nnSAM: Plug-and-play Segment Anything Model Improves nnUNet Performance

Yunxiang Li[1], Bowen Jing[1], Xiang Feng[2], Zihan Li[3], Yihe Zhang[1], Yongbo He[2], Jing Wang[1], You Zhang[1](✉)

[1] University of Texas Southwestern Medical Center, Dallas, USA
[2] Hangzhou Dianzi University, Hangzhou, China
[3] University of Washington, Seattle, USA
you.zhang@utsouthwestern.edu

Abstract. Recently, foundation models in computer vision, such as the Segment Anything Model (SAM), have shown immense promise in scalable image segmentation across diverse tasks. Besides, the field of medical image segmentation has benefited significantly from specialized neural networks like nnUNet, which adaptively configure for specific segmentation challenges. Herein, we present nnSAM, which synergistically integrates the SAM and nnUNet tailored for medical image segmentation. This model leverages the powerful feature extraction capabilities of SAM while harnessing the adaptive configuration capabilities of nnUNet. With few-shot fine-tuning, nnSAM offers a robust and adaptable solution in medical image segmentation. Our comprehensive evaluation on different numbers of 2D training samples highlights the few-shot superior performance of nnSAM, which makes it more valuable when training data is scarce. By melding the strengths of both its predecessors, nnSAM positions itself as a new benchmark in medical image segmentation, offering a tool that amalgamates broad applicability with specialized efficiency. The code is available at https://github.com/Kent0n-Li/nnSAM.

1 Introduction

The efficient and accurate segmentation of medical images has been a crucial step in the modern clinical workflow including disease diagnosis, treatment planning and monitoring, and disease prognosis [1]. While medical image segmentation is a very time-consuming task, the advent of deep learning-based segmentation technology has significantly reduced the time and cost of manual segmentation performed by radiologists [2]. Among the deep learning architectures designed for biomedical image segmentation, U-Net stands out for its effectiveness and efficiency [3]. Immediately after that, a large number of researchers designed different UNet-based network architectures for different tasks [4]. For example, TransUNet integrates the advantages of U-Net and Transformers to redefine the benchmark of medical image segmentation [5]. By utilizing the global context capability of Transformers and the precise localization property of U-Net,

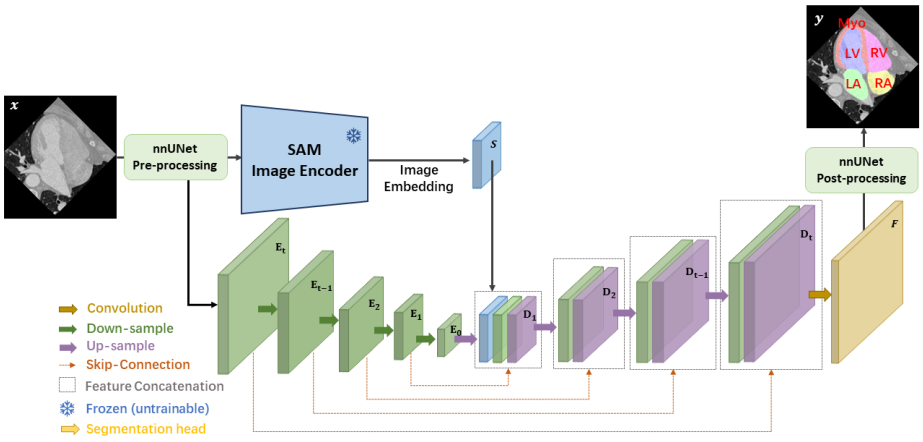

Fig. 1. The architecture of nnSAM, which integrates nnUNet's encoder and the pre-trained SAM encoder. The concatenated embeddings are input for nnUNet's decoder. (LV: left ventricle; RV: right ventricle; LA: left atrium; RA: right atrium; Myo: myocardium of LV)

this innovative fusion promises to capture long-range dependencies while maintaining the meticulous segmentation necessary for clinical applications. Another example is UNet++ [6], which is designed to enhance the performance of medical image segmentation and bridge the semantic gap between the encoder and decoder feature maps by incorporating deeply-supervised encoder-decoder networks interlinked with nested, dense skip pathways. SwinUNet [7] introduces a pure Transformer-based approach to medical image segmentation, leveraging the U-shaped Encoder-Decoder architecture and skip-connections for enhanced local-global semantic feature learning. This model exhibits superior performance over traditional convolution-based methods and mixed transformer-convolution architectures.

