Encoder-Decoder Architectures for Clinically Relevant Coronary Artery Segmentation

João Lourenço Silva[1], Miguel Nobre Menezes[2], Tiago Rodrigues[2], Beatriz Silva[2], Fausto J. Pinto[2], and Arlindo L. Oliveira[1]

[1] INESC-ID / Instituto Superior Técnico, University of Lisbon
{joao.lourenco.silva,arlindo.oliveira}@tecnico.ulisboa.pt
[2] Cardiology Department, CAML, CCUL, Lisbon School of Medicine

Abstract. Coronary X-ray angiography is a crucial clinical procedure for the diagnosis and treatment of coronary artery disease, which accounts for roughly 16% of global deaths every year. However, the images acquired in these procedures have low resolution and poor contrast, making lesion detection and assessment challenging. Accurate coronary artery segmentation not only helps mitigate these problems, but also allows the extraction of relevant anatomical features for further analysis by quantitative methods. Although automated segmentation of coronary arteries has been proposed before, previous approaches have used non-optimal segmentation criteria, leading to less useful results. Most methods either segment only the major vessel, discarding important information from the remaining ones, or segment the whole coronary tree, based mostly on contrast information, producing a noisy output that includes vessels that are not relevant for diagnosis. We adopt a better-suited clinical criterion and segment vessels according to their clinical relevance. Additionally, we simultaneously perform catheter segmentation, which may be useful for diagnosis due to the scale factor provided by the catheter's known diameter, and is a task that has not yet been performed with good results. To derive the optimal approach, we conducted an extensive comparative study of encoder-decoder architectures trained on a combination of focal loss and a variant of generalized dice loss. Based on the EfficientNet and the UNet++ architectures, we propose a line of efficient and high-performance segmentation models using a new decoder architecture, the EfficientUNet++, whose best-performing version achieves a generalized dice score of 0.9202 ± 0.0356, and artery and catheter class dice scores of 0.8858 ± 0.0461 and 0.7627 ± 0.1812.

1 Introduction

Coronary arteries are the blood vessels that carry oxygen and nutrient-rich blood to the heart tissue. Coronary Artery Disease (CAD), also known as Coronary Heart Disease or Ischemic Heart Disease, is a disease caused by the partial or complete blockage of the coronary arteries, which leads to limited or even ceased blood flow to the heart tissue and consequent myocardial dysfunction, being the cause of roughly 16% of global deaths every year [34].

X-ray coronary angiography (CAG) is one of the main procedures for CAD diagnosis and treatment. Traditionally, physicians use CAG images to assess the presence of stenosis, i.e., artery narrowing, through visual inspection. However, this method's subjectivity and potential unreliability led to the development of Quantitative Coronary Angiography (QCA), a diagnostic support tool that uses semi-automatic edge-detection algorithms to report vessel diameters at user-specified locations and the point of stenosis. Nevertheless, the low contrast and resolution of CAG images, the uneven contrast agent distribution, and the presence of artifacts, such as pacemakers, the spine, and the catheter itself, make this task very challenging. Thus, QCA still often requires manual correction of vessel boundaries. Furthermore, QCA only allows the analysis of a small vessel section at a time, limiting its use in clinical practice, in which the severity of stenosis is still assessed visually in most cases, rather than with QCA software.

Recently, deep learning methods have significantly improved coronary artery segmentation performance in CAG images, promising to overcome the faults of QCA's edge-detection algorithms. Most of them either segment only the major vessel [21,45,46,42] or try to segment the whole coronary tree, based primarily on contrast differences [8,35,49,52]. These criteria, however, may not be clinically optimal. The former discards potentially damaged vessels whose lesions may not be negligible, and the latter includes secondary vessels that may not be relevant for either diagnostic or therapeutic purposes, potentially distracting physicians from the important ones. We circumvent these shortcomings by adopting a better-suited clinical criterion developed in collaboration with expert cardiologists, in which a vessel is only segmented if it is 2 mm or wider at its origin. Since thinner vessels have higher risks of poor intervention outcomes, these are usually approached conservatively. Furthermore, the minimum diameter of commercially available revascularization devices is 2 mm [36,14]. Therefore, collateral vessels with diameters below 2 mm at their origin are generally deemed inadequate for revascularization and, when interpreting angiograms, physicians tend to ignore them.

With more complex lesion assessment and anatomical feature extraction in mind, we also segment the catheter, whose known diameter provides a scale factor that may help models determine vessel width and be important for diagnostic. To the best of our knowledge, simultaneous catheter and coronary artery segmentation in CAG images has only been performed in one previous work [39], reporting dice score coefficients (DSCs) far inferior to ours.

