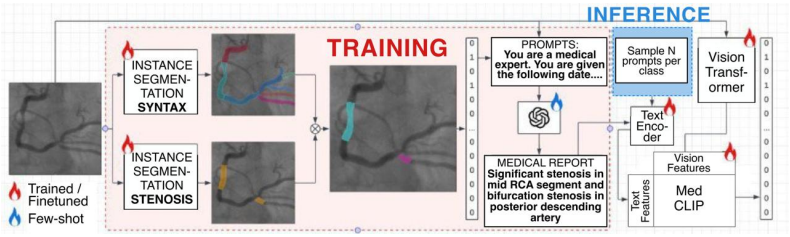
ROBT310 Final Project Team 2: Advancing Coronary Angiography Vessel Image Analysis via Guided Self-Supervised Masking Pre-training Strategy

Shakhnazar Zhumabekov, Dilnaz Ualiyeva,
Kamila Assylbek, Taikozha Turdyakyn

***Abstract*— Coronary angiography plays a crucial role in diagnosing coronary artery disease, though it remains a complex, time-intensive process that heavily relies on physician expertise. Recent advances in deep learning offer promise for faster, more automated analysis, yet these methods are hindered by the need for large, well-annotated datasets, which are costly and labor-intensive to obtain. In response, our project introduces a self-supervised learning approach using a Frangi filtering for Guided Masked Autoencoder framework tailored specifically for coronary angiography data. By training on unannotated datasets through a pretext task of unmasking angiography patches, the model learns valuable visual representations without labeled data. Once trained, these model backbones can be adapted to segmentation tasks with minimal labeled data, achieving performance on par with fully supervised models while enhancing diagnostic consistency and clinical efficiency. This approach could help streamline coronary artery analysis, bringing faster and more accessible diagnostic support to clinical settings.**



***Index Terms*— coronary angiograms, self-supervised learning, masked autoencoders, deep learning, pre-text tasks, frangi filter.**

I. Introduction

Coronary artery disease (CAD) remains one of the leading causes of mortality worldwide, underscoring the need for early and accurate diagnostic tools [1]. The current gold standard for diagnosing CAD is coronary angiography, an imaging procedure that uses X-rays and contrast agents to visualize coronary arteries. This technique provides detailed insights into vessel structure, allowing for the detection of stenosis and other abnormalities; however, it heavily relies on the expertise of physicians and is subject to interobserver variability, which can lead to inconsistent diagnoses [2].

Recent advancements in machine learning have introduced the potential for automating coronary artery segmentation, aiming to enhance diagnostic accuracy and streamline clinical workflows. Several researchers applied novel supervised Deep Learning methods such as [3], [4], [5]. Self-supervised learning approaches have become particularly valuable in medical imaging, as they reduce reliance on annotated data while achieving robust performance on downstream tasks such as segmentation. For instance, self-supervised methods like the Diffusion Adversarial Representation Learning (DARL) model [6] and the adversarial learning-based framework proposed by Ma et al. [7] have shown that adversarial training can generate effective vessel representations and segmentation masks without extensive manual annotation. These models leverage large volumes of unannotated data to learn vascular structures through adversarial cycles, making them well-suited for complex and variable coronary imaging conditions.

In addition, ensemble learning approaches, such as those employed by Bilal et al. [8], have demonstrated success in combining multiple model predictions to improve segmentation accuracy in coronary artery analysis. This multi-model approach helps to handle variations in imaging quality and patient anatomy, a critical factor in real-world applications. Transformer-based architectures like Mask2Former [9] have also contributed to the field by creating a universal segmentation framework that generalizes across multiple segmentation tasks, making it particularly useful for the intricate challenges of coronary vessel analysis.

The proper Coronary Artery Disease detection and localization requires not only binary segmentation of the vessels from the background, but also accurate segmentation of the each anatomical vessel structure in separate class, creating a task of the instance segmentation. This is useful for numerical calculations of the stenosis related to Coronary Artery Disease [10]. There is a very limited amount of research done in automatic instance segmentation of the vessel regions and structures, some of which is highlighted by [11] where CNN-

Submitted November 2024.
Corresponding authors: Students.
Shakhnazar Zhumabekov with Nazarbayev University, Nur-Sultan, Kazakhstan, School of Engineering and Digital Sciences, Kabanbay Batyr ave. 53, (e-mail: shakhnazar.zhumabekov@nu.edu.kz).
Dilnaz Ualiyeva with Nazarbayev University, Nur-Sultan, Kazakhstan, School of Sciences and Humanities, Kabanbay Batyr ave. 53, (e-mail: dilnaz.ualiyeva@nu.edu.kz).
Kamila Assylbek with Nazarbayev University, Nur-Sultan, Kazakhstan, School of Sciences and Humanities, Kabanbay Batyr ave. 53, (e-mail: kamila.assylbek@nu.edu.kz).
Taikozha Turdyakyn with Nazarbayev University, Nur-Sultan, Kazakhstan, School of Sciences and Humanities, Kabanbay Batyr ave. 53, (e-mail: taikozha.turdyakyn@nu.edu.kz).

based approach is used for segmentation of 11 different vessel structures and [12] which utilizes adversarial training setup to separate 20 different anatomic vessel structures present in coronary angiography image. Yet, the models trained with these approaches are trained in supervised manner and on a private datasets, so the reproducibility of their results is not possible on open benchmarks for this particular task. In recent years, works by [10], and [13] attempt to approach instance segmentation using semi-supervised or ensemble approaches, highlighting the fact that there is a very small amount of labeled data for instance segmentation of all particular anatomical structures present in coronary angiography image to make supervised approaches effective. However, there is no work of adapting any well-established Self-Supervised Learning paradigm such as that with Masked Autoencoding [14], Contrastive Learning [15], BYOL [16] or any other pre-training paradigm highlighted by Kolesnikov et al. [17] for Vessel Structure segmentation. Building on these advancements, our project seeks to develop a self-supervised learning framework capable of learning robust visual representations under constraints of extremely small amount of data for being useful in medical imaging analysis tasks such as instance vessel segmentation. By leveraging large-scale unannotated datasets, and examining guided masking autoencoding and reconstruction as a self-supervised paradigm, we aim to create a reliable diagnostic tool that matches or surpasses human-level performance in coronary artery image analysis, ultimately enhancing CAD diagnosis in clinical settings. The key findings of this work include the following:

- We present a novel Guided Masking Strategy based on Frangi Filtering and Image processing techniques for masking specifically image patches containing vessels. The following approach allows encoder-decoder models to focus more on learning visual features from image patches containing predominantly vessels rather than on patches containing background, making procedure of pre-training more controllable in setting of low signal to noise ratio present in CAD image. We argue that this method is more suitable in domain of CAD imaging compared to Random Masking strategies due to unequal distribution of the useful information to the background noise in the image.
- We showcase that lower masking ratios of 35-25% are more favorable for medical image analysis, compared to high masking ratios of 75% initially proposed by authors of [14].
- We develop intuition over utilizing Frangi Filter for proper generation of vessel probability map after applying it to get binary segmentation mask of vessel structure present in the image. We present the range of sigma parameters which is most useful for vessels with appropriate thickness to be distinguished from the background.
- We develop own Frangi Filter implementation based on tensor computations for computation on batches of images efficiently, as previous implementations were adapted for numpy arrays.