Historically, deep learning models for medical image segmentation were often tailor-made for specific datasets or applications, making it challenging to generalize their effectiveness across various segmentation tasks. Acknowledging these challenges, the nnUNet framework [8] was proposed. The nnUNet framework, a "no-new-Net", takes a unique approach by abstaining from proposing new network architectures. Instead, it refocuses efforts on methodological, architectural search, and data preprocessing steps to yield optimal performance (Isensee et al., 2020). This ingenious strategy posits that with appropriate preprocessing and postprocessing, even a basic network architecture can achieve state-of-the-art performance across a wide variety of medical segmentation tasks. This robust framework emerged as an all-in-one solution, automating significant parts of the UNet-based medical image segmentation pipeline - right from preprocessing, and network architecture selection, to post-processing.

While the emergence of models like nnUNet signifies a transition to more flexible approaches in medical image segmentation, the quality of results still relies on ample training data. Acquiring large volumes of labeled medical images is not only costly but also challenging. For medical image segmentation tasks with limited amount of training data, a few-shot solution is more practical and of greater interest. Besides, the advent of the segment anything model (SAM) [9,10,11] may be able to address this problem even more effectively. Leveraging the vastness and variety of the SA-1B training dataset, SAM presents the promising potential of wide-ranging usability across diverse image categories. However, its reliance on prompts raises questions about its seamless integration in fully automated workflows. This aspect, although a boon for adaptability, may pose challenges in high-throughput medical scenarios, demanding real-time or uninterrupted operations.

In the wake of this, we introduce nnSAM, a novel plug-and-play solution designed to enhance medical image segmentation, particularly in settings with limited labeled data. nnSAM synergizes the powerful feature extraction and generalization capabilities of the Segment Anything Model [9] with the adaptive capabilities of nnUNet [8]. By leveraging the image encoder of the Segment Anything Model and seamlessly integrating it with nnUNet's architecture, nnSAM produces a latent space representation that serves as the foundation for enhanced segmentation. This fusion ensures that even in scenarios where training data is scant, high-quality medical image segmentation is achievable. Our method achieves an intriguing balance between precision and adaptability, shaping the future trajectory of medical image segmentation.

The main contributions of this paper are summarised as follows:

- We introduce nnSAM, a novel fusion of the Segment Anything Model (SAM) and nnUNet. By amalgamating the powerful feature extraction capabilities of SAM with the adaptable architecture of nnUNet, nnSAM ensures enhanced segmentation quality, even with minimal training data.
- Our comprehensive evaluation illuminates the superior performance of nnSAM, providing a new baseline for medical image segmentation.

2 Method

2.1 Architecture Overview

The architecture of the proposed nnSAM is depicted in Fig. 1. The model is designed to combine the strengths of nnUNet [8] and the Segment Anything Model (SAM) [9]. Specifically, nnSAM consists of two parallel encoders: the nnUNet encoder and the SAM encoder, which is a pre-trained Vision Transformer (ViT) [12]. The embeddings from these encoders are concatenated and then input nnUNet's decoder to produce the final segmentation map. Furthermore, the SAM encoder is used as a plug-and-play plugin whose parameters are all frozen during training. That is, we only update the encoder and decoder of nnUNet during training.