To determine the best architecture for this task, we conducted an extensive comparative study of existing encoders and decoders, which provided insights into the best architectural patterns for this and, presumably, other medical image segmentation problems. Based on our findings, we propose a new computationally efficient and high-performing decoder architecture, the EfficientUNet++. Combined with an EfficientNet-B5 [38] encoder, the EfficientUNet++ achieved a generalized dice score (GDS) of 0.9202 ± 0.0356, and DSCs of 0.8858 ± 0.0461 and 0.7627 ± 0.1812, for the artery and catheter classes, respectively.

Overall, the main contributions of this paper are as follows:

1. We propose a new and better clinically-suited criterion for catheter and coronary artery segmentation in CAG images, in which vessels are only labeled as such if they are deemed relevant for diagnostic and therapeutic purposes;
2. We perform an extensive quantitative and qualitative comparison of the performance of existing encoders and decoders, which may provide valuable insights for other medical image segmentation tasks;
3. Based on the findings of our study, we propose a line of models with the best performance-computation trade-offs, from which practitioners can choose according to the available hardware and clinical needs.

2 Related Work

2.1 Major Vessel Segmentation

Previous work has shown that major vessel segmentation can be improved by replacing the U-Net's encoder with popular image classification backbones, either pre-trained on ImageNet [45,46] or trained from scratch on a relatively small dataset composed of 3200 CAG images [42]. Additionally, it has also been shown that using a modified generalized dice loss function with class weights to offset class imbalance and a tunable penalty for false positives and false negatives could further improve performance [45]. In the sequence of these findings, we train our models using a combination of the proposed loss function and Focal Loss (FL), and compare the performance of different state-of-the-art encoders.

Other authors have proposed a U-Net-based nested encoder-decoder architecture, the T-Net [21]. To simplify optimization, the authors replaced the U-Net's blocks with residual ones. In addition, to enable feature reuse, they arranged the pooling and up-sampling operations to make all the feature maps extracted by the encoder available to every layer of equal or greater depth of the decoder, in a DenseNet-like [17] fashion. These modifications, which enhance information flow through the network and enable it to outperform a standard U-Net, are also present in the UNet++ decoder [51] tested in this work.

2.2 Full Coronary Tree Segmentation

One of the main challenges of coronary artery segmentation is distinguishing vessels from artifacts. Since the former are only visible in the presence of contrast, previous work has used images acquired before contrast injection as a second-channel input to help a U-Net discern between vessels and background [8]. However, to be effective, this approach must be coupled with an image alignment algorithm that compensates for the motion caused by heartbeat and respiration. Furthermore, it requires the entire angiographic sequence to be acquired with minimal table motion, which can be hard to achieve, as standard clinical practice involves moving the patient table to follow the flow of dye within the vessels.

In line with what we propose in this paper, some authors have also attempted to use different decoder architectures to achieve better performance than what is possible with the commonly used U-Net. Specifically, they have used a pretrained PSPNet [50,52], and a UNet++ combined with a feature pyramid network to improve multi-scale feature detection [49]. Other proposals include a deeply supervised encoder-decoder network with Gaussian convolutions, explicitly designed for vessel segmentation [35].

2.3 Catheter and Full Coronary Tree Segmentation

To the best of our knowledge, simultaneous catheter and coronary tree segmentation has only been addressed once [39]. Using a U-Net-based Siamese architecture trained on multi-class labels generated from low-level binary segmentation and optical flow, the authors obtained DSCs of 0.54 and 0.69 for the artery and catheter classes, respectively, far below the 0.8858 and 0.7627 DSCs we obtain.

2.4 Other Segmentation Criteria

As far as we know, an intermediate segmentation criterion, in which arteries with diameters inferior to 1 mm at their origin are not segmented, has only been proposed once [19]. However, the rationale behind this criterion is not explained, and it still includes many vessels that are ineligible for revascularization. To perform this task, the authors coupled a trainable preprocessing network with the U-Net and DeepLabV3+ architectures [3]. Even though their results demonstrate that the preprocessing module improves performance, that is not the focus of this work, and we leave the study of preprocessing methods for future work.

