guidance	$\mathcal{L}_{\mathrm{cons.}}$	ARCADE	CAXF
–	–	68.00	83.15
–	✓	68.45	84.03
✓	–	68.30	83.96
✓	✓	**68.85**	**84.49**

Table 3: Ablation study of the proposed anatomy-guided masking strategy and anatomical consistency loss. DSC is reported in this table.

Ablation Study

Ablation studies are conducted on ARCADE and CAXF with default settings highlighted in gray. Each model is fine-tuned with three random seeds. Table 3 shows the results of sequentially adding each proposed component to the vanilla baseline, *i.e.*, MAE ($\gamma = 0.5$).

Effectiveness of Anatomy-Guided Masking Strategy. Simply using the anatomy-guided masking strategy yields a clear performance boost ($+0.30\%$ DSC on ARCADE and $+0.81\%$ DSC on CAXF). To better understand this improvement, we measure the proportion of vessel-containing

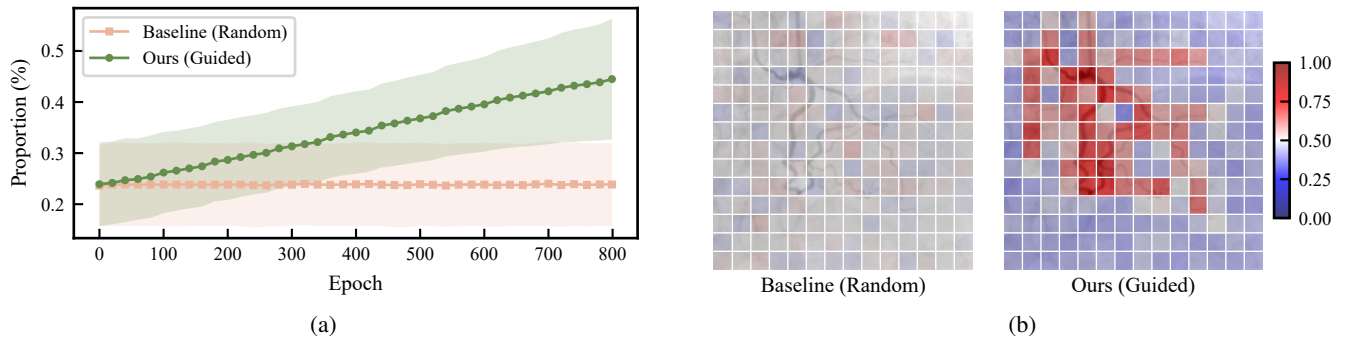


Figure 3: Some evidence of anatomy-guided masking strategy. (a) Proportion of vessel-containing patches in the masked patches during pre-training. (b) Patch-wise masking ratio over the pre-training process, *i.e.*, $\frac{1}{E}\sum_{j=1}^{E}\mathbb{I}\,(\text{Patch } x_i \text{ is masked in epoch } j)$.

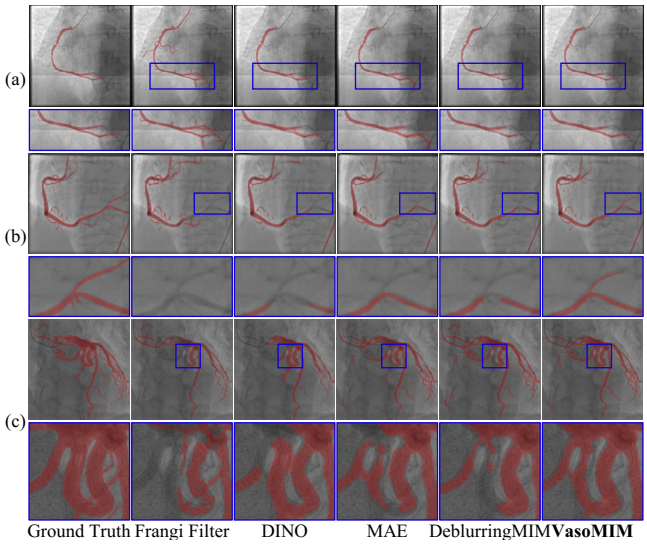


Figure 4: Qualitative results on (a) ARCADE, (b) CAXF, and (c) XCAV. Details are zoomed in within blue boxes.

Setting	SS ($\times 10^{-2}$) ↑	CHI ↑	DBI ↓
w/o $\mathcal{L}_{\text{cons.}}$	−4.19	17.11	25.32
w/ $\mathcal{L}_{\text{cons.}}$	**0.54**	**607.24**	**4.03**

SS: Silhouette Score; CHI: Calinski-Harabasz Index; DBI: Davies-Bouldin Index.

