Revisiting Model Scaling with a U-Net Benchmark for 3D Medical Image Segmentation

Ziyan Huang[1,2,3], Jin Ye[3,4], Haoyu Wang[1,2,3], Zhongying Deng[5], Zhikai Yang[6], Yanzhou Su[3], Jie Liu[7], Tianbin Li[3], Yun Gu[1], Shaoting Zhang[3], Yu Qiao[3], Lixu Gu[1,2,8,*], and Junjun He[3,*]

[1]Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, 200240, China
[2]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[3]Shanghai Artificial Intelligence Laboratory, Shanghai, 200232, China
[4]Department of Data Science and AI, Monash University, Melbourne, 3800, Australia
[5]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, CB3 0WA, United Kingdom
[6]Department of Biomedical Engineering and Health, KTH Royal Institute of Technology, Stockholm, 14152, Sweden
[7]Accumed Technology Co., Ltd., Shanghai, 200240, China
[8]Pittsburgh Institute, Sichuan University, Chengdu, 610207, China
[*]Corresponding authors: gulixu@sjtu.edu.cn; hejunjun@pjlab.org.cn

S1 Detailed Results on 42 Datasets

This section provides detailed segmentation results of all 18 U-Net model variants across 42 public 3D medical image segmentation datasets. For each dataset, we report the mean and standard deviation of Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD). Each subsection includes a brief dataset description, evaluation tables for all model configurations, and violin plots for visual comparison.

S1.1 MSD Task01 Brain

The MSD Brain[1] dataset comprises 484 publicly labeled cases, which include the segmentation of three tumor subregions: edema, enhancing tumor, and non-enhancing tumor. The utilized sequences are native T1-weighted (T1), Gadolinium (Gd)-enhanced T1-weighted (T1-Gd), native T2-weighted (T2), and T2 fluid-attenuated inversion recovery (FLAIR). This dataset includes the same cases as the 2016 and 2017 Brain Tumor Segmentation (BraTS) challenges. We employed a 5-fold cross-validation for our experimental results. The target spacing is $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$, and the crop size for training and inference is $128 \times 128 \times 128$.

Table S1. DSC results of 18 U-Net models on the MSD Task01 Brain dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8376 ± 0.1276	0.8344 ± 0.1279	0.8389 ± 0.1254	0.8370 ± 0.1257	0.8388 ± 0.1295	0.8396 ± 0.1216
W32	0.8413 ± 0.1267	0.8412 ± 0.1236	0.8420 ± 0.1262	0.8421 ± 0.1244	0.8425 ± 0.1273	0.8425 ± 0.1227
W64	$\mathbf{0.8439 \pm 0.1243}$	0.8438 ± 0.1221	$\mathbf{0.8439 \pm 0.1226}$	0.8409 ± 0.1295	0.8432 ± 0.1245	0.8438 ± 0.1249

Table S2. NSD results of 18 U-Net models on the MSD Task01 Brain dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7946 ± 0.1404	0.7881 ± 0.1435	0.7959 ± 0.1411	0.7922 ± 0.1399	0.7975 ± 0.1434	0.7942 ± 0.1375
W32	0.8018 ± 0.1388	0.8013 ± 0.1357	0.8029 ± 0.1405	0.8025 ± 0.1373	0.8053 ± 0.1396	0.8029 ± 0.1373
W64	$\mathbf{0.8066 \pm 0.1375}$	0.8049 ± 0.1354	0.8061 ± 0.1376	0.8024 ± 0.1430	0.8057 ± 0.1396	0.8064 ± 0.1380

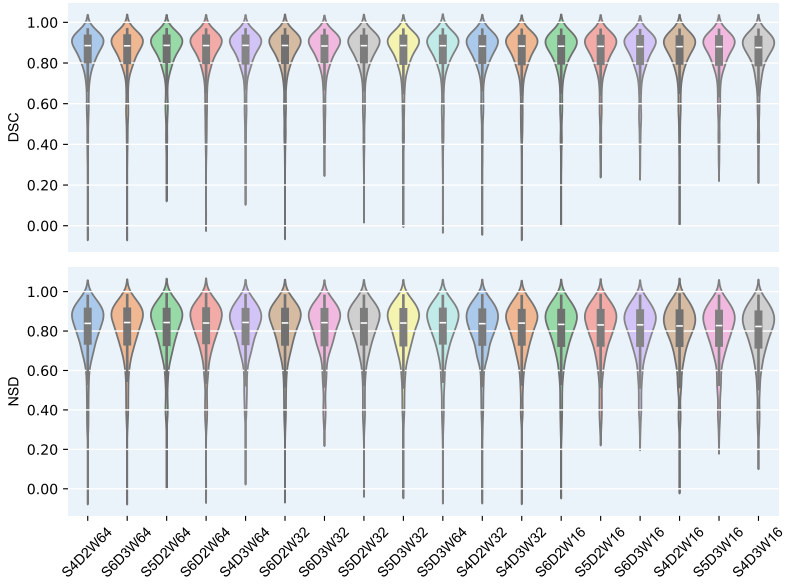

Figure S1. Violin plots of DSC and NSD for MSD Task01 Brain dataset.

S1.2 MSD Task02 Heart

The MSD Task02 Heart[1] dataset comprises 20 publicly labeled cases, aimed at segmenting the left atrium from MR images. We employed a 5-fold cross-validation for our experimental results. The target spacing is 1.25 mm × 1.25 mm × 1.37 mm, and the crop size for training and inference is 80 × 192 × 160.

Table S3. DSC results of 18 U-Net models on the MSD Task02 Heart dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9318 ± 0.0189	0.9318 ± 0.0183	0.9325 ± 0.0194	0.9325 ± 0.0173	0.9325 ± 0.0189	0.9312 ± 0.0179
W32	0.9323 ± 0.0191	0.9325 ± 0.0170	0.9326 ± 0.0181	0.9323 ± 0.0175	0.9321 ± 0.0187	**0.9329 ± 0.0167**
W64	0.9319 ± 0.0178	0.9317 ± 0.0181	0.9322 ± 0.0177	0.9323 ± 0.0168	0.9325 ± 0.0177	0.9321 ± 0.0177

Table S4. NSD results of 18 U-Net models on the MSD Task02 Heart dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6247 ± 0.0785	0.6252 ± 0.0799	0.6281 ± 0.0847	0.6264 ± 0.0776	0.6301 ± 0.0817	0.6250 ± 0.0727
W32	**0.6313 ± 0.0777**	0.6290 ± 0.0724	0.6281 ± 0.0747	0.6257 ± 0.0760	0.6278 ± 0.0773	0.6286 ± 0.0719
W64	0.6280 ± 0.0770	0.6271 ± 0.0763	0.6287 ± 0.0782	0.6262 ± 0.0749	0.6317 ± 0.0709	0.6275 ± 0.0764

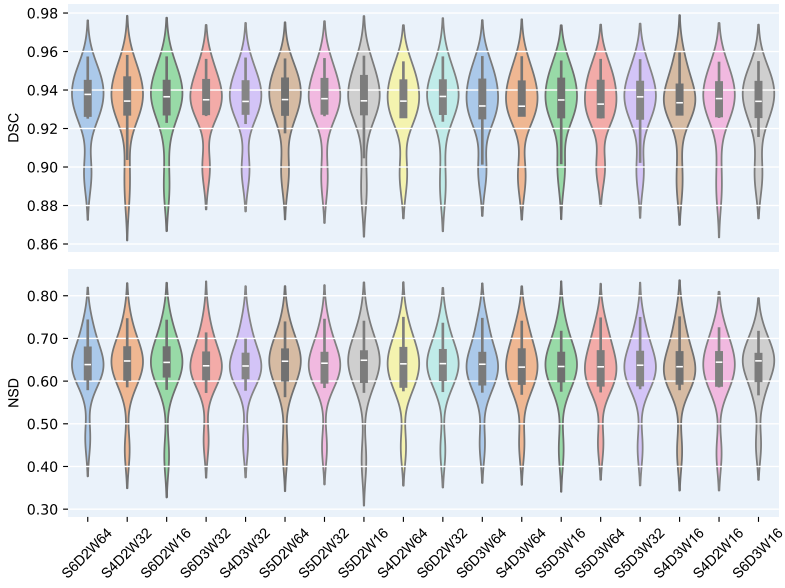

Figure S2. Violin plots of DSC and NSD for MSD Task02 Heart dataset.

S1.3 MSD Task03 Liver

The MSD Task03 Liver[1] dataset comprises 131 publicly labeled cases, aimed at segmenting the liver and liver tumor from CT images. We employed a 5-fold cross-validation for our experimental results. The target spacing is $0.77\,\text{mm} \times 0.77\,\text{mm} \times 1\,\text{mm}$, and the crop size for training and inference is $128 \times 128 \times 128$.

Table S5. DSC results of 18 U-Net models on the MSD Task03 Liver dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7867 ± 0.1605	0.7932 ± 0.1552	0.7892 ± 0.1692	0.7990 ± 0.1518	0.7918 ± 0.1693	0.8022 ± 0.1511
W32	0.7928 ± 0.1533	0.8059 ± 0.1483	0.7955 ± 0.1658	0.8124 ± 0.1452	0.7950 ± 0.1673	0.7964 ± 0.1637
W64	0.8005 ± 0.1528	$\mathbf{0.8145 \pm 0.1442}$	0.8008 ± 0.1510	0.8093 ± 0.1496	0.8019 ± 0.1651	0.8091 ± 0.1595

Table S6. NSD results of 18 U-Net models on the MSD Task03 Liver dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6336 ± 0.1530	0.6420 ± 0.1515	0.6412 ± 0.1590	0.6515 ± 0.1469	0.6439 ± 0.1576	0.6540 ± 0.1460
W32	0.6441 ± 0.1505	0.6575 ± 0.1480	0.6517 ± 0.1577	0.6681 ± 0.1460	0.6519 ± 0.1546	0.6522 ± 0.1530
W64	0.6541 ± 0.1540	$\mathbf{0.6690 \pm 0.1471}$	0.6578 ± 0.1458	0.6671 ± 0.1474	0.6602 ± 0.1570	0.6663 ± 0.1571

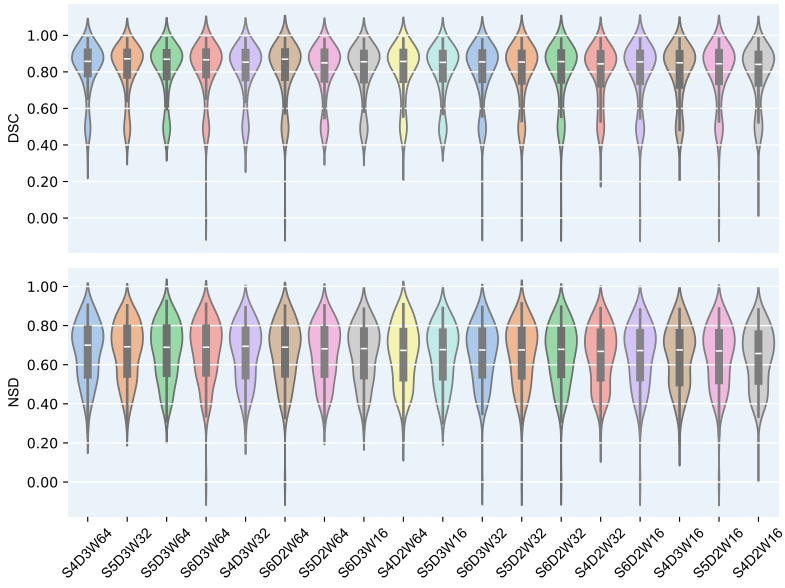

Figure S3. Violin plots of DSC and NSD for MSD Task03 Liver dataset.

S1.4 MSD Task04 Hippocampus

The MSD Task04 Hippocampus[1] dataset contains 260 publicly labeled cases and focuses on segmenting the anterior and posterior hippocampus from MR images. We employed a 5-fold cross-validation for our experimental results. The target spacing is $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$, and the crop size for training and inference is $40 \times 56 \times 40$.

Table S7. DSC results of 18 U-Net models on the MSD Task04 Hippocampus dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8898 ± 0.0340	0.8899 ± 0.0343	0.8889 ± 0.0340	0.8901 ± 0.0342	0.8889 ± 0.0351	$\mathbf{0.8903 \pm 0.0337}$
W32	0.8886 ± 0.0344	0.8894 ± 0.0344	0.8892 ± 0.0340	0.8893 ± 0.0340	0.8890 ± 0.0336	0.8897 ± 0.0344
W64	0.8883 ± 0.0335	0.8886 ± 0.0341	0.8888 ± 0.0334	0.8886 ± 0.0337	0.8889 ± 0.0337	0.8892 ± 0.0342

Table S8. NSD results of 18 U-Net models on the MSD Task04 Hippocampus dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9735 ± 0.0293	0.9741 ± 0.0292	0.9733 ± 0.0292	$\mathbf{0.9752 \pm 0.0286}$	0.9731 ± 0.0305	0.9745 ± 0.0290
W32	0.9732 ± 0.0301	0.9742 ± 0.0292	0.9736 ± 0.0294	0.9741 ± 0.0286	0.9743 ± 0.0290	0.9747 ± 0.0285
W64	0.9742 ± 0.0286	0.9743 ± 0.0287	0.9747 ± 0.0288	0.9750 ± 0.0282	0.9744 ± 0.0288	0.9748 ± 0.0286

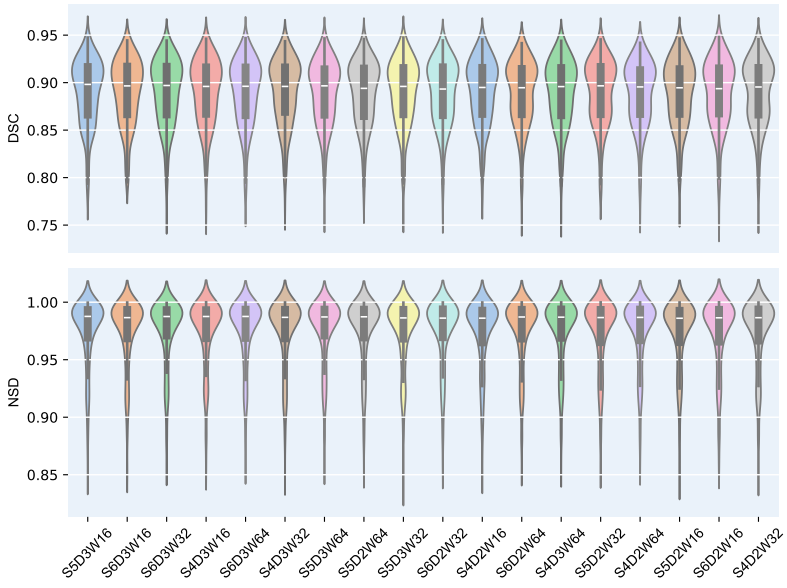

Figure S4. Violin plots of DSC and NSD for MSD Task04 Hippocampus dataset.

S1.5 MSD Task05 Prostate

The MSD Task05 Prostate[1] dataset comprises 32 publicly labeled cases, aimed at segmenting the Peripheral Zone and Transition Zone of the prostate in T2 and ADC MR images. We employed a 5-fold cross-validation for our experimental results. The target spacing is 0.63 mm × 0.63 mm × 3.6 mm, and the crop size for training and inference is 20 × 320 × 256.

Table S9. DSC results of 18 U-Net models on the MSD Task05 Prostate dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7076 ± 0.1524	0.7223 ± 0.1536	0.7403 ± 0.1550	0.7420 ± 0.1552	0.7475 ± 0.1461	0.7508 ± 0.1512
W32	0.7269 ± 0.1421	0.7225 ± 0.1538	0.7468 ± 0.1466	0.7539 ± 0.1508	0.7463 ± 0.1498	0.7525 ± 0.1543
W64	0.7341 ± 0.1373	0.7372 ± 0.1481	0.7555 ± 0.1449	0.7503 ± 0.1489	**0.7577 ± 0.1451**	0.7563 ± 0.1508

Table S10. NSD results of 18 U-Net models on the MSD Task05 Prostate dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.5586 ± 0.1482	0.5772 ± 0.1430	0.5956 ± 0.1520	0.5976 ± 0.1474	0.6014 ± 0.1427	0.6130 ± 0.1431
W32	0.5837 ± 0.1398	0.5837 ± 0.1489	0.6105 ± 0.1420	0.6212 ± 0.1476	0.6046 ± 0.1477	0.6213 ± 0.1458
W64	0.5956 ± 0.1403	0.6008 ± 0.1386	0.6256 ± 0.1434	0.6160 ± 0.1460	**0.6282 ± 0.1426**	0.6273 ± 0.1451

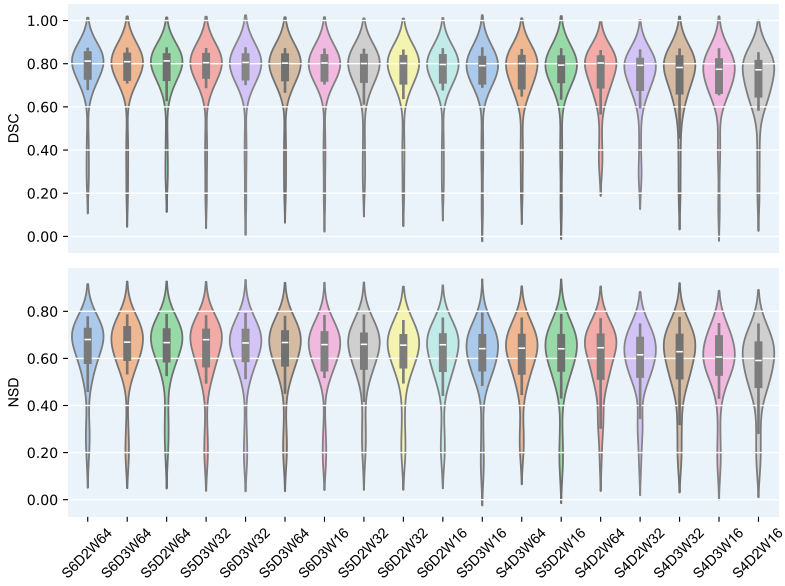

Figure S5. Violin plots of DSC and NSD for MSD Task05 Prostate dataset.