3 Model Evaluation Metrics

The segmentation quality of each class is measured using the DSC overlap metric. Let $\mathcal{G}$ and $\mathcal{P}$ be the sets of points belonging to a ground-truth segmentation mask and the segmentation mask predicted by a model, then, mathematically,

$$\mathrm{DSC} = 2\frac{|\mathcal{G} \cap \mathcal{P}|}{|\mathcal{G}| + |\mathcal{P}|}. \tag{1}$$

Overall segmentation quality is measured using the GDS, another overlap metric. Let C be the number of classes, N the number of pixels, $g_{ci} \in \{0,1\}$ denote whether pixel i belongs to class c or not, and $p_{ci} \in [0,1]$ represent the probability of pixel i belonging to class c assigned by the model. Then,

$$\mathrm{GDS} = 2\frac{\sum_{c=1}^{C} w_c \sum_{i=1}^{N} g_{ci} p_{ci}}{\sum_{c=1}^{C} w_c \sum_{i=1}^{N} g_{ci} + p_{ci}}, \tag{2}$$

where $w_c = 1/(\sum_{i=1}^{N} g_{ci})^2$ is the weight assigned to class c. These weights correct each class's contribution to the score by the inverse of its volume, reducing the correlation between region size and score [4]. Consequently, only models with good segmentation performance across all classes can achieve high GDS.

4 Loss Function

The problem we aim to solve can be interpreted as the combination of a macro-level and a micro-level one. The former consists of identifying the arteries and catheters, distinguishing them from each other and from artifacts, and determining which arteries have diameters equal or larger than 2 mm at their origin. The latter concerns the precise delineation of class contours, which are crucial to perform accurate anatomical measurements and reliable diagnoses.

To address these problems we trained the models on a loss function composed of a variant of GDL [37], named Penalty Generalized Dice Loss (pGDL) [45], which conveys information on global segmentation quality, and FL, which provides a pixel-wise evaluation focused on hard pixels, which are usually the ones that belong to less common classes and are near class boundaries and artifact regions. Using the notation defined in Section 3, the loss function can be defined as

$$\text{Loss} = \text{pGDL} + \lambda \text{FL} = 1 - \frac{\text{GDS}}{1 + k(1 - \text{GDS})} - \lambda\alpha(1 - p_{ci})^{\gamma} \log(p_{ci}), \quad (3)$$

where λ controls the relative weight of each term, k defines the magnitude of an additional penalty for false positives and false negatives, α balances the importance of positive and negative examples, and $(1 - p'_{ci})^{\gamma}$ is a modulating factor controlled by $\gamma \geq 0$ that down-weights easy examples and forces the model to focus on and learn from hard ones.

For simplicity, we use $\lambda = 1$ in all experiments. The hyperparameter k is set to $k = 0.75$, since we verified, in informal experiments, that this value works well for all models and leads to better performance than $k = 0$ for most. Following Lin et al. [28], we set the FL's hyperparameters to $\alpha = 0.25$ and $\gamma = 2$.

5 Architecture

Given the profusion of high-performing segmentation models, we started by conducting an extensive comparative study of existing encoders and decoders, aiming to determine the best models for this task and to identify the key components behind their performances. As a result, we propose a new computationally efficient and high-performing decoder architecture, the EfficientUNet++.

5.1 Encoder Comparison

Previous work [42,45,46] has shown that the U-Net's performance can be enhanced by replacing its encoder with more sophisticated backbones, both when pre-trained on a large dataset, like ImageNet [45,46], and when trained from scratch on a relatively small dataset composed of 3200 CAG images [42].

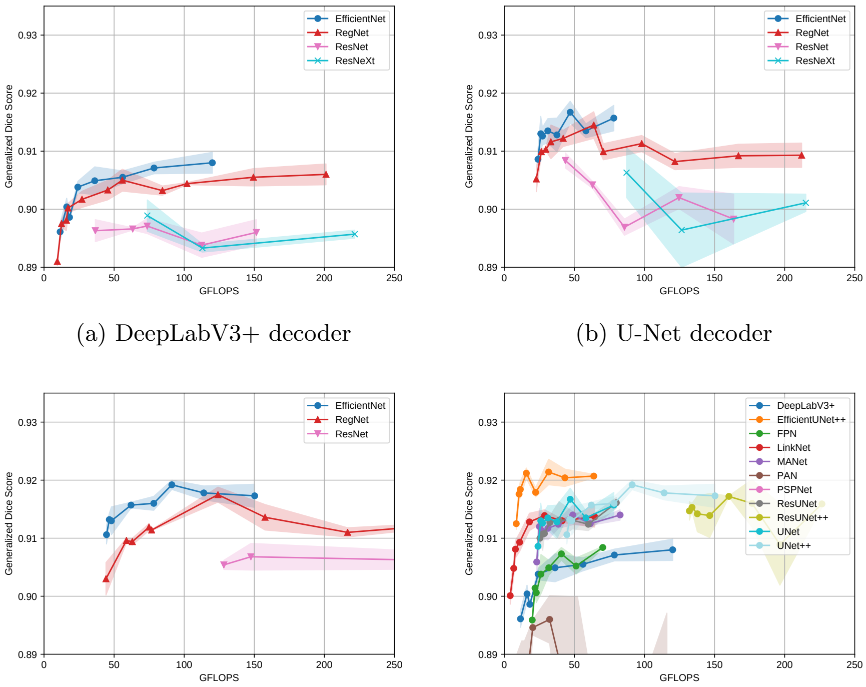
Inspired by these findings, we tested multiple encoders on the coronary artery segmentation task. To avoid overfitting the encoders to our small dataset and evaluate the quality of the visual representations learned from ImageNet, we froze their weights during decoder training, which also shortened network training time and lowered GPU memory use during training. To investigate the existence of synergies between specific encoder-decoder pairs, we trained each backbone with multiple decoders: the U-Net [32], commonly used for medical image segmentation; the UNet++ [51], which has been shown to outperform the U-Net in multiple medical image segmentation tasks; and the DeepLabV3+ [3], a state-of-the-art semantic segmentation architecture. Figures 1a, 1b and 1c display segmentation performance as a function of FLOPS, when using encoders from the EfficientNet [38], RegNetY [31], ResNeXt [43] and ResNet [13] families.