Table 4: Results of clustering metrics on XCAD.

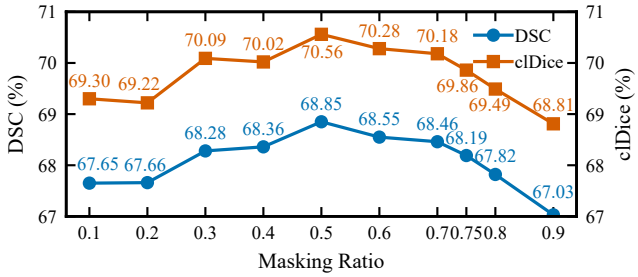


Figure 5: In-depth analysis of the masking ratio γ on ARCADE. γ is set to 0.5 in our default settings.

patches in the masked patches during the pre-training process. As shown in Fig. 3(a), our strategy masks a significantly larger number of vessel-containing patches, whereas the baseline masks a small and relatively constant proportion. We further visualize an example in Fig. 3 (b), where the masking ratio of each patch over pre-training is colored. Our strategy clearly favors masking patches rich in vascular anatomy, whereas the baseline shows no such preference.

Role of Anatomical Consistency Loss. Adopting the anatomical consistency loss $\mathcal{L}_{\text{cons.}}$ to the baseline also produces a notable gain (*e.g.*, $+0.88\%$ DSC on CAXF). This suggests that $\mathcal{L}_{\text{cons.}}$ enables the model to learn more discriminative vascular representations. To further verify the impact of $\mathcal{L}_{\text{cons.}}$, we first split patches into vessel-containing and background-only groups, then use UMAP (Healy and McInnes 2024) to project their representations from the model pre-trained *w/* and *w/o* $\mathcal{L}_{\text{cons.}}$ into a low-dimensional space, and evaluate how well these representations cluster. As shown in Table 4, the model pre-trained *w/* $\mathcal{L}_{\text{cons.}}$ achieves much higher SS and CHI, and a far lower DBI than the model pre-trained *w/o* $\mathcal{L}_{\text{cons.}}$, indicating more compact and well-separated clusters of vascular representations.

In-Depth Analysis

We next present a detailed analysis of the proposed VasoMIM. Default settings are highlighted in gray. All results are averaged over three random seeds.

Masking Ratio. As illustrated in Fig. 5, we observe that a moderate masking ratio (*i.e.*, 0.5) yields better performance. This result differs slightly from prior work, *e.g.*, MAE adopts a higher ratio of $\gamma = 0.75$. We hypothesize that this discrepancy stems from the unique characteristics of X-ray angiograms. Unlike natural images, which contain objects of varying sizes, vessels in X-ray angiograms occupy a relatively small portion of images. Using a larger masking ratio tends to mask background-only patches that carry little informative content. This may limit the model's ability to learn useful vascular representations.

Case	β_0	β_E	ARCADE	CAXF
Random	0	0	68.45	84.03
Weak-to-Strong	0	0.5	**68.85**	**84.49**
	0	1	68.52	84.24
	1	1	65.36	81.17
Strong-to-Weak	0.5	0	67.81	83.41

Table 5: Effects of different masking strategies. DSC is reported in this table.

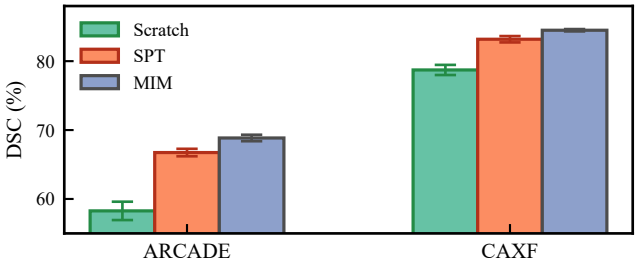


Figure 6: SPT *vs.* MIM. Our MIM (*i.e.*, VasoMIM) yields average $+2.11\%$ and $+1.31\%$ DSC improvements on ARCADE and CAXF, respectively.