II. Related Works

A. Motivation behind Self-Supervised Learning for Coronary Angiography

Research on automated vessel segmentation image analysis in coronary angiography has advanced significantly, particularly through the integration of deep learning and self-supervised learning methods. Traditional vessel instance segmentation and classification techniques often required extensive annotated datasets and were constrained by the complexities of medical imaging, such as low signal to noise ratio, and significant amount of motion and occlusion artifacts, background noise and low contrast, which make coronary vessels challenging to distinguish and annotate. This led to the exploration of SSL methods for vessel segmentation, which aim to alleviate the need for large labeled datasets while still achieving high accuracy [18].

B. Self-Supervised Visual Representation Learning

In the Self-Supervised Learning paradigm, the visual representations are learned not from direct labeled data, but by solving pretext tasks that uncover patterns and structures within the unannotated data itself. Unlike traditional supervised learning, which relies on large quantities of labeled images to guide feature extraction, self-supervised learning enables models to capture meaningful features autonomously, making it ideal for scenarios with limited labeled data [15]. Numerous pre-text tasks were designed by researchers, for instance, in the rotation task, models learn to recognize visual features by predicting the correct orientation of rotated images, which helps them understand object shapes and structures [17]. The exemplar task involves distinguishing between augmented versions of the same image, teaching the model to recognize distinct instances despite transformations. The jigsaw task, where an image is split into patches and shuffled, requires the model to rearrange them, promoting an understanding of spatial relationships [17].

More advanced methods include the Masked Autoencoder (MAE), which masks out parts of an image and tasks the model with reconstructing the missing regions, allowing it to learn contextual information [14]. SimCLR leverages contrastive learning by maximizing agreement between different augmented views of the same image, forcing the model to capture essential visual features [15]. Similarly, BYOL (Bootstrap Your Own Latent) uses two neural networks to create and compare different representations of the same image without requiring negative samples, helping the model learn meaningful representations purely from image data [16]. Together, these pretext tasks allow self-supervised models to capture rich, transferable features without labeled data, making them ideal for medical imaging applications where labeled data is scarce.

C. Self-Supervised Learning in Medical Image analysis

In the domain of Medical Image Analysis, predominant downstream task for the backbone model trained on Self-Supervised Learning paradigm is Semantic or Instance Seg-

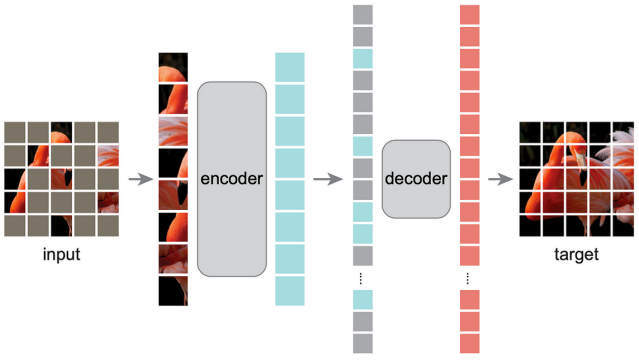


Fig. 1: This figure illustrates an MAE (Masked Autoencoder) architecture for self-supervised learning, where a random subset of image patches (e.g., 75 %) is masked out during pre-training [14]. The encoder processes only the visible patches, while the decoder reconstructs the full image, using both encoded patches and masked tokens. After pre-training, the decoder is removed, allowing the encoder to effectively recognize uncorrupted images. This technique enables feature extraction from partial data, a promising approach for scenarios with limited labeled images.

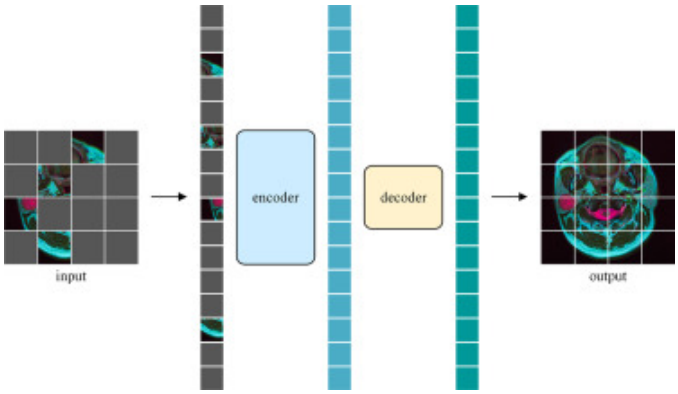


Fig. 2: This figure illustrated SwinMAE masking pretraining paradigm for brain tumor images as shown in the work by [20].

mentation of a particular lesion in the medical image. Therefore, in the datasets for medical imaging there is no availability of the contrasting classes and images, most of the images are comprising of the single class with segmentation classes representing a particular lesion. Due to such feature of the medical imaging, the applications of the SimCLR [15] or BYOL [16] are very limited as there is no labels to contrast among and apply contrastive loss. Recent advancements in SSL have shown a shift towards masked autoencoding strategies, which are label-agnostic and enable the recovery of essential visual features from raw image data through a reconstruction pre-task. Notably, AnatoMask [19] introduces a masked image modeling (MIM) approach that dynamically identifies and masks anatomically significant regions within 3D medical images, enhancing pretraining efficiency and ultimately improving performance in segmentation tasks across various imaging modalities.

Moreover, the Swin MAE framework has been adapted for different applications within medical imaging. One variant [20] focuses on brain tumor segmentation, effectively leveraging masked autoencoding to learn from limited medical data. This general-purpose approach demonstrates the capacity to extract meaningful semantic features, allowing for improved performance in downstream tasks without relying heavily on labeled datasets. Another variant of Swin MAE is specifically tailored for dental applications [21], employing a masked autoencoder framework with the Swin Transformer backbone to address the challenges of limited annotated dental radiographs. This method effectively extracts semantic features from small datasets, achieving performance comparable to supervised approaches in tasks such as teeth numbering and restoration detection. Additionally, a convolutional masked image modeling [22] technique has been proposed for boosting downstream dense prediction tasks in pathology images. This method utilizes Mask tokens to facilitate information propagation and imposes transformation constraints to achieve affine and color-invariant embeddings, leading to significant improvements in transfer learning performance on standard benchmark datasets.

Together, these studies underscore the potential of self-supervised learning methodologies, particularly masked autoencoding, to address the challenges posed by the scarcity of labeled data in medical imaging. By focusing on the recovery of critical visual features and enhancing model robustness, these approaches pave the way for more effective segmentation and analysis of medical images, including applications in brain tumors, dental imaging, and pathology.