S1.6 MSD Task06 Lung

The MSD Task06 Lung[1] dataset comprises 63 publicly labeled cases, aimed at segmenting the lung tumor in CT images. We employed a 5-fold cross-validation for our experimental results. The target spacing is 0.79 mm × 0.79 mm × 1.2 mm, and the crop size for training and inference is 80 × 192 × 160.

Table S11. DSC results of 18 U-Net models on the MSD Task06 Lung dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6625 ± 0.2581	0.6979 ± 0.2230	0.6902 ± 0.2572	0.6935 ± 0.2232	0.6949 ± 0.2363	0.6956 ± 0.2071
W32	0.6819 ± 0.2377	0.6695 ± 0.2451	0.6772 ± 0.2551	0.6971 ± 0.2294	0.7033 ± 0.2341	0.6790 ± 0.2398
W64	0.6578 ± 0.2613	0.6593 ± 0.2459	0.6804 ± 0.2588	**0.7184 ± 0.2051**	0.6980 ± 0.2331	0.7017 ± 0.2129

Table S12. NSD results of 18 U-Net models on the MSD Task06 Lung dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.5364 ± 0.2725	0.5731 ± 0.2608	0.5659 ± 0.2795	0.5642 ± 0.2449	0.5724 ± 0.2559	0.5687 ± 0.2296
W32	0.5486 ± 0.2645	0.5355 ± 0.2583	0.5533 ± 0.2744	0.5750 ± 0.2529	0.5780 ± 0.2679	0.5537 ± 0.2523
W64	0.5384 ± 0.2772	0.5364 ± 0.2648	0.5490 ± 0.2775	**0.5910 ± 0.2449**	0.5689 ± 0.2631	0.5740 ± 0.2364

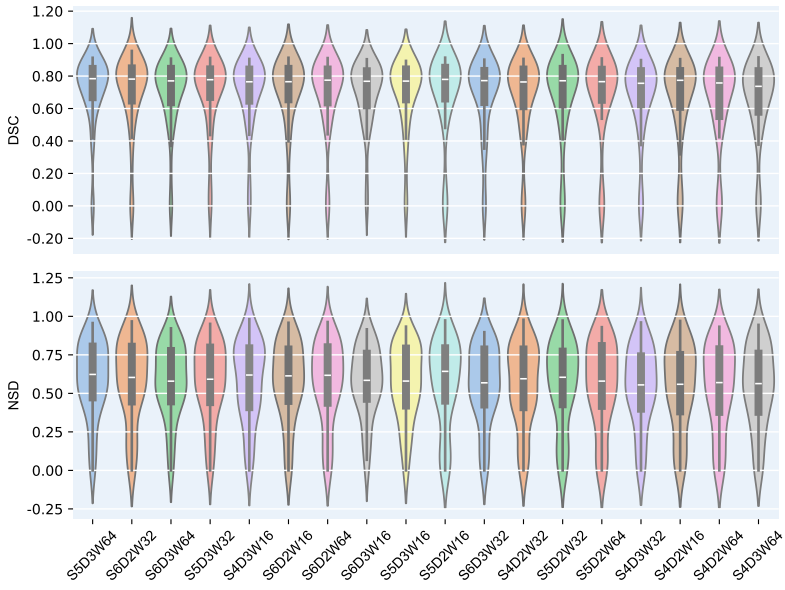

Figure S6. Violin plots of DSC and NSD for MSD Task06 Lung dataset.

S1.7 MSD Task07 Pancreas

The MSD Task07 Pancreas[1] dataset comprises 281 publicly labeled cases, aimed at segmenting the pancreas and pancreatic tumor in CT images. We employed a 5-fold cross-validation for our experimental results. The target spacing is 0.8 mm × 0.8 mm × 2.5 mm, and the crop size for training and inference is 40 × 224 × 224.

Table S13. DSC results of 18 U-Net models on the MSD Task07 Pancreas dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6635 ± 0.1686	0.6698 ± 0.1716	0.6769 ± 0.1732	0.6814 ± 0.1676	0.6657 ± 0.1735	0.6729 ± 0.1743
W32	0.6611 ± 0.1730	0.6721 ± 0.1715	0.6666 ± 0.1762	**0.6840 ± 0.1695**	0.6724 ± 0.1710	0.6824 ± 0.1734
W64	0.6635 ± 0.1749	0.6714 ± 0.1747	0.6632 ± 0.1756	0.6788 ± 0.1732	0.6689 ± 0.1781	0.6772 ± 0.1721

Table S14. NSD results of 18 U-Net models on the MSD Task07 Pancreas dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.5240 ± 0.1717	0.5338 ± 0.1747	0.5384 ± 0.1769	0.5460 ± 0.1705	0.5330 ± 0.1739	0.5399 ± 0.1767
W32	0.5261 ± 0.1749	0.5390 ± 0.1770	0.5321 ± 0.1805	0.5475 ± 0.1743	0.5365 ± 0.1738	**0.5489 ± 0.1780**
W64	0.5309 ± 0.1770	0.5389 ± 0.1798	0.5301 ± 0.1787	0.5471 ± 0.1787	0.5362 ± 0.1800	0.5408 ± 0.1756

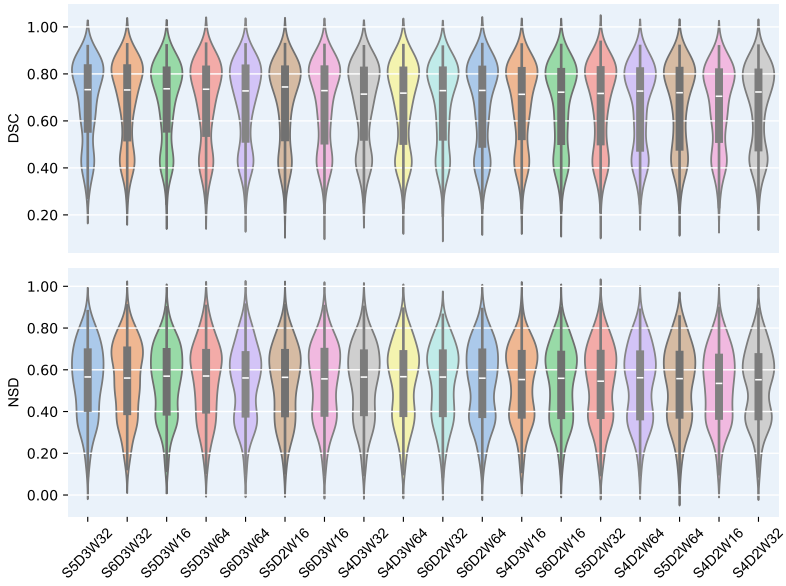

Figure S7. Violin plots of DSC and NSD for MSD Task07 Pancreas dataset.

S1.8 MSD Task08 Hepatic Vessels

The MSD Task08 Hepatic Vessels[1] dataset comprises 303 publicly labeled cases, aimed at segmenting the hepatic vessels and tumor in CT images. We employed a 5-fold cross-validation for our experimental results. The target spacing is 0.8 mm × 0.8 mm × 5.0 mm, and the crop size for training and inference is 60 × 192 × 192.

Table S15. DSC results of 18 U-Net models on the MSD Task08 Hepatic Vessels dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6745 ± 0.1370	0.6798 ± 0.1336	0.6814 ± 0.1330	0.6816 ± 0.1337	0.6804 ± 0.1316	0.6772 ± 0.1333
W32	0.6769 ± 0.1347	0.6852 ± 0.1321	0.6816 ± 0.1342	0.6847 ± 0.1306	0.6828 ± 0.1324	0.6840 ± 0.1282
W64	0.6758 ± 0.1360	0.6852 ± 0.1347	0.6857 ± 0.1309	**0.6869 ± 0.1290**	0.6842 ± 0.1319	0.6860 ± 0.1294

Table S16. NSD results of 18 U-Net models on the MSD Task08 Hepatic Vessels dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6174 ± 0.1264	0.6250 ± 0.1235	0.6267 ± 0.1248	0.6273 ± 0.1236	0.6235 ± 0.1216	0.6218 ± 0.1244
W32	0.6211 ± 0.1261	0.6304 ± 0.1229	0.6276 ± 0.1280	0.6316 ± 0.1247	0.6293 ± 0.1268	0.6293 ± 0.1229
W64	0.6195 ± 0.1290	0.6324 ± 0.1264	0.6313 ± 0.1277	**0.6346 ± 0.1242**	0.6315 ± 0.1279	0.6335 ± 0.1250

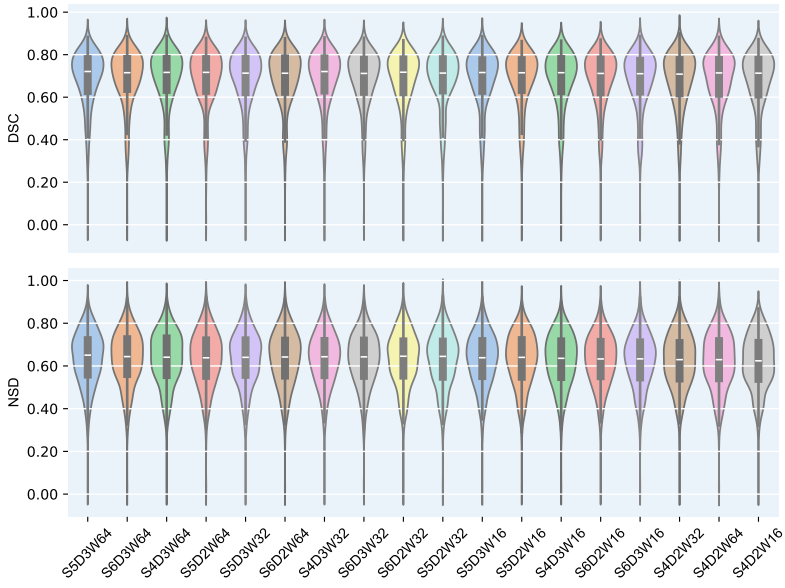

Figure S8. Violin plots of DSC and NSD for MSD Task08 Hepatic Vessels dataset.

S1.9 MSD Task09 Spleen

The MSD Task09 Spleen[1] dataset comprises 61 publicly labeled cases, aimed at segmenting the spleen in CT images. We employed a 5-fold cross validation for our experimental results. The target spacing is 0.79 mm × 0.79 mm × 5.0 mm, and the crop size for training and inference is 60 × 192 × 160.

Table S17. DSC results of 18 U-Net models on the MSD Task09 Spleen dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9534 ± 0.0418	0.9548 ± 0.0517	0.9641 ± 0.0382	0.9664 ± 0.0317	0.9461 ± 0.0990	0.9430 ± 0.0835
W32	0.9593 ± 0.0298	0.9600 ± 0.0441	0.9668 ± 0.0256	0.9636 ± 0.0368	0.9630 ± 0.0447	0.9571 ± 0.0527
W64	0.9609 ± 0.0349	0.9604 ± 0.0361	0.9648 ± 0.0351	0.9570 ± 0.0988	**0.9675 ± 0.0235**	0.9566 ± 0.0540

Table S18. NSD results of 18 U-Net models on the MSD Task09 Spleen dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9000 ± 0.0811	0.9090 ± 0.0908	0.9245 ± 0.0772	0.9292 ± 0.0686	0.9096 ± 0.1188	0.8980 ± 0.1000
W32	0.9107 ± 0.0677	0.9200 ± 0.0777	0.9294 ± 0.0595	0.9283 ± 0.0664	0.9261 ± 0.0812	0.9182 ± 0.0843
W64	0.9141 ± 0.0747	0.9199 ± 0.0699	0.9291 ± 0.0713	0.9259 ± 0.1141	**0.9313 ± 0.0569**	0.9169 ± 0.0867

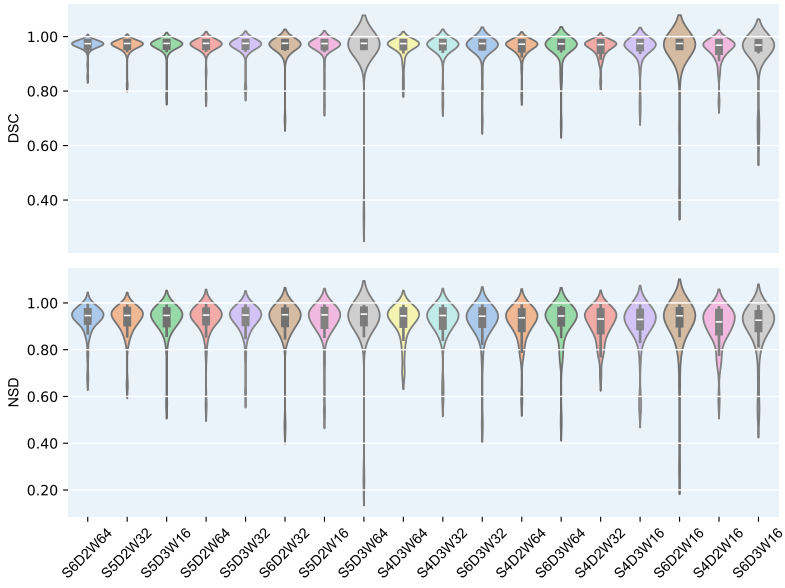

Figure S9. Violin plots of DSC and NSD for MSD Task09 Spleen dataset.

S1.10 MSD Task10 Colon

The MSD Task10 Colon[1] dataset comprises 126 publicly labeled cases, aimed at segmenting the colon tumors in CT images. We employed a 5-fold cross validation for our experimental results. The target spacing is 0.78 mm × 0.78 mm × 5.0 mm, and the crop size for training and inference is 56 × 192 × 160.

Table S19. DSC results of 18 U-Net models on the MSD Task10 Colon dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.4212 ± 0.3692	0.4365 ± 0.3772	0.4631 ± 0.3691	0.4569 ± 0.3551	0.4541 ± 0.3779	0.4472 ± 0.3661
W32	0.4229 ± 0.3724	0.4542 ± 0.3700	0.4675 ± 0.3770	0.4431 ± 0.3700	0.4585 ± 0.3728	0.4657 ± 0.3679
W64	0.4373 ± 0.3716	0.4596 ± 0.3796	**0.4677 ± 0.3726**	0.4657 ± 0.3694	0.4592 ± 0.3849	0.4503 ± 0.3774

Table S20. NSD results of 18 U-Net models on the MSD Task10 Colon dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.3836 ± 0.3335	0.3971 ± 0.3450	0.4239 ± 0.3355	0.4141 ± 0.3216	0.4149 ± 0.3453	0.4076 ± 0.3311
W32	0.3849 ± 0.3382	0.4149 ± 0.3385	0.4287 ± 0.3452	0.4069 ± 0.3394	0.4227 ± 0.3415	0.4243 ± 0.3383
W64	0.3996 ± 0.3406	0.4208 ± 0.3500	**0.4324 ± 0.3443**	0.4300 ± 0.3424	0.4232 ± 0.3552	0.4142 ± 0.3476

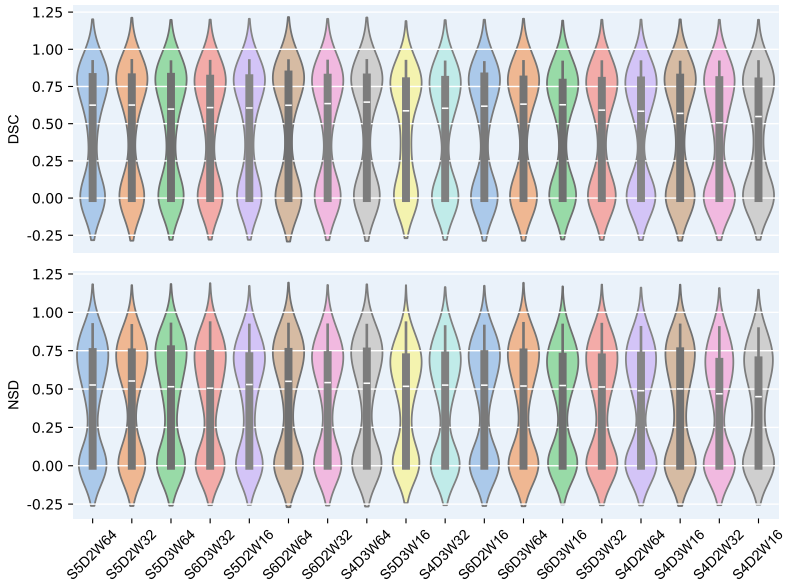

Figure S10. Violin plots of DSC and NSD for MSD Task10 Colon dataset.

S1.11 TotalSegmentator

The TotalSegmentator[2] dataset comprises 104 full-body anatomical structures. We trained our models on the official training set of 1082 samples and evaluated them on the official validation set of 57 samples. The target spacing is 1.5mm × 1.5mm × 1.5mm, with a crop size of 128 × 128 × 128 for both training and inference.