2.2 Auto-configured nnUNet Architecture

Integrating nnUNet into the nnSAM architecture offers a distinct advantage of automated configuration, making it highly adaptable to the unique needs of various medical imaging datasets. This adaptive capability starts from a configuration process that automatically molds the encoder's architecture to suit specific characteristics like the dimensions of the medical images, the number of channels, and the diversity of classes involved in the segmentation task. Additionally, nnUNet provides an automated preprocessing phase, which includes normalizing the data and applying optional data augmentation techniques such as rotations, scaling, and elastic deformations. These preprocessing and augmentation steps are crucial for improving the robustness and generalization performance of the model. Beyond these, nnUNet also has the capability to select the most fitting loss function and optimizer settings based on the dataset's inherent attributes. For example, if there's a noticeable class imbalance within the dataset, a weighted version of loss may be automatically chosen to better train the model. This is further supplemented by nnUNet's hyperparameter tuning process that involves a grid search over a predefined set of crucial hyperparameters like learning rate and batch size. This comprehensive suite of automatic configuration features allows the nnSAM architecture to ensure that its encoder component is optimally set up for each specific medical imaging task, thus enhancing both its efficiency and effectiveness. In this context, the network architecture self-adjusts based on the dataset, aiming for optimal performance by dynamically modifying parameters such as layer count and convolutional kernel size. This eliminates the need for manual design, accelerating development timelines. Since the number of layers of the network is determined by the specific dataset, in Fig. 1 we number the encoder layers as E_t to E_0 and decoder layers as D_1 to D_t.

2.3 SAM Encoder

The SAM encoder is a pre-trained Vision Transformer model that has been trained on the extensive SA-1B segmentation dataset. Due to its training on a large dataset, the SAM encoder excels at feature extraction for segmentation tasks. However, its segmentation ability is highly prompt-dependent, meaning it struggles to independently identify what should be segmented. Therefore, we only utilize the SAM encoder for its feature extraction strengths, while leaving the task of identifying the region of interest (ROI) for segmentation to nnUNet. For an input image $x \in \mathbb{R}^{H \times W \times C}$, where $H \times W$ are the spatial dimensions and C is the number of channels, the SAM encoder needs the input $H \times W$ to be of size 1024×1024. To meet this requirement, we resize it to 1024×1024 using linear interpolation after the pre-processing of nnUNet. The SAM encoder produces an image embedding S with dimensions 64×64. We then resize this embedding S to match the dimensions of feature D_1 and concatenate them. Besides, in order to balance the inference speed of SAM with the powerful segmentation capability, we use MobileSAM [11,13], which is more than 60 times smaller than the original SAM, but has comparable performance. It is a lightweight version

of SAM obtained by distillation, where the knowledge from the original image encoder is transferred into the lightweight image encoder.

Table 1. DICE and ASD of different methods on different training sample sizes.

		4	**8**	**12**	**16**	**20**
UNet	DICE	59.31 ± 22.30	62.46 ± 20.37	66.86 ± 21.13	71.86 ± 11.27	76.52 ± 9.43
	ASD	19.54 ± 7.59	20.82 ± 7.16	17.86 ± 6.90	19.79 ± 6.08	18.15 ± 7.52
SwinUNet	DICE	81.24 ± 17.58	80.82 ± 14.06	83.15 ± 9.21	84.37 ± 6.91	86.88 ± 5.13
	ASD	4.79 ± 3.01	5.29 ± 4.29	3.61 ± 2.57	4.12 ± 2.94	4.17 ± 4.12
TransUNet	DICE	81.23 ± 6.62	82.34 ± 5.98	84.82 ± 4.80	87.05 ± 4.60	87.11 ± 3.99
	ASD	4.18 ± 1.90	8.79 ± 3.29	9.77 ± 2.94	11.03 ± 3.84	11.99 ± 3.30
AutoSAM	DICE	65.10 ± 23.62	65.89 ± 20.58	67.63 ± 21.77	77.55 ± 7.55	78.53 ± 8.30
	ASD	16.55 ± 7.83	19.60 ± 5.78	16.98 ± 5.96	16.73 ± 6.02	15.92 ± 6.36
nnUNet	DICE	81.77 ± 13.68	84.45 ± 18.72	88.36 ± 13.00	92.35 ± 7.55	93.15 ± 7.86
	ASD	6.97 ± 4.87	4.90 ± 6.08	3.15 ± 4.74	1.56 ± 2.26	1.40 ± 2.20
nnSAM	DICE	**84.67 ± 13.52**	**86.36 ± 16.19**	**90.74 ± 11.89**	**93.20 ± 5.53**	**93.75 ± 5.35**
	ASD	**3.87 ± 5.04**	**3.29 ± 5.15**	**2.18 ± 3.97**	**1.43 ± 1.69**	**1.23 ± 1.64**