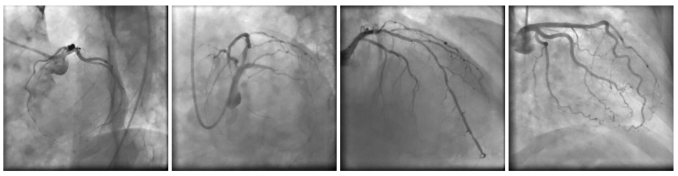
D. Self-Supervised Learning for Instance Segmentation

One key approach involves leveraging self-supervised learning frameworks. Techniques such as Mask DINO and Masked-Attention Transformers, for example, unify object detection and segmentation, supporting both instance and semantic segmentation without requiring task-specific adjustments. These models demonstrate that self-supervised frameworks can enhance segmentation performance across various applications by generalizing learned representations from unannotated data [9], [23].

E. Self-Supervised Learning applied to Coronary Angiography Segmentation

In coronary angiography specifically, several methods have adopted adversarial learning for vessel segmentation, such as the self-supervised model proposed by Ma et al. [7], which leverages an adversarial framework to improve vessel delineation in noisy angiograms. By generating synthetic vessels, this method enables the model to learn fine vessel structures and distinguish them from complex backgrounds. Diffusion adversarial representation learning is another emerging method, enhancing the ability to identify vessel structures through a denoising diffusion probabilistic model, allowing for robust segmentation without the extensive manual annotations that supervised methods typically require [6]. Lastly, [18] use Generative Adversarial Networks for training Deep subtraction algorithm applied to the semantic segmentation task on the vessel image data.

Further advancements have also focused on coronary artery stenosis localization, a critical component for diagnosing coronary artery disease (CAD). Ensemble learning techniques,



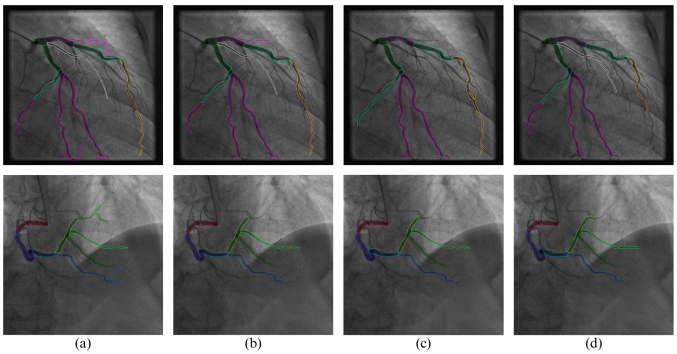
(a) LAO 45° CAU 35° (b) LAO 45° CAU 25° (c) RAO 30° CRA 20° (d) RAO 28° CAU 15°

Fig. 3: The following figure illustrates sample unlabeled images from CADICA dataset presenting coronary angiography which will be used for model pre-training [24].

such as those presented in recent ARCADE challenges, provide multi-phase and multi-view segmentation approaches that boost segmentation precision and enable accurate identification of stenotic regions and vessel segments. These methods segment the coronary artery tree by combining models like YOLO-based frameworks for vessel enhancement, addressing inter-observer variability and reducing errors in clinical analysis [8], [10], [13]. With the accuracies of only 86.3% in task of segment classification, supervised approaches developed by [13] showcase the importance of developing models with higher abilities to learn from limited amount of data, such as those developed in Self-Supervised Learning paradigm. The task of segment classification is of particular importance as it is necessary for structural understanding of the vessel anatomy and understanding of the stenosis levels [8].

F. Coronary Angiography datasets for pre-training and benchmarking

In collaborative effort between medical community and deep learning researchers, several datasets were created for benchmarking segmentation models using well-annotated coronary data. Researchers in [7] collected a 1621 coronary angiography masks and images with a release of XCAD dataset. Another dataset, CADICA was formed at Hospital Universitario Virgen de la Victoria, Málaga, Spain [24].The invasive coronary angiography (ICA) videos were acquired as Digital Imaging and Communication in Medicine (DICOM) files recorded at 10 frames per second and with different duration (4-8 seconds) depending on the projection used, but they were converted to PNG images for effortless management. The frame size of each video is 512 x 512 pixels, while the length of the videos varies from 1 to 151 frames. It contains 40 000 image data and segmentation masks for coronary angiography. Lastly, authors of [18] utilize 38 210 live CAD images represented in LM-CAD dataset designed for semantic segmentation of the vessels. From these datasets, we take 55 000 unlabaled images for our tasks, with skipping most of the images from CADICA dataset, taking only 15169 from it. While mentioned datasets are significant in sizes, they don't contain segmentation masks for instance segmentation of the anatomical structure of the vessels, containing only binary segmentation. Therefore, in addition to binary segmentation datasets, benchmark datasets like ARCADE have been instrumental in evaluating and advancing instance segmentation models. This dataset provides small-scale annotated X-ray coronary angiograms, allowing researchers to pre-train their models on previously mentioned datasets, and apply pre-trained models to the downstream task of instance segmentation by training and benchmarking on this dataset [2]. As a result, for instance segmentation, only ARCADE is a suitable benchmark dataset, but there is a notable amount of data which can used for SSL approaches.



(a) (b) (c) (d)

Fig. 4: This figure illustrated sample instance segmentation masks along with an images of the coronary angiography sampled from ARCADE dataset [2].

Together, these advancements underscore the promise of self-supervised and adversarial learning frameworks for achieving accurate, efficient coronary artery segmentation. By reducing dependency on annotated data and improving adaptability to real-world clinical settings, these methods offer a path forward for more reliable and accessible CAD diagnostics.

III. Methodology

In this project, we implement a self-supervised learning model designed to automate feature detection and segmentation in coronary angiograms, thereby reducing the need for manual interpretation and annotation. Our approach builds on recent advancements in applying Masked Autoencoder pretraining strategy for medical image analysis highlighted by the works of [19], [22], [21], [20] to the large corpus of the unlabeled CAD data present in CADICA, XCAD, LM-CAD datasets for later adoption of the trained backbone to the instance segmentation as to the downstream task on the ARCADE dataset.

A. Data Collection and Preprocessing

Given unlabeled amount of data from 3 main datasets for Masked Pre-training, we are keeping 55 000 images in one folder for feeding into pretraining backbone network. This data allows the model to learn vessel representations effectively while minimizing the need for extensive manual annotation [2], [18]. Pre-processing involves contrast enhancement and noise reduction to improve vessel visibility, critical for training in noisy environments with low-contrast images [10]. For input standardization, we apply normalization techniques to ensure uniformity across different imaging conditions and patient-specific variations in the angiograms [7]. It is also crucial to maintain same sizes across the datasets, they are kept to be 512 by 512 pixels, and reduced later to 224 by 224 in order to have equal patches during masked pre-training of the backbone models.

