A Convolutional-Transformer Model for FFR and iFR Assessment From Coronary Angiography

Raffaele Mineo, F. Proietto Salanitri, G. Bellitto, I. Kavasidis, O. De Filippo, M. Millesimo, G. M. De Ferrari, M. Aldinucci, D. Giordano, S. Palazzo, F. D'Ascenzo, and C. Spampinato

***Abstract*— The quantification of stenosis severity from X-ray catheter angiography is a challenging task. Indeed, this requires to fully understand the lesion's geometry by analyzing dynamics of the contrast material, only relying on visual observation by clinicians. To support decision making for cardiac intervention, we propose a hybrid CNN-Transformer model for the assessment of angiography-based non-invasive fractional flow-reserve (FFR) and instantaneous wave-free ratio (iFR) of intermediate coronary stenosis. Our approach predicts whether a coronary artery stenosis is hemodynamically significant and provides direct FFR and iFR estimates. This is achieved through a combination of regression and classification branches that forces the model to focus on the cut-off region of FFR (around 0.8 FFR value), which is highly critical for decision-making. We also propose a spatio-temporal factorization mechanisms that redesigns the transformer's self-attention mechanism to capture both local spatial and temporal interactions between vessel geometry, blood flow dynamics, and lesion morphology. The proposed method achieves state-of-the-art performance on a dataset of 778 exams from 389 patients. Unlike existing methods, our approach employs a single angiography view and does not require knowledge of the key frame; supervision at training time is provided by a classification loss (based on a threshold of the FFR/iFR values) and a regression loss for direct estimation. Finally, the analysis of model interpretability and calibration shows that, in spite of the complexity of angiographic imaging data, our method can robustly identify the location of the stenosis and correlate prediction uncertainty to the provided output scores.**

***Index Terms*— Attention methods, coronary angiography, coronary stenosis quantification.**

I. Introduction

QUANTIFICATION of severity of stenosis occlusion is the first evaluation step during coronary angiography to decide whether to perform a stent intervention [1], [2] or not. However, visual assessment of stenosis severity may lead to inter-observer variability depending on the experience of operators, clinical presentation of patients and attitude to perform or not intracoronary assessment with imaging or with functional data [3]. To overcome these limitations, an established guideline method to grade coronary lesions or multi-vessel diseases consists in invasive coronary physiology assessment either using fractional flow reserve (FFR) or, more recently, instantaneous wave-free ratio (iFR) [1], [2]. However, FFR and iFR assessment through coronary pressure wires presents a few limitations: from the time required to conduct the measurements to the cost of the diagnostic procedure to the low, but not negligible, risk of complications due to the invasive nature of the exam.

In spite of these drawbacks, FFR/iFR quantification has been increasingly used to guide revascularization strategies in multivessel disease. Indeed, studies show that FFR values below 0.80 are indicative of hemodynamically-significant stenoses [1], [4], [5], [6], and that iFR values have an analogous meaning below a threshold of 0.89 [1], [7], [8]; instead, patients with FFR/iFR values above threshold do not benefit more from revascularization than from optimal treatment alone [1], [9].

Thus, automated, fast and reliable estimation of FFR/iFR values (or, equivalently, of hemodynamical stenosis significance based on those) would provide essential support to clinicians in making correct decisions, as well as to reduce the procedures patients have to undergo. In the last decade, convolutional neural networks have been widely applied to a variety of medical image analysis tasks (e.g., organ segmentation [10], [11], diagnosis [12], [13], genomics [14], [15], etc.).

Manuscript received 22 January 2024; revised 15 March 2024; accepted 20 March 2024. Date of current version 1 August 2024. The work of Raffaele Mineo, who has contributed to the development and evaluation of the hybrid CNN-Transformer model and its evaluation, has been supported by MUR PRIN 2020, project: "LEGO.AI: LEarning the Geometry of knOwledge in AI systems", n. 2020TA3K9N, CUP: E63C20011250001. The work of F. Proietto Salanitri, G. Bellitto, who have contributed to the definition of the factorized spatio-temporal transformer mechanism, has been supported by MUR PNRR project PE0000013-FAIR. Raffaele Mineo is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma. An Italian patent application no. 102024000014434 has been filed, covering the method proposed in this work. *(Raffaele Mineo and F. Proietto Salanitri contributed equally to this work.) (Equal supervision by F. D'Ascenzo and C. Spampinato.) (Corresponding author: F. Proietto Salanitri.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Comitato Etico Interaziendale A.O.U. Città della Salute e della Scienza – A.O. Ordine Mauriziano Di Torino – A.S.L. Città di Torino.

Raffaele Mineo, F. Proietto Salanitri, G. Bellitto, I. Kavasidis, D. Giordano, S. Palazzo, and C. Spampinato are with the Department of Electrical, Electronics and Computer Engineering, University of Catania, 95123 Catania, Italy (e-mail: raffaele.mineo@phd.unict.it; federica.proiettosalanitri@unict.it; giovanni.bellitto@unict.it; kavasidis@dieei.unict.it; daniela.giordano@unict.it; simone.palazzo@unict.it; concetto.spampinato@unict.it).

O. De Filippo, M. Millesimo, G. M. De Ferrari, and F. D'Ascenzo are with the Department of Medical Sciences, University of Turin, 10124 Turin, Italy (e-mail: ovidio.defilippo@gmail.com; michele.millesimo@unito.it; gaetanomaria.deferrari@unito.it; fabrizio.dascenzo@unito.it).

M. Aldinucci is with the Department of Computer Science, University of Turin, 10124 Turin, Italy (e-mail: marco.aldinucci@unito.it).

Digital Object Identifier 10.1109/TMI.2024.3383283

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

Recently, vision transformers [16] have further advanced the state of the art, while showing significant properties in terms of decision interpretability and robustness [17], [18], [19].

As a consequence of the success of deep learning in medical image analysis, a variety of methods have been proposed to support cardiologists in cardiovascular imaging analysis and risk assessment [20], [21], as well as for automated/semi-automated quantification of artery stenosis assessment from coronary angiography [22], [23], [24]. Among the latter, the most promising strategies [22], [23] employ multiple angiography views together with a *key frame*, i.e., the video frame with the best image quality, full-contrast agent penetration, and clearly contrasted vessel borders. Though effective, these solutions have two main drawbacks: they require multiple exams on the patients and an extra effort by cardiologists, who have to manually identify the key frame for each exam.

In order to overcome these limitations, we propose an approach for assessing stenosis severity from angiography videos through both direct and indirect estimation of FFR/iFR values. Our method requires neither the collection of multiple views nor the selection of a key frame; instead, we combine and leverage the specific peculiarities of both CNN and transformer architectures to extract meaningful spatio-temporal features for physics-based modeling of contrast flow. In particular, we exploit the powerful inductive bias of CNNs to learn local spatio-temporal features by means of 3D convolutions; then, we feed 3D local features to a transformer encoder, which employs factorized self-attention to extract global spatial and temporal features in two separate stages, with the aim to capture long-range dependencies (in both space and time) in the input video and helping the model to focus mostly on flow dynamics thus solving the typical visual inspection ambiguities. Moreover, we employ the learned spatio-temporal features as a shared backbone in a multi-task setting, by introducing multiple output branches that tackle stenosis severity assessment from: 1) a classification standpoint (as under or over significant clinical thresholds), and 2) a regression perspective, by directly predicting FFR and iFR values. This formulation encourages the learning of more robust and generalizable features around the cut-off clinical threshold (as shown in the results) allowing for an improved assessment of FFR to support personalized intervention. It also provides a simple but effective way to employ heterogeneously-labeled datasets, thus making it possible to supervise the model with inputs having the corresponding targets expressed in terms of either a discrete categorization or direct FFR/iFR scores.

We tested the proposed approach on a dataset collected from multiple Italian hospitals, which includes 778 angiographic exams from 389 patients (much more than those used in recent works, e.g., [22], [23], [24]). Our method yields good and reliable performance both on classification and on regression. Extensive comparison with recent state-of-the-art approaches indicates also that the proposed approach outperforms significantly existing methods employing multiple views or key frame information. To summarize, the main contributions of this paper are:

- We propose a hybrid convolutional-transformer model that factorizes spatio-temporal feature extraction for learning complex interactions between vessel geometry, blood flow dynamics, and lesion morphology in order to quantify FFR/iFR
- We introduce a multi-branch architecture to support different assessment modalities (classification of hemodynamical stenosis significance and FFR/iFR regression), encouraging to learn robust features around a predefined clinical cut-off as well as enabling the training on heterogeneously-labeled datasets;
- We carry out an extensive experimental analysis to validate the proposed method, showing its effectiveness and its performance, compared to the state of the art;
- We carry out interpretability and confidence calibration analysis, confirming the reliability of the model's decisions, with a higher probability of correctness and lower uncertainly of our approach than existing methods.