3 Experimental Setting

We evaluated our model using the MM-WHS dataset [14]. The preprocessed data from CFDnet [15] was utilized, resulting in a collection of 212 cardiac CT images at a resolution of 240x220. For the purpose of assessing performance with varying few-shot training data sizes (ranging from 4 to 20 samples), we partitioned the dataset into 112 images for testing and 80 for validation. This allowed us to study how the performance of our model scales with the size of the labeled data available. The dataset contains labels for five different anatomical cardiac classes, facilitating a multi-class segmentation task. Specifically, the classes include the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), and myocardium of LV (Myo). Each of these classes presents its own unique challenges for segmentation, making the dataset particularly well-suited for testing the robustness and versatility of our model. Our experiments with different few-shot scenarios aim to simulate real-world clinical settings where labeled data might be expensive to obtain. Through this experimental setup, we hope to shed light on the trade-off between the quantity of training data and the performance of the model, particularly in the complex task of cardiac image segmentation involving multiple anatomical structures.

For SwinUNet [7], TransUNet [5], UNet [3], and nnUNet [8], we have taken widely used public codes, while for AutoSAM [16], since there is no official open-source code, we have reproduced it as much as possible based on the article descriptions. All methods were trained and tested on A100 80G. For the evaluation metric, we used Average Symmetric Surface Distance (ASD) and the Dice

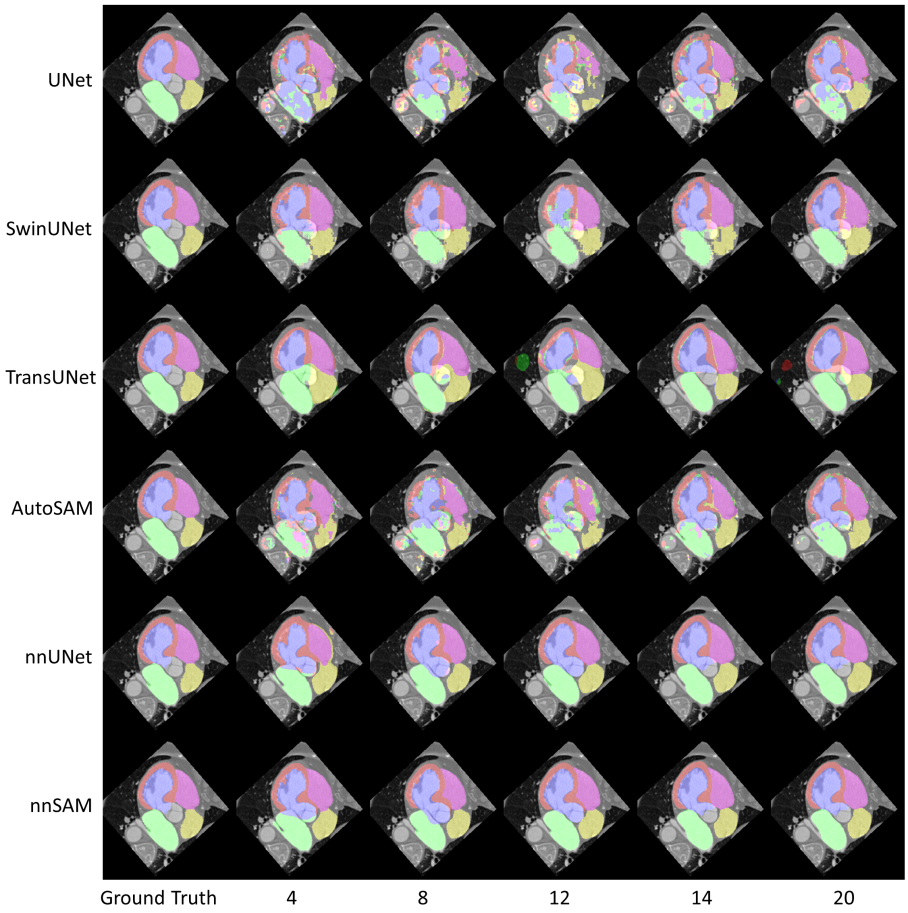

Fig. 2. Example 1 of segmentation visualization results for different methods on different numbers of training samples.