B. Self-Supervised Learning Framework

The core of our approach involves self-supervised representation learning using an masked autoencoding setup. We explore several MAE approaches with default Encoder-Decoder design architectures provided by their authors in [14] and [20]. The encoder in the network is responsible for encoding the images into corresponding tokens and mapping them to a high-dimensional semantic space, while decoder will restore images based on latent space representations. The loss function used for training calculates the mean square error of the original images and the reconstructed images. Only the mask patches are involved in the loss calculation, which makes the network training focus on predicting the mask patches [20] [14]. **for the classical MAE architecture**

- **Encoder:** the whole image is divided into regular non-overlapping patches of 16×16 size in the patch partition layer, following the original design of the Vision Transformer (ViT). Each patch is then flattened into a 1D vector and subsequently passed through a linear embedding layer, which maps the patch features to a fixed embedding size (e.g., 768 dimensions, consistent with the ViT-B/16 configuration). These embedded tokens are further augmented with positional encodings to retain spatial information.
- **Decoder:**in a regular MAE, the decoder is designed with flexibility and efficiency in mind, as its primary goal is to reconstruct the missing regions of the input from the latent representations produced by the encoder. Typically, decoder weights are discarded after training, and only the encoder is retained for transfer learning, so the decoder prioritizes computational simplicity. The MAE decoder receives two inputs: the latent tokens output by the encoder and the mask tokens corresponding to the masked patches. The mask tokens are learnable parameters that act as placeholders for the missing regions.

For the Swin MAE architecture

- **Swin Encoder:** the whole image is divided into regular non-overlapping patches of 4×4 size in the patch partition layer, and the length of tokens is later mapped to 96 by the linear embedding layer, which is the value used in Swin-T. The tokens are then continued into the subsequent four swin Transformer blocks in total, with no patch merging layer after the last block, consistent with the encoder of SwinUnet [25] used in the downstream task.
- **Swin Decoder:** in general, the decoder weights are dropped after training, and only the encoder weights are used in transfer learning for downstream tasks, so the decoder can be designed flexibly. A lightweight decoder design will, on the one hand, reduce computation and memory usage, which can reduce training time and allow using a larger batch size, and on the other hand, make the network training more focused on the encoder, which leads to better transfer results on downstream tasks. We use decoder architecture largely inspired by [25] but with few key differences. Instead of using the final patch expanding layer to restore the original image dimensions, a predictor projection layer is directly used to map the dimensions of the tokens to 48 as MAE does. In addition, no skip connection layer is added between the encoder and decoder.

Both architecture designs of MAE paradigms are going to be experimented for reconstruction task.In both scenarios, the reconstruction task involves predicting the **masked regions** of the input (e.g., image patches) using the visible regions processed by the encoder. The decoder reconstructs the input, and the **reconstruction loss** is computed using **Mean Squared Error (MSE)**, focusing only on the masked patches:

$$\text{MSE Loss} = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2$$

where x_i is the true value, $\hat{x}_i$ is the reconstruction, and N is the number of masked patches. This loss ensures the model learns meaningful representations for transfer learning, leveraging self-supervised learning to train on large unlabeled datasets.

C. Classical Masking methods and Masking ratio experiments

During the pre-training, according to the [14] and [20] standard masking ratio for images was deduced empirically to be 75%, yet it is not clear how such significant masking ratio can affect and help in learning to reconstruct very thin visual features present in coronary angiography images. We argue that such masking ratio is not suitable for coronary angiography imaging as in this case model will learn to reconstruct predominantly redundant background noise which is a large portion of an image, and will be large portion of a masked patches too. Therefore under conditions of the random masking the initial masking ratio is chosen to be initially as low as about 20% then moved to the 35% incrementally moving to the 75%. The initial masking method is going to be random masking introduced by [14] and followed across other MAE paradigms including [19], [21] and [22]. It can be formulated as in MAE, **random masking** is a preprocessing step where a subset of input tokens (e.g., image patches) is randomly selected and removed (masked). Let the input data be represented as a matrix $X \in \mathbb{R}^{P\times D}$, where P is the number of patches, and D is the feature dimension of each patch. A predefined **masking ratio** $r \in [0, 1]$ determines the fraction of patches to mask. For each patch i, a mask value $M_i \in \{0, 1\}$ is sampled from a Bernoulli distribution:

$$M_i \sim \text{Bernoulli}(1 - r),$$

where $M_i = 1$ indicates the patch is visible and $M_i = 0$ indicates it is masked. The masked input X_m is then obtained by retaining only the visible patches using an element-wise product:

$$X_m = M \odot X,$$

where $\odot$ denotes the Hadamard (element-wise) product. Masked patches are set to zero or replaced by learnable *mask tokens*. This randomness ensures the model generalizes well by

exposing it to diverse masked patterns during training, while the masking ratio r controls the difficulty of the reconstruction task. Masking is applied only during training; the full input is used during inference.

The training will undergo for 200 epochs with a batch size of 96 for all of the masking ratio and method experiments. Everything in this part will be used on composition of LM-CAD, CADICA, XCAD datasets with a distribution of 95% being in train and 5% of the data belonging to the test set.

D. Guided Masking method and Masking Ratio experiments

In our guided masking strategy, we introduce the goal to focus the masking on areas of an image that are likely to contain vessel-like structures. In this way it is assumed to provide better representation learning of a vessel features, since encoder and decoder both will be now adapted more to reconstruct and learn patches containing vessels instead of a background sampled from random masking of the patches. The process starts by applying **Gaussian smoothing** to the input image $I \in \mathbb{R}^{H \times W}$, where H and W are the height and width of the image, respectively. The Gaussian smoothing operation is defined as:

$$I_{\text{smoothed}}(x, y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} I(x+i, y+j) \cdot \mathcal{G}(i, j, \sigma),$$

where $\mathcal{G}(i, j, \sigma)$ is the Gaussian kernel, typically defined as:

$$\mathcal{G}(i, j, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{i^2 + j^2}{2\sigma^2}\right),$$

with σ controlling the level of smoothing. After smoothing, a **Frangi filter** is applied to enhance vessel-like structures. The Frangi filter is designed to highlight the linear structures in the image, and it is typically defined by the eigenvalues of the Hessian matrix H, computed at each pixel:

$$H = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{bmatrix},$$

where I_{xx}, I_{xy}, I_{yy} are the second-order image derivatives. The eigenvalues λ_1 and λ_2 of the Hessian matrix represent the local curvature in the principal directions. For vessel-like structures, the Frangi filter uses the ratio of eigenvalues:

$$F(x, y) = \left(\frac{\lambda_1}{\lambda_2}\right) \cdot \exp\left(-\frac{|\lambda_1|^2 + |\lambda_2|^2}{2\sigma_{\text{filter}}^2}\right),$$

where σ_{filter} is a scale parameter. The result of applying the Frangi filter is a **vessel probability map** $V \in \mathbb{R}^{H \times W}$, where each pixel value $V(x, y)$ represents the likelihood of that pixel being part of a vessel.