Table S21. DSC results of 18 U-Net models on the dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7570 ± 0.1108	0.7938 ± 0.1110	0.8118 ± 0.1094	0.8214 ± 0.1259	0.8202 ± 0.1102	0.8163 ± 0.1173
W32	0.8329 ± 0.1129	0.8499 ± 0.1013	0.8457 ± 0.1216	0.8584 ± 0.1151	0.8532 ± 0.1276	0.8591 ± 0.1172
W64	0.8561 ± 0.1139	0.8655 ± 0.1266	0.8693 ± 0.1182	**0.8714 ± 0.1141**	0.8679 ± 0.1292	0.8710 ± 0.1235

Table S22. NSD results of 18 U-Net models on the TotalSegmentator dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6313 ± 0.1174	0.6879 ± 0.1192	0.7030 ± 0.1217	0.7235 ± 0.1331	0.7184 ± 0.1210	0.7148 ± 0.1310
W32	0.7375 ± 0.1217	0.7626 ± 0.1175	0.7625 ± 0.1336	0.7776 ± 0.1272	0.7732 ± 0.1359	0.7788 ± 0.1283
W64	0.7780 ± 0.1231	0.7947 ± 0.1326	0.7996 ± 0.1287	0.8025 ± 0.1238	0.8009 ± 0.1364	**0.8042 ± 0.1326**

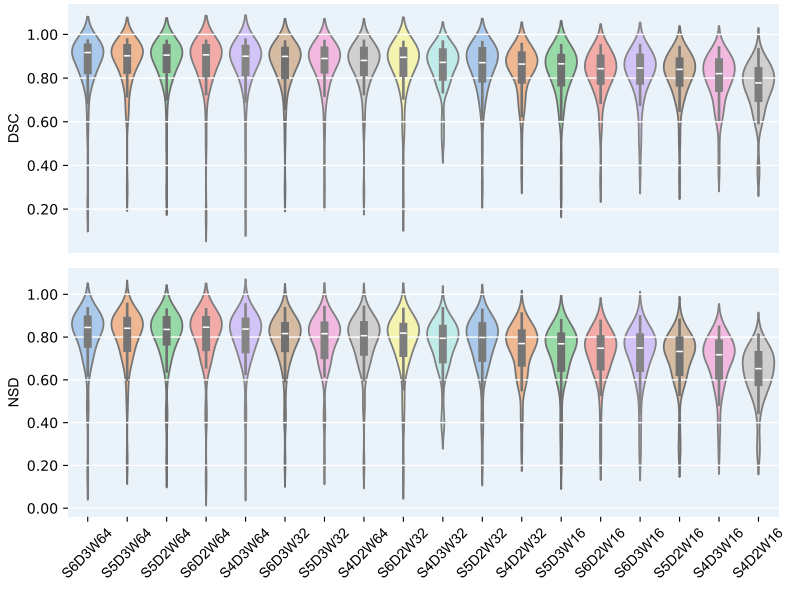

Figure S11. Violin plots of DSC and NSD for TotalSegmentator dataset.

S1.12 TotalSegmentator MRI

The TotalSegmentator MRI[3] dataset includes 298 MR images and provides segmentation annotations for 56 commonly used anatomical structures. We used the official split of 268 cases for training and 30 cases for validation. The target spacing is 1.03mm $\times$ 1.18mm $\times$ 3.0mm, with a crop size of 160 $\times$ 160 $\times$ 48 for both training and inference.

Table S23. DSC results of 18 U-Net models on the TotalSegmentator MRI dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.3890 ± 0.2123	0.4184 ± 0.2156	0.5502 ± 0.2357	0.5413 ± 0.2352	0.5671 ± 0.2516	0.5445 ± 0.2552
W32	0.4450 ± 0.2394	0.5594 ± 0.2663	0.5348 ± 0.2597	0.5862 ± 0.2403	0.5624 ± 0.2574	0.5726 ± 0.2552
W64	0.4805 ± 0.2519	0.5398 ± 0.2621	0.5469 ± 0.2556	0.5478 ± 0.2681	$\mathbf{0.5923 \pm 0.2569}$	0.5508 ± 0.2571

Table S24. NSD results of 18 U-Net models on the TotalSegmentator MRI dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.2530 ± 0.1641	0.2882 ± 0.1691	0.4130 ± 0.2234	0.4084 ± 0.2009	0.4437 ± 0.2330	0.4201 ± 0.2312
W32	0.3221 ± 0.1918	0.4367 ± 0.2517	0.4161 ± 0.2289	0.4673 ± 0.2316	0.4429 ± 0.2359	0.4543 ± 0.2404
W64	0.3645 ± 0.2177	0.4234 ± 0.2347	0.4346 ± 0.2244	0.4342 ± 0.2423	$\mathbf{0.4804 \pm 0.2403}$	0.4400 ± 0.2215

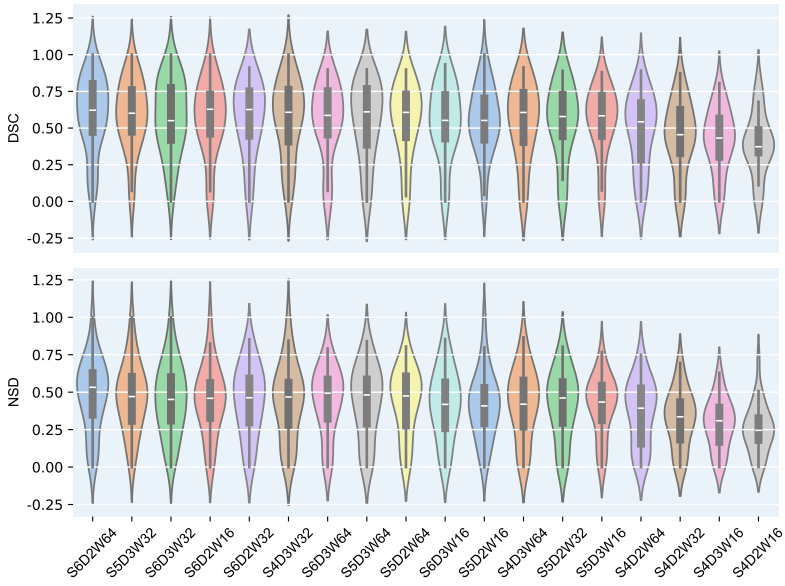

Figure S12. Violin plots of DSC and NSD for TotalSegmentator MRI dataset.

S1.13 AutoPET

The AutoPET[4] is a large-scale PET/CT image dataset focused on whole-body tumor segmentation, comprising 1,014 images from 900 patients, with 501 images containing tumor annotations. We randomly split these annotated images into 398 for training and 103 for testing. The target spacing is 2mm × 2mm × 3mm, with a crop size of 128 × 128 × 128 for both training and inference.

Table S25. DSC results of 18 U-Net models on the AutoPET dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7244 ± 0.2257	0.6746 ± 0.2629	0.6754 ± 0.2659	0.6600 ± 0.2584	0.6664 ± 0.2699	0.6616 ± 0.2663
W32	**0.7422 ± 0.2123**	0.6981 ± 0.2511	0.6964 ± 0.2477	0.6612 ± 0.2730	0.6922 ± 0.2522	0.6861 ± 0.2569
W64	0.7347 ± 0.2204	0.7151 ± 0.2390	0.7057 ± 0.2511	0.6702 ± 0.2612	0.6972 ± 0.2521	0.6932 ± 0.2508

Table S26. NSD results of 18 U-Net models on the AutoPET dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.5706 ± 0.2115	0.5181 ± 0.2249	0.5178 ± 0.2228	0.5056 ± 0.2173	0.5073 ± 0.2239	0.5041 ± 0.2248
W32	**0.5922 ± 0.2084**	0.5452 ± 0.2242	0.5410 ± 0.2164	0.5073 ± 0.2314	0.5373 ± 0.2163	0.5290 ± 0.2193
W64	0.5846 ± 0.2073	0.5623 ± 0.2189	0.5582 ± 0.2206	0.5117 ± 0.2220	0.5397 ± 0.2212	0.5333 ± 0.2209

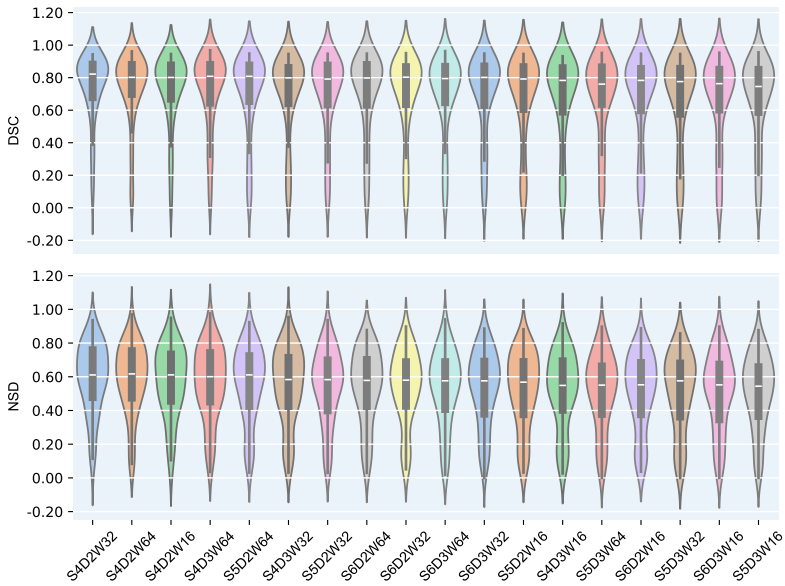

Figure S13. Violin plots of DSC and NSD for AutoPET dataset.

S1.14 CT-ORG

The CT-ORG[5] is a CT multi-organ segmentation dataset, providing annotations for six organs: lungs, bones, liver, kidneys, bladder, and brain, and includes 140 CT scans. We used the official split of 119 cases for training and 21 cases for testing. The target spacing is 0.78mm $\times$ 0.78mm $\times$ 1.0mm, with a crop size of 128 $\times$ 128 $\times$ 128 for both training and inference.

Table S27. DSC results of 18 U-Net models on the CT-ORG dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9242 ± 0.0484	0.9267 ± 0.0408	0.9280 ± 0.0467	0.9363 ± 0.0218	0.9280 ± 0.0399	0.9354 ± 0.0229
W32	0.9285 ± 0.0468	0.9362 ± 0.0227	0.9352 ± 0.0225	0.9299 ± 0.0458	$\mathbf{0.9378 \pm 0.0226}$	0.9370 ± 0.0217
W64	0.9204 ± 0.0632	0.9316 ± 0.0464	0.9299 ± 0.0475	0.9312 ± 0.0462	0.9315 ± 0.0469	0.9320 ± 0.0462

Table S28. NSD results of 18 U-Net models on the CT-ORG dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7861 ± 0.0618	0.7976 ± 0.0667	0.8011 ± 0.0571	0.8085 ± 0.0504	0.8044 ± 0.0593	0.8101 ± 0.0501
W32	0.7969 ± 0.0573	0.8122 ± 0.0489	0.8108 ± 0.0460	0.8081 ± 0.0516	$\mathbf{0.8152 \pm 0.0451}$	0.8136 ± 0.0441
W64	0.7923 ± 0.0667	0.8134 ± 0.0513	0.8116 ± 0.0528	0.8132 ± 0.0532	0.8130 ± 0.0510	0.8127 ± 0.0503

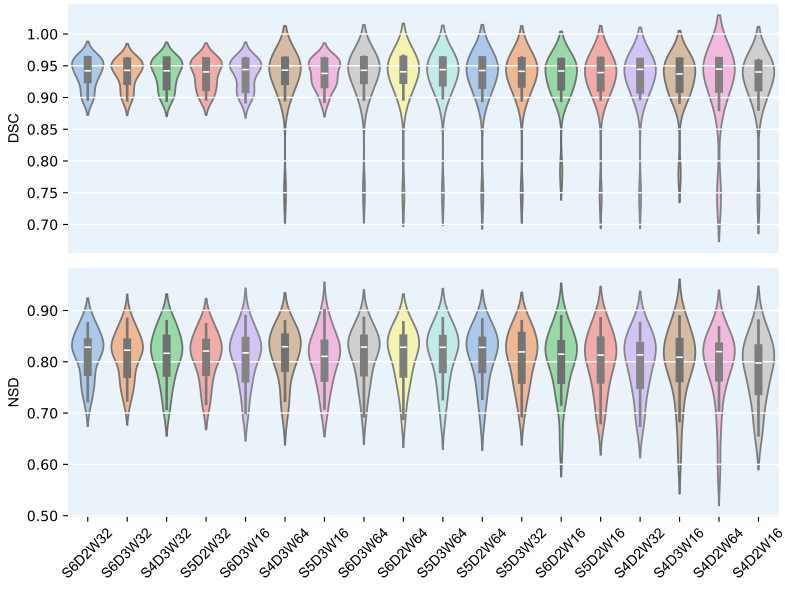

Figure S14. Violin plots of DSC and NSD for CT-ORG dataset.

S1.15 SegRap Task1

The SegRap 2023[6] dataset provides precise segmentations of Organs-At-Risk (OAR) and Gross Tumor Volume (GTV) in non-contrast and contrast-enhanced CT images of nasopharyngeal carcinoma patients. Task 1 involves detailed segmentation of 45 types of OAR, including small or complex head and neck organs, with a total of 120 cases. We randomly split these into 96 for training and 24 for testing. The target spacing is 0.54mm $\times$ 0.54mm $\times$ 3.0mm, with a crop size of $192 \times 192 \times 24$ for both training and inference.

Table S29. DSC results of 18 U-Net models on the SegRap2023 Task1 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.5784 ± 0.0861	0.4757 ± 0.0504	0.6609 ± 0.1045	0.6548 ± 0.1010	0.6442 ± 0.0983	0.6394 ± 0.0852
W32	0.6346 ± 0.1307	0.6363 ± 0.1210	0.6762 ± 0.1252	0.7208 ± 0.1093	0.7277 ± 0.1043	0.7004 ± 0.1090
W64	0.6607 ± 0.1431	0.6649 ± 0.1413	0.7061 ± 0.1336	0.7192 ± 0.1211	$\mathbf{0.7500 \pm 0.1150}$	0.6918 ± 0.1143

Table S30. NSD results of 18 U-Net models on the SegRap2023 Task1 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.5768 ± 0.0635	0.4683 ± 0.0455	0.6612 ± 0.0961	0.6588 ± 0.0876	0.6391 ± 0.0965	0.6427 ± 0.0745
W32	0.6302 ± 0.1127	0.6398 ± 0.1105	0.6818 ± 0.1188	0.7341 ± 0.1017	0.7363 ± 0.0965	0.7044 ± 0.0992
W64	0.6550 ± 0.1383	0.6652 ± 0.1332	0.7104 ± 0.1328	0.7264 ± 0.1150	$\mathbf{0.7603 \pm 0.1149}$	0.7102 ± 0.1136

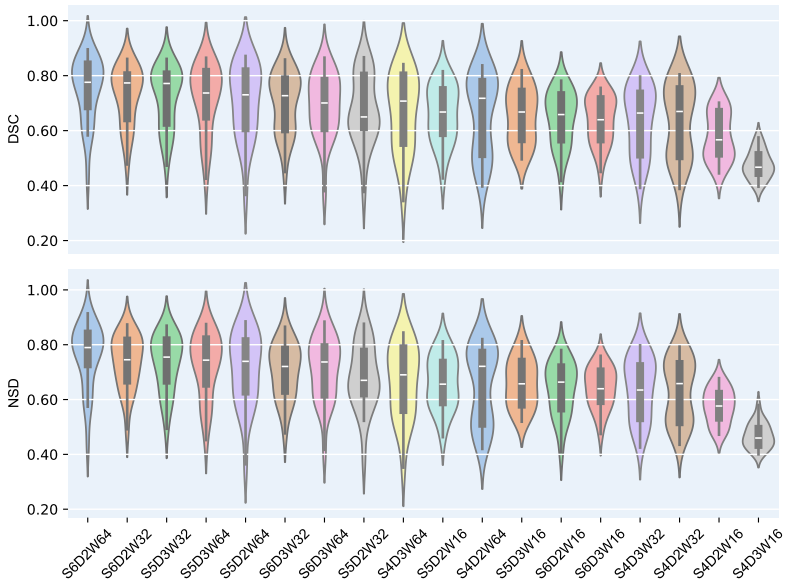

Figure S15. Violin plots of DSC and NSD for SegRap2023 Task1 dataset.

S1.16 SegRap Task2

The SegRap 2023 Task 2[6] dataset includes the same both non-contrast and contrast-enhanced CT images as Task1. Task 2 focuses on the segmentation of the Gross Tumor Volume (GTV), specifically targeting nasopharyngeal carcinoma and its associated lymph nodes, with a total of 120 cases. We randomly split these into 96 for training and 24 for testing. The target spacing is 0.54mm $\times$ 0.54mm $\times$ 3.0mm, with a crop size of 192 $\times$ 192 $\times$ 24 for both training and inference.

Table S31. DSC results of 18 U-Net models on the SegRap2023 Task2 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7381 ± 0.0675	0.7491 ± 0.0656	0.7493 ± 0.0636	0.7526 ± 0.0724	0.7495 ± 0.0686	0.7543 ± 0.0665
W32	0.7478 ± 0.0661	0.7530 ± 0.0672	0.7483 ± 0.0680	0.7510 ± 0.0739	0.7486 ± 0.0683	0.7528 ± 0.0679
W64	0.7478 ± 0.0670	0.7493 ± 0.0696	0.7480 ± 0.0704	$\mathbf{0.7574 \pm 0.0710}$	0.7444 ± 0.0723	0.7533 ± 0.0709

Table S32. NSD results of 18 U-Net models on the SegRap2023 Task2 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.4938 ± 0.0778	0.5078 ± 0.0796	0.5102 ± 0.0785	0.5137 ± 0.0772	0.5124 ± 0.0814	0.5142 ± 0.0775
W32	0.5085 ± 0.0790	0.5203 ± 0.0838	0.5145 ± 0.0832	0.5159 ± 0.0826	0.5136 ± 0.0808	0.5163 ± 0.0821
W64	0.5140 ± 0.0824	0.5194 ± 0.0882	0.5166 ± 0.0921	$\mathbf{0.5226 \pm 0.0835}$	0.5100 ± 0.0873	0.5176 ± 0.0871

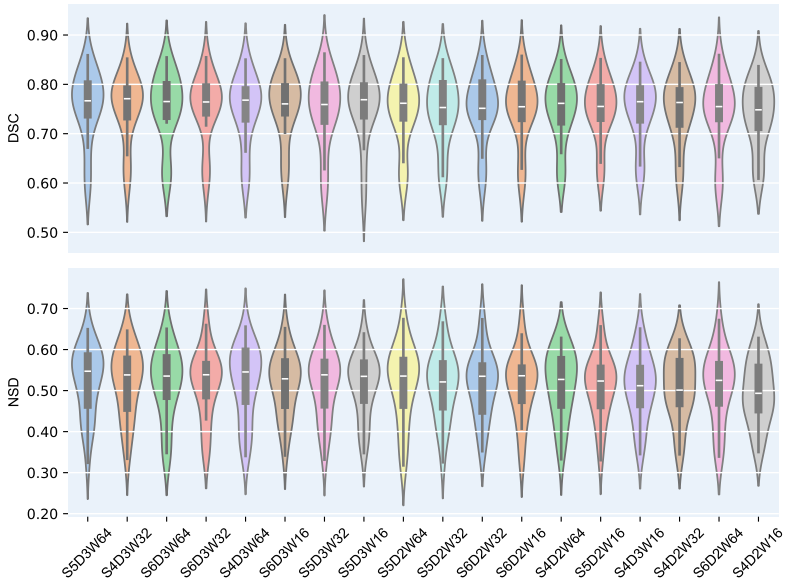

Figure S16. Violin plots of DSC and NSD for SegRap2023 Task2 dataset.