Similarity Coefficient (DICE) [17]. The ASD is a metric that quantifies the average distance between the surfaces of two segmented objects. DICE evaluates the similarity between two segmented objects, considering the volume overlap between the two objects.

4 Results

Table 1 demonstrates the model performance with different number of training data from 20 to 4 images. The proposed nnSAM outperforms other medical image segmentation model benchmarks in terms of DICE and ASD for all sample sizes. When trained with 20 labeled images, nnSAM achieves a Dice score of 93.75% and an ASD of 1.23. The segmentation accuracy of nnUNet, which is

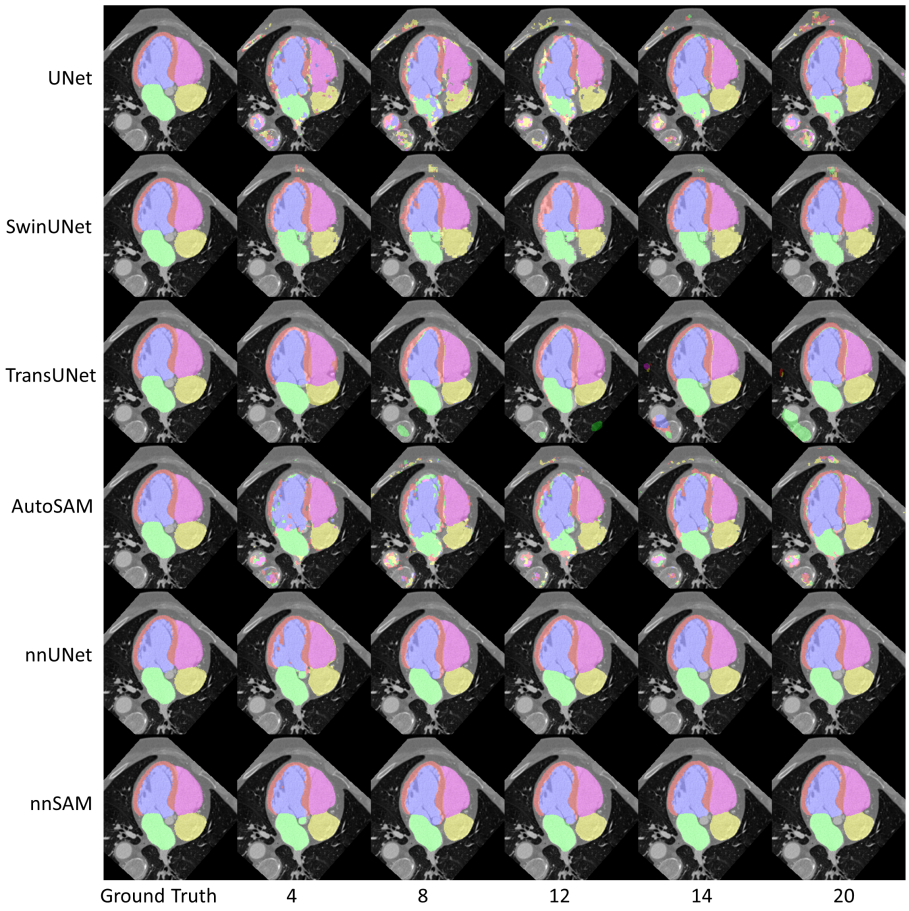

Fig. 3. Example 2 of segmentation visualization results for different methods on different numbers of training samples.

recognized as one of the best segmentation models, is also substantially ahead of the other methods, but slightly lower than that of nnSAM. Other methods, including SwinUNet and TransUNet, seem to have much lower accuracies, with a DICE below 90% and an average surface distance above 4. The worst performance of UNet may be due to the fact that both SwinUNet and TransUNet are based on strong pre-trained models that perform better on small samples, while UNet is trained from scratch. In addition, as the number of training samples gradually decreases, the advantage of nnSAM over other methods is further emphasized. In particular, when trained with only 4 labeled images, nnSAM outperforms the second-place nnUNet's DICE by 2.9%, and outperforms SwinUNet and TransUNet by more than 3%. Overall, nnSAM provides higher segmenta-

Table 2. DICE of different methods on different anatomical cardiac classes.