Next, we **normalize** the vessel probability map to a range of $[0, 1]$. The normalization is done by subtracting the minimum value and dividing by the range:

$$V_{\text{normalized}}(x, y) = \frac{V(x, y) - \min(V)}{\max(V) - \min(V)}.$$

Following normalization, the image is divided into non-overlapping patches of size 16×16, creating a total of $P = \frac{H}{16} \times \frac{W}{16}$ patches. The image is restructured into a set of patches $P = \{p_1, p_2, \ldots, p_P\}$, where each patch p_i contains 16x16 pixel values from the image. For each patch p_i, the **average intensity** $I_{\text{avg}}(p_i)$ is calculated by averaging the values in the vessel probability map for the pixels inside the patch:

$$I_{\text{avg}}(p_i) = \frac{1}{16^2} \sum_{(x,y) \in p_i} V_{\text{normalized}}(x, y).$$

This gives a scalar value for each patch representing the likelihood of that patch containing a vessel. Once the average intensities are computed for all patches, they are sorted in descending order, such that the patches with the highest average intensities are those most likely to contain vessel structures.

Finally, **guided masking** is performed by selecting the patches with the highest intensities for masking. For a given masking ratio r, we mask the top $r \times P$ patches, where r is the fraction of patches to be masked. For example, if $r = 0.35$, 35% of the patches will be selected for masking. To ensure a balance between vessel and background, the next step is to select a second fraction from the least likely patches (background patches). Suppose that 60% of the masked patches are chosen from the patches with the highest average intensity (most likely to contain vessels), and the remaining 40% are selected from patches with the lowest intensities (background). Mathematically, we can define the masking procedure as follows:

1) **Masking for vessel regions**: Select the top $0.6 \times 0.35 \times P$ patches from the sorted list with the highest average intensities $I_{\text{avg}}(p_i)$.
2) **Masking for background regions**: Select the next $0.4 \times 0.35 \times P$ patches with the lowest $I_{\text{avg}}(p_i)$.

These selected patches are then masked in the image, where their pixel values are set to zero or replaced with learnable mask tokens. This guided masking strategy ensures that the model focuses on reconstructing vessel-related regions while also considering some background context.

E. Evaluation Metrics

We plan to evaluate reconstruction quality for both random and guided masking methods by looking into dynamics of convergence and MSE loss values shown in previous subsection, same applied to the masking ratios. Additionally, we plan to look on visual quality of a reconstruction under the same masking ratio on which it was trained.

IV. Experimental Platform

The models and methods for masking were evaluated by training on an NVIDIA RTX 6000 GPU, which is equipped with 24 GB of GDDR6 memory, offering high computational power suitable for deep learning tasks. The RTX 6000 leverages the NVIDIA Turing architecture, which includes advanced features like Tensor Cores for optimized matrix operations, making it particularly efficient for training large

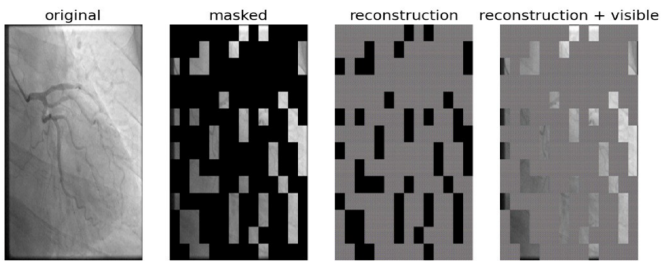


Fig. 5: Swin MAE random reconstruction results trained for 50 epochs on a dataset of 55 000 images with a masking ratio of 75% proposed by authors as best for medical imaging [20]

neural networkprovided the necessary computational resources for efficient training. The deep learning framework used was PyTorch, and the Frangi filter was initially implemented using the skimage library for vessel enhancement. The training environment was supported by essential libraries such as NumPy for numerical operations, and Matplotlib for visualization. CUDA 11.3 and cuDNN 8.1 were utilized for GPU acceleration. The training process employed a batch size of 96 and the Adam optimizer, with the standard learning rate of 0.001 which then decreased gradually. The model was trained for a total of 200 epochs. Additional libraries included torchvision and openCV.

V. Experiments and Results

A. Experimental Results on Swin MAE Compared to Classical MAE architecture

Firstly we started with Swin MAE masking method with random masking of 75% and 35% 25% with a training for 200 epochs with our dataset. Below we showcase the results for training Swin MAE for 50 and 100 epochs. The results on Swin MAE showed fast convergence and decrease in MSE loss, however the reconstruction was of very poor quality for all 75% and 35% masking ratios. It can be evident from both 5 and 6, which show that SwinMAE learnt to reconstruct only shallow visual representations such as background on 50 epochs, and color dynamics and orientation in the image shown on a 100 epochs. It captured overall pixel intensity trend, but it can't reconstruct fine-grained vessel structures represented in tubular features. In contrast, classical MAE presented much more robust learning of a visual representation over the same training setups with batch size being 96 images, dataset containing 55 000 images and number of epochs being 200 on a same computing infrastructure. On both sets of images for 25% and 35% masking ratio for the classical MAE in figures 8, 7 we can notice significant progress in reconstruction quality within only first 100 epochs. It is visually clear that model learned not only to reconstruct tubular thickness and its dark color arrangement for the vessels, but also the direction in which vessel structure is propagated within the masked patch. Same applies to elongation of the vessel or encircling of the vessel in the image. It is evident that model learned the spatial relationship in the image and the inter-relationships between

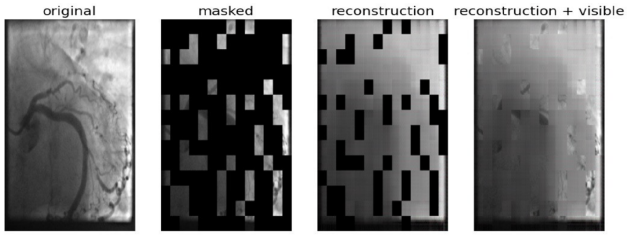


Fig. 6: Swin MAE random reconstruction results trained for 100 epochs on a dataset of 55 000 images with a masking ratio of 75%

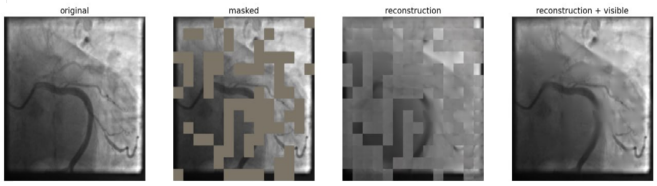


Fig. 7: Classical MAE random reconstruction results trained for 100 epochs on a dataset of 55 000 images with a masking ratio of 35%

closely located patches by looking at which it was able to deduce the structure and formation of the vessel in the masked patch. In this task Classical MAE gained significantly better reconstruction results compared to Swin MAE, even though their training pattern was very similiar, with loss converging already at 100 epoch as seen in figure 9.