S1.17 HaN-Seg

The HaN-Seg[7] dataset is a publicly available head and neck (HaN) risk organ segmentation dataset, comprising CT and MR images from 42 patients. It features annotations for 30 organs at risk (OARs). We randomly divided the dataset into 33 images for training and 9 for testing, using only the CT images for both phases. The target spacing is 0.64mm $\times$ 0.64mm $\times$ 2.0mm, with a crop size of $192 \times 224 \times 40$ for both training and inference.

Table S33. DSC results of 18 U-Net models on the HaN-Seg dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.5826 ± 0.0625	0.5837 ± 0.0718	0.5867 ± 0.0909	0.6205 ± 0.0773	0.5923 ± 0.0666	0.6203 ± 0.0597
W32	0.5787 ± 0.0626	0.6112 ± 0.0980	0.5843 ± 0.0891	0.6311 ± 0.0619	0.6043 ± 0.0684	0.6209 ± 0.0502
W64	0.5720 ± 0.0780	0.5794 ± 0.0661	0.6061 ± 0.0607	$\mathbf{0.6443 \pm 0.0805}$	0.5869 ± 0.0703	0.6216 ± 0.0717

Table S34. NSD results of 18 U-Net models on the HaN-Seg dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.5946 ± 0.0623	0.5918 ± 0.0828	0.5929 ± 0.0932	0.6347 ± 0.0748	0.6032 ± 0.0639	0.6354 ± 0.0640
W32	0.5940 ± 0.0706	0.6271 ± 0.1093	0.6014 ± 0.0855	0.6485 ± 0.0682	0.6181 ± 0.0753	0.6369 ± 0.0482
W64	0.5832 ± 0.0792	0.5969 ± 0.0805	0.6262 ± 0.0698	$\mathbf{0.6633 \pm 0.0853}$	0.6058 ± 0.0702	0.6408 ± 0.0684

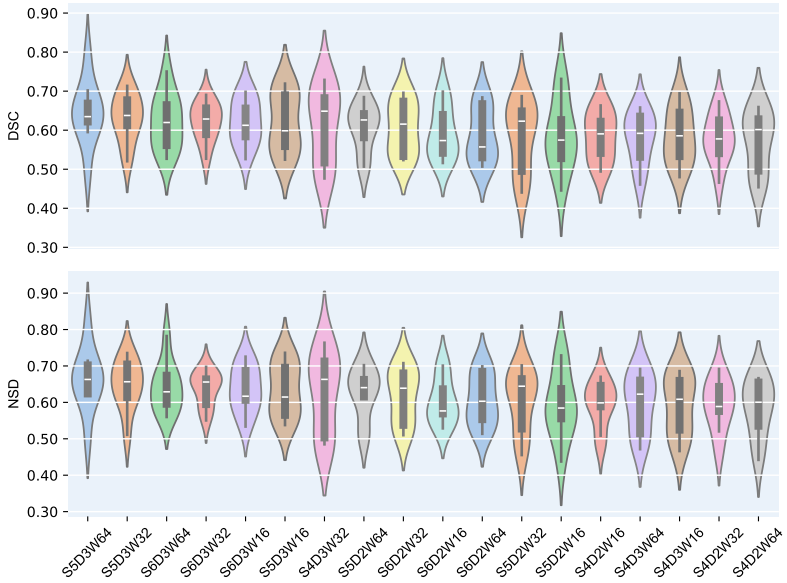

Figure S17. Violin plots of DSC and NSD for HaN-Seg dataset.

S1.18 FeTA 2022

The FeTA 2022[8] dataset from the Fetal Tissue Annotation and Segmentation Challenge includes T2-weighted fetal brain MRI reconstruction data from 120 cases, with manual segmentation labels for seven different brain tissues. We randomly divided the dataset into 96 training cases and 24 testing cases. The target spacing is 1mm $\times$ 1mm $\times$ 1mm, with a crop size of $128 \times 128 \times 128$ for both training and inference.

Table S35. DSC results of 18 U-Net models on the FeTA 2022 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8467 ± 0.0622	0.8496 ± 0.0582	$\mathbf{0.8547 \pm 0.0527}$	0.8507 ± 0.0592	0.8515 ± 0.0566	0.8485 ± 0.0664
W32	0.8490 ± 0.0583	0.8491 ± 0.0592	0.8482 ± 0.0606	0.8479 ± 0.0617	0.8478 ± 0.0622	0.8491 ± 0.0609
W64	0.8504 ± 0.0577	0.8513 ± 0.0544	0.8491 ± 0.0574	0.8468 ± 0.0608	0.8492 ± 0.0576	0.8492 ± 0.0583

Table S36. NSD results of 18 U-Net models on the FeTA 2022 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9050 ± 0.0784	0.9095 ± 0.0707	$\mathbf{0.9140 \pm 0.0680}$	0.9114 ± 0.0726	0.9118 ± 0.0703	0.9068 ± 0.0837
W32	0.9110 ± 0.0690	0.9109 ± 0.0705	0.9106 ± 0.0722	0.9097 ± 0.0724	0.9118 ± 0.0720	0.9111 ± 0.0729
W64	0.9138 ± 0.0676	0.9142 ± 0.0634	0.9128 ± 0.0674	0.9108 ± 0.0693	0.9127 ± 0.0680	0.9126 ± 0.0678

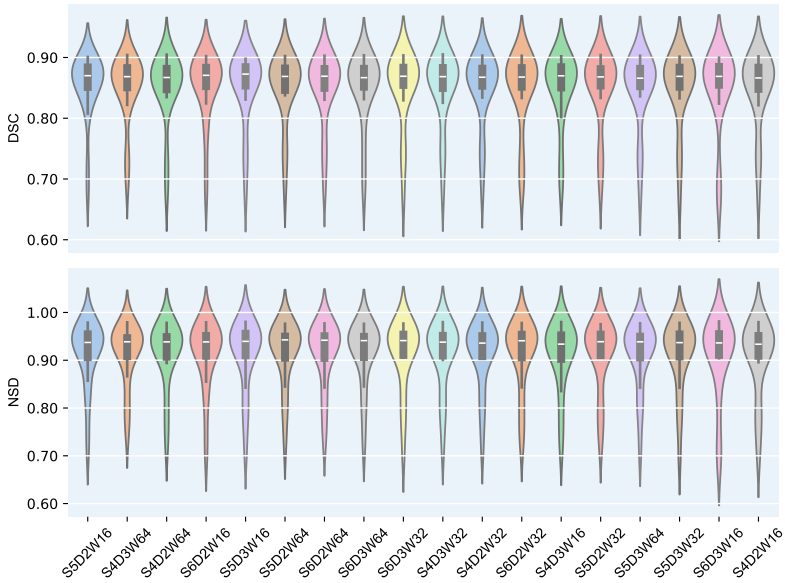

Figure S18. Violin plots of DSC and NSD for FeTA 2022 Task2 dataset.

S1.19 BraTS21

The BraTS21[9] is a large multimodal MR brain tumor segmentation dataset, with a publicly available annotated training set comprising 1,251 cases. It includes four image modalities: T1, T1Gd, T2, and T2-FLAIR. Annotations cover the enhanced tumor (ET), peritumoral edema (ED), and necrotic tumor core (NCR). The dataset is randomly divided into 1,000 cases for training and 251 for testing. The target spacing is 1mm × 1mm × 1mm, with a crop size of 128 × 128 × 128 for both training and inference.

Table S37. DSC results of 18 U-Net models on the BraTS21 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9084 ± 0.1129	0.9076 ± 0.1111	0.9061 ± 0.1209	0.9075 ± 0.1169	0.9065 ± 0.1217	0.9060 ± 0.1188
W32	0.9070 ± 0.1191	0.9071 ± 0.1168	0.9080 ± 0.1184	0.9062 ± 0.1208	0.9076 ± 0.1191	0.9087 ± 0.1180
W64	0.9070 ± 0.1233	0.9058 ± 0.1264	0.9063 ± 0.1266	**0.9090 ± 0.1219**	0.9072 ± 0.1224	0.9067 ± 0.1239

Table S38. NSD results of 18 U-Net models on the BraTS21 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8989 ± 0.1355	0.8984 ± 0.1338	0.8997 ± 0.1393	0.8989 ± 0.1388	0.8986 ± 0.1422	0.8987 ± 0.1381
W32	0.8993 ± 0.1400	0.8999 ± 0.1369	0.9004 ± 0.1404	0.8991 ± 0.1402	0.9001 ± 0.1411	0.9012 ± 0.1398
W64	0.9003 ± 0.1422	0.8981 ± 0.1454	0.9003 ± 0.1426	**0.9019 ± 0.1413**	0.9006 ± 0.1408	0.8993 ± 0.1442

Figure S19. Violin plots of DSC and NSD for BraTS21 dataset.

S1.20 CAS 2023

The CAS 2023[10] dataset is a segmentation dataset for MR modality. It consists of a public training set with 100 cases, which includes annotations for the cerebral arteries. We randomly split the dataset into 80 cases for training and 20 cases for testing. The target spacing is 1mm × 1mm × 1mm, with a crop size of 128 × 128 × 128 for both training and inference.

Table S39. DSC results of 18 U-Net models on the CAS 2023 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8672 ± 0.0578	0.8663 ± 0.0559	0.8681 ± 0.0550	0.8652 ± 0.0540	0.8667 ± 0.0558	0.8639 ± 0.0536
W32	**0.8680 ± 0.0547**	0.8670 ± 0.0552	0.8654 ± 0.0558	0.8670 ± 0.0538	0.8657 ± 0.0576	0.8664 ± 0.0543
W64	0.8670 ± 0.0558	0.8674 ± 0.0534	0.8679 ± 0.0555	0.8671 ± 0.0545	0.8677 ± 0.0552	0.8666 ± 0.0551

Table S40. NSD results of 18 U-Net models on the CAS 2023 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9095 ± 0.0431	0.9086 ± 0.0423	0.9104 ± 0.0412	0.9072 ± 0.0406	0.9099 ± 0.0416	0.9079 ± 0.0413
W32	**0.9117 ± 0.0415**	0.9099 ± 0.0410	0.9090 ± 0.0407	0.9097 ± 0.0409	0.9089 ± 0.0409	0.9090 ± 0.0407
W64	0.9102 ± 0.0420	0.9098 ± 0.0409	0.9114 ± 0.0400	0.9109 ± 0.0415	0.9115 ± 0.0414	0.9103 ± 0.0422

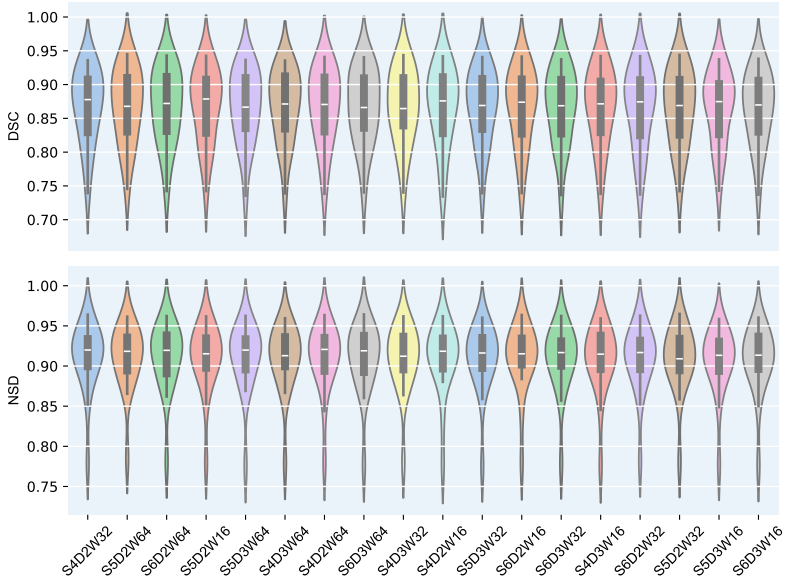

Figure S20. Violin plots of DSC and NSD for CAS 2023 dataset.

S1.21 INSTANCE 2022

The INSTANCE 2022[11] dataset focused on intracranial hemorrhage (ICH) segmentation using non-contrast head CT (NCCT) scans. The dataset contains 100 public labeled cases. We randomly split the dataset into 80 cases for training and 20 cases for testing. The target spacing is 0.45mm $\times$ 0.45mm $\times$ 5.0mm, with a crop size of $320 \times 320 \times 16$ for both training and inference.

Table S41. DSC results of 18 U-Net models on the INSTANCE 2022 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6849 ± 0.2346	0.6859 ± 0.2260	0.6918 ± 0.2267	0.6806 ± 0.2382	0.6854 ± 0.2431	0.6935 ± 0.2247
W32	0.6826 ± 0.2417	0.6799 ± 0.2444	0.6689 ± 0.2552	$\mathbf{0.6960 \pm 0.2156}$	0.6852 ± 0.2356	0.6770 ± 0.2411
W64	0.6755 ± 0.2521	0.6866 ± 0.2234	0.6822 ± 0.2496	0.6906 ± 0.2356	0.6767 ± 0.2398	0.6695 ± 0.2406

Table S42. NSD results of 18 U-Net models on the INSTANCE 2022 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7055 ± 0.1981	0.7139 ± 0.1736	$\mathbf{0.7234 \pm 0.1829}$	0.7068 ± 0.1938	0.7130 ± 0.2047	0.7206 ± 0.1800
W32	0.7050 ± 0.2139	0.7099 ± 0.2076	0.6861 ± 0.2164	0.7131 ± 0.1770	0.7133 ± 0.1901	0.6974 ± 0.2003
W64	0.7023 ± 0.2185	0.7132 ± 0.1827	0.7164 ± 0.2176	0.7176 ± 0.2038	0.7085 ± 0.1917	0.6876 ± 0.2043

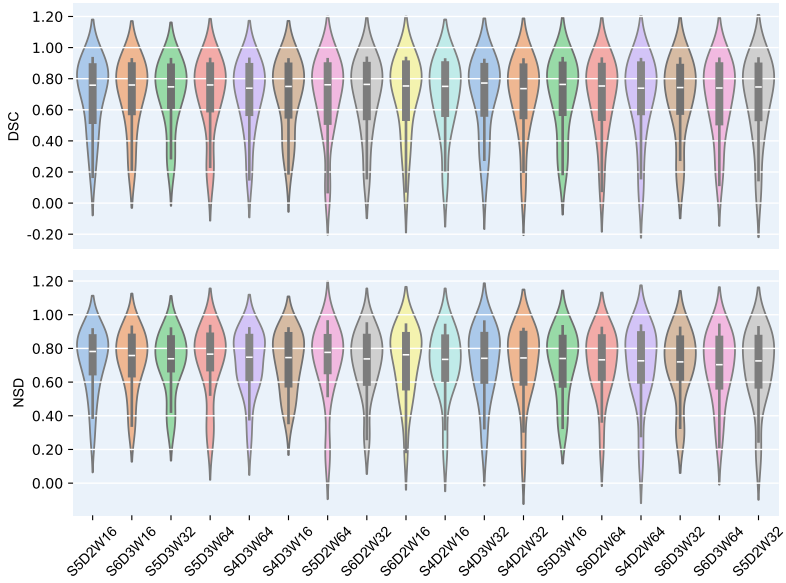

Figure S21. Violin plots of DSC and NSD for INSTANCE 2022 dataset.

S1.22 ISLES22

The ISLES22 (Ischemic Stroke Lesion Segmentation)[12] dataset aims to automatically segment acute to subacute ischemic stroke lesions using multimodal MRI images, including FLAIR, DWI, and ADC. The dataset is officially divided into a training set of 250 cases and a testing set of 150 cases. We randomly split the 250 publicly available training cases into 200 for training and 50 for testing. The target spacing is 2.0mm $\times$ 2.0mm $\times$ 2.0mm, with a crop size of $80 \times 96 \times 80$ for both training and inference.

Table S43. DSC results of 18 U-Net models on the ISLES22 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7916 ± 0.1581	0.7838 ± 0.1748	0.7852 ± 0.1764	0.7922 ± 0.1677	0.7820 ± 0.1764	0.7886 ± 0.1709
W32	0.7926 ± 0.1706	0.7957 ± 0.1523	0.7837 ± 0.1809	$\mathbf{0.7968 \pm 0.1553}$	0.7798 ± 0.1815	0.7955 ± 0.1524
W64	0.7888 ± 0.1698	0.7906 ± 0.1694	0.7798 ± 0.1839	0.7953 ± 0.1548	0.7848 ± 0.1733	0.7846 ± 0.1791

Table S44. NSD results of 18 U-Net models on the ISLES22 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7706 ± 0.1501	0.7630 ± 0.1654	0.7605 ± 0.1668	0.7707 ± 0.1603	0.7588 ± 0.1653	0.7695 ± 0.1594
W32	0.7739 ± 0.1586	0.7751 ± 0.1494	0.7609 ± 0.1696	$\mathbf{0.7746 \pm 0.1487}$	0.7566 ± 0.1712	0.7714 ± 0.1432
W64	0.7647 ± 0.1607	0.7685 ± 0.1589	0.7554 ± 0.1719	0.7725 ± 0.1465	0.7583 ± 0.1623	0.7616 ± 0.1677

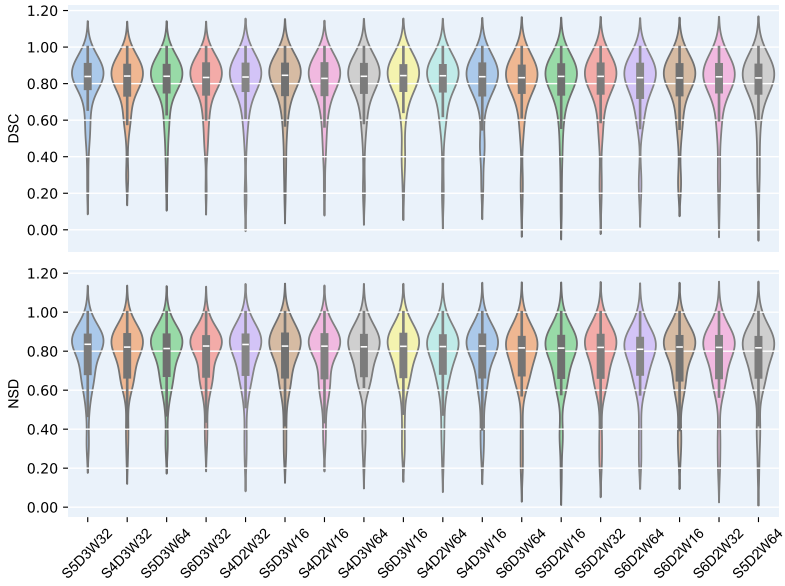

Figure S22. Violin plots of DSC and NSD for ISLES22 dataset.