		Myo	LA	LV	RA	RV
Unet	4	50.65 ± 20.23	64.91 ± 25.10	68.33 ± 23.92	55.53 ± 24.40	57.10± 30.68
	8	49.51 ± 18.57	62.49 ± 19.24	74.21 ± 23.22	59.35 ± 22.65	66.75 ± 29.56
	12	51.15 ± 20.35	66.16 ± 21.71	78.47 ± 21.80	61.06 ± 23.48	77.47 ± 24.21
	16	61.95 ± 10.22	70.84 ± 13.12	80.60± 9.46	71.05 ± 17.22	74.85 ± 20.90
	20	63.02 ± 13.21	73.44 ± 11.32	84.57 ± 9.91	75.65 ± 13.60	85.94 ± 8.11
SwinUNet	4	67.56 ± 21.19	83.07 ± 20.08	**90.26 ± 9.21**	79.29 ± 22.43	86.01 ± 17.54
	8	64.56 ± 17.34	83.43 ± 14.34	88.02 ± 9.57	81.18 ± 17.75	86.92 ± 14.92
	12	65.15 ± 17.08	85.97 ± 10.08	88.38 ± 7.82	85.35 ± 8.84	90.9 ± 5.71
	16	73.19 ± 10.44	84.55 ± 7.78	90.88 ± 4.51	83.27 ± 9.72	89.94 ± 6.05
	20	76.14 ± 9.55	87.23 ± 6.87	92.26 ± 3.81	87.07 ± 6.79	91.69 ± 4.0
TransUNet	4	67.30 ± 9.79	86.08 ± 11.45	89.71 ± 4.06	**81.0 ± 9.79**	82.08 ± 6.55
	8	68.16 ± 8.19	84.53 ± 8.04	87.56 ± 4.59	83.52 ± 9.80	87.94 ± 5.85
	12	70.24 ± 8.0	85.17 ± 9.71	89.43 ± 4.68	88.52 ± 6.11	90.76 ± 3.86
	16	76.61 ± 5.95	87.51 ± 9.34	91.63 ± 3.38	88.08 ± 6.65	91.42 ± 3.16
	20	77.83 ± 5.55	85.32 ± 10.48	92.41 ± 3.57	88.74 ± 4.56	91.23 ± 2.33
AutoSAM	4	54.17 ± 22.87	63.56 ± 22.45	78.28 ± 26.70	57.39 ± 26.15	72.09 ± 27.96
	8	52.98 ± 19.73	61.86 ± 18.86	78.82 ± 22.03	58.03 ± 24.23	77.76 ± 25.14
	12	53.53 ± 20.31	64.71 ± 21.73	78.96 ± 22.67	63.83 ± 26.64	77.15 ± 25.15
	16	65.35 ± 10.58	77.33 ± 9.33	86.75 ± 6.73	72.44 ± 14.53	85.88 ± 7.07
	20	67.22 ± 11.85	77.04 ± 10.33	86.77 ± 7.76	74.85 ± 14.22	86.76 ± 7.67
nnUNet	4	72.24 ± 13.01	83.69 ± 16.91	88.43 ± 11.24	78.10± 18.60	86.38 ± 15.40
	8	75.31 ± 19.25	88.12 ± 18.91	87.9 ± 19.63	82.37 ± 22.17	88.57 ± 19.31
	12	82.78 ± 12.02	93.37 ± 5.10	91.23 ± 16.82	84.41 ± 20.33	90.01 ± 14.40
	16	88.66 ± 5.01	94.95 ± 4.05	94.0 ± 13.62	90.79 ± 10.28	93.35 ± 10.04
	20	89.88 ± 4.74	96.03 ± 1.66	94.53 ± 14.29	91.69 ± 9.85	93.62 ± 11.34
nnSAM	4	**77.05 ± 14.47**	**88.67 ± 10.53**	89.93 ± 11.97	80.86 ± 21.13	**86.83 ± 14.75**
	8	**76.45 ± 17.03**	**91.48 ± 14.61**	**89.68 ± 18.05**	**84.29 ± 19.73**	**89.9 ± 16.92**
	12	**86.40 ± 9.69**	**94.89 ± 4.71**	**92.20 ± 16.65**	**88.76 ± 16.44**	**91.45 ± 14.27**
	16	**89.76 ± 3.12**	**95.44 ± 4.95**	**94.78 ± 10.85**	**92.26 ± 7.79**	**93.74 ± 9.72**
	20	**90.04 ± 3.46**	**96.08 ± 2.05**	**95.43 ± 9.70**	**92.69 ± 7.23**	**94.53 ± 8.39**

tion accuracy compared to other methods when the amount of training data is limited.