B. Experimental Results on Masking Ratio

Comparing masking ratios we can see that training on high masking ratios as suggested by authors of [20] and original [14] doesn't yield good reconstruction results as seen by the following figures 10, 11. It can be seen that just like with Swin MAE training, classical MAE is loosing its advantages in image reconstruction in high masking ratio, being able to reconstruct very low level image features such as general color intensity of the image. On both examples masked patches corresponding to the vessels are reconstructed with poor reconstruction quality, being unable to capture fine-grained tubular structures. This experimental finding is amplified in its importance in results showcased by a figure 12. The model being trained on a small masking ratio was able to reconstruct image with a 75% masking ratio in a much better way than models trained on high masking ratio initially. The model was able to reconstruct image with detailed visual attributes of the vessels including the overall direction of the vessel segments, thickness, color intensity and orientation in the image. The

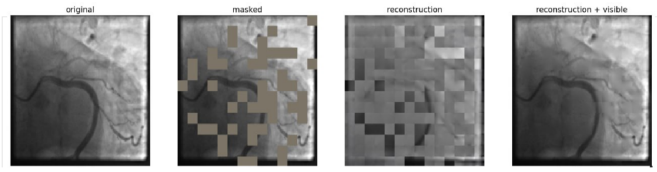


Fig. 8: Classical MAE random reconstruction results trained for 100 epochs on a dataset of 55 000 images with a masking ratio of 25%

Fig. 9: Classical MAE random reconstruction loss graph for training on 200 epochs on wandb. It can be seen that training converges already close to 200 epoch

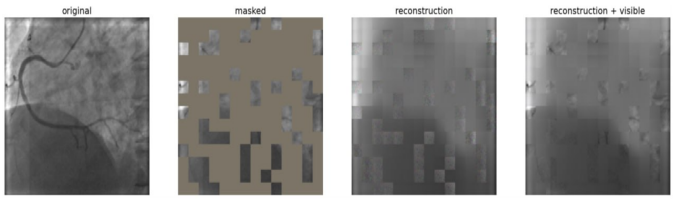


Fig. 10: Classical MAE random reconstruction results trained for 100 epochs on a dataset of 55 000 images with a masking ratio of 75%

empirical tests therefore present the one of the key findings **for a model to achieve proper reconstruction of a thin vessel image attributes it is essential to keep low masking ratios**.

C. Experimental Implementation of the Guided Masking

Yet, the problem with low masking ratios comes to the fact that under random masking most of the masked patches will belong to the background rather to the vessel patch, so reconstruction will not be focused on learning vessel features. Consequently, we proceeded with experiments and

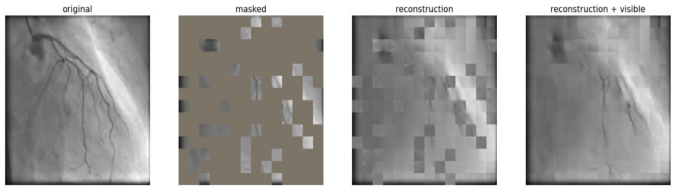


Fig. 11: One more example of Classical MAE random reconstruction results trained for 100 epochs on a dataset of 55 000 images with a masking ratio of 75%

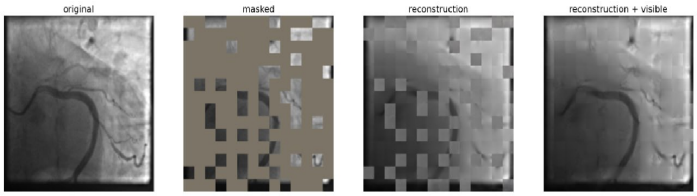


Fig. 12: Example showcasing reconstruction results of a Classical MAE trained on 25% masking ratio for 100 epochs being applied to a reconstruction of a masking in 75%

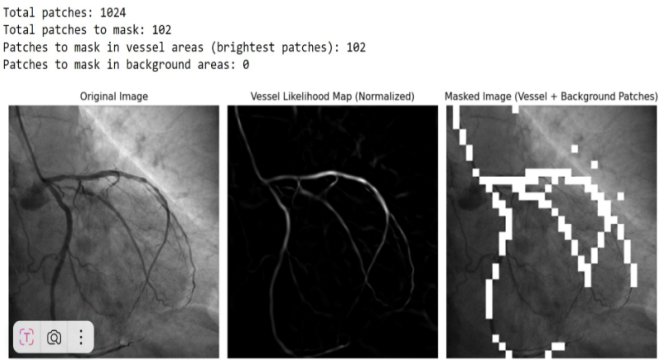


Fig. 13: Guided Masking Strategy which relies on Frangi filter to mask vessels with highest normalized pixel intensities. In this example we show original image on the left, normalized vessel map probability on the center and masked image with vessels on the right.

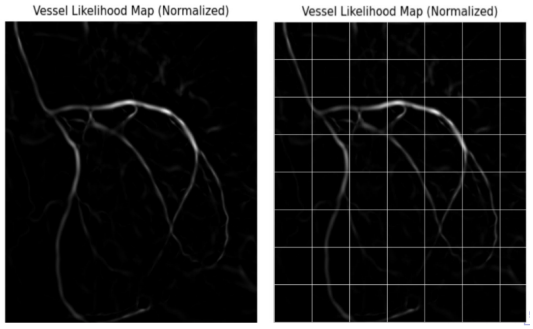


Fig. 14: A closer look into Guided Masking Strategy: In contrast to the Random Masking, here we firstly divide Vessel Probability Map created by Frangi Filtering into patches shown by a visualized grid. These patches are then calculated on their average intensity and used for masking of a brightest patches which correspond to the vessel structures in original CAD image.

implementations of our proposed Guided Masking strategy using Frangi fitering outlined in subsection D in Methodology. As seen in figure 13, the guided masking strategy calculates total amount of patches, receives masking ratio and based on it selects amount of patches to mask, containing ratio of the background and vessels to mask. In this example, we are making sure that within masking ratio vessels patches are 100% of the portion. Frangi filtering here is applied in range of 1-6 for its σ_{filter} in a loop to extract all tubular structures which we find important in the image to form vessel probability map. The rest procedure follows exact methodology for guided masking presented in Methodology. The results of applying guided masking based on this vessel probability map are very appropriate for vessel structures as it is evident from the image. In this way, the encoder-decoder model can learn to reconstruct precisely vessel patches, and increase its understanding of the necessary features compared to the previous random masking approach. The grid on a vessel probability map shown in figure 14 and consequent patch intensity calculations give us very well formulated masking strategy for hiding vessels, yet it yields computational challenges which had to be addressed and solved.