II. Related Work

Coronary stenosis is one of the major causes for heart failure, and occurs when the vessel narrows and blood cannot flow normally. According to the severity of a stenosis, cardiologists decide whether to treat it pharmaceutically or surgically [1]. In the last decade, a variety of deep learning methods for stenosis detection and severity classification, stenosis detection from imaging data have been proposed. These methods can be mainly categorized in two groups: *2D approaches* analyze individual frames from angiography videos and then carry out either late fusion or voting for final prediction; *2D+t models*, which have been less explored, directly extract spatio-temporal features from the entire video.

Most of the 2D classification methods either perform stenosis classification grading, generally using two or three severity levels, or classify stenosis as hemodynamically-significant by thresholding FFR/iFR values. These approaches are generally based on the automatic identification of a key frame, either using CNN architectures [25], [26] or through a combination of convolutional and recurrent networks [27], [28], [29], [30], [31] for feature extraction, followed by a final stenosis classification module. Other methods, instead, enforce the classifier to focus only on blood vessels [32], [33] by adding a pre-processing segmentation step that aims at reducing the area under analysis and avoiding visual artifacts due, for instance, to the pacemaker.

Deep learning has been extensively employed also for stenosis detection on individual frames. The vast majority of these methods adopt a standard pipeline consisting of (manual or automatic) key frame identification followed by object detection models for stenosis localization [32], [34], [35], [36]. A comprehensive benchmark of state-of-the-art object detection models for coronary stenoses is presented in [37].

Another line of 2D methods, instead, aims at analyzing the shape and visual appearance of blood vessels on the key frame to locate stenoses [38], [39], [40]. For example, Zhao et al. [38] first perform automatic segmentation of vessels, followed by keypoint extraction and classification for identifying the segments with the highest likelihood of stenosis. Finally, other 2D methods exploit intepretability approaches on frame-based stenosis classification models to

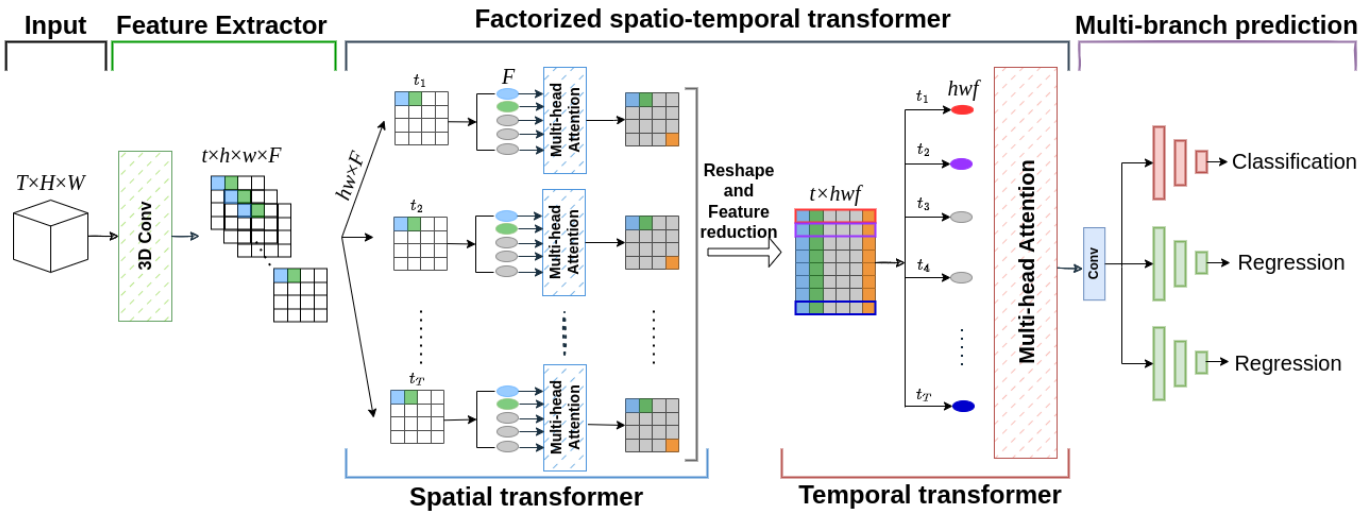


Fig. 1. **Architecture of the proposed approach for stenosis significance assessment**. Input angiography videos are first processed by a pre-trained 3D convolutional model for local feature extraction. Then, self-attention layers based on transformers are employed to capture intra-relations in space and time, and the resulting intermediate representation is fed to three output branches, providing predictions as either a significance class or a regression estimate of FFR/iFR values.

generate activation maps, used to guide the stenosis detection process [25], [27], [29].

Recently, a few 2D+t models operating on the entire angiography videos have been proposed [22], [23], [24] for quantitative coronary analysis and for stenosis detection [41]. References [22] and [23] are the most relevant to our work as they perform quantitative coronary analysis (QCA) of stenoses, through the regression of multiple clinical indices (e.g., minimum lumen diameter, proximal and distal reference vessel diameters, etc.), by using one main angiography view plus an additional side view and a manually-selected key frame. More specifically, [22] and [23] employ a shared 3D convolutional backbone (whose features are processed by an attention layer in [23]) for processing the two angiography views, and 2D dilated residual convolutions for extracting features from the key frame. The two sets of features are then processed by hierarchical self-attention for final QCA regression. Unlike the above methods, our approach requires neither multiple views nor a manually-selected key frame, significantly reducing the burden on patients and cardiologists. Instead, we combine a hybrid convolutional-transformer model (leveraging the recent advantages of vision transformers in terms of performance, robustness and interpretability [42], [43], [44]) with a multi-task formulation of stenosis severity quantification, encouraging the learning of more general features and supporting supervision through both discrete class labels and continuous FFR/iFR scores.

Our formulation follows a recent research trend on *hybrid* models that aim at combining convolutional layers with transformer/attention blocks, in the attempt to leverage the CNNs' inductive bias for feature extraction and hierarchical representation learning, while also harnessing the power of attention mechanisms to focus on salient regions within an image. A seminal work in this direction is described in [45], where non-local blocks (which can be implemented by means of self-attention) are introduced in standard convolutional architectures and achieve improved performance in image and video understanding tasks. CvT [46] improves vision transformers by introducing a convolutional token embedding at the beginning of each transformer layer and a convolutional projection which replaces the linear projection within the transformer module, to capture local spatial context. A more conceptual kind of hybrid architecture is represented by Swin Transformers [47], [48], which take inspiration from convolutional layers in their shifted windowing scheme (akin to the overlapping receptive fields in CNNs) and in the hierarchical representations produced by gradual patch merging.

Overall, our approach is the first hybrid CNN-Transformer for FFR quantification, featuring a spatio-temporal factorization technique to capture spatial and temporal dependencies for the representation of contrast flow within vessels.

III. Method

A. Overview

The proposed model, depicted in Fig. 1, consists of a cascade of spatio-temporal feature extraction blocks, based on a sequence of 3D convolution and spatio-temporal transformer layers, followed by multiple output branches, predicting a binary class on the hemodynamical significance of a stenosis (based on the thresholds suggest by the established literature [1], [5], [6], [7], i.e., 0.80 for FFR and 0.89 for iFR) and direct estimates of FFR and iFR values. The proposed architecture is designed to make use of convolutional feature extractors, which benefit from a strong inductive bias that helps in the extraction of local visual features, and of transformer layers based on self-attention, which are instead better at finding global correlations between regions of the input data and have been shown to improve model explainability. In the following, we introduce and describe in detail each component of the model.

B. Feature Extraction Through 3D Convolutions

In spite of the wider and wider diffusion of transformer architectures for vision tasks, one of their main limitations stands on the lack of a strong inductive bias that can take advantage of the regular structures of real-world images. On the contrary, convolutional architectures can quickly and effectively learn distinctive and reusable patterns from visual data, but they lack a principled way to extract global features without resorting to lossy downsampling operators. For these reasons, the first processing block of our model consists of a spatio-temporal feature extractor based on 3D convolutions, able to capture meaningful patterns from video sequences, that can be later refined by more complex processes.

Given an input video sample $\mathbf{X} \in \mathbb{R}^{T \times H \times W}$, with T, H and W being, respectively, the number of frames and the height and width of each frame, we initially extract local spatio-temporal features by feeding it into a 3D convolutional neural network $\mathcal{F}$. The output of the feature extractor is a set of F feature maps $\mathbf{H} \in \mathbb{R}^{t \times h \times w \times F}$, with t, h and w respectively smaller than T, H and W due to the downscaling effect of convolutional encoders.