Different Masking Strategies. We evaluate several masking strategies, with results summarized in Table 5. Simply increasing the degree of anatomical guidance does not consistently yield better performance, suggesting that a degree of randomness in the masking process is essential. Specifically, the configuration $\beta_0 = 0$, $\beta_E = 0.5$ achieves the best results, striking a balance between stronger anatomical guidance ($\beta_0 = \beta_E = 1$) and greater randomness ($\beta_0 = \beta_E = 0$). This outcome is quite intuitive. Aggressively masking patches with the highest anatomical relevance leads to most vessel-containing patches being masked. In such cases, the model is forced to reconstruct vascular anatomy solely from background-only patches with little to no semantic cues.

Strong-to-Weak Anatomical Guidance. We further investigate a reversed approach, adopting a strong-to-weak strategy during pre-training. As shown in Table 5, this approach leads to a clear performance drop, even compared to using a random masking strategy. This finding underscores the importance of our weak-to-strong masking design.

Segmentor-based Pre-Training *vs.* MIM. We compare a segmentor-based pre-training (SPT) approach with our MIM using the same segmentor (*i.e.*, U-Net). In SPT, the segmentor is pre-trained in a fully supervised manner using pseudo-labels generated by Frangi filter on the pre-training dataset and subsequently fine-tuned on downstream datasets. Fig. 6 demonstrates that both pre-training paradigms substantially outperform the model trained from scratch. Moreover, MIM further boosts DSC from 66.74% to 68.85% ($+2.11\%$) on ARCADE and from 83.18% to 84.49% ($+1.31\%$) on CAXF. Overall, these results indicate that our MIM learns more robust and transferable representations than SPT.

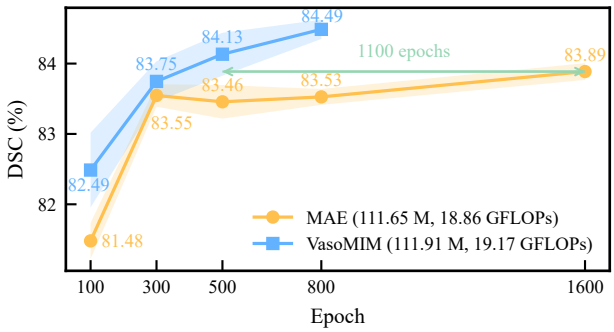


Figure 7: Fine-tuning on CAXF with models pre-trained for various epochs. GFLOPs are measured on an NVIDIA A6000 GPU using a single 224×224 masked RGB image.

Method	Ratio (%)		
	25	50	100
Scratch	38.09	50.67	58.27
MAE	56.58	63.04	68.17
VasoMIM	**56.93**	**63.64**	**68.85**

Table 6: Fine-tuning with 25%, 50%, and 100% training data on ARCADE. DSC is reported in this table.

Pre-Training and Data Efficiency. VasoMIM's advantages become even more pronounced under resource-limited scenarios. First, we fine-tune MAE and VasoMIM pre-trained for 100, 300, 500, and 800 epochs on CAXF. As shown in Fig. 7, VasoMIM consistently outperforms MAE at every pre-training duration. Notably, after only 300 epochs of pre-training ($\sim$ 6 hours), VasoMIM achieves a DSC of 83.75% on CAXF, slightly above MAE's 83.53%, which requires 800 epochs ($\sim$ 12 hours). Even when we extend MAE's pre-training to $1,600$ epochs ($\sim$ 24 hours) for a further performance boost, VasoMIM pre-trained for only 500 epochs ($\sim$ 10 hours) still performs better. Next, to test data efficiency, we fine-tune models on ARCADE using only 25%, 50%, or 100% of the training labels. Table 6 shows that VasoMIM outperforms MAE in all scenarios.

Conclusion

In this paper, we propose VasoMIM, a vascular anatomy-aware MIM framework specifically tailored for X-ray angiograms. By incorporating anatomical guidance into the masking strategy, our model can focus on vessel-relevant regions. Additionally, the proposed anatomical consistency loss significantly enhances the discriminability of learned vascular representations. Extensive experiments indicate that the proposed framework yields superior performance on three benchmarks. This work provides new perspectives on integrating anatomical knowledge into SSL frameworks.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC2415100, in part by the National Natural Science Foundation of China under Grant 62222316, Grant 62373351, Grant 82327801, Grant 62303463, in part by the Chinese Academy of Sciences Project for Young Scientists in Basic Research under Grant No. YSBR-104, in part by the Beijing Natural Science Foundation under Grant F252068, Grant 4254107, in part by Beijing Nova Program under Grant 20250484813, in part by China Postdoctoral Science Foundation under Grant 2024M763535, in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20251170.

References

Adams, R.; and Bischof, L. 1994. Seeded region growing. *IEEE TPAMI*, 16(6): 641–647.