Table 2 and Table 3 show the performance of DICE and ASD specifically under each category of labeling, from which it can be seen that our nnSAM achieves optimal results in the vast majority of cases. However, there are some data that are more counterintuitive, such as SwinUNet and TransUNet on the LA category where the ASD becomes larger as the sample data increases. The reasons may be as shown in Fig. 2 and Fig. 3, we found that TransUNet and SwinUNet show more false positive results as the sample size increases, and all of these results are far away from the correct segmentation position, leading to anomalous results in ASD. Besides, UNet and AutoSAM generate poor segmentation results, the myocardium of LV (Myo) is almost unrecognizable when the amount of training data is limited, suggesting that they do not cope well with the small

Table 3. ASD of different methods on different anatomical cardiac classes.

		Myo	LA	LV	RA	RV
UNet	4	16.46 ± 6.95	17.07 ± 7.85	20.83 ± 8.08	27.20 ± 11.72	16.11 ± 9.29
	8	19.33 ± 6.43	23.09 ± 9.35	17.35 ± 7.02	28.35 ± 10.84	16.0 ± 10.98
	12	13.92 ± 5.43	20.47 ± 8.74	12.81 ± 5.18	28.9 ± 12.17	13.20 ± 9.9
	16	16.52 ± 5.73	28.38 ± 9.26	16.67 ± 4.24	25.13 ± 11.38	12.25 ± 8.91
	20	16.36 ± 5.85	22.09 ± 9.34	13.39 ± 5.67	26.44 ± 12.58	12.48 ± 11.49
SwinUNet	4	4.72 ± 2.89	5.58 ± 3.99	**2.78 ± 2.01**	6.89 ± 5.20	4.0 ± 4.46
	8	3.41 ± 2.21	5.47 ± 6.88	3.45 ± 2.77	10.07 ± 10.25	4.04 ± 3.65
	12	2.85 ± 1.72	5.12 ± 5.45	2.99 ± 1.88	5.07 ± 3.80	**2.03 ± 1.55**
	16	2.51 ± 1.63	5.19 ± 6.24	2.39 ± 1.13	8.24 ± 6.32	2.28 ± 1.57
	20	2.28 ± 1.54	7.72 ± 13.93	2.49 ± 1.16	6.17 ± 5.34	2.18 ± 1.93
TransUNet	4	5.27 ± 1.98	4.77 ± 5.30	2.86 ± 1.01	**3.81 ± 2.42**	4.21 ± 1.79
	8	5.43 ± 2.66	14.43 ± 7.15	6.83 ± 5.17	14.54 ± 3.11	2.73 ± 1.89
	12	6.61 ± 3.41	21.03 ± 8.27	6.63 ± 2.9	12.06 ± 3.18	2.51 ± 2.38
	16	8.22 ± 5.75	19.16 ± 7.39	10.27 ± 5.60	15.08 ± 2.39	2.40 ± 3.14
	20	7.62 ± 3.78	24.38 ± 7.94	7.84 ± 4.73	18.17 ± 3.81	1.91 ± 1.01
AutoSAM	4	14.18 ± 7.62	23.89 ± 11.01	10.33 ± 7.67	20.66 ± 11.12	13.68 ± 10.81
	8	19.14 ± 6.86	26.87 ± 9.33	15.10± 5.75	25.18 ± 11.23	11.72 ± 7.97
	12	16.31 ± 6.92	23.55 ± 8.9	11.91 ± 6.02	19.42 ± 9.73	13.72 ± 9.98
	16	14.02 ± 5.23	26.80 ± 9.51	11.24 ± 5.82	22.45 ± 9.80	9.15 ± 7.64
	20	12.24 ± 4.69	26.81 ± 10.99	9.11 ± 4.43	23.80 ± 10.81	7.62 ± 7.49
nnUNet	4	8.19 ± 3.39	4.26 ± 4.77	3.98 ± 4.11	15.38 ± 12.83	3.06 ± 4.50
	8	3.06 ± 4.41	5.41 ± 9.40	3.97 ± 5.63	8.84 ± 11.64	3.23 ± 5.70
	12	2.09 ± 4.13	2.36 ± 4.40	2.43 ± 4.19	6.56 ± 11.39	2.33 ± 3.66
	16	1.12 ± 1.72	1.17 ± 2.22	**1.47 ± 2.43**	2.37 ± 3.99	1.64 ± 2.95
	20	1.01 ± 1.60	**0.78 ± 0.27**	1.50± 3.31	2.08 ± 3.27	1.60± 3.53
nnSAM	4	**2.73 ± 3.37**	**3.43 ± 4.46**	3.47 ± 4.22	6.61 ± 11.05	**3.13 ± 4.48**
	8	**2.77 ± 3.88**	**2.63 ± 6.43**	**3.21 ± 4.88**	**5.29 ± 8.45**	**2.57 ± 4.97**
	12	**1.71 ± 3.22**	**1.07 ± 0.84**	**2.34 ± 4.21**	**3.65 ± 8.38**	2.16 ± 4.31
	16	**1.03 ± 1.17**	**1.10± 2.96**	1.65 ± 2.46	**1.87 ± 2.93**	**1.49 ± 2.73**
	20	**0.9 ± 1.00**	**0.78 ± 0.34**	**1.25 ± 2.20**	**1.77 ± 2.86**	**1.48 ± 3.22**