In our implementation, we employ a ResNet3D [49] model as a feature extractor, pre-trained on the Kinetics-400 dataset [50] for video action recognition. This choice is in line with the motivation for employing convolutional operators only at the first stage of the model: in spite of the strong domain shift between natural videos and coronary angiographies, our usage of 3D convolutions as local feature extractors takes advantage of the generalizability of low-level features learned by pre-trained models; higher-level semantics and analysis are then carried out by the transformer blocks that follow the initial feature extraction stage. The weights of the feature extractor are fine-tuned along with the entire network at training time. In our preliminary experiments, alternative strategies such as training from scratch or freezing pre-trained weights led to low performance. Note that in order to adapt the pre-trained ResNet3D model from RGB to X-ray inputs, we simply squash the first-layer convolutional kernels by averaging over the channel dimensions.

C. Factorized Spatio-Temporal Transformers

Local 3D features extracted by the convolutional backbone are then processed by a factorized spatio-temporal transformer, with the objective of finding relations between non-local regions within and between frames. While this operation can be performed by a purely-convolutional model through pooling layers, we argue that this approach badly fits the nature of angiographic videos for the specific task of stenosis quantification, for two main reasons. First, the spatial extent of the stenotic region often covers a small fraction of the visible area: excessive spatial downsampling inevitably results in the loss of details, which negatively affects stenosis localization (either implicit or explicit) by the model. Second, while initial temporal down-sampling helps in reducing computational complexity and may take advantage of the video periodicity introduced by the cardiac cycle, temporal dimensional collapse would be needed to capture the spreading pattern of the contrast agent, whose speed may provide essential clues on both blood flow and vessel geometry; however, collapsing the temporal dimension to capture global patterns makes it impossible to extract time-varying dynamics.

To overcome these limitations, we employ a spatio-temporal transformer to process local 3D features and extract higher-level task-specific features, while retaining sufficient spatial and temporal resolution. Architecturally, we factorize the transformer into separate spatial and temporal modules, for more efficient computation [51].

Both modules internally apply scaled dot-product multi-head self-attention [52]. Let $\mathbf{T} \in \mathbb{R}^{N \times F}$ be a set of feature vectors, packed as matrix rows (F being feature dimensionality). A transformer layer produces a new set of feature $\mathbf{U} \in \mathbb{R}^{N \times F}$ as:

$$\mathbf{T}' = \mathrm{LN}\left(\mathrm{MA}\left(\mathbf{T}\right) + \mathbf{T}\right) \tag{1}$$

$$\mathbf{U} = \mathrm{LN}\left(\mathrm{MLP}\left(\mathbf{T}'\right) + \mathbf{T}'\right) \tag{2}$$

where LN is layer normalization [53], MLP is a two-layer multi-layer perception, with input and output sizes equal to F and a hidden layer with d_{MLP} neurons with ReLU activation, and MA is the multi-head self-attention operator, defined in the following. Note that the input and output dimensions of the transformer layer are unchanged (equal to $N \times F$), allowing for arbitrary sequences of layers. In each layer, the self-attention function A computes query, key, and value vectors for each feature vector in $\mathbf{T}$ by linear projection into $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{N \times d}$, with d as the feature dimension. The projection matrices $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V \in \mathbb{R}^{F \times d}$ are learnable. The output for each $\mathbf{T}$ row is a linear combination of $\mathbf{V}$ rows, weighted by dot-product similarity calculated as $\mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)$. The multi-head self-attention function MA receives h sets of projection matrices $\{\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V\}_{i=1}^h$. Outputs from the self-attention function are concatenated and linearly projected using $\mathbf{W}^O \in \mathbb{R}^{F \times F}$ to produce the final output. The attention feature dimensionality d is simplified to F/h.

In our model, the spatial transformer module receives convolutional features $\mathbf{H} \in \mathbb{R}^{t \times h \times w \times F}$, and processes each frame $\mathbf{H}_i \in \mathbb{R}^{h \times w \times F}$ independently. The height and width dimensions are grouped, in order to reshape each $\mathbf{H}_i$ as $\mathbf{T}_i \in \mathbb{R}^{N_{\mathrm{space}} \times F}$ (with $N_{\mathrm{space}} = wh$), with dimensions suitable for multi-head self-attention. The outputs of all spatial transformers are then reshaped and concatenated back as new features $\mathbf{H}_{\mathrm{space}} \in \mathbb{R}^{t \times h \times w \times F}$.

The temporal transformer module applies a similar operation on $\mathbf{H}_{\mathrm{space}}$, with the difference that attention is computed between concatenated frame-level feature vectors. First, in order to reduce feature dimensionality (which would explode due to concatenation), we project each spatial feature vector from F to a lower size f; then, all features within each frame are concatenated, resulting in a tensor $\mathbf{H}'_{\mathrm{space}} \in \mathbb{R}^{N_{\mathrm{time}} \times hwf}$ (with $N_{\mathrm{time}} = t$), which can be fed to the temporal transformer layers, producing the output tensor $\mathbf{H}_{\mathrm{time}} \in \mathbb{R}^{t \times h \times w \times f}$ (after reshaping).

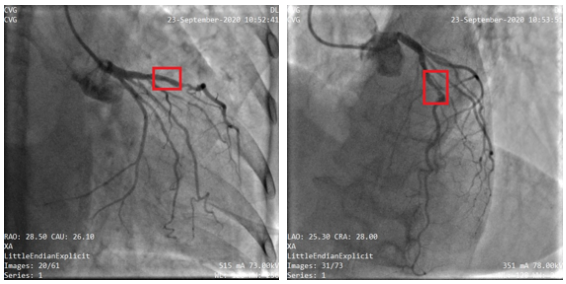


Fig. 2. **Examples of the two views per subject**. Bounding boxes indicate major stenoses as identified by two expert radiologists.

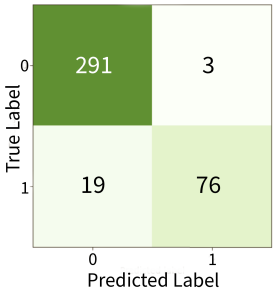


Fig. 3. Cumulative confusion matrix over the five tested folds.

In our implementation, both spatial and temporal transformers include two layers each, with $h = 4$ attention heads per layer.

D. Multi-Branch Prediction

The output $\mathbf{H}_{\text{time}}$ of the temporal transformer undergoes a dimensionality reduction step by means of two layers of 3D convolutions with spatio-temporal stride equal to 2, thus reducing each dimension by a factor of four: the resulting tensor is then flattened into a vector $\mathbf{V} \in \mathbb{R}^{fthw/64}$.

The multi-branch prediction module receives $\mathbf{V}$ and feeds it into three separate two-layer MLPs, with batch normalization and ReLU activations. The hidden layer of each MLP halves the number of input neurons (from 512 to 256) and the output layer projects to a single scalar (from 256 to 1). Thus, the final output of the layer is a set of scalars $\hat{y}_c$ (constrained between 0 and 1 through a sigmoid activation), $\hat{y}_{\text{FFR}}$ and $\hat{y}_{\text{iFR}}$, respectively predicting a binary class of stenosis significance and the direct estimates of FFR and iFR values. Given ground-truth values y_c, y_{FFR} and y_{iFR}, we define a training objective $\mathcal{L}$ as:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{FFR}} + \mathcal{L}_{\text{iFR}} \\ &= -y_c \log \hat{y}_c - (1 - y_c) \log \left(1 - \hat{y}_c\right) \\ &+ \lambda \left| y_{\text{FFR}} - \hat{y}_{\text{FFR}} \right| + \lambda \left| y_{\text{iFR}} - \hat{y}_{\text{iFR}} \right|, \end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{BCE}}$ is the binary cross-entropy classification loss, $\mathcal{L}_{\text{FFR}}$ and $\mathcal{L}_{\text{iFR}}$ are L_1 regression losses; λ is a weighing factor between classification and regression terms.

IV. Experimental Results

A. Dataset

We use a private dataset comprising 778 coronary angiographies, retrospectively collected between January 2020 and January 2022. The dataset encompasses two examinations (two examples are reported in Fig. 2) from 389 patients diagnosed with Chronic Coronary Syndrome (CCS) and Acute Coronary Syndrome (ACS), distributed across five hospitals as per Tab. II (first two columns). Patients underwent invasive physiological assessment of intermediate single coronary stenoses (diameter of stenosis $\geq 30\%$ and $\leq 90\%$ at angiographic visual estimation) through the measurement of FFR, iFR or both. Among the 389 patients, 303 were men and 86 women, and the mean age was 67.9 $\pm$ 9.61. IRB protocol number is 0092163.

Each angiography was evaluated by two expert cardiologists (for each hospital) through invasive physiological assessment of intermediate coronary stenosis with iFR, FFR, or both. In particular, FFR values are provided for 251 patients (64.5%), iFR values for 228 patients (58.6%); a set of 90 patients (23.1%) include both. For each exam, the major stenosis was identified by radiologists and labelled as hemodynamically significant if FFR is lower than 0.80 [1], [4], [5], [6] or if iFR is lower than 0.89 [1], [7], [8]. As a result, 94 patients (24.4%) are labeled as positive, while the remaining cases are negative. Coronary angiography and physiological measurements were performed following standardized clinical practice. Key frames were annotated by two expert cardiologists.