Azizi, S.; et al. 2021. Big self-supervised models advance medical image classification. In *Proc. CVPR*, 3478–3488.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEIT: BERT pre-training of image transformers. In *Proc. ICLR*.

Caron, M.; et al. 2021. Emerging properties in self-supervised vision transformers. In *Proc. ICCV*, 9650–9660.

Chaitanya, K.; Erdil, E.; Karani, N.; and Konukoglu, E. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Proc. NeurIPS*, volume 33, 12546–12558.

Chen, J.; et al. 2024. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *MedIA*, 97: 103280.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, 1597–1607.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proc. CVPR*, 9640–9649.

Danilov, V. V.; et al. 2021. Real-time coronary artery stenosis detection based on modern neural networks. *Sci. Rep.*, 11(1): 7582.

Dosovitskiy, A.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*.

Esteva, A.; et al. 2019. A guide to deep learning in healthcare. *Nat. Med.*, 25(1): 24–29.

Frangi, A. F.; Niessen, W. J.; Vincken, K. L.; and Viergever, M. A. 1998. Multiscale vessel enhancement filtering. In *Proc. MICCAI*, 130–137.

Fu, L.; et al. 2025. Rethinking patch dependence for masked autoencoders. *TMLR*.

Gui, J.; et al. 2024. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE TPAMI*, 46(12): 9052–9071.

Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H. R.; and Xu, D. 2021. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Proc. MICCAIW*, 272–284.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proc. CVPR*, 16000–16009.

Healy, J.; and McInnes, L. 2024. Uniform manifold approximation and projection. *Nat. Rev. Methods Primers*, 4(1): 82.

Hinojosa, C.; Liu, S.; and Ghanem, B. 2024. ColorMAE: Exploring data-independent masking strategies in masked autoencoders. In *Proc. ECCV*, 432–449.

Huang, D.-X.; et al. 2024. SPIRONet: Spatial-frequency learning and topological channel interaction network for vessel segmentation. *arXiv*.

Huang, D.-X.; et al. 2025. Real-time 2D/3D registration via CNN regression and centroid alignment. *IEEE TASE*, 22: 85–98.

Jiménez-Partinen, A.; et al. 2024. CADICA: A new dataset for coronary artery disease detection by using invasive coronary angiography. *Expert Syst.*, 41(12): e13708.

Kakogeorgiou, I.; et al. 2022. What to hide from your students: Attention-guided masked image modeling. In *Proc. ECCV*, 300–318.

Kang, Q.; et al. 2024. Deblurring masked image modeling for ultrasound image analysis. *MedIA*, 97: 103256.

Kheiri, B.; Simpson, T. F.; Osman, M.; German, D. M.; Fuss, C. S.; and Ferencik, M. 2022. Computed tomography vs invasive coronary angiography in patients with suspected coronary artery disease: A meta-analysis. *JACC: Cardiovasc. Imaging*, 15(12): 2147–2149.

Li, G.; Zheng, H.; Liu, D.; Wang, C.; Su, B.; and Zheng, C. 2022a. SemMAE: Semantic-guided masking for learning masked autoencoders. In *Proc. NeurIPS*, volume 35, 14290–14302.

Li, R.-Q.; Bian, G.-B.; Zhou, X.-H.; Xie, X.; Ni, Z.-L.; and Hou, Z. 2020. CAU-net: A novel convolutional neural network for coronary artery segmentation in digital substraction angiography. In *Proc. ICONIP*, 185–196.

Li, X.; Wang, W.; Yang, L.; and Yang, J. 2022b. Uniform masking: Enabling MAE pre-training for pyramid-based vision transformers with locality. *arXiv*.

Li, Y.; Luan, T.; Wu, Y.; Pan, S.; Chen, Y.; and Yang, X. 2024. AnatoMask: Enhancing medical image segmentation with reconstruction-guided self-masking. In *Proc. ECCV*, 146–163.

Liu, Z.; Gui, J.; and Luo, H. 2023. Good helper is around you: Attention-driven masked image modeling. In *Proc. AAAI*, volume 37, 1799–1807.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. In *Proc. ICLR*.

Ma, Y.; et al. 2021. Self-supervised vessel segmentation via adversarial learning. In *Proc. ICCV*, 7536–7545.

Mahmoudi, S. S.; et al. 2025. X-ray coronary angiogram images and SYNTAX score to develop machine-learning algorithms for CHD diagnosis. *Sci. Data*, 12(1): 471.

