Table 2. Ablation of Auto3DSeg SegResNet[1,28] following tutorial [2] using fold 0 of KiTS. All experiments were executed on A100 GPUs with 40GB of VRAM. Following the authors recommendation to increase compute resources (manual increase of epochs and patch size), A3DS SegResNet yielded significantly improved results, achieving 87.77% while taking approx. 240 GPU hours (80x30h) and a total of 320GB VRAM (8x40GB). Although this result now surpasses the standard nnU-Net (9h, 7GB VRAM; 86.25%) it is still outperformed by nnU-Net ResEnc M (11h, 9GB; 87.91%) and its larger cousins.

Model	GPU hours	VRAM (GPUs × MB)	Epochs	Batch Size	Patch Size	Spacing	KiTS Fold 0 DSC [%]
nnU-Net (org.)	8.88	1 × 6901	-	2	128 × 128 × 128	1 × 0.78 × 0.78	86.25
nnU-Net ResEnc M	11.39	1 × 8805	-	2	128 × 128 × 128	1 × 0.78 × 0.78	87.91
nnU-Net ResEnc L	35.28	1 × 24223	-	2	160 × 224 × 192	1 × 0.78 × 0.78	88.60
A3DS SegResNet	39.72	1 × 20267	300	2	144 × 224 × 224	1 × 0.78 × 0.78	83.73
A3DS SegResNet	61.28	8 × 20267	300	8 × 2	144 × 224 × 224	1 × 0.78 × 0.78	76.81
A3DS SegResNet	136.64	8 × 20267	600	8 × 2	144 × 224 × 224	0.78 × 0.78 × 0.78	85.60
A3DS SegResNet	247.44	8 × 39873	900	8 × 2	224 × 256 × 256	0.78 × 0.78 × 0.78	87.77

Table 3. Not all datasets can be recommended to develop and compare architectures. We report the standard deviation of DSC scores across folds of the same method (**intra** method SD). We compare this to the standard deviation computed over the average DSC scores across all methods on the dataset (**inter** method SD). The greater the ratio $\frac{inter_{SD}}{intra_{SD}}$, the more suitable is a dataset for separating methods. "SD": Standard Deviation

	BTCV	**ACDC**	**LiTS**	**BraTS2021**	**KiTS2023**	**AMOS2022**
nnU-Net (org.)	2.6%	0.8%	3.5%	0.62%	2.0%	0.43%
nnU-Net ResEnc M	2.4%	0.62%	2.6%	0.67%	2.2%	0.57%
nnU-Net ResEnc L	2.7%	0.6%	2.4%	0.57%	1.3%	0.59%
nnU-Net ResEnc XL	2.7%	0.51%	2.4%	0.62%	1.2%	0.43%
MedNeXt L k3	2.1%	0.26%	2.3%	0.66%	0.94%	0.43%
MedNeXt L k5	2.0%	0.2%	2.4%	0.59%	1.2%	0.43%
STU-Net S	2.2%	0.6%	3.3%	0.72%	1.7%	0.42%
STU-Net B	2.3%	0.78%	3.6%	1.0%	1.9%	0.52%
STU-Net L	2.6%	0.85%	2.4%	0.62%	2.1%	0.45%
SwinUNETR	2.7%	0.65%	3.1%	0.75%	2.0%	0.44%
SwinUNETRV2	2.1%	0.51%	2.8%	0.55%	1.7%	0.56%
nnFormer	2.1%	0.21%	2.3%	0.52%	4.2%	0.5%
CoTr	2.8%	0.83%	2.8%	0.69%	1.4%	0.64%
No-Mamba Base	1.9%	0.51%	2.9%	0.55%	2.1%	0.32%
U-Mamba Bot	2.3%	0.59%	2.1%	0.71%	2.7%	0.43%
U-Mamba Enc	2.3%	0.47%	1.7%	0.64%	2.2%	0.5%
A3DS SegResNet	3.0%	0.33%	2.7%	0.52%	1.7%	0.48%
A3DS DiNTS	3.0%	2.2%	2.5%	0.79%	5.3%	1.3%
A3DS SwinUNETR	1.8%	3.6%	6.6%	0.69%	1.5%	0.64%
Averages						
Intra Method SD	2.39%	0.79%	2.89%	0.66%	2.07%	0.53%
Inter Method SD	2.24%	2.83%	3.80%	0.84%	9.03%	2.52%
Inter/Intra Ratio	94%	357%	132%	127%	435%	474%
Averages w/o A3DS						
Intra Method SD	2.35%	0.56%	2.66%	0.66%	1.93%	0.48%
Inter Method SD	1.52%	0.57%	1.68%	0.35%	3.14%	2.28%
Inter/Intra Ratio	65%	102%	63%	53%	163%	477%