Since the angiographies were collected with different machines and practices available in the 5 involved hospitals, their spatial sizes and frame rates vary, respectively, from 512×512 to 1024×1024 and from 15 fps to 60 fps. Given these differences, all samples were resized to 256×256 pixel ($H \times W$ from Sect. III-B) and 15 fps (downsampling when needed). All collected videos were cut to a length of 60 frames by selecting the 4 central seconds of the videos. This choice is motivated by the typical range of angiographies (about 3.5 s), in order to be strategically centered around the region of interest with the highest level of perfusion. In cases the angiography video was shorter than 4 seconds, we employ padding by replicating the first and the last frames up to 4 seconds.

In all the single-view experiments, we use only one random view per patient (either in training or validation/test) excluding the other remaining views.

B. Training and Evaluation Procedure

We train our model to minimize the multi-task loss $\mathcal{L}$ from (3), through gradient descent with the AdamW optimizer [54] (learning rate: 1e-5, batch size: 8), for a total of 300 epochs. Input X-ray angiography videos were standardized and augmented through: 1) 2D horizontal and vertical flipping and 90° rotations (identically applied to all frames of the same video), and 2) temporal sampling rate augmentation by reducing arbitrarily the frame rate for a more effective learning of flow velocity. Furthermore, to deal with the overrepresentation of negative cases, we employed random oversampling by randomly duplicating instances from the positive class.

At training time, the two regression branches are not necessarily both activated, since not all samples are provided with

TABLE I
CLASSIFICATION PERFORMANCE COMPARISON OF OUR MODEL USING DIFFERENT INPUT DATA MODALITIES

	Modality	Accuracy(%)	AUC	Sensitivity(%)	Specificity(%)
Single-view	3D	$\mathbf{89.4 \pm 7.2}$	$\mathbf{0.95 \pm 0.03}$	80.0 ± 13.9	$\mathbf{98.9 \pm 0.9}$
Single-view + key frame	2D+3D	86.4 ± 8.8	0.86 ± 0.09	79.4 ± 14.6	93.4 ± 7.4
Multi-view	3D	89.3 ± 7.0	0.95 ± 0.04	$\mathbf{80.9 \pm 13.9}$	97.2 ± 2.0
Multi-view + key frames	2D+3D	89.1 ± 6.1	0.94 ± 0.05	80.0 ± 11.6	96.2 ± 2.1

both FFR and iFR values: as a result, L_1 loss terms $\mathcal{L}_{\text{FFR}}$ and $\mathcal{L}_{\text{iFR}}$ are occasionally ignored when the corresponding target is not present. To cope with class imbalance, as classification loss, we employed balanced binary cross entropy$\mathcal{L}_{\text{BCE}}$ defined as follows:

$$BCE(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[\alpha \cdot y_i \cdot \log(p_i) + (1-\alpha) \cdot (1-y_i) \cdot \log(1-p_i)] \quad (4)$$

where y is the true target value (actual label), p_i is the predicted probability for the positive class. N is the number of samples in the dataset and α the weight or factor that balances the contribution of positive and negative classes. In our case, it is set to the ratio of the number of positive samples (less represented class) to the total number of samples, ensuring that both classes have equal influence on the loss.

BCE is optimized as follows: the target class, representing hemodynamical stenosis significance, is based on the availability of FFR and iFR values and on the corresponding thresholds. If both are available, FFR is used to set the classification target. The λ weighing factor in (3) is empirically set to 10, based on the relative magnitudes of cross-entropy and L_1 losses.

As mentioned in Sect. III-B, we employ a ResNet3D architecture for feature extraction. Given a $T \times H \times W = 60 \times 256 \times 256$ input, the output feature volume is $t \times h \times w \times F = 8 \times 16 \times 16 \times 128$. Feature size F is obtained by reducing ResNet3D's output features from 512 to 128 through a linear projection. Output features from the spatial transformer are further projected from $F = 128$ to $f = 16$ to prevent feature explosion due to temporal concatenation. The size of the input to the multi-branch module is therefore $ftwh/64 = 512$. Finally, the multi-branch prediction module comprises three identical blocks, each one composed by a first fully connected layer with 256 neurons followed by ReLU and dropout ($p = 0.1$) and a final classification layer with two neurons.

We carry out model evaluation for the proposed approach and for state-of-the-art methods on the classification task, reporting balanced accuracy (to account for class imbalance), area under the ROC curve (AUC), sensitivity and specificity. For our approach, we also report regression accuracy on FFR and iFR scores as mean squared error (MSE) and mean absolute error (MAE). Performance metrics of the proposed approach and the methods under comparison are computed through 5-fold stratified nested cross-validation. For each experiment fold, 60% of the dataset is used for training, 20% for validation and 20% for test. Validation accuracy is used to select the training epoch at which test performance is evaluated: performance metrics are then averaged over all test folds. Furthermore, the split of the dataset into training, validation, and test sets was performed on a per-patient basis, in order to guarantee that angiograms from a single patient are never unintentionally split between the training and test sets, thus reducing any potential overestimation of generation performance due to correlation among samples.

Experiments are carried out on two NVIDIA Tesla T4 GPUs using automatic mixed precision for training. The proposed approach was implemented in PyTorch and MONAI; all code is available at https://github.com/perceivelab/conv-transf-ffr-ifr-assessment.

C. Effect of Input Data Modality

Most of existing methods for stenosis quantification, such as [22], [23], and [24], employ multiple views and key frame information. Thus, our first battery of experiments aims at evaluating how input data modality affects model performance. In particular, we compare our method, as presented in Sect. III, with variants thereof, using: 1) an additional convolutional branch (consisting of a ResNet18 network) that extracts 2D features from the key frame and concatenates them, before multi-task prediction, with the features provided by the spatio-temporal transformer; 2) two views (as existing methods, i.e., [22], [23]): in this case, the model backbone (CNN and transformer) is shared across the views and feature concatenation is carried out before multi-task prediction; 3) two views in addition to two key frames (one for each view), with each keyframe processed as presented in item 1. Table I reports the results of this comparison: our model outperforms multi-view variants on three out of four metrics, while being very close on the fourth. It is interesting to notice that adding key frame information seems to hinder performance. This may be due to the fact that, as also shown by other experiments presented in the following, the model mostly focuses on temporal information for its predictions; spatial-only features from a single frame may thus introduce noise by increasing dimensionality with no informative contribution. Multi-view inputs, instead, do not seem to significantly harm performance, but the lack of improvements demonstrates our method's capability to perform well with a single view, avoiding extra burdens on physicians and patients.

We further delve into the performance of the single-view model by inspecting the cumulative confusion matrix among the five folds, shown in Fig. 3 and demonstrating very good specificity and satisfactory sensitivity. In order to investigate possible bias in the data and, consequently, in our approach, we also compute individual performance for each hospital. In particular, we use data from five different hospitals with

TABLE II

PER-HOSPITAL CLASSIFICATION PERFORMANCE YIEDLED BY THE PROPOSED APPROACH. PERFORMANCE IS CONSISTENT ACROSS THE FIVE CENTERS DESPITE THE HIGH DATA IMBALANCE

Hospital	N_Patients	TP	FN	FP	TN	Acc.(%)	Sens.(%)	Spec.(%)
1	90	10	3	0	77	88.46	76.92	100.00
2	121	8	2	1	110	89.55	80.00	99.10
3	58	33	8	1	16	87.30	80.50	94.11
4	80	14	4	1	61	88.08	77.78	98.39
5	40	11	2	0	27	92.31	84.61	100.00

TABLE III

CLASSIFICATION PERFORMANCE COMPARISON WITH THE STATE OF THE ART. FOR EACH MODEL, WE REPORT THE INPUT MODALITY AS WELL AS BALANCED ACCURACY (*Acc.*), AUC, SENSITIVITY (*Sens.*) AND SPECIFICITY (*Spec.*), COMPUTED THROUGH 5-FOLD NESTED CROSS-VALIDATION. *S* STANDS FOR SINGLE VIEW, I.E., ONLY ONE VIEW PER PATIENT IS USED, WHILE *M* STANDS FOR MULTIVIEW, I.E., AT LEAST TWO VIEWS PER PATIENT ARE EMPLOYED