Notably, for every decoder, the best performance is achieved using an EfficientNet backbone, suggesting that these models generalize and transfer better to new tasks. Furthermore, for the same performance, EfficientNet encoders are always more computationally efficient than other backbones. Due to their compound scaling, EfficientNet models are generally thinner at each scale than other encoders, i.e., use fewer channels to represent extracted features. Thus, since decoder computation scales with feature map dimension, EfficientNet backbones make for much more computationally efficient segmentation models than wider encoders. This is particularly evident when using complex decoders, such as the UNet++, that perform heavy processing on extracted features. Additionally, the efficient representations of EfficientNet encoders improve memory efficiency, allowing relatively larger training batches, which can be valuable when working with limited hardware resources.

The results indicate that, in general, better image classification architectures enable higher segmentation performance, which is in line with the widely accepted premise that there is a strong correlation between image classification performance and feature extraction capabilities. However, higher capacity encoders are not always better, and for some combinations of encoder family and decoder, performance starts degrading once the encoder exceeds a certain size. Since encoder weights are not updated during training and the same decoders converge to better solutions when combined with smaller encoder from the same family, the source of degradation appears to be decoder overfitting. This phenomenon is especially striking when using the UNet++, the decoder with more parameters and the one that applies more processing to encoder feature maps. Regularization techniques [24], a larger dataset, or both would probably mitigate or even avoid decoder overfitting. However, possibly due to their thin feature maps, EfficientNet encoders seem to have a regularizing effect on decoders too.

5.2 Decoder Comparison

Given the superior performance and efficiency of EfficientNet backbones, each decoder was trained with multiple encoders from the EfficientNet family. Figure 1d shows the segmentation performance as a function of FLOPS for the DeepLabV3+ [3], FPN [27,23], PSPNet [50], U-Net [32] and the U-Net-based LinkNet [2], MA-Net [26], PAN [25], ResUNet [48], ResUNet++ [20] and UNet++ [51] decoder architectures.



(a) DeepLabV3+ decoder

(b) U-Net decoder

(c) UNet++ decoder

(d) EfficientNet encoders

Fig. 1: Segmentation performance, measured by the GDS, as a function of FLOPS. Figures (a), (b) and (c) show the performance of different encoders combined with the (a) DeepLabV3+, (b) U-Net and (c) UNet++ decoders. Figure (d) shows the performance of different decoders combined with the EfficientNet B0 to B7 encoders. Each polygonal line corresponds to an encoder family. The markers represent the following models, in ascending order of FLOPS: EfficientNet - B0, B1, B2, B3, B4, B5, B6, B7; RegNet - Y2, Y4, Y6, Y8, Y16, Y32, Y40, Y64, Y80, Y120, Y160; ResNet - 18, 34, 50, 101, 152; ResNeXt - 50_32x4d, 101_32x4d, 101_32x8d. Above 250 GFLOPS, performance keeps degrading and is omitted. Models with GDS below 0.89 are also omitted.

The results denote the importance of the skip connections between encoder and decoder at every scale used by U-Net-based models. The UNet++ and the ResUNet++ achieve the best performance among all decoders, and the LinkNet, MA-Net, U-Net, and ResUNet obtain good results, similar to each others'. However, the PAN, which builds on the U-Net, preserving the skip connections and augmenting it with attention mechanisms, performs poorly, suggesting the specific attention mechanisms used are prejudicial for this task.

In fact, the role played by attention mechanisms is not very clear. While they seem to harm the PAN's performance, they appear to be beneficial in the

ResUNet++, and not affect the MA-Net, which performs very similarly to the U-Net, on which it is based. Also unclear is the importance of residual connections, which reduce the ResUNet's performance, compared to the U-Net, but work well in the ResUNet++. Since attention mechanisms and residual connections alone do not enhance the performances of the PAN, MA-Net, and ResUNet, it seems to be their combination that allows the ResUNet++ to perform so well.