Members, W. C.; et al. 2022. 2021 ACC/AHA/SCAI guideline for coronary artery revascularization: A report of the American College of Cardiology/American Heart Association Joint Committee on clinical practice guidelines. *JACC*, 79(2): e21–e129.

Pérez-García, F.; et al. 2025. Exploring scalable medical image encoders beyond text supervision. *Nat. Mach. Intell.*, 7: 119–130.

Popov, M.; et al. 2024. Dataset for automatic region-based coronary artery disease diagnostics using X-ray angiography images. *Sci. Data*, 11(1): 20.

Ronneberger, O.; et al. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, 234–241.

Ruan, J.; Li, J.; and Xiang, S. 2024. VM-UNet: Vision Mamba UNet for medical image segmentation. *arXiv*.

Shit, S.; et al. 2021. clDice: A novel topology-preserving loss function for tubular structure segmentation. In *Proc. CVPR*, 16560–16569.

Taghizadeh Dehkordi, M.; Doost Hoseini, A. M.; Sadri, S.; and Soltanianzadeh, H. 2014. Local feature fitting active contour for segmenting vessels in angiograms. *IET CV*, 8(3): 161–170.

Tang, F.; et al. 2025. MambaMIM: Pre-training Mamba with state space token interpolation and its application to medical image segmentation. *MedIA*, 103606.

Vaduganathan, M.; Mensah, G. A.; Turco, J. V.; Fuster, V.; and Roth, G. A. 2022. The global burden of cardiovascular diseases and risk: A compass for future health. *JACC*, 80(25): 2361–2371.

Valanarasu, J. M. J.; and Patel, V. M. 2022. UNeXt: MLP-based rapid medical image segmentation network. In *Proc. MICCAI*, 23–33.

Wang, C.; et al. 2020. Tensor-cut: A tensor-based graph-cut blood vessel segmentation method and its application to renal artery segmentation. *MedIA*, 60: 101623.

Wang, H.; et al. 2025. Bootstrap masked visual modeling via hard patch mining. *IEEE TPAMI*. DOI:10.1109/TPAMI.2025.3557001.

Wang, J.; Chen, J.; Chen, D.; and Wu, J. 2024. LKM-UNet: Large kernel vision Mamba UNet for medical image segmentation. In *Proc. MICCAI*, 360–370.

Wu, C.-H.; et al. 2025. DeNVeR: Deformable neural vessel representations for unsupervised video vessel segmentation. In *Proc. CVPR*, 15682–15692.

Wu, L.; Zhuang, J.; and Chen, H. 2024. VoCo: A simple-yet-effective volume contrastive learning framework for 3D medical image analysis. In *Proc. CVPR*, 22873–22882.

Xie, Y.; Gu, L.; Harada, T.; Zhang, J.; Xia, Y.; and Wu, Q. 2024. Rethinking masked image modelling for medical image representation. *MedIA*, 98: 103304.

Xie, Z.; et al. 2022. SimMIM: A simple framework for masked image modeling. In *Proc. CVPR*, 9653–9663.

Xu, Z.; Liu, Y.; Xu, G.; and Lukasiewicz, T. 2025. Self-supervised medical image segmentation using deep reinforced adaptive masking. *IEEE TMI*, 44(1): 180–193.

Yuan, J.; et al. 2023. HAP: Structure-aware masked image modeling for human-centric perception. In *Proc. NeurIPS*, volume 36, 50597–50616.

Yue, Y.; Wang, Y.; Tao, C.; Liu, P.; Song, S.; and Huang, G. 2025. CheXWorld: Exploring image world modeling for radiograph representation learning. In *Proc. CVPR*, 20778–20788.

Zhou, H.-Y.; Yu, S.; Bian, C.; Hu, Y.; Ma, K.; and Zheng, Y. 2020. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In *Proc. MICCAI*, 398–407.

Zhou, S. K.; et al. 2021. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE*, 109(5): 820–838.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2019. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE TMI*, 39(6): 1856–1867.

Zhuang, J.; Luo, L.; Wang, Q.; Wu, M.; Luo, L.; and Chen, H. 2025a. Advancing volumetric medical image segmentation via global-local masked autoencoders. *IEEE TMI*. DOI:10.1109/TMI.2025.3569782.

Zhuang, J.; et al. 2025b. MiM: Mask in mask self-supervised pre-training for 3D medical image analysis. *IEEE TMI*. DOI:10.1109/TMI.2025.3564382.