		View	Input	Acc. (%)	AUC	Sens. (%)	Spec. (%)
CNN	**S3D** [55]	S	2D+t	79.5 ± 5.5	0.93 ± 0.03	65.4 ± 14.2	93.6 ± 4.8
	ResNet3D [49]	S	2D+t	77.9 ± 5.1	0.85 ± 0.04	64.0 ± 9.1	91.8 ± 1.8
	MVCNN [56]	M	2D	84.5 ± 6.1	0.85 ± 0.07	76.8 ± 10.2	92.2 ± 2.2
	GVCNN [57]	M	2D	78.4 ± 4.5	0.87 ± 0.05	71.0 ± 7.3	85.8 ± 4.4
Transformer	**ViT-B_16** [16]	S	2D	76.8 ± 6.1	0.74 ± 0.11	69.0 ± 14.1	84.6 ± 5.9
	ViT-3D [17]	S	2D+t	73.7 ± 3.7	0.78 ± 0.06	63.0 ± 7.4	84.4 ± 1.2
	Swin Transformer [48]	S	2D	63.8 ± 9.7	0.70 ± 0.11	41.0 ± 27.2	86.6 ± 8.2
	Video Swin [58]	S	2D+t	81.2 ± 7.0	0.92 ± 0.06	65.2 ± 14.2	95.6 ± 0.5
Hybrid	**DMQCA** [22]	M	2D-(2D+t)	81.7 ± 4.2	0.82 ± 0.04	70.2 ± 9.1	93.2 ± 2.1
	DMTRL [24]	S	(2D+t)	79.1 ± 4.2	0.85 ± 0.06	67.0 ± 7.5	91.2 ± 3.9
	HEAL [23]	M	2D-(2D+t)	79.5 ± 5.9	0.84 ± 0.06	67.6 ± 10.6	91.4 ± 3.7
	Proposed method	S	2D+t	$\mathbf{89.4 \pm 7.2}$	$\mathbf{0.95 \pm 0.03}$	$\mathbf{80.0 \pm 13.9}$	$\mathbf{98.9 \pm 0.9}$

different number of patients per hospital and different acquisition modality and the obtained results are reported in Table II. It can be noted that, despite the unbalanced data, performance is consistent across hospitals, thus suggesting the no bias affects the yielded performance.

Furthermore, sensitivity is worse than sensitivity and this is possible due to the hard-defined clinical cut-off (0.8) for defining the positive and negative cases. This limitation is, however, mitigated by the low error (reported later) in terms of FFR regression that allows for personalized intervention.

D. Comparison With the State of the Art

We then compare our performance to that achieved by state-of-the-art models. Specifically, this evaluation includes the following architectures: *CNN-based models* — S3D [55], ResNet_3D-18 [49], GVCNN [57] and MVCNN [56]; *vision transformer models* — ViT-B_16 [16], ViT-3D [17] and Video Swin Transformer [58]; hybrid models, specifically designed for angiography video analysis, that combine features from both CNN and transformer architectures — DMTRL [24], DMQCA [22] and HEAL [23].[1]

Results, shown in Table III, report significantly higher performance by our model compared to the state of the art, on all metrics. Remarkably, our approach shows a gain of

[1]The authors of these approaches did not release source code, thus the results refer to our implementation thereof. For these methods, model selection was carried out through grid-search with cross-validation.

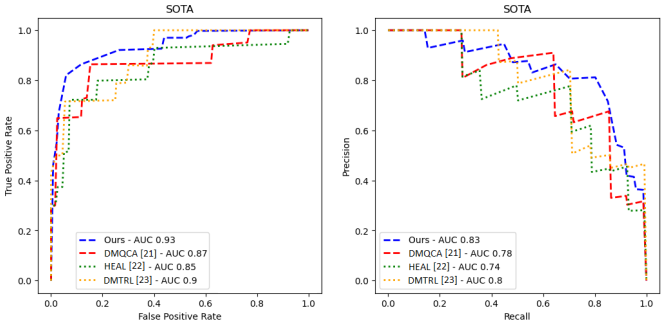


Fig. 4. **ROC (left) and Precision-Recall (right) curves** comparison between our approach and state-of-the-art methods.

over ten percent points on the sensitivity score, which is the most significant metric for clinical validation of automated angiography analysis tools [59]. Moreover, this analysis confirms that our method yields better performance with fewer input data, as it only requires a single-view video, while, for instance, DMQCA [22] and HEAL [23] employ two views (requiring physicians and patients to carry out two exams on patients) and the key frame (to be manually identified by physicians). Fig. 4 reports the comparison in terms of ROC and precision-recall curves between our approach and the state-of-the-art methods specifically designed for stenosis quantification, i.e., DMQCA [22], HEAL [23] and DMTRL [24].

We evaluate our approach also in terms of interpretability and calibration to assess the reliability of the model. As for

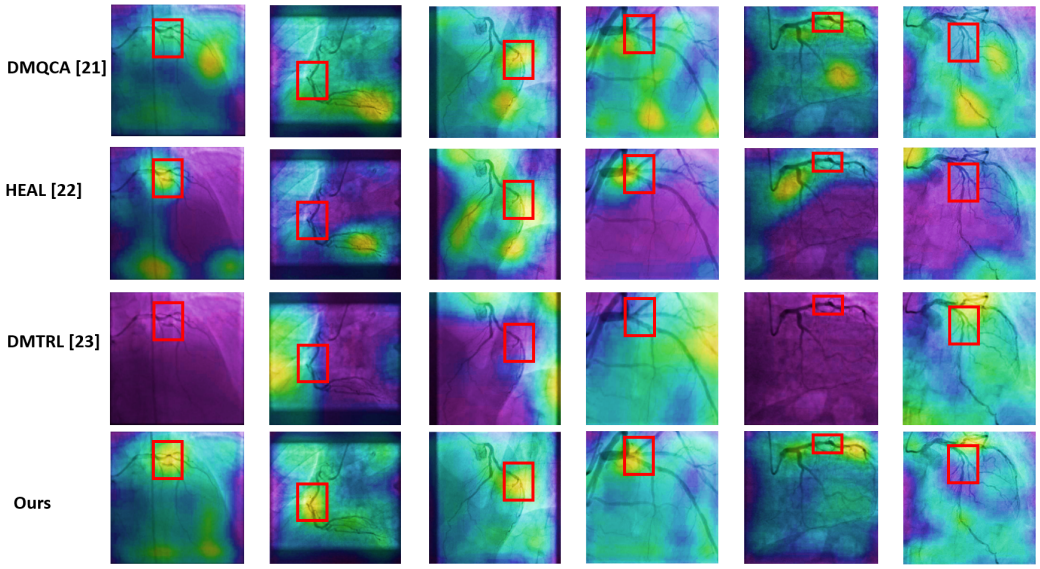


Fig. 5. **Interpretability maps**, computed by M3D-cam [60], [61], of our method (last row) and DMQCA [22], HEAL [23] and DMTRL [24] (first, second, and third row, respectively). For each image, we also report, as a red bounding box, the location of the major stenosis, as identified by cardiologists. In each map, the yellow parts are the most activated ones, while the purple parts are the least activated ones.

TABLE IV
CALIBRATION PERFORMANCE IN TERMS OF BRIER SCORE (THE LOWER, THE BETTER) BETWEEN OUR APPROACH AND STATE-OF-THE-ART-ONES

	Brier Score (↓)
DMQCA [22]	0.145 ± 0.032
HEAL [23]	0.125 ± 0.041
DMTRL [24]	0.154 ± 0.058
Proposed method	0.089 ± 0.047

the calibration, we compute the Brier Score and compare it with the ones obtained by other models performing stenosis quantification. Results in Table IV shows that our approach is more reliable for clinical applications than existing ones, as it is able to associate each prediction with a more accurate confidence score, reducing the probability of major mistakes (especially in case of false positives) and providing a more interpretable decision to physicians. To further enhance trust in our model, we also evaluate interpretability maps to assess whether our model focuses on the major stenosis, as identified by radiologists, to make its predictions. For this analysis, we employ M3D-Cam [60], [61] to compute interpretability maps for our model's and state-of-the-art models' decisions. Results, shown in Fig. 5, indicate that the proposed approach generally attends to the entire vessel structures, involved in the flow contrast over time, placing a particular emphasis on the major stenosis (as identified by cardiologists). In contrast, this level of comprehensive attention to vessel structures does not consistently manifest in the three methods under comparison. This observation underscores the capacity of the proposed approach to learn spatio-temporal features associated with flow dynamics, vessel geometry, and lesion morphology.

E. Ablation Studies

We then carry out ablation studies to validate our architectural design choices.

Starting from a baseline model, consisting of a 3D convolutional network (ResNet3D) with a single classification branch, we add — first individually, then together — the spatio-temporal transformer and the multi-task branch (which introduces FFR/iFR regression). Results, given in Table V, indicate how each block contributes to enhance performance. It is interesting to note that the major gain is obtained when including the multi-branch module with multi-task loss: in particular, sensitivity increases by more than five percent points with respect to the hybrid convolutional-transformer network alone.

Thus, FFR/iFR regression emerges as the optimal approach, delivering superior accuracy. Subsequently, we evaluate whether FFR or iFR yields the highest gain in classification accuracy. However, as detailed in Sect. III-D, the available dataset lacks both FFR and iFR values for all exams. Consequently, we perform this evaluation on a subset of 81 cases from the original dataset, where data includes both values. FFR and iFR regression employ factorized spatio-temporal transformers and multi-branch prediction, as previously presented. Due to the smaller size of this subset, we compute results based on 60/20/20 random splits for training/validation/test, and report test values at the lowest validation loss.