S1.23 WMH

The WMH (White Matter Hyperintensity)[13] dataset is a multimodal brain MRI collection designed for the automatic segmentation of WMH. It includes 60 training cases from various institutions, featuring T1-weighted and FLAIR images, along with expert-annotated white matter hyperintensities. We randomly split the 60 publicly available training cases into 48 for training and 12 for testing. The target spacing is 1.0mm $\times$ 0.98mm $\times$ 3.0mm, with a crop size of $192 \times 224 \times 40$ for both training and inference.

Table S45. DSC results of 18 U-Net models on the WMH dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8272 ± 0.0848	0.8276 ± 0.0833	0.8265 ± 0.0852	0.8266 ± 0.0835	0.8229 ± 0.0917	0.8258 ± 0.0893
W32	0.8191 ± 0.0916	0.8279 ± 0.0843	0.8220 ± 0.0929	0.8207 ± 0.0928	0.8247 ± 0.0853	0.8237 ± 0.0919
W64	0.8219 ± 0.0913	0.8242 ± 0.0873	0.8172 ± 0.1031	0.8234 ± 0.0924	0.8175 ± 0.0936	$\mathbf{0.8290 \pm 0.0899}$

Table S46. NSD results of 18 U-Net models on the WMH dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9120 ± 0.0562	0.9137 ± 0.0550	0.9139 ± 0.0566	0.9124 ± 0.0581	0.9098 ± 0.0622	0.9128 ± 0.0602
W32	0.9086 ± 0.0600	0.9158 ± 0.0550	0.9090 ± 0.0641	0.9094 ± 0.0619	0.9128 ± 0.0565	0.9126 ± 0.0636
W64	0.9104 ± 0.0612	0.9108 ± 0.0598	0.9077 ± 0.0687	0.9122 ± 0.0607	0.9095 ± 0.0631	$\mathbf{0.9157 \pm 0.0598}$

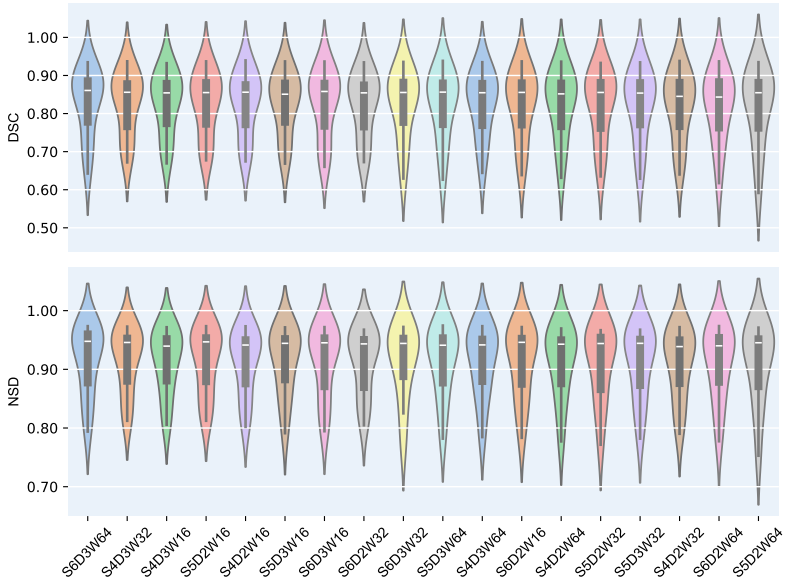

Figure S23. Violin plots of DSC and NSD for WMH dataset.

S1.24 ATM22

The ATM22 (Airway Tree Modeling Challenge 2022)[14] is a large-scale CT dataset for lung airway segmentation. This dataset includes 500 CT scans with complete lung airway annotations. The official release provides 300 cases for training, which we randomly divided into 240 cases for training and 60 cases for testing. The target spacing is 0.78mm $\times$ 0.78mm $\times$ 0.5mm, with a crop size of 112 $\times$ 128 $\times$ 160 for both training and inference.

Table S47. DSC results of 18 U-Net models on the ATM22 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9008 ± 0.1081	0.8999 ± 0.1081	0.9069 ± 0.1077	0.9034 ± 0.1091	0.9057 ± 0.1082	0.9031 ± 0.1092
W32	0.9042 ± 0.1079	0.9044 ± 0.1073	**0.9095 ± 0.1075**	0.9045 ± 0.1084	0.9040 ± 0.1088	0.9079 ± 0.1076
W64	0.9045 ± 0.1081	0.9052 ± 0.1076	0.9082 ± 0.1076	0.9088 ± 0.1080	0.9067 ± 0.1080	0.9067 ± 0.1078

Table S48. NSD results of 18 U-Net models on the ATM22 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9293 ± 0.1198	0.9294 ± 0.1199	0.9333 ± 0.1197	0.9315 ± 0.1207	0.9325 ± 0.1203	0.9302 ± 0.1204
W32	0.9327 ± 0.1187	0.9333 ± 0.1186	**0.9362 ± 0.1189**	0.9330 ± 0.1196	0.9335 ± 0.1198	0.9356 ± 0.1190
W64	0.9339 ± 0.1189	0.9353 ± 0.1184	0.9354 ± 0.1191	0.9363 ± 0.1192	0.9355 ± 0.1189	0.9353 ± 0.1192

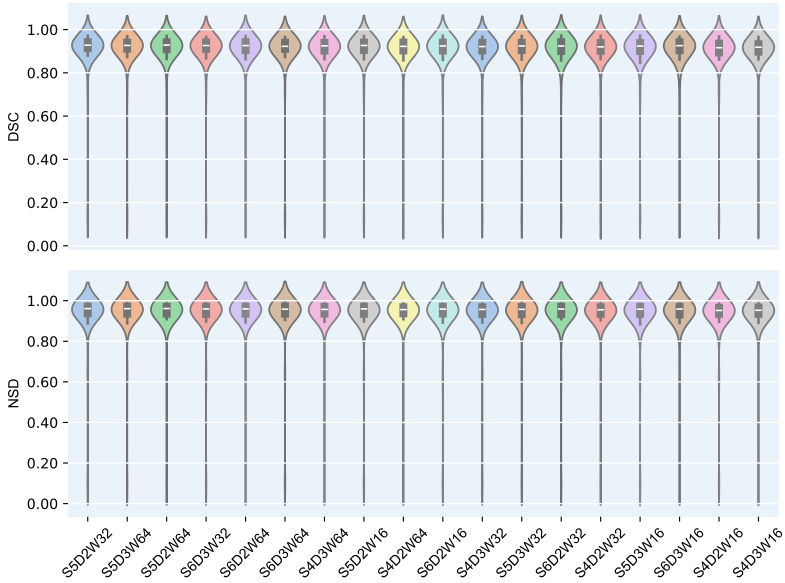

Figure S24. Violin plots of DSC and NSD for ATM22 dataset.

S1.25 Parse 2022

The Parse 2022 (Pulmonary Artery Segmentation)[15] is a CT dataset for pulmonary artery segmentation and a challenge in MICCAI 2022. This dataset includes 200 3D CT scans with finely annotated pulmonary arteries by 10 doctors. The official release provides 100 cases for training, which we randomly divided into 80 cases for training and 20 cases for testing. The target spacing is 0.65mm $\times$ 0.65mm $\times$ 1.0mm, with a crop size of $160 \times 160 \times 96$ for both training and inference.

Table S49. DSC results of 18 U-Net models on the Parse 2022 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8667 ± 0.0269	0.8680 ± 0.0275	0.8683 ± 0.0283	0.8681 ± 0.0283	0.8684 ± 0.0278	$\mathbf{0.8695 \pm 0.0285}$
W32	0.8638 ± 0.0292	0.8679 ± 0.0285	0.8689 ± 0.0283	0.8681 ± 0.0278	0.8679 ± 0.0294	0.8677 ± 0.0298
W64	0.8619 ± 0.0294	0.8660 ± 0.0291	0.8684 ± 0.0279	0.8662 ± 0.0279	0.8671 ± 0.0290	0.8670 ± 0.0283

Table S50. NSD results of 18 U-Net models on the Parse 2022 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9007 ± 0.0356	0.9044 ± 0.0365	0.9029 ± 0.0379	0.9054 ± 0.0369	0.9025 ± 0.0385	$\mathbf{0.9061 \pm 0.0379}$
W32	0.9033 ± 0.0354	0.9051 ± 0.0356	0.9043 ± 0.0373	0.9054 ± 0.0357	0.9046 ± 0.0365	0.9058 ± 0.0374
W64	0.9023 ± 0.0347	0.9042 ± 0.0382	0.9048 ± 0.0353	0.9051 ± 0.0353	0.9042 ± 0.0355	0.9044 ± 0.0362

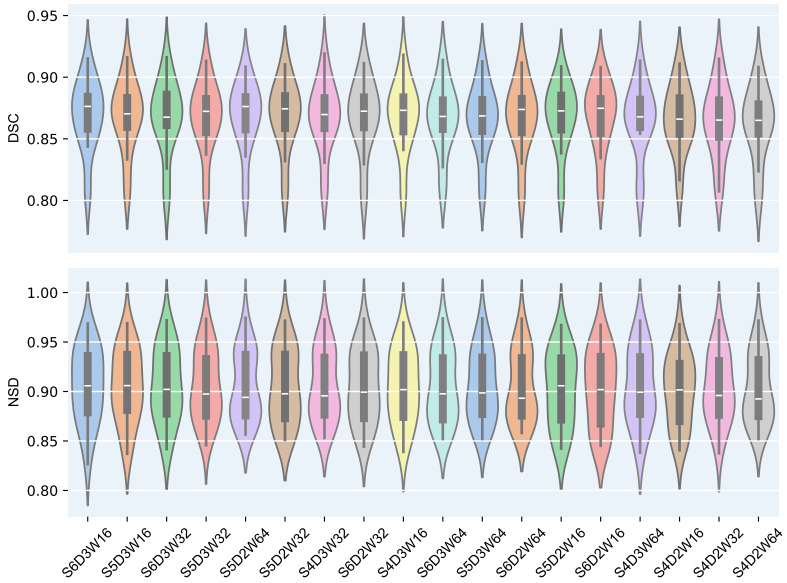

Figure S25. Violin plots of DSC and NSD for Parse 2022 dataset.

S1.26 SegTHOR

The SegTHOR (Segmentation of THoracic Organs at Risk)[16] is a CT dataset specifically for the segmentation of thoracic organs at risk (OARs), including the heart, trachea, aorta, and esophagus. This dataset comprises 60 CT images, with 40 cases for training and 20 cases for testing. We randomly divided the dataset into 32 cases for training and 8 cases for testing. The target spacing is 0.98mm $\times$ 0.98mm $\times$ 2.5mm, with a crop size of 160 $\times$ 192 $\times$ 64 for both training and inference.

Table S51. DSC results of 18 U-Net models on the SegTHOR dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9007 ± 0.0250	0.9089 ± 0.0175	0.9068 ± 0.0188	0.9109 ± 0.0164	0.9026 ± 0.0205	$\mathbf{0.9112 \pm 0.0178}$
W32	0.9035 ± 0.0234	0.9074 ± 0.0211	0.9097 ± 0.0179	0.9103 ± 0.0179	0.9066 ± 0.0212	0.9100 ± 0.0178
W64	0.9055 ± 0.0221	0.9100 ± 0.0181	0.9100 ± 0.0178	0.9107 ± 0.0173	0.9082 ± 0.0202	0.9105 ± 0.0168

Table S52. NSD results of 18 U-Net models on the SegTHOR dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7824 ± 0.0833	0.7977 ± 0.0786	0.7951 ± 0.0751	0.8008 ± 0.0761	0.7930 ± 0.0724	0.8011 ± 0.0782
W32	0.7863 ± 0.0801	0.7976 ± 0.0795	0.8011 ± 0.0718	0.8012 ± 0.0756	0.7973 ± 0.0752	0.8017 ± 0.0772
W64	0.7931 ± 0.0774	0.8021 ± 0.0740	0.8002 ± 0.0730	$\mathbf{0.8030 \pm 0.0703}$	0.8024 ± 0.0744	0.8016 ± 0.0696

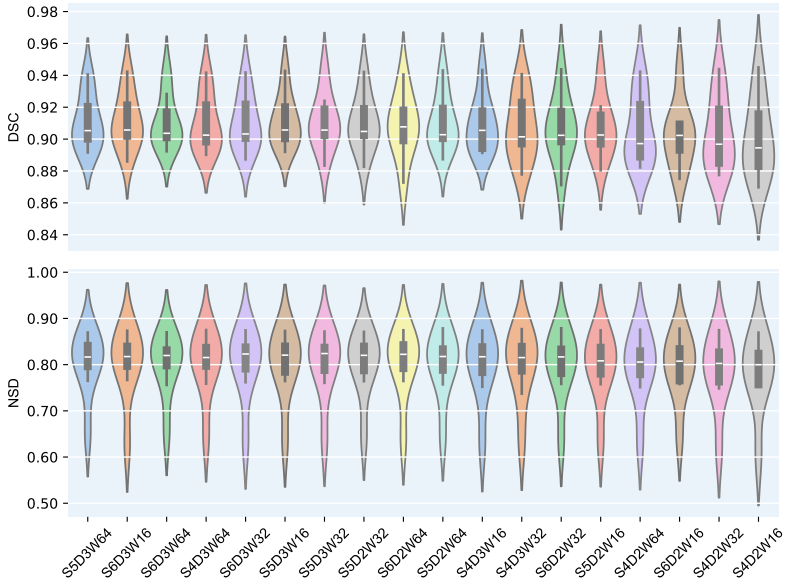

Figure S26. Violin plots of DSC and NSD for SegTHOR dataset.

S1.27 TDSC-ABUS 2023

The TDSC-ABUS 2023[17] dataset is for 3D breast ultrasound tumor segmentation. The dataset includes 100 public cases for training, which we randomly divided into 80 cases for training and 20 cases for testing. The target spacing is 1mm × 1mm × 1mm, with a crop size of 160 × 160 × 64 for both training and inference.

Table S53. DSC results of 18 U-Net models on the TDSC-ABUS 2023 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.4400 ± 0.2924	0.3662 ± 0.2691	0.4129 ± 0.3280	0.3717 ± 0.2665	0.4200 ± 0.2760	0.3232 ± 0.2291
W32	0.4566 ± 0.3479	0.3823 ± 0.2747	0.4237 ± 0.3257	0.3654 ± 0.2686	0.4364 ± 0.3272	0.4001 ± 0.3162
W64	**0.4751 ± 0.3347**	0.4093 ± 0.3058	0.4282 ± 0.3367	0.4002 ± 0.2713	0.4501 ± 0.3417	0.4091 ± 0.2953

Table S54. NSD results of 18 U-Net models on the TDSC-ABUS 2023 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.2941 ± 0.2083	0.2495 ± 0.1806	0.3055 ± 0.2403	0.2590 ± 0.1739	0.2940 ± 0.1858	0.2235 ± 0.1435
W32	0.3327 ± 0.2682	0.2669 ± 0.1816	0.3170 ± 0.2393	0.2510 ± 0.1590	0.3218 ± 0.2351	0.2884 ± 0.2254
W64	**0.3403 ± 0.2616**	0.2968 ± 0.2172	0.3210 ± 0.2545	0.2794 ± 0.1796	0.3295 ± 0.2534	0.3027 ± 0.2092

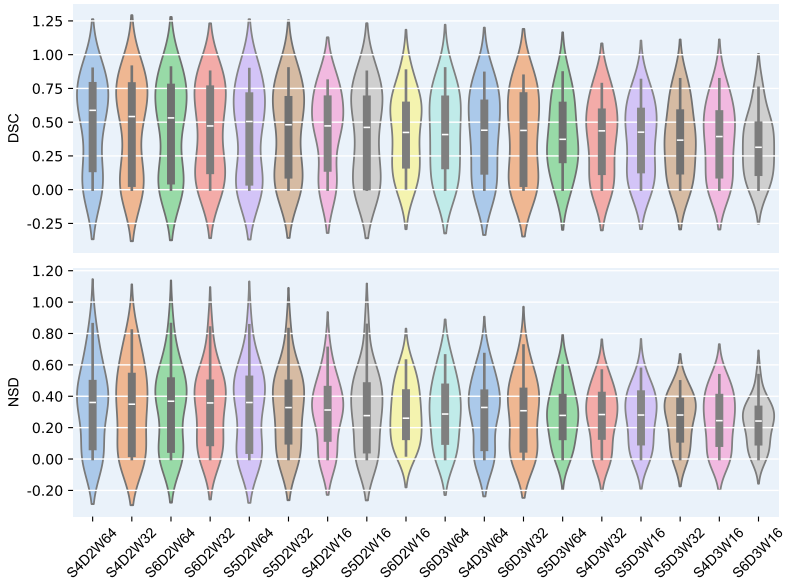

Figure S27. Violin plots of DSC and NSD for TDSC-ABUS 2023 dataset.

S1.28 FLARE 2022

The FLARE 2022[18] dataset focuses on the segmentation of 13 abdominal organs in CT images. We use the official public set of 50 training cases for training and 50 validation cases for testing. The target spacing is 0.74mm × 0.74mm × 2.5mm, with a crop size of 192 × 192 × 48 for both training and inference.