Interestingly, the UNet++ performs similarly to the ResUNet++ but more efficiently, both parameter and computation-wise, without using residual connections nor attention. Instead, it uses densely connected nested decoder subnetworks, which promote feature reuse and allow it to extract more information from the encoder's feature maps at each scale.

Architectures that do not leverage localization information from low-level encoder feature maps, like the DeepLabV3+, the FPN and the PSPNet, are the worst-performing. While the lack of localization accuracy allows them to do well in generic segmentation tasks, it harms their performance when applied to medical images, which require fine segmentation. In the case of the DeepLabV3+ and the FPN, which use skip connections from encoder feature maps at one-fourth of original image resolution, this only leads to a slight decrease in performance, compared to U-Net-based models. On the other hand, for the PSPNet, which relies on localization information from encoder feature maps at one-eighth of input resolution, it results in very poor performance.

5.3 EfficientUNet++ Architecture

When coupled with EfficientNet backbones, the UNet++ achieves high segmentation performance with reasonable parameter and computational efficiency. However, while the number of parameters is not a major concern, the computation required for inference can be prohibitive of widespread clinical use, as it requires expensive hardware to be run promptly for entire angiographic sequences, usually comprised of about a hundred frames.

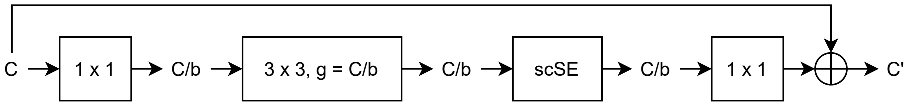


Fig. 2: EfficientUNet++'s convolutional block. Each convolution, except for the last, which is not activated, is followed by BN [18] and Hardswish activation [15]. C and C' are the numbers of input and output channels. Feature map height and width are not altered. We set the bottleneck ratio, b, to 1, and the number of convolution groups, g, to the number of input channels, making the 3 × 3 convolution depthwise. In the scSE block we use a squeeze ratio of 1.

To address this, we propose a new architecture, the EfficientUNet++. Building on the UNet++, it reduces computational complexity by replacing its blocks with residual inverted bottleneck blocks with depthwise separable convolutions, and enhances performance by processing feature maps with concurrent spatial and channel squeeze and excitation (scSE) blocks [33], which combine the channel attention of squeeze and excitation (SE) blocks [16] with spatial attention.

As shown in Figure 1d, when combined with EfficientNet encoders, the EfficientUNet++ establishes a line of efficient and high-performing segmentation models. Coupled with an EfficientNet-B5 encoder, it achieves an average GDS of 0.9202 and DSCs of 0.8858 and 0.7627 for the artery and catheter classes, respectively, outperforming all other models. Figures 3c and 3f display the segmentation masks produced by this model for a left coronary artery (LCA) and a right coronary artery (RCA), respectively.

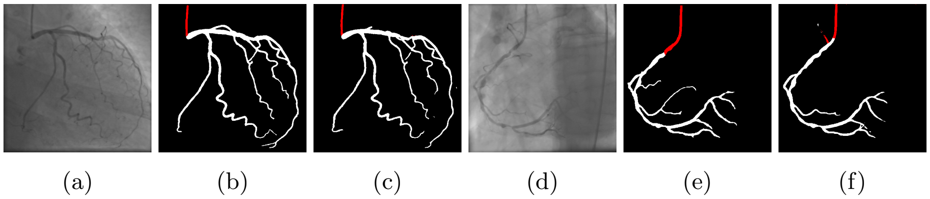


Fig. 3: Segmentation of (a-c) an LCA and (d-f) an RCA. Figures (b) and (e) corresponds to the ground-truth masks, and (c) and (f) are the masks produced by an EfficientNet-B5 encoder combined with an EfficientUNet++ decoder.

5.4 Performance vs. Computation Trade-Off

In clinical practice, depending on the available hardware resources and clinical needs, it may be necessary to make a trade-off between performance and computational efficiency. To help practitioners make that choice, we present, in Figure 4, the Pareto frontier of all tested models, where each model is the most efficient at each performance level, and the best-performing at each computation regime. Thus, all other models need not be considered when looking for the best trade-off between performance and computational efficiency. Notably, neither the U-Net, UNet++ nor DeepLabV3+ appear in this plot, implying that they do not obtain the best trade-offs.