The results, presented in Table VI, affirm the following key findings: 1) Employing regression on either FFR or iFR results in a noteworthy enhancement (exceeding 15 percentage points) compared to the baseline, even with a more limited dataset; 2) FFR emerges as more reliable than iFR; specifically, when iFR is utilized, performance slightly lags behind that of FFR but still surpasses the baseline by a considerable margin. Despite FFR's higher reliability, we opted to include iFR in the analysis for the entire dataset. This decision stems from the unavailability of FFR estimates for all exams and given the observed superior accuracy of iFR regression compared to the baselines.

TABLE V

ABLATION ON ARCHITECTURE MODULES. RESULTS REFER TO CLASSIFICATION PERFORMANCE

	Acc.(%)	AUC	Sens.(%)	Spec.(%)
3D CNN	77.9 ± 5.1	0.85 ± 0.04	64.0 ± 9.1	91.8 ± 1.8
+ multi-branch prediction	87.9 ± 6.5	0.94 ± 0.05	78.5 ± 14.7	96.4 ± 1.2
+ factorized spatio-temporal transf.	86.6 ± 8.1	0.94 ± 0.07	74.9 ± 15.4	96.0 ± 1.5
+ multi-branch prediction	$\mathbf{89.4 \pm 7.2}$	$\mathbf{0.95 \pm 0.03}$	$\mathbf{80.0 \pm 13.9}$	$\mathbf{98.9 \pm 0.9}$

TABLE VI

ABLATION ON FFR-IFR REGRESSION ON A SUBSET OF THE OVERALL DATASET. RESULTS REFER TO CLASSIFICATION PERFORMANCE

	Acc.(%)	AUC	Sens.(%)	Spec.(%)
Baseline - 3D CNN	76.6	0.71	50.0	97.2
+ FFR pred	93.3	0.93	90.9	95.7
+ iFR pred	89.8	0.92	81.8	97.8
+ FFR and iFR pred	89.9	0.92	81.8	97.9

TABLE VII

ABLATION ON NUMBER OF FRAMES. RESULTS REFER TO CLASSIFICATION PERFORMANCE

#	Accuracy(%)	AUC	Sensitivity(%)	Specificity(%)
15	80.8 ± 6.4	0.90 ± 0.09	71.3 ± 11.6	90.2 ± 3.1
30	81.5 ± 5.5	0.93 ± 0.05	71.0 ± 10.2	91.9 ± 6.1
45	84.6 ± 11.6	0.95 ± 0.04	73.2 ± 13.0	96.0 ± 1.2
60	$\mathbf{89.4 \pm 7.2}$	$\mathbf{0.95 \pm 0.03}$	$\mathbf{80.0 \pm 13.9}$	$\mathbf{98.9 \pm 0.9}$

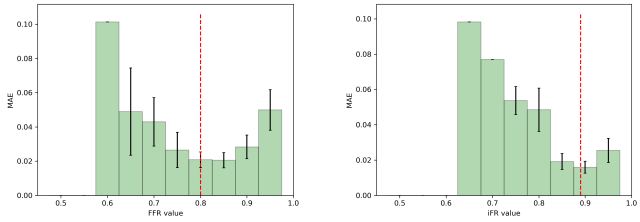


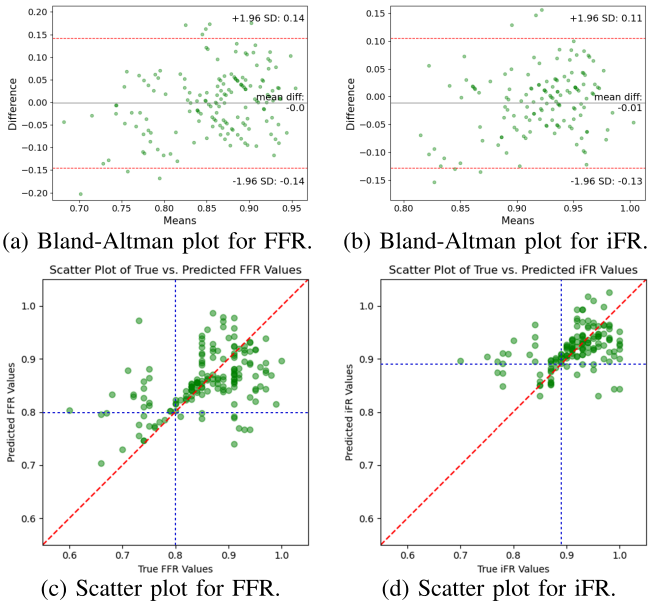
Fig. 6. **MAE at different FFR and iFR range intervals.** Blue bars measure the regression MAE computed at each interval of the FFR/iFR ranges. Standard deviation on each interval is reported as an error bar. Vertical red lines correspond to thresholds for FFR and iFR, below which a stenosis is considered to be hemodynamically significant.

TABLE VIII

FFR AND IFR REGRESSION PERFORMANCE OF THE PROPOSED MODEL

	MSE	MAE
FFR regression	0.036 ± 0.007	0.063 ± 0.008
iFR regression	0.025 ± 0.006	0.045 ± 0.008

The proposed model uses 60 frames to provide the network with enough data to learn flow dynamics. This choice aims to empower the model in capturing and effectively discerning subtle spatio-temporal patterns within the data. To underscore the significance of this design choice, we compare our model's performance under three different input additional video sequence lengths, namely, 15, 30, and 45 frames. As reported in Table VII, the impact of the frames number on performance is clear. With only 15 frames, the model exhibits significantly lower performance compared to the 60-frame configuration. We also observe a performance gain as we increase the number of frames, confirming that increasing the temporal context helps the model to learn the motion dynamics associated to vessel geometry and lesion morphology.



(a) Bland-Altman plot for FFR. (b) Bland-Altman plot for iFR.

(c) Scatter plot for FFR. (d) Scatter plot for iFR.

Fig. 7. (First row) **Bland-Altman plot** between the predicted and the reference values for FFR and iFR. There is agreement between the two sets of values with 0 mean, suggesting no systematic bias in the evaluation. (Second row) **Scatter plots** for FFR and iFR demonstrate that the lowest errors and the smaller standard deviations are consistently observed around clinically-relevant thresholds (0.8 for FFR, 0.89 for iFR, depicted in blue dashed lines).

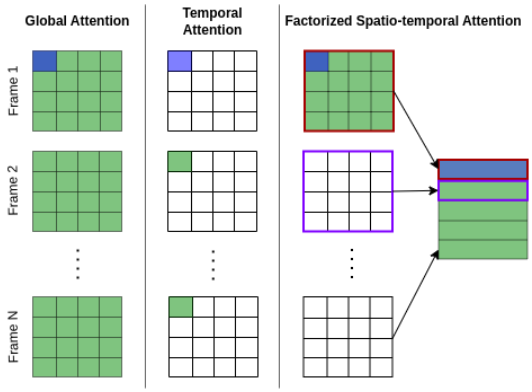


Fig. 8. **Attention mechanisms**. *Global attention*: each spatio-temporal location (in blue) attends to all other locations (in green). *Temporal attention*: each location (in blue) attends the same spatial location along the time dimension (in green). *Factorized spatio-temporal attention*: attention is first computed on each frame independently, and then along the temporal dimension over frame-level features.

The positive impact of multi-branch prediction to our model suggests that regressing continuous values of FFR and iFR

TABLE IX
ABLATION ON ATTENTION STRATEGIES

	Acc.(%)	AUC	Sens.(%)	Spec.(%)
Non-local block—based attention				
Global	61.9 ± 5.60	0.84 ± 0.11	25.2 ± 11.7	98.6 ± 0.8
Temporal	79.3 ± 1.18	0.89 ± 0.06	64.3 ± 19.6	94.3 ± 5.4
Factorized spatio-temporal	84.0 ± 8.40	0.91 ± 0.05	73.2 ± 17.8	94.8 ± 5.2
Transformer—based attention				
Global	83.5 ± 8.1	0.93 ± 0.03	74.6 ± 16.4	92.3 ± 1.6
Temporal	87.8 ± 7.4	0.95 ± 0.02	80.7 ± 13.8	94.8 ± 1.1
Factorized spatio-temporal	$\mathbf{89.4 \pm 7.2}$	$\mathbf{0.95 \pm 0.03}$	$\mathbf{80.0 \pm 13.9}$	$\mathbf{98.9 \pm 0.9}$