Table S55. DSC results of 18 U-Net models on the FLARE 2022 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7755 ± 0.1135	0.8293 ± 0.1003	0.8377 ± 0.0875	0.8616 ± 0.0798	0.8516 ± 0.0802	0.8646 ± 0.0754
W32	0.7962 ± 0.1197	0.8372 ± 0.1047	0.8463 ± 0.0959	0.8594 ± 0.0794	0.8601 ± 0.0751	0.8691 ± 0.0833
W64	0.8147 ± 0.1095	0.8444 ± 0.1023	0.8592 ± 0.0841	0.8656 ± 0.0846	0.8628 ± 0.0707	**0.8725 ± 0.0761**

Table S56. NSD results of 18 U-Net models on the FLARE 2022 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6890 ± 0.1270	0.7568 ± 0.1145	0.7667 ± 0.1066	0.7990 ± 0.1008	0.7849 ± 0.1028	0.8023 ± 0.0956
W32	0.7177 ± 0.1302	0.7677 ± 0.1205	0.7819 ± 0.1090	0.7978 ± 0.1007	0.7993 ± 0.0980	0.8113 ± 0.1022
W64	0.7396 ± 0.1224	0.7795 ± 0.1155	0.7982 ± 0.1001	0.8082 ± 0.1036	0.8043 ± 0.0930	**0.8175 ± 0.0948**

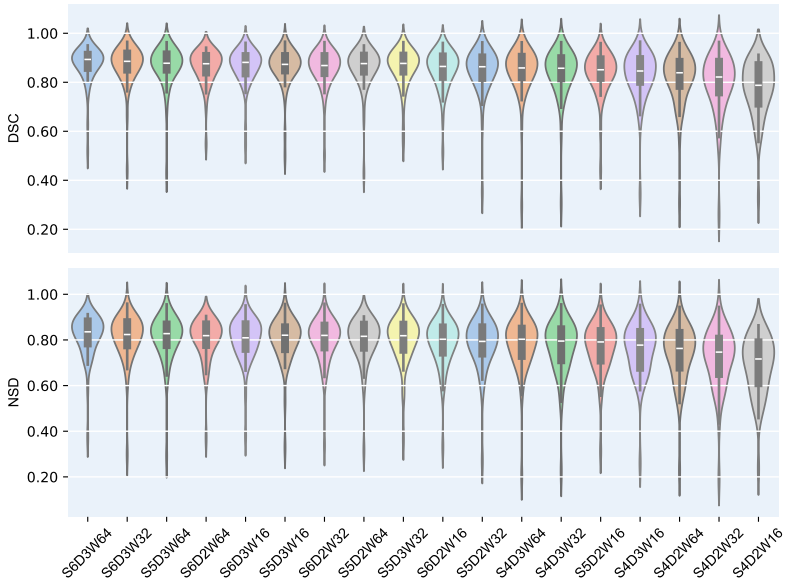

Figure S28. Violin plots of DSC and NSD for FLARE 2022 dataset.

S1.29 AMOS 2022 Task1

The AMOS 2022 Task 1 [19] dataset focuses on segmenting 15 abdominal organs in CT images. We use the official public set of 200 training cases for training and 100 validation cases for testing. The target spacing is 0.78mm $\times$ 0.78mm $\times$ 2.0mm, with a crop size of $192 \times 192 \times 48$ for both training and inference.

Table S57. DSC results of 18 U-Net models on the AMOS 2022 Task1 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8633 ± 0.0525	0.8843 ± 0.0445	0.8916 ± 0.0441	0.8959 ± 0.0418	0.8969 ± 0.0412	0.8968 ± 0.0429
W32	0.8857 ± 0.0462	0.8962 ± 0.0426	0.8999 ± 0.0425	0.9022 ± 0.0404	0.9032 ± 0.0412	0.9019 ± 0.0414
W64	0.8963 ± 0.0419	0.9006 ± 0.0418	0.9031 ± 0.0412	0.9044 ± 0.0410	0.9059 ± 0.0416	$\mathbf{0.9061 \pm 0.0400}$

Table S58. NSD results of 18 U-Net models on the AMOS 2022 Task1 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7999 ± 0.0742	0.8376 ± 0.0646	0.8483 ± 0.0646	0.8563 ± 0.0603	0.8565 ± 0.0597	0.8588 ± 0.0616
W32	0.8389 ± 0.0669	0.8580 ± 0.0630	0.8632 ± 0.0621	0.8670 ± 0.0592	0.8676 ± 0.0598	0.8674 ± 0.0607
W64	0.8572 ± 0.0624	0.8661 ± 0.0625	0.8695 ± 0.0612	0.8720 ± 0.0599	0.8732 ± 0.0602	$\mathbf{0.8739 \pm 0.0590}$

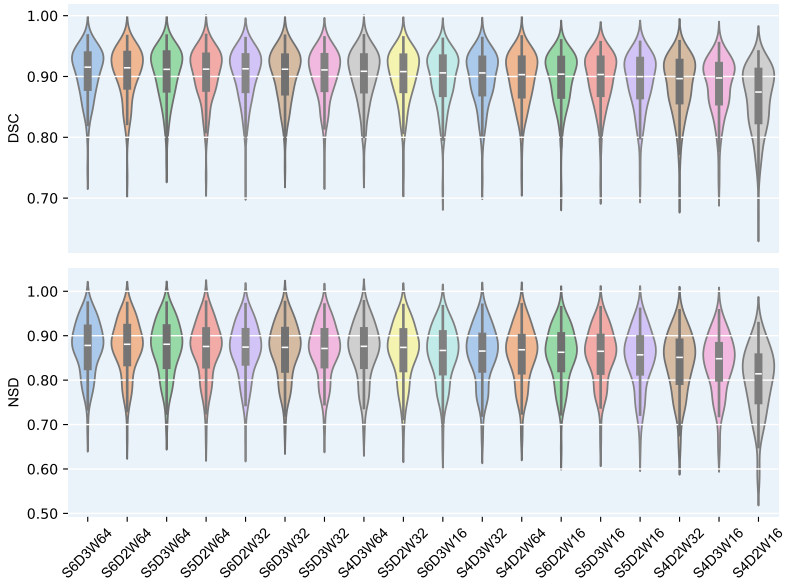

Figure S29. Violin plots of DSC and NSD for AMOS 2022 Task1 dataset.

S1.30 AMOS 2022 Task2

The AMOS 2022 Task 2[19] dataset focuses on segmenting 15 abdominal organs in both CT and MR images. The CT data is the same as in Task 1, with the addition of MR data. We use the official public set of 240 training cases (200 CT, 40 MR) for training and 120 validation cases (100 CT, 20 MR) for testing. The target spacing is 0.78mm $\times$ 0.78mm $\times$ 2.0mm, with a crop size of $160 \times 160 \times 64$ for both training and inference.

Table S59. DSC results of 18 U-Net models on the AMOS 2022 Task2 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8542 ± 0.0534	0.8785 ± 0.0460	0.8837 ± 0.0470	0.8904 ± 0.0435	0.8896 ± 0.0451	0.8939 ± 0.0427
W32	0.8796 ± 0.0480	0.8914 ± 0.0438	0.8940 ± 0.0437	0.8980 ± 0.0427	0.8972 ± 0.0445	0.8976 ± 0.0432
W64	0.8904 ± 0.0442	0.8971 ± 0.0424	0.8976 ± 0.0445	0.9004 ± 0.0425	0.9002 ± 0.0448	$\mathbf{0.9006 \pm 0.0427}$

Table S60. NSD results of 18 U-Net models on the AMOS 2022 Task2 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7775 ± 0.0819	0.8195 ± 0.0740	0.8280 ± 0.0733	0.8396 ± 0.0702	0.8387 ± 0.0717	0.8434 ± 0.0696
W32	0.8221 ± 0.0772	0.8428 ± 0.0704	0.8482 ± 0.0683	0.8540 ± 0.0675	0.8538 ± 0.0689	0.8538 ± 0.0687
W64	0.8423 ± 0.0712	0.8537 ± 0.0689	0.8563 ± 0.0693	0.8591 ± 0.0679	0.8598 ± 0.0687	$\mathbf{0.8602 \pm 0.0680}$

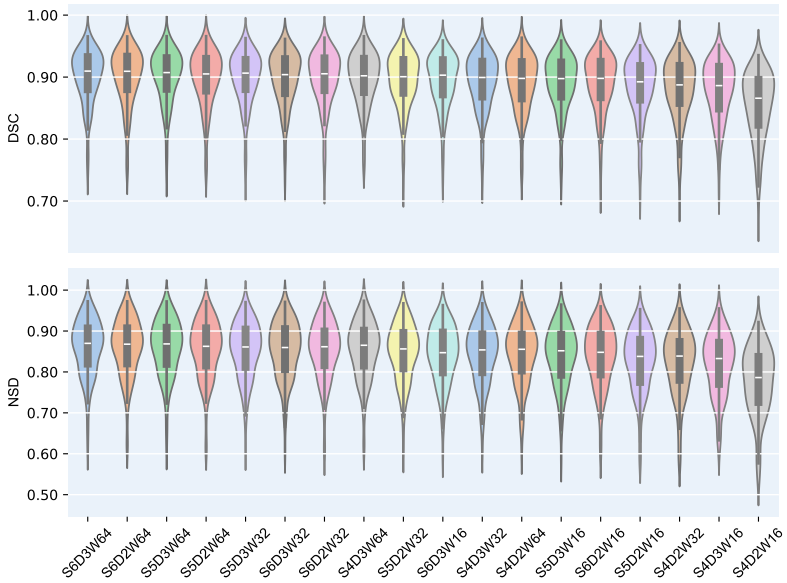

Figure S30. Violin plots of DSC and NSD for AMOS 2022 Task2 dataset.

S1.31 WORD

The WORD (Whole Abdominal ORgan Dataset)[20] is a large-scale CT dataset for abdominal organ segmentation, comprising 150 CT scans with detailed annotations for 16 organs. We split the publicly available 120 cases, using 96 cases for training and 24 cases for testing. The target spacing is 0.98mm × 0.98mm × 3.0mm, with a crop size of 160 × 160 × 64 for both training and inference.

Table S61. DSC results of 18 U-Net models on the WORD dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8086 ± 0.0517	0.8133 ± 0.0566	0.8158 ± 0.0600	0.8051 ± 0.0519	0.8034 ± 0.0510	0.8004 ± 0.0514
W32	0.8248 ± 0.0582	0.8337 ± 0.0707	0.8324 ± 0.0645	0.8320 ± 0.0627	0.8275 ± 0.0618	0.8308 ± 0.0598
W64	0.8337 ± 0.0623	0.8328 ± 0.0543	0.8329 ± 0.0603	0.8327 ± 0.0658	**0.8362 ± 0.0587**	0.8358 ± 0.0641

Table S62. NSD results of 18 U-Net models on the WORD dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7214 ± 0.0622	0.7366 ± 0.0641	0.7420 ± 0.0657	0.7269 ± 0.0606	0.7237 ± 0.0613	0.7417 ± 0.0616
W32	0.7548 ± 0.0668	0.7694 ± 0.0756	0.7650 ± 0.0692	0.7714 ± 0.0705	0.7594 ± 0.0700	0.7660 ± 0.0653
W64	0.7678 ± 0.0694	0.7722 ± 0.0639	0.7706 ± 0.0671	0.7738 ± 0.0725	0.7719 ± 0.0666	**0.7745 ± 0.0699**

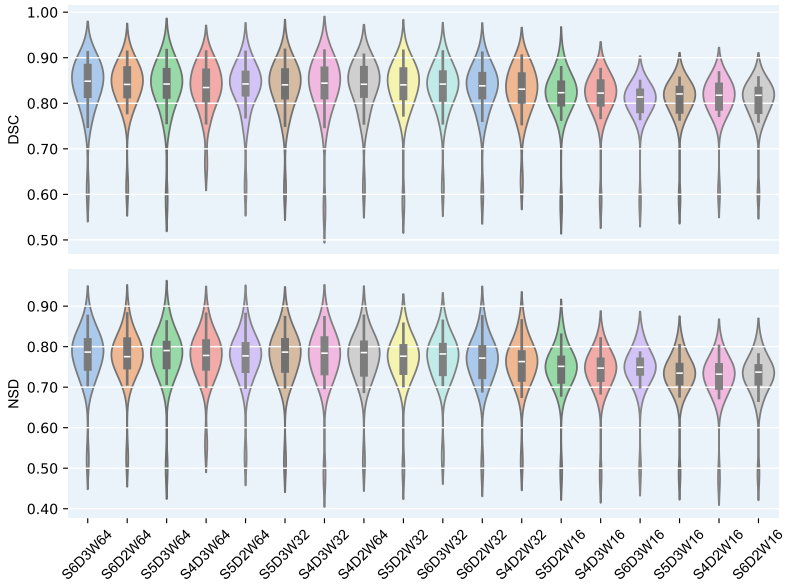

Figure S31. Violin plots of DSC and NSD forWORD dataset.

S1.32 AbdomenCT-1K

The AbdomenCT-1K[21] is a large-scale abdominal CT dataset that contains 1,000 public CT scans used for the segmentation of four abdominal organs, including the liver, kidneys, spleen, and pancreas. We selected 50 hard cases for the test set through five-fold cross-validation, and the remaining 950 cases were used for re-training. The target spacing is 0.79mm $\times$ 0.79mm $\times$ 2.5mm, with a crop size of $160 \times 160 \times 96$ for both training and inference.

Table S63. DSC results of 18 U-Net models on the AbdomenCT-1K dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7549 ± 0.1420	0.9006 ± 0.0648	0.9033 ± 0.0624	0.9120 ± 0.0571	0.9067 ± 0.0601	0.9150 ± 0.0515
W32	0.8997 ± 0.0657	0.9125 ± 0.0561	0.9144 ± 0.0581	0.9144 ± 0.0614	0.9169 ± 0.0530	0.9178 ± 0.0504
W64	0.9124 ± 0.0583	0.9148 ± 0.0568	0.9136 ± 0.0590	$\mathbf{0.9206 \pm 0.0474}$	0.9166 ± 0.0550	0.9177 ± 0.0525

Table S64. NSD results of 18 U-Net models on the AbdomenCT-1K dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6273 ± 0.1462	0.8042 ± 0.0789	0.8113 ± 0.0782	0.8231 ± 0.0766	0.8171 ± 0.0776	0.8239 ± 0.0720
W32	0.8038 ± 0.0805	0.8240 ± 0.0734	0.8244 ± 0.0746	0.8270 ± 0.0773	0.8297 ± 0.0736	0.8303 ± 0.0708
W64	0.8198 ± 0.0752	0.8256 ± 0.0717	0.8270 ± 0.0726	$\mathbf{0.8359 \pm 0.0672}$	0.8296 ± 0.0732	0.8325 ± 0.0694

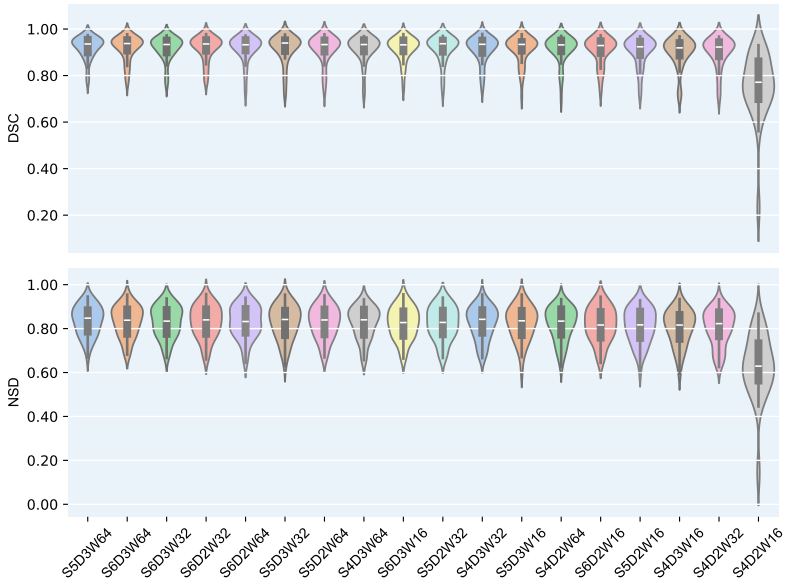

Figure S32. Violin plots of DSC and NSD for AbdomenCT-1K dataset.

S1.33 BTCV

The BTCV[22] dataset is an important benchmark for the task of abdominal organ segmentation, involving the segmentation of 13 types of abdominal organs from CT images. It provides 30 publicly available cases. We randomly divided the data into 24 cases for training and 6 cases for testing. The target spacing is 0.76mm $\times$ 0.76mm $\times$ 3.0mm, with a crop size of 192 $\times$ 192 $\times$ 48 for both training and inference.

Table S65. DSC results of 18 U-Net models on the BTCV dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7941 ± 0.0577	0.8252 ± 0.0515	0.8287 ± 0.0546	0.8377 ± 0.0551	0.8377 ± 0.0511	0.8383 ± 0.0500
W32	0.8028 ± 0.0527	0.8290 ± 0.0494	0.8301 ± 0.0531	0.8388 ± 0.0541	0.8361 ± 0.0513	0.8413 ± 0.0521
W64	0.8096 ± 0.0559	0.8248 ± 0.0579	0.8308 ± 0.0556	0.8349 ± 0.0548	$\mathbf{0.8418 \pm 0.0466}$	0.8398 ± 0.0525

Table S66. NSD results of 18 U-Net models on the BTCV dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7346 ± 0.0743	0.7770 ± 0.0641	0.7851 ± 0.0651	0.7947 ± 0.0638	0.7935 ± 0.0619	0.7958 ± 0.0584
W32	0.7499 ± 0.0663	0.7849 ± 0.0599	0.7868 ± 0.0629	0.7980 ± 0.0596	0.7913 ± 0.0592	0.7999 ± 0.0597
W64	0.7628 ± 0.0692	0.7822 ± 0.0689	0.7894 ± 0.0646	0.7954 ± 0.0638	0.8001 ± 0.0551	$\mathbf{0.8004 \pm 0.0605}$

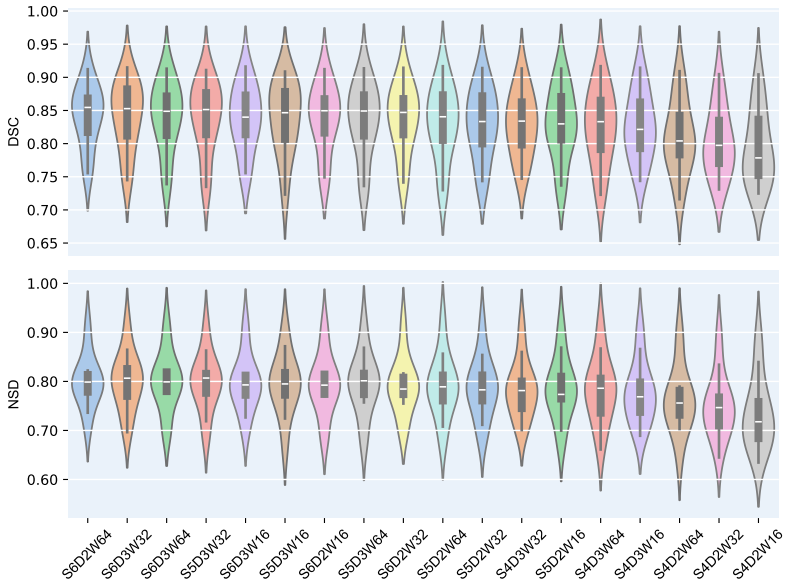

Figure S33. Violin plots of DSC and NSD for BTCV dataset.