In Figure 4, the lower a point is, the less computation it requires for inference, and the more to the left it is, the better its performance. Therefore, the gentler the slope between a model A and a better-performing model B, the more significant the relative merit of B compared to A. Considering that and the low performance of PSPNet decoders, the slightly more computationally demanding LinkNet-based architectures are probably the best choice when in the presence of a restrictive upper bound on computation. When less constrained by the computation budget, the EfiicientUNet++ decoder offers the best performance, being slightly better than the previously best-performing UNet++ and requiring only about a third of the computation (see Figure 1d).

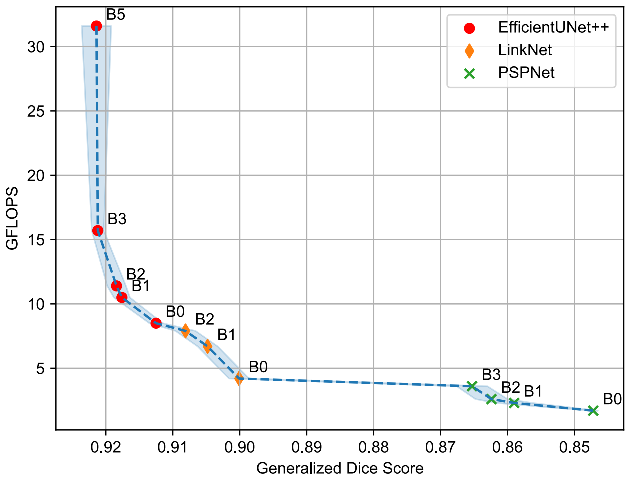
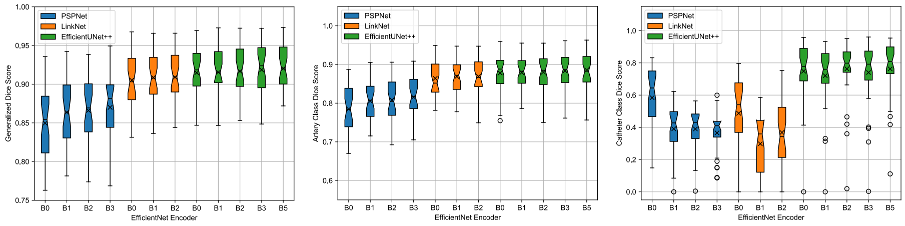


Fig. 4: FLOPS as a function of performance, measured by GDS. The dashed polygonal line corresponds to the Pareto frontier. Each marker represents a model: the text labels indicate the EfficientNet backbone, from B0 to B7, and the colors denote the decoder architecture.

6 Experimental Results

Figure 5 displays the GDS and DSC boxplots of the Pareto-efficient models determined in the previous section. Models using the PSPNet decoder are the worst-performing, only obtaining GDSs above 0.90 in 25% or less of the cases. LinkNet and EfficientUNet++ decoders obtain similar score distributions, with the latter achieve slightly larger mean and quartile values.



(a) GDS

(b) Artery class DSC

(c) Catheter class DSC

Fig. 5: Boxplot of the Pareto-efficient models' GDS and class DSCs. Each box corresponds to a model composed of an encoder denoted by the label in the x-axis and a decoder denoted by the color of the box. For each box, the line inside it and the cross represent the median and mean score of the respective model.

Despite the use of a loss function designed to handle class imbalance, artery segmentation performance is significantly superior to that of the lower volume catheter class. Also, being the most frequent class in the dataset, the artery class is the one that influences the GDS the most, and thus its DSC and the GDS follow similar trends. On the other hand, the DSC of the rarer catheter exhibits a more irregular behavior. For the EfficientUNet++ decoder, performance is high and consistent across cases, tending to improve when higher capacity encoders are used. However, for the remaining decoders, performance is rather inconsistent, and using larger scale encoders seems to harm performance instead of enhancing it, which may be an indicator of overfitting.

7 Implementation Details

7.1 Training Methodology

Encoders were pre-trained on ImageNet [6] and had their weights frozen during decoder training. Decoder parameters in hidden and output layers were initialized using Kaiming [12] and Xavier initialization [10], respectively. Each model was trained for 150 epochs using Adam [22] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a mini-batch size of 8, no weight decay, and an initial learning rate of 0.001, divided by ten at the 50^{th} and 100^{th} epochs. Most experiments were run using public PyTorch [30] implementations [44] under an MIT license. To keep comparisons fair, the repository was extended with ResUNet, ResUNet++ and EfficientUNet++ implementations. To obtain average scores and standard deviations, each model was trained and tested three times.

7.2 Dataset

Our dataset comprises 270 monochromatic 512 × 512 anonymized CAG images collected at a single center, between 2017 and 2019. The images were acquired from multiple viewing angles of LCA and RCA of 47 random patients, over the age of eighteen, who underwent CAG and invasive physiological assessment, i.e., fractional flow reserve (FFR), instantaneous wave-free ratio (iFR) or other indices' measurement. Each patient's angiograms were annotated by one of three expert cardiologists, blinded to the patient's identity, demographic information and medical history. Approximately one third of the patients underwent revascularization during or after the diagnostic angiography.