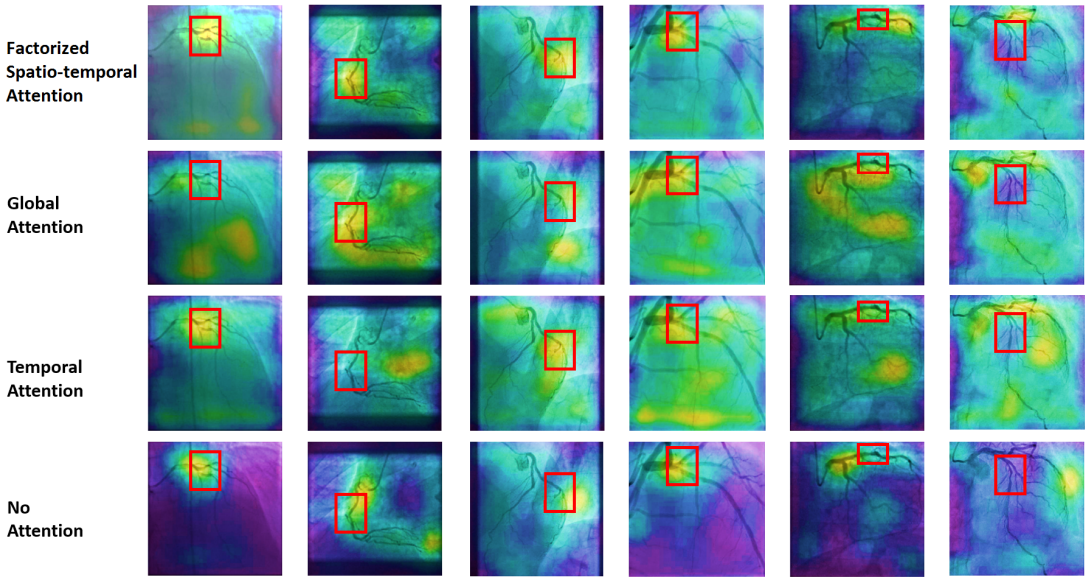


Fig. 9. **Interpretability maps**, computed through M3D-cam [60], [61], when using different attention strategies. For each image, we also report, as a red bounding box, the major stenosis as identified by clinicians. In each part, the yellow parts are the most activated ones, while the purple areas, the least activated ones.

tends to better regularize our model, compared to using classification alone (though based on the same FFR/iFR values). To quantify how well our model predicts FFR and iFR, we compute MAE and MSE metrics, reported in Table VIII. In order to better investigate the behavior of the model, Fig. 6 plots MAE with standard deviation values computed on different intervals of the FFR and iFR ranges, showing that the model is more precise around thresholds for hemodynamical significance for both iFR and FFR. This directly reflects on improved classification of hemodynamical stenosis significance. Furthermore, we also conduct an analysis of our model's performance through Bland-Altman and scatter plots, shown in Fig. 7. The Bland-Altman analysis indicates a strong agreement between the values predicted by our model and those measured with a mean difference of zero, indicating that there is no systematic bias. Scatter plots, instead, demonstrate that the lowest errors and the smaller standard deviations are consistently observed around clinically-relevant thresholds (0.8 for FFR, 0.89 for iFR). These findings collectively substantiate our claim on the role of multitask learning in forcing the model to improve the assessment around the clinical cut-off for FFR, as it is critical for personalized intervention.

Finally, we investigate the impact of the attention mechanism employed within the spatio-temporal transformer encoder. In particular, we evaluate the performance of our model when using a) *global attention*, i.e., each location in the feature volume attends to all other locations in space and time; b) *temporal attention*, i.e., each location attends the same spatial location along the time dimension; c) *factorized spatio-temporal attention* as explained in Sect. III-C, i.e., attention is first computed on each frame independently, overall spatial locations, and then along the temporal dimension over frame-level features. The different attention strategies are illustrated in Fig. 8. We also compare the multi-head attention in transformers with other attention mechanisms, such as the one based on non-local blocks [45]. Also in this case, we developed the three strategies (global, temporal and factorized spatio-temporal) mentioned earlier.

Table IX shows our transformer-based method yields better results than the ones obtained with non-local blocks. Among the transformer-based attention strategies, temporal attention appears to be the main responsible for improving performance; spatial information, while marginally improving accuracy when combined to temporal features in the factorized approach, is actually detriment in the global attention setting, which may be due to a difficulty by the model in learning correlation patterns over the whole spatio-temporal volume. This result, demonstrating the critical importance of

temporal information in coronary angiographies, also supports our original motivation for including transformer layers for high-level analysis, in order to avoid temporal downsampling which would lead to the loss of essential information. The quantitative improvement in performance also reflects on the interpretability maps, as illustrated in Fig. 9.

V. Discussion and Conclusion

Diagnosing coronary artery conditions, especially in terms of predicting functional metrics like FFR or iFR, presents significant challenges for clinicians. In this work, we present a novel hybrid convolutional-transformer method for coronary stenosis quantification through the estimation of FFR and iFR values. Unlike other methods from the established literature, our approach requires neither multiple angiography views as input, nor the identification of a key frame, thus reducing the burden on both medical staff and patients. Instead, we show that the combination of convolutional and transformer-based architectures, together with a multi-branch prediction approach for estimating stenosis hemodynamical significance (posed as a binary classification problem) as well as direct FFR/iFR scores, allows our method to outperform state-of-the-art approaches that require multiple views and/or key frame information. The method was tested on a dataset with 389 cases, which represents the largest dataset employed so far in works for automated FFR assessment, reaching an accuracy of about 90%, a sensitivity of 80% and a specificity over 95%.

These standard performance metrics coupled with Brier Score for probabilistic predictions and model's interpretability maps affirm the model's reliability and its alignment with clinically relevant features. The extensive evaluation enhances trust and demonstrates, despite the limited dataset size, the effectiveness of the devised countermeasures to discourages overfitting.

We argue that the main reason for the success of our approach stands in its independence from a key frame. In contrast to relying on key frames that capture specific moments in the cardiac cycle, our method learns blood flow dynamics throughout the entire cycle. This enforces the model to capture subtle flow pattern changes that may go missed by the existing morphology-centric methods or may not be apparent to a naked eye (by clinicians). Indeed, coronary artery conditions require a holistic understanding of complex interactions between factors like vessel geometry, blood flow dynamics, and lesion morphology. Thanks to its spatio-temporal focus, the proposed method appears at understanding these interactions effectively, while existing methods as well as physicians struggle to capture these complex relationships.

Moreover, relying primarily on the visual assessment of stenosis morphology, as human clinicians do, is the main cause for inter-observer variability [3] and for reduced reproducibility, while providing quantitative measures on flow-related parameters ensures consistency in the assessment.

In future research endeavors, we plan to carry out a more extensive assessment including more data from different geographical areas to enhance the applicability of the proposed approach. Furthermore, our approach is validated on single coronary lesions (despite multiple minor lesions are already present in our dataset), since the value of FFR/iFR on diffuse diseases has still limited practical inferential use [62].

We also plan to employ specific synthetic data generation, such as [63], both as a form of augmentation and to share data in a privacy-preserving way.

Finally, we would like to highlight that our approach is thought for conventional angiography (using 2D+t data). It enables the estimation of blood fluid dynamics, while it is not able to reconstruct the geometry of arterial structures, for which FFR-CT procedure is required. We thus aim also to extend our approach with CT data to complement our findings, thereby offering a comprehensive approach to coronary artery disease assessment that can work on both conventional angiography via coronary catheterization and three-dimensional computational modeling using CT scans.

References

[1] F.-J. Neumann et al., "2018 ESC/EACTS Guidelines on myocardial revascularization," *Eur. Heart J.*, vol. 40, no. 2, pp. 87–165, 2018.

[2] J. Knuuti and V. Revenco, "2019 ESC guidelines for the diagnosis and management of chronic coronary syndromes," *Eur. Heart J.*, vol. 41, no. 3, pp. 407–477, 2019.

[3] M. J. Grundeken et al., "Visual estimation versus different quantitative coronary angiography methods to assess lesion severity in bifurcation lesions," *Catheterization Cardiovascular Intervent.*, vol. 91, no. 7, pp. 1263–1270, 2018.

[4] N. P. Johnson et al., "Repeatability of fractional flow reserve despite variations in systemic and coronary hemodynamics," *JACC, Cardiovascular Intervent.*, vol. 8, no. 8, pp. 1018–1027, 2015.

[5] P. A. Tonino et al., "Fractional flow reserve versus angiography for guiding percutaneous coronary intervention," *New England J. Med.*, vol. 360, no. 3, pp. 213–224, 2009.

[6] B. De Bruyne et al., "Fractional flow reserve–guided PCI versus medical therapy in stable coronary disease," *New England J. Med.*, vol. 367, no. 11, pp. 991–1001, 2012.

[7] S. Baumann et al., "Instantaneous wave-free ratio (iFR) to determine hemodynamically significant coronary stenosis: A comprehensive review," *World J. Cardiol.*, vol. 10, no. 12, p. 267, 2018.