Table 4. Normalized Surface Distance (NSD) with tolerance 2 mm for all methods and datasets. Reported values are averages over the five-fold cross-validation. NSD was computed using https://github.com/google-deepmind/surface-distance. Relative performance between methods is consistent with the observations based on Dice alone (see Table 1 and Results section).

Architecture	LiTS	BTCV	ACDC	BraTS2021	KiTS2023	AMOS2022
nnU-Net (org.)	78.26	85.53	94.93	93.64	82.91	91.49
nnU-Net ResEnc M	79.96	86.01	95.50	93.71	84.10	91.72
nnU-Net ResEnc L	80.39	86.08	95.11	93.59	85.93	92.35
nnU-Net ResEnc XL	79.64	85.89	94.90	93.61	86.49	92.64
MedNeXt L k3	81.07	87.78	96.07	93.85	86.29	92.72
MedNeXt L k5	81.26	88.18	96.09	94.04	85.67	92.86
STU-Net S	76.20	85.13	94.27	93.26	81.08	90.81
STU-Net B	77.33	85.30	94.59	93.54	83.08	91.28
STU-Net L	78.85	85.81	95.12	93.66	83.02	92.30
SwinUNETR	73.06	79.79	94.12	93.16	75.91	85.13
SwinUNETRV2	75.38	82.52	95.15	93.15	80.11	88.47
nnFormer	74.66	82.29	95.83	93.22	69.43	82.93
CoTr	77.25	84.10	93.74	93.49	80.92	90.75
No-Mamba Base	78.88	86.14	95.26	93.64	83.56	92.08
U-Mamba Bot	78.91	86.40	95.40	93.65	83.27	92.00
U-Mamba Enc	78.60	84.60	94.33	93.21	83.64	91.25
A3DS SegResNet	76.46	82.01	93.88	93.40	75.61	89.85
A3DS DiNTS	62.49	77.30	83.67	90.36	58.74	82.75
A3DS SwinUNETR	61.16	74.59	83.94	92.00	46.37	86.93

Table 5. Ablation of average DSC when using isotropic spacing for the nnU-Net ResEnc L on ACDC and BTCV instead of the default anisotropic spacing. Results indicate that MedNeXt's performance can in part be explained by its target spacing selection and is thus not exclusively linked to the better architecture. "DSC": Dice similarity coefficient.

Dataset	Method	Patch Size	Spacing	Batch Size	DSC
BTCV	nnU-Net ResEnc L	80x256x256	3x0.76x0.76	2	83.35
	nnU-Net ResEnc L (iso)	192x192x192	1x1x1	2	84.01
	MedNeXt L k3	128x128x128	1x1x1	2	84.70
ACDC	nnU-Net ResEnc L	20x256x224	5x1.56x1.56	10	91.69
	nnU-Net ResEnc L (iso)	96x256x256	1x1x1	3	92.64
	MedNeXt L k3	128x128x128	1x1x1	2	92.65
AMOS	nnU-Net ResEnc L	96x224x224	2x0.71x0.71	2	89.40
	nnU-Net ResEnc L (iso)	192x192x192	1x1x1	2	89.60
	MedNeXt L k3	128x128x128	1x1x1	2	89.62

Table 6. Since STU-Net was presented as a model for transfer learning, we fine-tuned a STU-Net L network, that was pre-trained on the totalsegmentator dataset [35] for 4000 epochs, on the other datasets, analogous to the corresponding publication [20]. Fine-tuning on BraTS did not converge using the default fine-tuning learning rate of 0.001.

	BTCV n=30	ACDC n=200	LiTS n=131	BraTS n=1251	KiTS n=489	AMOS n=360	VRAM [GB]	RT [h]	Arch.	nnU
STU-Net L [20]	83.36	91.31	80.31	91.26	85.84	89.34	26.50	51	CNN	Yes
STU-Net L pretrained [20]	84.28	91.53	81.57	0	88.32	89.46	26.50	51*	CNN	Yes

*Fine-tuning runtime only. Pre-training takes about 4 times longer.