S1.34 BTCV Cervix

The BTCV Cervix[22] dataset is a CT segmentation dataset for radiotherapy planning in patients with cervical cancer, annotated with four types of organs: bladder, uterus, rectum, and small intestine. It provides 30 publicly available cases. We randomly divided the data into 24 cases for training and 6 cases for testing. The target spacing is 0.98mm $\times$ 0.98mm $\times$ 2.5mm, with a crop size of $160 \times 192 \times 64$ for both training and inference.

Table S67. DSC results of 18 U-Net models on the BTCV Cervix dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6642 ± 0.1062	0.6614 ± 0.1013	0.6696 ± 0.0976	0.6730 ± 0.0985	0.6677 ± 0.1088	0.6659 ± 0.0925
W32	0.6596 ± 0.1148	0.6631 ± 0.0988	0.6715 ± 0.0992	0.6632 ± 0.0965	0.6669 ± 0.1023	0.6649 ± 0.0942
W64	0.6635 ± 0.1219	0.6717 ± 0.0985	0.6667 ± 0.1058	0.6635 ± 0.1028	$\mathbf{0.6747 \pm 0.0939}$	0.6666 ± 0.0978

Table S68. NSD results of 18 U-Net models on the BTCV Cervix dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.4160 ± 0.1426	0.4273 ± 0.1403	0.4223 ± 0.1413	0.4308 ± 0.1415	0.4195 ± 0.1417	0.4348 ± 0.1438
W32	0.4147 ± 0.1448	0.4265 ± 0.1432	0.4262 ± 0.1468	0.4305 ± 0.1425	0.4230 ± 0.1448	0.4263 ± 0.1378
W64	0.4225 ± 0.1562	0.4305 ± 0.1442	0.4310 ± 0.1543	$\mathbf{0.4342 \pm 0.1497}$	0.4260 ± 0.1428	0.4316 ± 0.1420

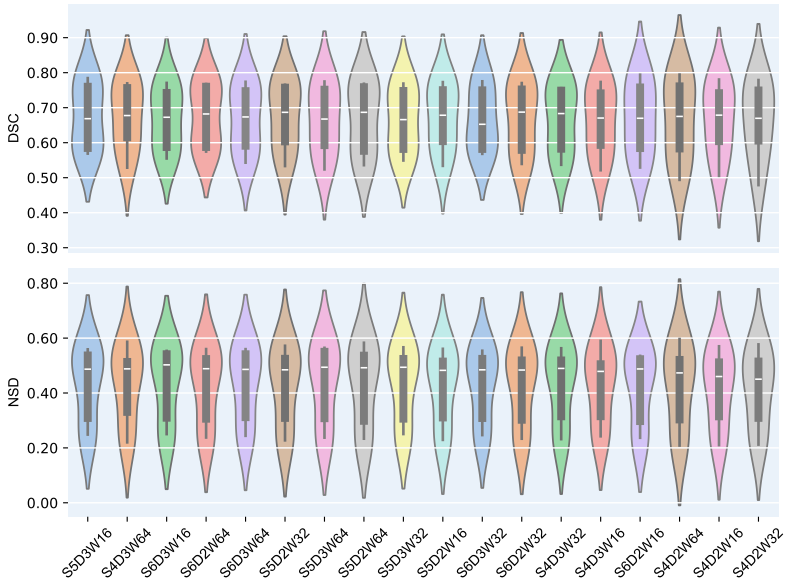

Figure S34. Violin plots of DSC and NSD for BTCV Cervix dataset.

S1.35 KiTS21

The KiTS21 (Kidney Tumor Segmentation)[23] is a CT dataset for segmenting kidneys and the tumors and cysts on them. The KiTS21 dataset includes 300 publicly available training cases. We randomly divided the data into 240 cases for training and 60 cases for testing. The target spacing is 0.78mm × 0.78mm × 3.0mm, with a crop size of 128 × 128 × 128 for both training and inference.

Table S69. DSC results of 18 U-Net models on the KiTS21 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8683 ± 0.1409	0.8933 ± 0.1105	0.8887 ± 0.1183	0.8945 ± 0.1202	0.8954 ± 0.1138	0.8720 ± 0.1355
W32	0.8697 ± 0.1516	0.8921 ± 0.1205	0.8919 ± 0.1215	0.8958 ± 0.1165	0.8910 ± 0.1222	0.8931 ± 0.1214
W64	0.8715 ± 0.1460	0.8871 ± 0.1241	0.8890 ± 0.1359	0.8871 ± 0.1351	**0.8968 ± 0.1130**	0.8892 ± 0.1208

Table S70. NSD results of 18 U-Net models on the KiTS21 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7743 ± 0.1496	0.8100 ± 0.1072	0.8153 ± 0.1086	0.8239 ± 0.1132	0.8232 ± 0.1072	0.7998 ± 0.1195
W32	0.7857 ± 0.1491	0.8179 ± 0.1238	0.8202 ± 0.1157	0.8233 ± 0.1098	0.8193 ± 0.1134	0.8225 ± 0.1119
W64	0.7886 ± 0.1466	0.8109 ± 0.1296	0.8222 ± 0.1254	0.8198 ± 0.1246	**0.8271 ± 0.1058**	0.8185 ± 0.1141

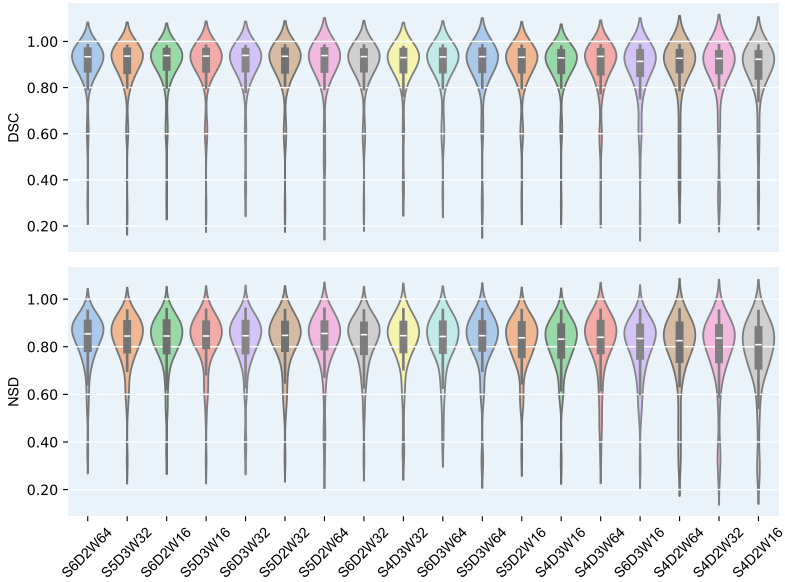

Figure S35. Violin plots of DSC and NSD for KiTS21 dataset.

S1.36 ACDC

The ACDC (Automatic Cardiac Diagnosis Challenge)[24] dataset focuses on segmenting the left ventricle (LV), right ventricle (RV), and myocardium (Myo) in cardiac cine-MRI images during diastole (ED) and systole (ES). It consists of 150 cases, divided into 100 training cases and 50 testing cases. We followed the official split for training and evaluation. The target spacing is 1.5mm $\times$ 1.5mm $\times$ 10mm, with a crop size of $224 \times 256 \times 20$ for both training and inference.

Table S71. DSC results of 18 U-Net models on the ACDC dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9132 ± 0.0357	0.9178 ± 0.0310	0.9136 ± 0.0338	0.9184 ± 0.0300	0.9149 ± 0.0295	0.9182 ± 0.0306
W32	0.9137 ± 0.0308	0.9160 ± 0.0398	0.9163 ± 0.0304	0.9176 ± 0.0301	0.9153 ± 0.0310	0.9167 ± 0.0313
W64	0.9134 ± 0.0317	0.9175 ± 0.0306	0.9157 ± 0.0315	$\mathbf{0.9189 \pm 0.0290}$	0.9153 ± 0.0315	0.9187 ± 0.0300

Table S72. NSD results of 18 U-Net models on the ACDC dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8179 ± 0.0620	0.8262 ± 0.0588	0.8176 ± 0.0632	0.8249 ± 0.0576	0.8163 ± 0.0604	0.8258 ± 0.0579
W32	0.8187 ± 0.0580	0.8241 ± 0.0606	0.8218 ± 0.0590	0.8243 ± 0.0573	0.8200 ± 0.0608	0.8235 ± 0.0581
W64	0.8182 ± 0.0592	0.8249 ± 0.0562	0.8218 ± 0.0600	$\mathbf{0.8280 \pm 0.0563}$	0.8217 ± 0.0595	0.8280 ± 0.0579

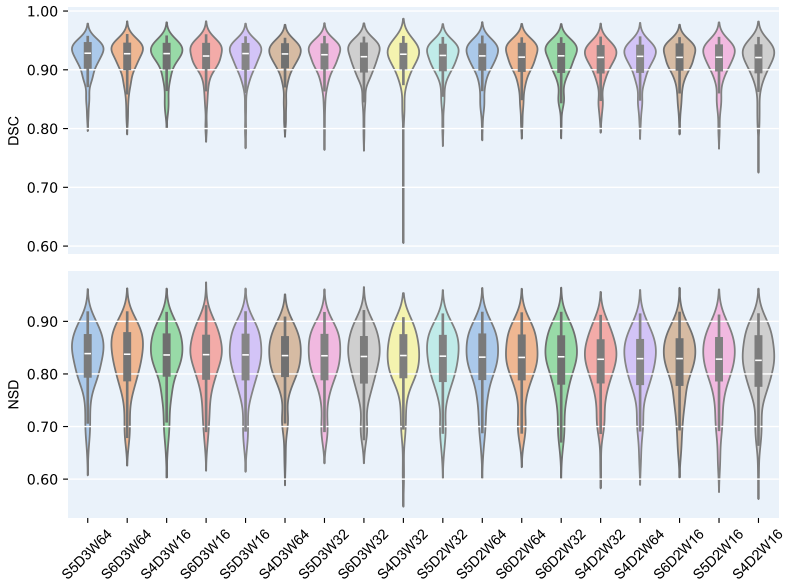

Figure S36. Violin plots of DSC and NSD for ACDC dataset.

S1.37 LAScarQS 2022 Task1

The LAScarQS 2022 Task1[25] dataset focuses on segmenting the left atrium (LA) and LA scar in post-ablation LGE MRI scans of atrial fibrillation patients. It comprises 60 public training cases and 34 private test cases. We randomly divided the data into 48 training cases and 12 test cases for our study. The target spacing is 0.625mm $\times$ 0.625mm $\times$ 2.5mm, with a crop size of $228 \times 228 \times 20$ for both training and inference.

Table S73. DSC results of 18 U-Net models on the dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7227 ± 0.0609	0.7214 ± 0.0646	0.7203 ± 0.0601	0.7217 ± 0.0620	0.7219 ± 0.0593	$\mathbf{0.7275 \pm 0.0588}$
W32	0.7120 ± 0.0672	0.7122 ± 0.0632	0.7171 ± 0.0588	0.7160 ± 0.0585	0.7174 ± 0.0587	0.7126 ± 0.0607
W64	0.7078 ± 0.0643	0.7099 ± 0.0598	0.7126 ± 0.0573	0.7104 ± 0.0589	0.7129 ± 0.0586	0.7101 ± 0.0581

Table S74. NSD results of 18 U-Net models on the dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7711 ± 0.1051	0.7771 ± 0.0979	0.7812 ± 0.0962	0.7854 ± 0.0960	0.7833 ± 0.0896	$\mathbf{0.7901 \pm 0.0875}$
W32	0.7712 ± 0.1087	0.7773 ± 0.1014	0.7853 ± 0.0931	0.7844 ± 0.0922	0.7854 ± 0.0935	0.7825 ± 0.0920
W64	0.7715 ± 0.1062	0.7785 ± 0.0952	0.7839 ± 0.0877	0.7817 ± 0.0892	0.7826 ± 0.0914	0.7807 ± 0.0906

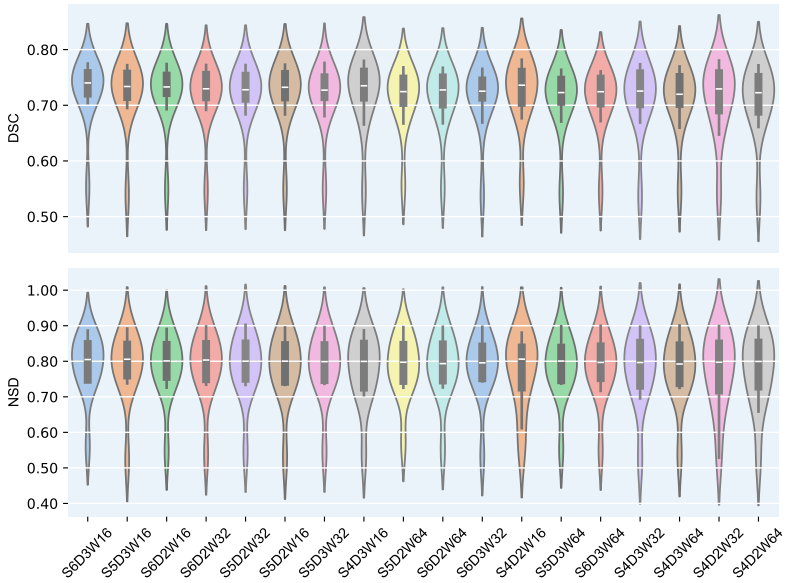

Figure S37. Violin plots of DSC and NSD for LAScarQS 2022 Task1 dataset.

S1.38 LAScarQS 2022 Task2

The LAScarQS 2022 Task2[25] dataset focuses solely on left atrium (LA) segmentation in LGE MRI scans. It includes both pre- and post-ablation images, with 130 training cases and 64 test cases. We randomly divided the data into 104 training cases and 26 test cases for our study. The target spacing is 0.625mm $\times$ 0.625mm $\times$ 2.5mm, with a crop size of $224 \times 256 \times 40$ for both training and inference.

Table S75. DSC results of 18 U-Net models on the LAScarQS 2022 Task2 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.9144 ± 0.0220	0.9171 ± 0.0237	0.9183 ± 0.0213	0.9186 ± 0.0213	0.9180 ± 0.0223	0.9174 ± 0.0235
W32	0.9180 ± 0.0208	$\mathbf{0.9206 \pm 0.0218}$	0.9193 ± 0.0234	0.9190 ± 0.0221	0.9190 ± 0.0244	0.9185 ± 0.0223
W64	0.9163 ± 0.0234	0.9191 ± 0.0226	0.9200 ± 0.0225	0.9196 ± 0.0218	0.9182 ± 0.0247	0.9187 ± 0.0230

Table S76. NSD results of 18 U-Net models on the LAScarQS 2022 Task2 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.7428 ± 0.0786	0.7532 ± 0.0741	0.7562 ± 0.0755	0.7595 ± 0.0695	0.7570 ± 0.0720	0.7570 ± 0.0761
W32	0.7523 ± 0.0781	$\mathbf{0.7633 \pm 0.0749}$	0.7586 ± 0.0802	0.7597 ± 0.0729	0.7597 ± 0.0842	0.7610 ± 0.0716
W64	0.7476 ± 0.0827	0.7572 ± 0.0816	0.7609 ± 0.0804	0.7607 ± 0.0770	0.7597 ± 0.0816	0.7615 ± 0.0754

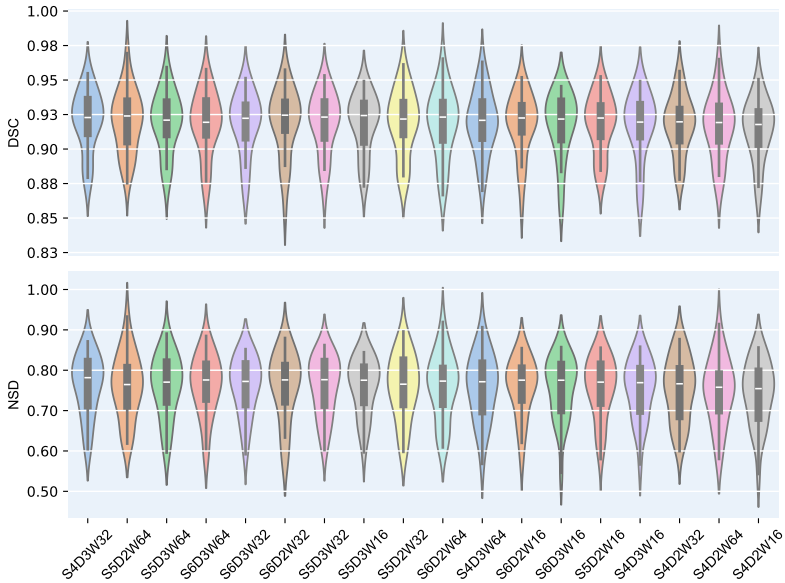

Figure S38. Violin plots of DSC and NSD for LAScarQS 2022 Task2 dataset.