D. Implementation Challenges of the Guided Masking

1) Random Masking and Patch Indexing in Classical MAE:

In random masking, patches are selected uniformly at random

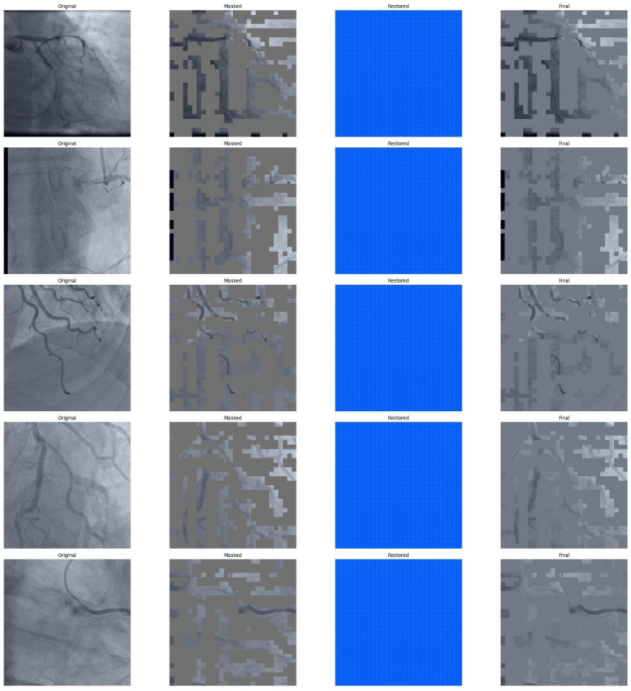


Fig. 15: A failed reconstruction caused by patch index problems induced by extracting highest intensity patches from the vessel probability map. Decoder durings its training receives wrong order of the patches and learns to reconstruct blue noise.

without considering their content. The image is divided into non-overlapping patches, and the patch indices are shuffled randomly. From this shuffled list, a subset of $r \times P$ patches is selected for masking, where r is the masking ratio and P is the total number of patches. This approach ensures that every patch has an equal probability of being masked, making the masking process content-agnostic and straightforward to implement.

2) Guided Masking and Patch Indexing in Classical MAE: In guided masking, patch selection is driven by the content of the image. After preprocessing the image using Gaussian smoothing and the Frangi filter, a vessel probability map is generated and normalized to the range $[0, 1]$. The image is then divided into patches, and the average intensity of the probability map is calculated for each patch. Patches are ranked by their intensities, and the top $\alpha \times (r \times P)$ patches with the highest intensities (likely containing vessels) and the bottom $(1 - \alpha) \times (r \times P)$ patches with the lowest intensities (likely background) are selected for masking. This content-driven approach prioritizes masking patches based on their likelihood of containing vessel structures.

3) Fixing patch indexing at Guided Masking: In this approach, the image is first processed with a **Frangi filter** to highlight vessel-like structures, generating a response map $\mathbf{R}$ that measures the likelihood of each pixel belonging to such structures. The image is then divided into patches, and the average Frangi response for each patch is computed as:

$$\mathbf{R}_{\text{patch}} = \frac{1}{N_{\text{patch}}} \sum_{\text{patch}} \mathbf{R}$$

Patches are ranked based on their response values, and a binary mask $\mathbf{M}$ is created, where the top $\alpha \times (r \times P)$ patches are kept (set to 0), and the bottom $(1 - \alpha) \times (r \times P)$ are masked (set to 1). This ensures that patches containing vessel structures are retained, while irrelevant patches are masked out. The final mask is applied to the patches in the encoding process, maintaining the integrity of the patch indexing and avoiding the misalignment problems seen in previously guided masking approaches.

E. Frangi filtering issues:deducing its sigma value ranges and making it adaptable for GPU Tensor computation

Previously, as shown in figure 14, vessel map probability was less intense and vessel structures were weakly highlighted, with bleak imaging. This is due to the fact that Frangi filter sigma values were in the range of 1-6 which are adapted for capturing thin vessels which are on the level of thickness and intensity are comparable to background. That is why it was hard to sometimes contrast vessel structures from the background in normalized vessel probability map, which had its own implications on selecting brightest patches and most visible and stout vessel segments. In order to mitigate this, we experimented with Frangi filter sigma values and found out that range of 4-10 is most suitable for us in visual context of the dataset. As it is evident from the figure 16, the new vessel probability still captures thin vessels, but is able to modulate higher intensities and overall brightness to the vessel map probability. The vessels are bright enough in filtered image to differentiate them from a background. Moreover, former Frangi Filter implementation based on skimage was using numpy array based processing which was extremely slow on GPU for batch of 96 images. To overcome it, Tensor based computation was introduced in Frangi filtering as a result of which computation of one batch of 96 images was reduced from 17 seconds to 0.4 seconds, significantly aiding to reducing training time. Figures 17 and 18 show preliminary results of training 100 epochs of Classical MAE model in this fashion. To a high degree model learned to mask significant amount of vessel patches based on a masking ratio input, and reconstruct vessels to some degree already. Its comparison with classical MAE is not as straightforward as it seems. It can be divided into several aspects:

- In terms of a guided masking, our approach masks significant portion of a vessel compared to a random masking which masks most of a backround given a masking ratio as an input.
- Yet, Classical MAE with random masking seems to produce for a while better reconstructions of a vessels, so it requires further comprehensive review of the guiding masking strategy. Moreover, reconstructive ability should be evaluated along with adapting trained backbones of these MAE models under guided and random masking to the downstream tasks of a vessel segmentation.

VI. Discussion and Conclusion

This study completed examination of the best masking ratios for vessel images. showing that it is more appropriate

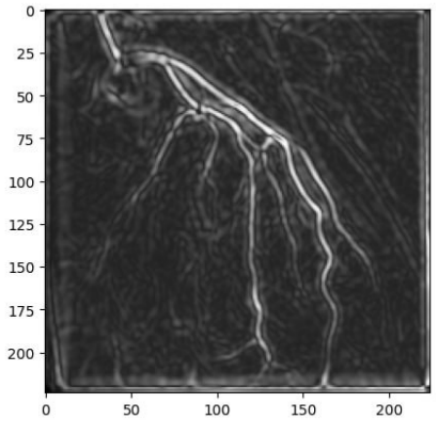


Fig. 16: New vessel probability map with Frangi filter applied in range of 4-10 sigmas

Fig. 17: Guided Masking with proper patch indexing after Frangi filtering. The image is preprocessed to be of a brown color with vessels be less seen as a result of applying Frangi filtering in different range (4-10) compared to (1-6) previously

to use small masking ratios of 25-35% for CAD images. It also evaluated existing MAE architecture, outlining empirical advantage of the Classical MAE over Swin MAE. Finally it presented the development of the Guided Masking Strategy. At the end of this study, Guided Masking showcased ability to target and mask significant portions of the vessel based on a guidance from Frangi filtering with Average patch intensity calculations. Patch indexing problems, Frangi sigma range and computational dynamics during training were highlighted as a key challenges of this study with in depth analysis on how they were overcomed. Yet, there is still significant room for research to finalize this project, including

- Comprehensive evaluation of the Guided Masking Strategy on various CAD images with test cases for Frangi filtering. We noticed that sometimes the edge of a CAD image may not be a vessel but a an edge of an image which represents a tubular black frame which is often confused by a Frangi Filter as a vessel and goes to the masking, even though it is basically an artifact. To our knowledge, such images present 20% of our data, and it is crucial to approach preprocessing step which can reduce effect of this artifact on a Frangi filter and its guidance masking.
- We need to apply backbones trained on Random masking and Guided masking for Vessel segmentation as a downstream task to get quantitative metrics on what approach is better for Clinical contexts of applying Deep Learning models. Currently, we relied solely on training convergence and on quality of reconstruction images, which is not comprehensive review of the models. They should be empirically applied to a target medical task in which exists a significant lack of data.
- If downstream task will provide positive results, it is interesting to apply this Self-Supervised Guided Masking Strategy to other domains with vessel structures such as Retina vessels and Brain vessels. In these domains might exist a downstream clinical tasks to which such pre-training can be useful.
- We are interested in incorporating models which went such pre-training to the medical reporting domain later by adapting their backbones to the instance segmentation task and connecting it with LLM models for generation of the medical report on CAD diseases. In this way, models trained on a guided masking approach can have lasting impact in medical settings.