[8] J. E. Davies et al., "Use of the instantaneous wave-free ratio or fractional flow reserve in PCI," *New England J. Med.*, vol. 376, no. 19, pp. 1824–1834, 2017.

[9] W. F. Fearon et al., "Economic evaluation of fractional flow reserve–guided percutaneous coronary intervention in patients with multivessel disease," *Circulation*, vol. 122, no. 24, pp. 2545–2550, 2010.

[10] F. P. Salanitri, G. Bellitto, I. Irmakci, S. Palazzo, U. Bagci, and C. Spampinato, "Hierarchical 3D feature learning forpancreas segmentation," in *Proc. MLMI*, 2021, pp. 238–247.

[11] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "TGANet: Text-guided attention for improved polyp segmentation," in *Proc. MICCAI*, 2022, pp. 151–160.

[12] X. Li, M. Jia, M. T. Islam, L. Yu, and L. Xing, "Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4023–4033, Dec. 2020.

[13] M. Shorfuzzaman and M. S. Hossain, "MetaCOVID: A Siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients," *Pattern Recognit.*, vol. 113, May 2021, Art. no. 107700.

[14] C. Pino et al., "Interpretable deep model for predicting gene-addicted non-small-cell lung cancer in ct scans," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 891–894.

[15] R. Dias and A. Torkamani, "Artificial intelligence in clinical and genomic diagnostics," *Genome Med.*, vol. 11, no. 1, p. 70, Dec. 2019.

[16] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[17] F. P. Salanitri et al., "Neural transformers for intraductal papillary mucosal neoplasms (IPMN) classification in MRI images," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 475–479.

[18] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Proc. MICCAI*, Cham, Switzerland. Springer, 2021, pp. 36–46.

[19] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2021, pp. 14–24.
[20] G. Litjens et al., "State-of-the-art deep learning in cardiovascular image analysis," *JACC, Cardiovascular Imag.*, vol. 12, no. 8, pp. 1549–1565, Aug. 2019.
[21] M. Motwani et al., "Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis," *Eur. Heart J.*, vol. 38, Jun. 2016, Art. no. ehw188.
[22] D. Zhang, G. Yang, S. Zhao, Y. Zhang, H. Zhang, and S. Li, "Direct quantification for coronary artery stenosis using multiview learning," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2019, pp. 449–457.
[23] D. Zhang et al., "Direct quantification of coronary artery stenosis through hierarchical attentive multi-view learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4322–4334, Dec. 2020.
[24] W. Xue, G. Brahm, S. Pandey, S. Leung, and S. Li, "Full left ventricle quantification via deep multitask relationships learning," *Med. Image Anal.*, vol. 43, pp. 54–58, Jan. 2018.
[25] J. H. Moon et al., "Automatic stenosis recognition from coronary angiography using convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 198, Jan. 2021, Art. no. 105819.
[26] D. L. Rodrigues, M. Nobre Menezes, F. J. Pinto, and A. L. Oliveira, "Automated detection of coronary artery stenosis in X-ray angiography using deep neural networks," 2021, *arXiv:2103.02969*.
[27] C. Cong, Y. Kato, H. D. Vasconcellos, J. Lima, and B. Venkatesh, "Automated stenosis detection and classification in X-ray angiography using deep neural network," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 1301–1308.
[28] H. Ma, P. Ambrosini, and T. V. Walsum, "Fast prospective detection of contrast inflow in X-ray angiograms with convolutional neural network and recurrent neural network," in *Proc. MICCAI*, 2017, pp. 453–461.
[29] C. Cong, Y. Kato, H. D. Vasconcellos, M. R. Ostovaneh, J. A. Lima, and B. Ambale-Venkatesh, "Deep learning-based end-to-end automated stenosis classification and localization on catheter coronary angiography," *Frontiers Cardiovascular Med.*, vol. 10, 2023, Art. no. 944135.
[30] K. Antczak and Ł. Liberadzki, "Stenosis detection with deep convolutional neural networks," in *Proc. MATEC Web Conf.*, vol. 210, 2018, Paper 04001.
[31] E. Ovalle-Magallanes, J. G. Avina-Cervantes, I. Cruz-Aceves, and J. Ruiz-Pinales, "Hybrid classical–quantum convolutional neural network for stenosis detection in X-ray coronary angiography," *Expert Syst. Appl.*, vol. 189, Mar. 2022, Art. no. 116112.
[32] W. Wu, J. Zhang, H. Xie, Y. Zhao, S. Zhang, and L. Gu, "Automatic detection of coronary artery stenosis by convolutional neural network with temporal constraint," *Comput. Biol. Med.*, vol. 118, Mar. 2020, Art. no. 103657.
[33] B. Au et al., "Automated characterization of stenosis in invasive coronary angiography images with convolutional neural networks," 2018, *arXiv:1807.10597*.
[34] V. V. Danilov et al., "Real-time coronary artery stenosis detection based on modern neural networks," *Sci. Rep.*, vol. 11, no. 1, p. 7582, Apr. 2021.
[35] T. Du, X. Liu, H. Zhang, and B. Xu, "Real-time lesion detection of cardiac coronary artery using deep neural networks," in *Proc. Int. Conf. Netw. Infrastruct. Digit. Content (IC-NIDC)*, Aug. 2018, pp. 150–154.
[36] K. Pang, D. Ai, H. Fang, J. Fan, H. Song, and J. Yang, "Stenosis-DetNet: Sequence consistency-based stenosis detection for X-ray coronary angiography," *Computerized Med. Imag. Graph.*, vol. 89, Apr. 2021, Art. no. 101900.
[37] V. Danilov, O. Gerget, K. Klyshnikov, E. Ovcharenko, and A. Frangi, "Comparative study of deep learning models for automatic coronary stenosis detection in X-ray angiography," in *Proc. GraphiCon*, vol. 2744, 2020, pp. 1–11.
[38] C. Zhao et al., "Automatic extraction and stenosis evaluation of coronary arteries in invasive coronary angiograms," *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104667.
[39] C. Zhao et al., "A new approach to extracting coronary arteries and detecting stenosis in invasive coronary angiograms," 2021, *arXiv:2101.09848*.
[40] T. Wan, H. Feng, C. Tong, D. Li, and Z. Qin, "Automated identification and grading of coronary artery stenoses with X-ray angiography," *Comput. Methods Programs Biomed.*, vol. 167, pp. 13–22, Dec. 2018.
[41] T. Han et al., "Coronary artery stenosis detection via proposal-shifted spatial–temporal transformer in X-ray angiography," *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106546.
[42] W. Nam, S. Gur, J. Choi, L. Wolf, and S.-W. Lee, "Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 3, 2020, pp. 2501–2508.
[43] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 782–791.
[44] M. Springenberg, A. Frommholz, M. Wenzel, E. Weicken, J. Ma, and N. Strodthoff, "From CNNs to vision transformers—A comprehensive evaluation of deep learning models for histopathology," 2022, *arXiv:2204.05044*.
[45] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. CVPR*, Jun. 2018, pp. 7794–7803.
[46] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.
[47] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[48] Z. Liu et al., "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11999–12009.
[49] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.
[50] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.
[51] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6816–6826.
[52] A. Vaswani et al., "Attention is all you need," in *Proc. NIPS*, vol. 30, 2017.
[53] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[54] I. L. and F. H., "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.
[55] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. ECCV*, 2018, pp. 305–321.
[56] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 945–953.
[57] Y. Feng, Z. Zhang, X. Zhao, R. Ji, and Y. Gao, "GVCNN: Group-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 264–272.
[58] Z. Liu et al., "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201.
[59] G. Ravipati et al., "Comparison of sensitivity, specificity, positive predictive value, and negative predictive value of stress testing versus 64-multislice coronary computed tomography angiography in predicting obstructive coronary artery disease diagnosed by coronary angiography," *Amer. J. Cardiol.*, vol. 101, no. 6, pp. 774–775, Mar. 2008.
[60] K. Gotkowski, C. Gonzalez, A. Bucher, and A. Mukhopadhyay, "M3D-CAM: A PyTorch library to generate 3D data attention maps for medical deep learning," in *Proc. German Workshop Med. Image Comput.* Regensburg, Germany: Springer, Mar. 2021, pp. 217–222.
[61] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.
[62] C. Collet et al., "Measurement of hyperemic pullback pressure gradients to characterize patterns of coronary atherosclerosis," *J. Amer. College Cardiol.*, vol. 74, no. 14, pp. 1772–1784, Oct. 2019.
[63] M. Pennisi, F. Proietto Salanitri, G. Bellitto, S. Palazzo, U. Bagci, and C. Spampinato, "A privacy-preserving walk in the latent space of generative models for medical applications," in *Proc. MICCAI*, 2023, pp. 422–431.

Open Access funding provided by 'Università degli Studi di Catania' within the CRUI CARE Agreement