S1.39 MyoPS 2020

The MyoPS 2020[26] dataset focuses on myocardial pathology segmentation using multi-sequence CMR. It includes aligned T2, bSSFP, and LGE MRI data from 45 patients, manually annotated for scar, edema, normal myocardium, and ventricular blood pools. The dataset comprises 25 training and 20 test cases. We randomly split the public training set into 20 training and 5 cases for testing. The target spacing is 0.73mm $\times$ 0.73mm $\times$ 14mm, with a crop size of $512 \times 512 \times 6$ for both training and inference.

Table S77. DSC results of 18 U-Net models on the MyoPS 2020 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6453 ± 0.1350	0.6799 ± 0.1052	0.7091 ± 0.0619	0.7178 ± 0.0511	0.7316 ± 0.0545	0.7359 ± 0.0444
W32	0.6421 ± 0.1288	0.6829 ± 0.0994	0.7267 ± 0.0518	0.7285 ± 0.0341	0.7271 ± 0.0502	0.7274 ± 0.0498
W64	0.6429 ± 0.1483	0.6774 ± 0.1046	0.7137 ± 0.0558	0.7252 ± 0.0444	0.7359 ± 0.0572	$\mathbf{0.7400 \pm 0.0405}$

Table S78. NSD results of 18 U-Net models on the MyoPS 2020 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6216 ± 0.1464	0.6528 ± 0.1328	0.6798 ± 0.0760	0.6930 ± 0.0678	0.7022 ± 0.0658	0.7003 ± 0.0596
W32	0.6147 ± 0.1421	0.6686 ± 0.1158	0.6962 ± 0.0723	0.7051 ± 0.0561	0.7026 ± 0.0616	0.7054 ± 0.0705
W64	0.6160 ± 0.1706	0.6659 ± 0.1261	0.6904 ± 0.0722	0.7082 ± 0.0562	0.7076 ± 0.0750	$\mathbf{0.7205 \pm 0.0555}$

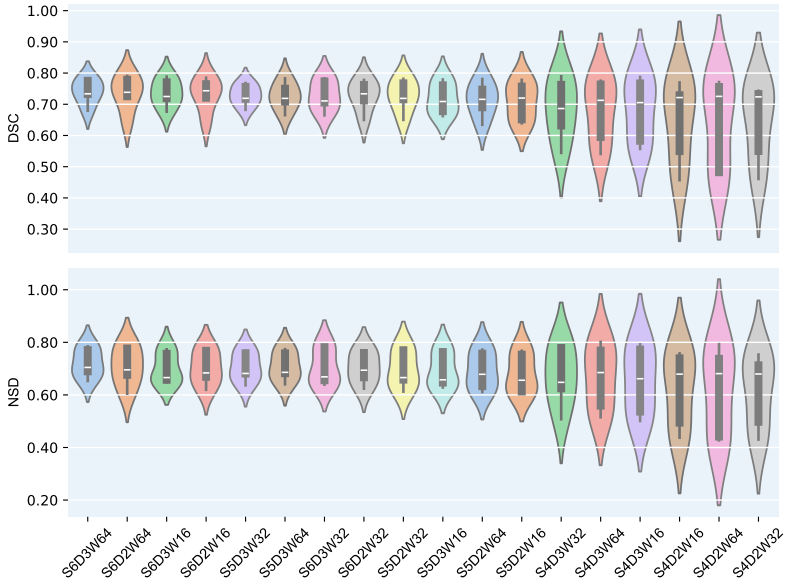

Figure S39. Violin plots of DSC and NSD for MyoPS 2020 dataset.

S1.40 MM-WHS

The MM-WHS (Multi-Modality Whole Heart Segmentation)[27] dataset provides manual segmentation for seven key cardiac substructures: left and right ventricles, left and right atria, left ventricular myocardium, ascending aorta, and pulmonary artery. The public training set includes 40 cases (20 CT and 20 MRI). For our study, we randomly selected 32 cases for training (16 CT and 16 MRI) and 8 cases (4 CT and 4 MRI) for testing. The target spacing is 0.44mm $\times$ 0.44mm $\times$ 0.625mm, with a crop size of 160 $\times$ 192 $\times$ 80 for both training and inference.

Table S79. DSC results of 18 U-Net models on the MM-WHS dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.8600 ± 0.0829	0.8743 ± 0.0787	0.8995 ± 0.0374	$\mathbf{0.9033 \pm 0.0345}$	0.8983 ± 0.0387	0.9028 ± 0.0347
W32	0.8628 ± 0.0874	0.8782 ± 0.0729	0.8975 ± 0.0413	0.8992 ± 0.0385	0.8997 ± 0.0381	0.8971 ± 0.0402
W64	0.8600 ± 0.0918	0.8816 ± 0.0681	0.8926 ± 0.0509	0.9009 ± 0.0374	0.8989 ± 0.0389	0.8989 ± 0.0405

Table S80. NSD results of 18 U-Net models on the MM-WHS dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6515 ± 0.1638	0.6834 ± 0.1458	0.7115 ± 0.1166	0.7205 ± 0.1120	0.7156 ± 0.1162	0.7199 ± 0.1091
W32	0.6645 ± 0.1556	0.6928 ± 0.1363	0.7134 ± 0.1180	0.7167 ± 0.1141	$\mathbf{0.7206 \pm 0.1114}$	0.7174 ± 0.1094
W64	0.6688 ± 0.1549	0.6968 ± 0.1357	0.7104 ± 0.1248	0.7194 ± 0.1132	0.7195 ± 0.1128	0.7191 ± 0.1145

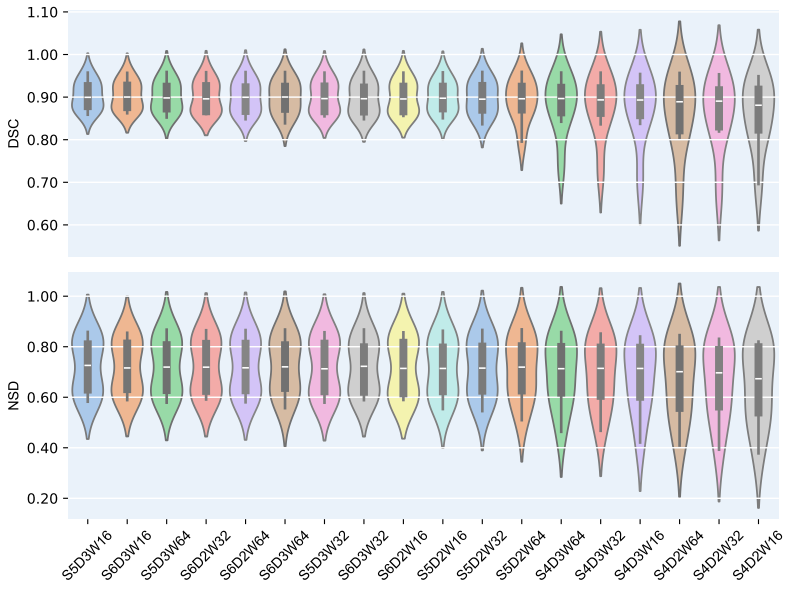

Figure S40. Violin plots of DSC and NSD for MM-WHS dataset.

S1.41 VerSe 2020

The VerSe 2020[28] dataset consists of 319 cases from 300 patients, providing a large-scale, multi-device, multi-center CT image dataset for spinal segmentation. It includes 26 annotated vertebral categories, including C1-C7, T1-T12, L1-L5, T13, and L6, while vertebrae that are only partially visible at the top or bottom of the scans are not annotated. We randomly split the dataset into 248 cases for training and 71 cases for testing. The target spacing is 1.0mm $\times$ 1.0mm $\times$ 0.93mm, with a crop size of $112 \times 112 \times 160$ for both training and inference.

Table S81. DSC results of 18 U-Net models on the VerSe 2020 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6782 ± 0.1497	0.7166 ± 0.1489	0.7157 ± 0.1562	0.7269 ± 0.1581	0.7306 ± 0.1495	0.7268 ± 0.1468
W32	0.7069 ± 0.1443	0.7191 ± 0.1533	0.7210 ± 0.1532	0.7307 ± 0.1623	0.7345 ± 0.1504	0.7370 ± 0.1386
W64	0.7225 ± 0.1489	0.7318 ± 0.1513	0.7227 ± 0.1642	0.7353 ± 0.1519	$\mathbf{0.7377 \pm 0.1486}$	0.7311 ± 0.1519

Table S82. NSD results of 18 U-Net models on the VerSe 2020 dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.6816 ± 0.1483	0.7316 ± 0.1467	0.7345 ± 0.1543	0.7479 ± 0.1567	0.7494 ± 0.1477	0.7463 ± 0.1479
W32	0.7201 ± 0.1421	0.7377 ± 0.1511	0.7412 ± 0.1496	0.7522 ± 0.1607	0.7554 ± 0.1463	0.7565 ± 0.1388
W64	0.7381 ± 0.1468	0.7503 ± 0.1506	0.7439 ± 0.1628	0.7564 ± 0.1511	$\mathbf{0.7594 \pm 0.1462}$	0.7519 ± 0.1521

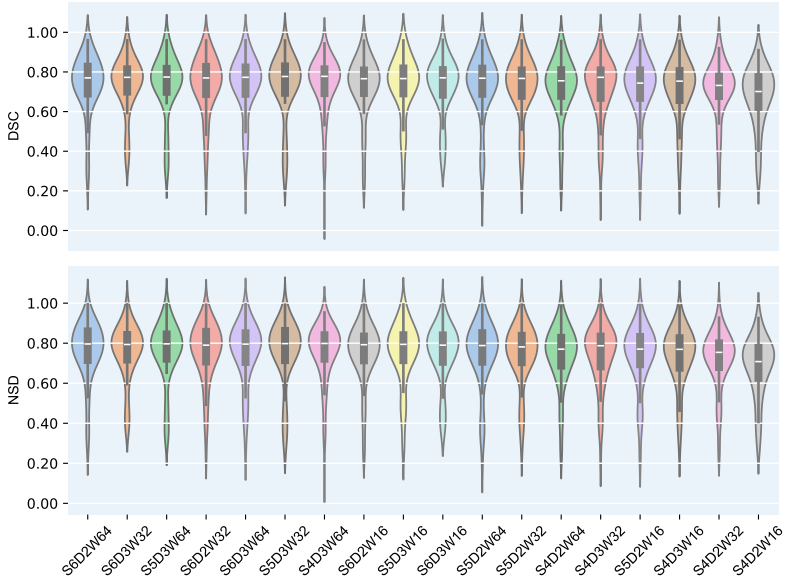

Figure S41. Violin plots of DSC and NSD for VerSe 2020 dataset.

S1.42 CTSpine1K

The CTSpine1K[29] is a large-scale CT dataset specifically designed for spinal segmentation, comprising 1005 CT scans collected from four publicly available datasets. It includes annotations for 25 vertebral classes, spanning C1-C7, T1-T12, and L1-L6. We randomly split the dataset into 804 samples for training and 201 samples for testing. The target spacing is 0.76mm $\times$ 0.76mm $\times$ 0.8mm, with a crop size of 128 $\times$ 128 $\times$ 128 for both training and inference.

Table S83. DSC results of 18 U-Net models on the CTSpine1K dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.5120 $\pm$ 0.4214	0.5291 $\pm$ 0.4348	0.5466 $\pm$ 0.4450	0.5446 $\pm$ 0.4431	0.5433 $\pm$ 0.4421	0.5440 $\pm$ 0.4443
W32	0.5334 $\pm$ 0.4353	0.5469 $\pm$ 0.4446	0.5510 $\pm$ 0.4484	0.5511 $\pm$ 0.4476	0.5515 $\pm$ 0.4482	0.5472 $\pm$ 0.4439
W64	0.5427 $\pm$ 0.4427	0.5516 $\pm$ 0.4486	**0.5555 $\pm$ 0.4514**	0.5532 $\pm$ 0.4492	0.5518 $\pm$ 0.4496	0.5530 $\pm$ 0.4493

Table S84. NSD results of 18 U-Net models on the CTSpine1K dataset.

	S4		S5		S6	
	D2	D3	D2	D3	D2	D3
W16	0.5004 $\pm$ 0.4271	0.5248 $\pm$ 0.4463	0.5457 $\pm$ 0.4594	0.5438 $\pm$ 0.4574	0.5421 $\pm$ 0.4559	0.5427 $\pm$ 0.4587
W32	0.5280 $\pm$ 0.4456	0.5450 $\pm$ 0.4582	0.5501 $\pm$ 0.4627	0.5511 $\pm$ 0.4624	0.5509 $\pm$ 0.4629	0.5472 $\pm$ 0.4586
W64	0.5404 $\pm$ 0.4556	0.5509 $\pm$ 0.4632	**0.5554 $\pm$ 0.4667**	0.5536 $\pm$ 0.4643	0.5520 $\pm$ 0.4646	0.5532 $\pm$ 0.4647

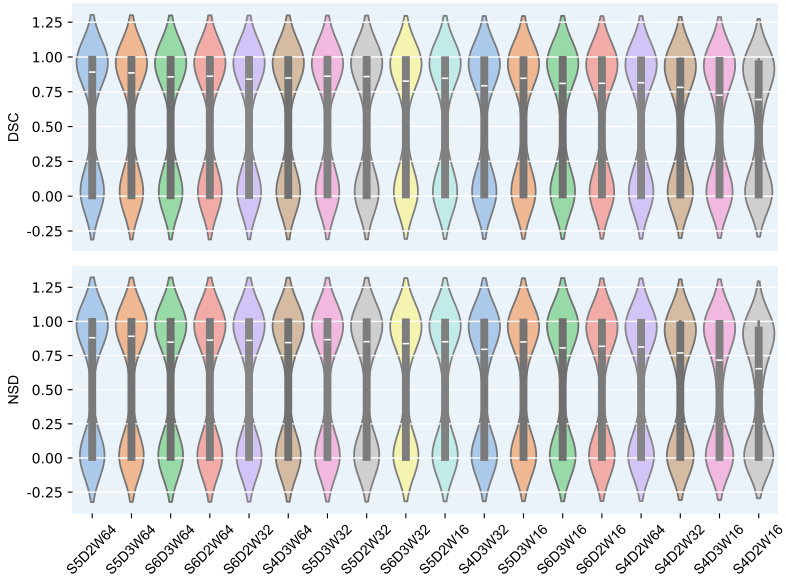

Figure S42. Violin plots of DSC and NSD for CTSpine1K dataset.

References

1. Antonelli, M. *et al.* The medical segmentation decathlon. *Nat. Commun.* **13**, 4128 (2022).
2. Wasserthal, J. *et al.* Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiol. Artif. Intell.* **5**, e230024 (2023).
3. D'Antonoli, T. A. *et al.* Totalsegmentator mri: Robust sequence-independent segmentation of multiple anatomic structures in mri. *Radiology* (2025).
4. Gatidis, S. *et al.* A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Sci. Data* **9**, 601 (2022).
5. Rister, B., Yi, D., Shivakumar, K., Nobashi, T. & Rubin, D. L. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Sci. Data* **7**, 381 (2020).
6. Luo, X. *et al.* Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. *Med. Image Anal.* **101**, 103447 (2025).
7. Podobnik, G., Strojan, P., Peterlin, P., Ibragimov, B. & Vrtovec, T. Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. *Med. Phys.* **50**, 1917–1927 (2023).
8. Payette, K. *et al.* An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Sci. Data* **8**, 167 (2021).
9. Baid, U. *et al.* The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. Preprint at https://arxiv.org/abs/2107.02314 (2021).
10. Chen, H. *et al.* Cerebral artery segmentation challenge. (2023).
11. Li, X. *et al.* The state-of-the-art 3d anisotropic intracranial hemorrhage segmentation on non-contrast head ct: the instance challenge. Preprint at https://arxiv.org/abs/2301.03281 (2023).
12. Hernandez Petzsche, M. R. *et al.* Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Sci. Data* **9**, 762 (2022).
13. Kuijf, H. J. *et al.* Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE Trans. Med. Imaging* **38**, 2556–2568 (2019).
14. Zhang, M. *et al.* Multi-site, multi-domain airway tree modeling. *Med. Image Anal.* **90**, 102957 (2023).
15. Luo, G. *et al.* Efficient automatic segmentation for multi-level pulmonary arteries: the parse challenge. Preprint at https://arxiv.org/abs/2304.03708 (2023).
16. Lambert, Z., Petitjean, C., Dubray, B. & Kuan, S. Segthor: Segmentation of thoracic organs at risk in ct images. In *Proc. Int. Conf. Image Process. Theory Tools Appl. (IPTA)*, 1–6 (2020).
17. Zhou, Y. *et al.* Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. *Med. Image Anal.* **70**, 101918 (2021).
18. Ma, J. *et al.* Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the flare22 challenge. *Lancet Digit. Heal.* **6**, e815–e826 (2024).
19. Ji, Y. *et al.* Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022).
20. Luo, X. *et al.* Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Med. Image Anal.* **82**, 102642 (2022).
21. Ma, J. *et al.* Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 6695–6714 (2022).
22. Landman, B. *et al.* Multi-atlas labeling beyond the cranial vault: workshop and challenge. (2017).
23. Heller, N. *et al.* The kits21 challenge: automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. Preprint at https://arxiv.org/abs/2307.01984 (2023).
24. Bernard, O. *et al.* Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans. Med. Imaging* **37**, 2514–2525 (2018).
25. Li, L., Zimmer, V. A., Schnabel, J. A. & Zhuang, X. Atrialjsqnet: A new framework for joint segmentation and quantification of left atrium and scars incorporating spatial and shape information. *Med. Image Anal.* **76**, 102303 (2022).

26. Qiu, J. *et al.* Myops-net: Myocardial pathology segmentation with flexible combination of multi-sequence cmr images. *Med. Image Anal.* **84**, 102694 (2023).

27. Zhuang, X. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 2933–2946 (2019).

28. Sekuboyina, A. *et al.* Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. *Med. Image Anal.* **73**, 102166 (2021).

29. Deng, Y. *et al.* Ctspine1k: A large-scale dataset for spinal vertebrae segmentation in computed tomography. Preprint at https://arxiv.org/abs/2105.14711 (2021).