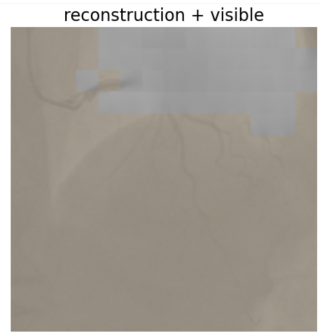


Fig. 18: Reconstruction results for 100 epochs of training with Guided Masking after solving patch indexing issue

Concluding all of these findings, that while significant advancements were made in understanding of a key masking concepts applied to CAD images such as masking ratio, frangi filter indexes and MAE paradigm, there is a still room for improvement. It is important to note that further studies are needed to conduct evaluation of the novel Guided Masking method in downstream tasks and medical settings.

REFERENCES

[1] E. K. Jonathan C. Brown, Thomas E. Gerhardt, *Risk Factors for Coronary Artery Disease*. StatPearls, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:219879508

[2] M. Popov, A. Amanturdieva, N. Zhaksylyk *et al.*, "Dataset for automatic region-based coronary artery disease diagnostics using x-ray angiography images," *Scientific Data*, vol. 11, p. 20, 2024.

[3] K. Iyer, C. P. Najarian, A. A. Fattah, and others., "Angionet: a convolutional neural network for vessel segmentation in x-ray angiography," *Scientific Reports*, 2021.

[4] Z. Gao, L. Wang, R. Soroushmehr, A. Wood, J. Gryak, B. K. Nallamothu, and K. Najarian, "Vessel segmentation for x-ray coronary angiography using ensemble methods with deep learning and filter-based features," *BMC Medical Imaging*, vol. 22, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246018251

[5] X. Zhu, Z. Cheng, S. Wang, X. Chen, and G. Lu, "Coronary angiography image segmentation based on pspnet," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105897, 2021.

[6] B. Kim, Y. Oh, and J. C. Ye, "Diffusion adversarial representation learning for self-supervised vessel segmentation," 2023. [Online]. Available: https://arxiv.org/abs/2209.14566

[7] Y. Ma, Y. Hua, H. Deng, T. Song, H. Wang, Z. Xue, H. Cao, R. Ma, and H. Guan, "Self-supervised vessel segmentation via adversarial learning," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7516–7525.

[8] M. Bilal, D. Martinho, R. Sim, A. Qayyum, H. Vohra, M. Caputo, T. Akinosho, S. Abioye, Z. Khan, W. Niaz, and J. Qadir, "Multivessel coronary artery segmentation and stenosis localisation using ensemble learning," 2023. [Online]. Available: https://arxiv.org/abs/2310.17954

[9] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," 2022. [Online]. Available: https://arxiv.org/abs/2112.01527

[10] T. Liu, H. Lin, A. K. Katsaggelos, and A. Kline, "Yolo-angio: An algorithm for coronary anatomy segmentation," 2023. [Online]. Available: https://arxiv.org/abs/2310.15898

[11] R. Avram, J. E. Olgin, A. Wan, Z. Ahmed, L. Verreault-Julien, S. Abreau, D. Wan, J. E. Gonzalez, D. Y. So, K. Soni, and G. H. Tison, "Cathai: fully automated coronary angiography interpretation and stenosis estimation," *Nature*, 2023.

[12] T. Du, L. Xie, H. Zhang, X. W. X. Liu, D. Chen, Y. Xu, Z. Sun, W. Zhou, L. Song, C. Guan, A. J. Lansky, and B. Xu, "2021," *EuroIntervention*, 2021.

[13] J. Ku, Y.-H. Lee, J. Shin, I. K. Lee, and H.-W. Kim, "Mpseg : Multi-phase strategy for coronary artery segmentation," 2023. [Online]. Available: https://arxiv.org/abs/2311.10306

[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020.

[16] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *NeurIPS*, 2020.

[17] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Y. Zeng, H. Liu, J. Hu, Z. Zhao, and Q. She, "Pretrained subtraction and segmentation model for coronary angiograms," *Scientific Reports*, vol. 14, p. 19888, 2024.

[19] Y. Li, T. Luan, Y. Wu, S. Pan, Y. Chen, and X. Yang, "Anatomask: Enhancing medical image segmentation with reconstruction-guided self-masking," in *European Conference on Computer Vision (ECCV)*, 2024.

[20] Z. Xu, Y. Dai, F. Liu, W. Chen, Y. Liu, L. Shi, S. Liu, and Y. Zhou, "Swin mae: Masked autoencoders for small datasets," *COMPUTERS IN BIOLOGY AND MEDICINE*, vol. 161, JUL 2023.

[21] A. Almalki and L. J. Latecki, "Self-supervised learning with masked image modeling for teeth numbering, detection of dental restorations, and instance segmentation in dental panoramic radiographs," in *Winter Conference on Computer Vision (WACV)*, 2023.

[22] Y. Yang, L. Pan, L. Liu, and E. A. Stone, "Convolutional masked image modeling for dense prediction tasks on pathology images," in *Winter Conference on Computer Vision (WACV)*, 2024.

[23] F. Li, H. Zhang, H. xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," 2022. [Online]. Available: https://arxiv.org/abs/2206.02777

[24] A. Jiménez-Partinen, M. A. Molina-Cabello, K. Thurnhofer-Hemsi, E. Palomo, J. Rodríguez-Capitán, A. I. Molina-Ramos, and M. Jiménez-Navarro, "Cadica: a new dataset for coronary artery disease," 2024.

[25] H. Cao, Y. Wang, D. J. J. Chen, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv:2105.05537, 2021. III-B, IV*, 2021.

Shakhnazar Zhumabekov Contribution to project: guided masking implementation, development of Frangi filter for tensor computation, main training conduction.

Dilnaz Ualiyeva Contribution to project: dataset collection and preprocessing, result documentation, report formatting and presentation preparation.

Kamila Assylbek Contribution to project: gaussian smoothing application, literature review, report proofreading

Taikozha Turdyakyn Contribution to project: classical MAE and Swin MAE implementation and conducting masking ratio experiments