3D Arterial Segmentation via Single 2D Projections and Depth Supervision in Contrast-Enhanced CT Images Supplementary Material

Table 1: **Data augmentation and input preprocessing during training.** The input volume V is pre-cropped to $(300, 300, 150)$ around the ostia of the superior mesenteric artery.

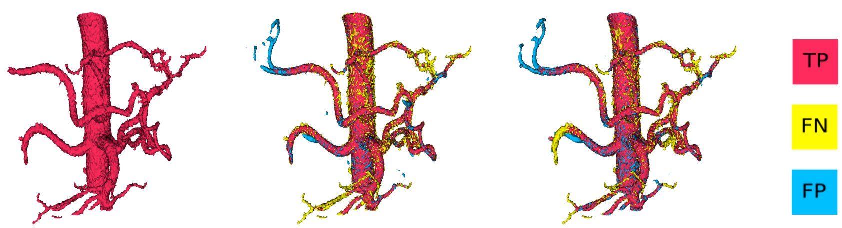
Transformation	Parameters	Probability
Rescale	[0, 1]	$p = 1$
Gaussian blur	$\sigma \in [0, 2]$	$p = 0.75$
Gamma change	$\lambda \in [-3, 3]$	$p = 0.75$
	scaling $s \in [0.8, 0.2]$	$p = 0.9$
Affine transformation	rotations of up to $15°$	$p = 0.9$
	translation of up to $(30, 30, 3)$ in each axis	$p = 0.9$
Crop volume	from $(300, 300, 150)$ to $(256, 256, 128)$	$p = 1$

Table 2: **Processing operations for different stages of our pipeline**: rib extraction, maximum intensity projection of the input, and postprocessing of the depth maps. • $M_{y_1|y_2} \in \mathbb{R}^{300\times300\times150}$ denotes a cropping mask $M(x, y, z) = 1$ if $y < y_1$ or $y > (300 - y_2)$, else 0. • $V * C_a \leq b$ denotes a convolution operation of a volume V with a cube of size a, followed by a thresholding operation at threshold b and re-binarization.

Stage	Operation	Parameters
Rib extraction	threshold and binarize	$V < 300 \to 0, \quad V \geq 300 \to 1$
	connected components (CCs)	connectivity of 6 (no diagonal pixels)
	mask out	CCs smaller than 100000 pixels
	mask out	CCs c where $\exists(x, y, z) \in c$ such that $M_{80\|30}(x, y, z) = 0$
	binary dilation	structuring element: cube of size 5
Input MIP	apply mask	ribs, vertebrae, $M_{80\|100}$
	clip volume	$[150, 255]$
	crop volume	$(256, 256, 128)$
	rotate: Euler angles $(\alpha_x, \alpha_y, \alpha_z)$	$(0, 0, 0)$, $(-90, 0, 0)$ or $(0, 0, -90)$
	resample	$(256, 256, 128)$
	rescale	$[150, 255] \to [0, 255]$
	maximum intensity projection	y axis
Depth map	compute depth map	intensity fluctuation $th = 0.1$ (Step 3 of depth map generation)
	remove disconnected pixels	$D * C_3 \leq 1$
	remove very sparse areas	$D * C_{11} \leq 9$

Table 3: **Ablation experiment on training with fixed viewpoints (VPs)**. We train models using 2D projections on **fixed** viewpoints (same for each training sample). The 3 viewpoints considered are: coronal projection $\rightarrow c$, axial projection $\rightarrow a$, sagittal projection $\rightarrow s$. Each experiment is averaged over 5 cross-validation folds in accordance with our experimental design.

# VPs	Viewpoints	Dice	Precision	Recall	Skeleton Recall	MSD
3	c, a, s	90.78 ± 1.30	90.66 ± 1.30	91.18 ± 3.08	81.77 ± 2.13	1.16 ± 0.13
2	c, a	88.97 ± 1.11	85.26 ± 1.72	93.43 ± 1.35	83.78 ± 2.15	1.22 ± 0.09
	a, s	91.01 ± 0.65	90.20 ± 2.70	92.14 ± 2.09	81.56 ± 2.91	1.13 ± 0.05
	c, s	90.68 ± 0.44	89.03 ± 0.95	92.64 ± 0.96	81.20 ± 1.16	1.07 ± 0.03
	avg	90.22 ± 1.19	88.16 ± 2.86	92.74 ± 1.63	82.18 ± 2.47	1.14 ± 0.09
1	c	77.59 ± 1.91	68.15 ± 2.76	91.17 ± 3.04	79.89 ± 1.69	2.18 ± 0.21
	a	32.82 ± 23.33	24.61 ± 23.13	92.44 ± 3.45	80.73 ± 2.39	4.38 ± 5.17
	s	71.86 ± 3.62	58.66 ± 4.99	93.94 ± 1.91	82.94 ± 1.91	2.32 ± 0.27
	avg	60.76 ± 24.14	50.47 ± 23.21	92.52 ± 3.09	81.19 ± 2.39	2.96 ± 3.15



(a) Ground truth (b) 3D supervision (c) Ours (d) Legend

Fig. 1: **Qualitative results**. 3D rendering of the predicted segmentation of one of our models (c) compared to a model trained using full 3D supervision (b).

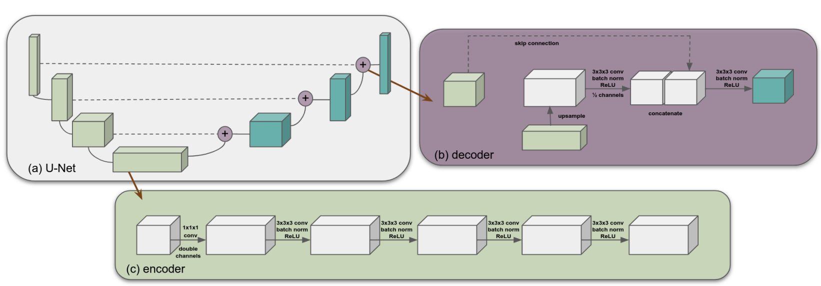


Fig. 2: **Network architecture** Our U-Net has 4 layers. Between each encoder layer we perform a 2× max pooling operation and double the output channels. The number of output channels at each layer are: 16, 32, 64, 128.