sample task. The segmentation obtained by TransUNet looks comparable to that obtained by SwinUNet, and the boundary of TransUNet is smoother than that obtained with SwinUNet obtains smoother boundaries, which may be attributed to the structure of the Shifted window characteristic of SwinUNet. nnUNet has the most nnSAM results, but specifically in the details of the segmentation results, e.g., Myo in Fig. 2, nnUNet has more false positives. These results suggest that nnSAM provides superior accuracy in segmenting challenging targets with only a small number of training samples, which may be attributed to the strong generality of the SAM encoder and the power of nnUNet's auto-configuration framework.

5 Conclusion

We introduce nnSAM, a novel, few-shot learning solution for medical image segmentation that melds the strengths of the Segment Anything Model (SAM) and nnUNet. Our extensive evaluation across different numbers of 2D training samples sets a new benchmark in medical image segmentation, especially in scenarios where training data is scarce. The results also highlight the robustness and adaptability of nnSAM, making it a promising tool for future research and practical applications in medical imaging.

References

1. Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021.
2. Berkman Sahiner, Aria Pezeshk, Lubomir M Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H Cha, Ronald M Summers, and Maryellen L Giger. Deep learning in medical imaging and radiation therapy. *Medical physics*, 46(1):e1–e36, 2019.
3. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.
4. Yunxiang Li, Shuai Wang, Jun Wang, Guodong Zeng, Wenjun Liu, Qianni Zhang, Qun Jin, and Yaqi Wang. Gt u-net: A u-net like group transformer network for tooth root segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 386–395. Springer, 2021.
5. Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
6. Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.
7. Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.
8. Fabian Isensee, Paul F Jaeger, Simon A A Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 17(2):203–211, 2020.
9. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
10. Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.
11. Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.
12. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

13. Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision*, pages 68–85. Springer, 2022.
14. Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58:101537, 2019.
15. Fuping Wu and Xiahai Zhuang. Cf distance: a new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):4274–4285, 2020.
16. Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370*, 2023.
17. Yunxiang Li, Guodong Zeng, Yifan Zhang, Jun Wang, Qun Jin, Lingling Sun, Qianni Zhang, Qisi Lian, Guiping Qian, Neng Xia, et al. Agmb-transformer: Anatomy-guided multi-branch transformer network for automated evaluation of root canal therapy. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1684–1695, 2021.