Measuring each vessel's diameter would have been impractical. Therefore, the catheter, whose diameter is known and varies between 1.8 mm and 2 mm, was used as a proxy to determine whether each vessel should be segmented or not. While this may have led to erroneous segmentation of some vessels with diameters close to 2 mm, it significantly simplified the annotation process. To diminish the quantity and impact of these and other errors, all annotations were reviewed multiple times by all three physicians.

We split the dataset at the patient level into a training, a validation and a test set, composed of 144/63, 21/9 and 23/10 images of the LCA/RCA, respectively. The images of each set were carefully chosen to keep it representative of the original one, having approximately identical distributions regarding the observed arteries, viewing angles and number of images annotated by each physician.

7.3 Data Augmentation

Our augmentation policy consists of the sequential application of the following random transformations: 1) $-20°$ to $20°$ rotation; 2) -10% to 10% horizontal and vertical translation; 3) -10% to 10% zoom; 4) -40% to 40% brightness variation, to account for brightness variability across devices. Following Fort et al. [9], we perform online augmentation and draw multiple augmentation samples per image in a growing batch regime. However, we keep the original images in the batch, as we observed this to improve performance slightly. Specifically, each batch is composed of two original images and three augmentations of each.

8 Discussion and Future Work

In this work, we propose a new and better clinically-suited criterion for catheter and artery segmentation in CAG images, developed in collaboration with expert cardiologists. Whereas most previous approaches either segment only the major vessel or the whole coronary tree, based mostly on contrast information, we only segment vessels relevant for diagnostic and therapeutic purposes.

To determine the best approach for the task, we conducted a comprehensive comparison of encoder and decoder architectures, whose results may prove useful for other medical image segmentation tasks. We found the EfficientNet [38] and the UNet++ [51] to be the best-performing encoder and decoder architectures, respectively. Due to their compound scaling, EfficientNet backbones are not only computationally and parameter efficient, but also in the way they represent features, generally using fewer channels at each scale than other models. This has a threefold benefit: 1) requires less computation from decoders; 2) seems to have a regularizing effect on decoders; 3) reduces memory use during training.

The performance of the UNet++ is related to its structure, whose densely connected nested decoders operating at different scales promote feature reuse and information flow through the network, and allow better multi-scale and overall processing of the features extracted by the encoder. Based on the UNet++, we propose a new computationally efficient and high-performing decoder architecture, the EfficientUNet++, which simultaneously increases its computationally efficiency and performance, through the use of light-weight depthwise separable convolutions and scSE spatial and channel attention blocks [33].

In the future, we plan to further improve our models through the use of self and semi-supervised learning techniques that allow us to take advantage of the tens of thousands of available unlabeled CAG images. Furthermore, we intend to test the EfficientUNet++ with attention mechanisms other than scSE, such as the CBAM [40] and Triplet Attention [29], and explore new architectures, such as Vision Transformers [7], whose features contain explicit semantic segmentation information when self-trained [1], and hybrid models combining convolutions and self-attention [5,41,11,47], which possess long-range modeling capabilities that may be crucial in the coronary artery segmentation task. Finally, we plan to investigate the clinical usefulness of our models, which is not necessarily determined by the DSCs and GDS they obtain.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e Tecnologia (FCT), under the project with reference UIDB/50021/2020.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294 (2021)
2. Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP). pp. 1–4. IEEE (2017)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
4. Crum, W.R., Camara, O., Hill, D.L.: Generalized overlap measures for evaluation and validation in medical image analysis. IEEE transactions on medical imaging **25**(11), 1451–1461 (2006)
5. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. arXiv preprint arXiv:2106.04803 (2021)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Fan, J., Yang, J., Wang, Y., Yang, S., Ai, D., Huang, Y., Song, H., Hao, A., Wang, Y.: Multichannel fully convolutional network for coronary artery segmentation in x-ray angiograms. Ieee Access **6**, 44635–44643 (2018)
9. Fort, S., Brock, A., Pascanu, R., De, S., Smith, S.L.: Drawing multiple augmentation samples per image during training efficiently decreases test error. arXiv preprint arXiv:2105.13343 (2021)
10. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
11. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: a vision transformer in convnet's clothing for faster inference. arXiv preprint arXiv:2104.01136 (2021)
12. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

