S1. Assessing the impact of factorized 3D self-attention

Using fully 3D self-attention involves significant costs in both GPU memory and computation time, making our full length pipeline with 100 pre-training epochs and 50 fine-tuning epochs unfeasible. As such, comparing both approaches requires us to reduce both the size of the pre-training data and the amount of epochs. Specifically, we do pre-training only on the 1,391 high resolution scans and we run 25 pre-training epochs/20 fine-tuning epochs. We do this only for the best pre-training strategy (in practice, this is equivalent to model **C** from Table 3 in the main text, but with shorter training times). For the sake exhaustiveness, we also compare the impact of using full 3D self-attention without any kind of pre-training. The results from this experiment can be found in Table S1. Overall, we can see that models S_a and S_c perform significantly worse than their factorized counterparts on two of the O.O.D. datasets, suggesting that models using fully-3D self-attention may generalize poorly (at least when trained on smaller datasets).

Table S1. Detection performance using an IoU threshold $t_{IoU} = 0.3$ and assuming a fixed FPr=0.5, reported for models using full versus factorized attention (attn.) with and without pre-training (PT).

Mod.	**Attn.**	**PT**	**Se ↑ (%)**			
			Int.	**Ext.**	**CMHA**	**Priv.**
S_a	Full	✗	66.7	77.2	54.0	31.1
S_b	Fact.	✗	70.6	79.2	63.0	54.2
S_c	Full	✓	81.7	85.1	73.0	38.7
S_d	Fact.	✓	79.4	87.1	79.0	57.5

S2. Measuring the effect of segmentation accuracy

Due to the scale of our combined pre-training, fine-tuning, and evaluation datasets, running other deep learning-based segmentation pipelines can have a significant computational overhead. Neverthless, it is important to measure whether having a good segmentation pipeline can have a significant impact on our pipeline. As a proxy for comparing lower-versus higher-performing, we simulate a low segmentation performance scenario using intensity based thresholding. Specifically, we exploit the fact vessels are bright and use the 80% HU intensity percentile (computed while ignoring background pixels, which we remove with a brain mask [35]) in every scan to get a rough segmentation mask. Afterward, we use 3 mm dilation followed by 3 mm erosion to remove artifacts. Since a thresholding-based approach cannot discriminate between arteries and veins, it would be difficult to compare it with our artery-centric approach. Thus, we combine the artery mask from our segmentation model with its other output: the vein mask. This approach is similar to the one used in [9], albeit it has slightly inferior Sensitivity than the artery-only awareness used in this paper. Our results (see Table S2) suggest that training with low quality segmentation masks (V_a versus V_c and V_d) can negatively harm detection performance, even if inference is done on higher quality segmentation masks. Unsurprisingly, using deep learning-based segmentations for both training and inference results in the highest Sensitivity (V_d), albeit the performance hit from using thresholding-based masks at inference time (V_c) is not large. These findings suggest that, provided that training is done with high quality segmentations of vessels (or arteries specifically), the resulting model will be robust enough to use lower quality masks. Further experimentation is required to validate whether the use of higher quality deep learning-based segmentations can further improve model Sensitivity.

Table S2. Detection performance using an IoU threshold $t_{IoU} = 0.3$ and assuming a fixed FPr=0.5, reported for models thresholding-based (Thr.) versus deep learning-based (DL) vessel segmentation during training (encompassing both pre-training and fine-tuning) and/or inference (inf.)

Mod.	**Segmentation**		**Se ↑ (%)**			
	Train	**Inf.**				
			Int.	**Ext.**	**CMHA**	**Priv.**
V_a	Thr.	Thr.	77.8	84.2	70.0	53.3
V_b	Thr.	DL	74.6	86.1	67.0	44.8
V_c	DL	Thr.	82.5	90.0	83.0	68.9
V_d	DL	DL	86.5	95.0	88.0	75.5