14. van der Heijden, L.C., Kok, M.M., Danse, P.W., Schramm, A.R., Hartmann, M., Löwik, M.M., Linssen, G.C., Stoel, M.G., Doggen, C.J., von Birgelen, C.: Small-vessel treatment with contemporary newer-generation drug-eluting coronary stents in all-comers: Insights from 2-year dutch peers (twente ii) randomized trial. American heart journal **176**, 28–35 (2016)
15. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324 (2019)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
17. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015)
19. Iyer, K., Najarian, C.P., Fattah, A.A., Arthurs, C.J., Soroushmehr, S.R., Subban, V., Sankardas, M.A., Nadakuditi, R.R., Nallamothu, B.K., Figueroa, C.A.: Angionet: A convolutional neural network for vessel segmentation in x-ray angiography. medRxiv (2021)
20. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: Resunet++: An advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM). pp. 225–2255. IEEE (2019)
21. Jun, T.J., Kweon, J., Kim, Y.H., Kim, D.: T-net: Nested encoder–decoder architecture for the main vessel segmentation in coronary angiography. Neural Networks **128**, 216–233 (2020)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Kirillov, A., He, K., Girshick, R., Dollár, P.: Iccv_stuff_fair_final. http://presentations.cocodataset.org/COCO17-Stuff-FAIR.pdf, (Accessed on 10/06/2021)
24. Kukačka, J., Golkov, V., Cremers, D.: Regularization for deep learning: A taxonomy. arXiv preprint arXiv:1710.10686 (2017)
25. Li, H., Xiong, P., An, J., Wang, L.: Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180 (2018)
26. Li, R., Zheng, S., Duan, C., Zhang, C., Su, J., Atkinson, P.: Multi-attention-network for semantic segmentation of fine resolution remote sensing images. arXiv preprint arXiv:2009.02130 (2020)
27. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
28. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
29. Misra, D., Nalamada, T., Arasanipalai, A.U., Hou, Q.: Rotate to attend: Convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3139–3148 (2021)

30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
31. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10428–10436 (2020)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
33. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In: International conference on medical image computing and computer-assisted intervention. pp. 421–429. Springer (2018)
34. Rudd, K.E., Johnson, S.C., Agesa, K.M., Shackelford, K.A., Tsoi, D., Kievlan, D.R., Colombara, D.V., Ikuta, K.S., Kissoon, N., Finfer, S., et al.: Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. The Lancet **395**(10219), 200–211 (2020)
35. Samuel, P.M., Veeramalai, T.: Vssc net: vessel specific skip chain convolutional network for blood vessel segmentation. Computer Methods and Programs in Biomedicine **198**, 105769 (2021)
36. Sim, H.W., Ananthakrishna, R., Chan, S.P., Low, A.F., Lee, C.H., Chan, M.Y., Tay, E.L., Loh, P.H., Chan, K.H., Tan, H.C., et al.: Treatment of very small de novo coronary artery disease with 2.0 mm drug-coated balloons showed 1-year clinical outcome comparable with 2.0 mm drug-eluting stents. J Invasive Cardiol **30**(7), 256–261 (2018)
37. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 240–248. Springer (2017)
38. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)
39. Vlontzos, A., Mikolajczyk, K.: Deep segmentation and registration in x-ray angiography video. arXiv preprint arXiv:1805.06406 (2018)
40. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
41. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. arXiv preprint arXiv:2103.15808 (2021)
42. Xian, Z., Wang, X., Yan, S., Yang, D., Chen, J., Peng, C.: Main coronary vessel segmentation using deep learning in smart medical. Mathematical Problems in Engineering **2020** (2020)
43. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
44. Yakubovskiy, P.: Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2020)

45. Yang, S., Kweon, J., Kim, Y.H.: Major vessel segmentation on x-ray coronary angiography using deep networks with a novel penalty loss function. In: International Conference on Medical Imaging with Deep Learning–Extended Abstract Track (2019)
46. Yang, S., Kweon, J., Roh, J.H., Lee, J.H., Kang, H., Park, L.J., Kim, D.J., Yang, H., Hur, J., Kang, D.Y., et al.: Deep learning segmentation of major vessels in x-ray coronary angiography. Scientific reports **9**(1), 1–11 (2019)
47. Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., Wu, W.: Incorporating convolution designs into visual transformers. arXiv preprint arXiv:2103.11816 (2021)
48. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749–753 (2018)
49. Zhao, C., Tang, H., Tang, J., Zhang, C., He, Z., Wang, Y.P., Deng, H.W., Bober, R., Zhou, W.: Semantic segmentation to extract coronary arteries in fluoroscopy angiograms. medRxiv (2020)
50. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
51. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer (2018)
52. Zhu, X., Cheng, Z., Wang, S., Chen, X., Lu, G.: Coronary angiography image segmentation based on pspnet. Computer Methods and Programs in Biomedicine **200**, 105897 (2021)