TTT-UNet: Enhancing U-Net with Test-Time Training Layers for Biomedical Image Segmentation

Rong Zhou[1*] **Zhengqing Yuan**[3*] **Zhiling Yan**[1*]
Weixiang Sun[2*] **Kai Zhang**[1] **Yiwei Li**[4]
Yanfang Ye[3] **Xiang Li**[5] **Lifang He**[1] **Lichao Sun**[1]
[1]Lehigh University, [2]Northeastern University, [3]University of Notre Dame,
[4]University of Georgia, [5]Massachusetts General Hospital and Harvard Medical School

Abstract

Biomedical image segmentation is crucial for accurately diagnosing and analyzing various diseases. However, Convolutional Neural Networks (CNNs) and Transformers, the most commonly used architectures for this task, struggle to effectively capture long-range dependencies due to the inherent locality of CNNs and the computational complexity of Transformers. To address this limitation, we introduce TTT-UNet, a novel framework that integrates Test-Time Training (TTT) layers into the traditional U-Net architecture for biomedical image segmentation. TTT-UNet dynamically adjusts model parameters during the testing time, enhancing the model's ability to capture both local and long-range features. We evaluate TTT-UNet on multiple medical imaging datasets, including 3D abdominal organ segmentation in CT and MR images, instrument segmentation in endoscopy images, and cell segmentation in microscopy images. The results demonstrate that TTT-UNet consistently outperforms state-of-the-art CNN-based and Transformer-based segmentation models across all tasks. The code is available at `https://github.com/rongzhou7/TTT-UNet`.

1 Introduction

Image segmentation plays a crucial role in medical imaging, as it empowers medical professionals to identify biological structures and measure their morphology, aiding in the analysis and diagnosis of various diseases [1, 2]. In recent years, convolutional neural networks (CNNs) [3] have emerged as a promising approach in the field of biomedical image segmentation. Among various CNN-based techniques, U-Net [4] stands out for its straightforward structure and significant adaptability. Many enhancements and iterations [5–10] have been developed based on this U-shaped architecture, typically featuring a symmetric encoder-decoder design to capture multi-scale image features through convolutional operations. Leveraging this foundation, significant advancements have been achieved across a wide range of medical imaging applications [11–18]. These include cardiac segmentation in magnetic resonance (MR) imaging [19], multi-organ delineation in computed tomography (CT) scans [20], and others [21, 22].

Despite the remarkable representational capabilities of CNN-based models, their architectural design exhibits an inherent limitation in modeling long-range dependencies within images, because convolutional kernels are inherently local [23]. While skip connections in the U-Net architecture facilitate the merging of low-level details with high-level features, they mainly serve to directly merge local features, which does not substantially boost the network's ability to model long-range dependencies.

*Equal contribution. Contact emails: `roz322@lehigh.edu, zyuan2@nd.edu, zhy423@lehigh.edu, kennysweson@gmail.com`.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

This limitation becomes especially evident in scenarios with large inter-patient variations in shape, size, etc [23]. Such variability poses challenges to the CNN framework's ability to consistently and accurately capture information across extended spatial contexts, highlighting the need for innovative approaches to address this fundamental constraint.

Recognizing the limitations of CNNs in capturing long-range dependencies, the research community has shifted interest towards Transformer models for their ability to naturally understand global contexts [24]. This transition is evidenced in biomedical image segmentation, where approaches like TransUNet [25], UNETR [9], SwinUNETR [10] demonstrate the potential of integrating Transformers. These hybrid models that blend CNNs for high-resolution spatial detail and Transformers for global context emerge as a more effective strategy.

Despite their ability to capture global dependencies, Transformers are computationally intensive [26], especially in dense biomedical image segmentation tasks. Mamba [26], a state-space model designed for efficient sequence modeling, offers a more computationally efficient approach to long-range dependency modeling. Building on this, U-Mamba [27] integrates Mamba within U-Net, effectively combining high-resolution spatial detail with long-range dependency modeling to enhance biomedical image segmentation. Despite these advancements, U-Mamba and similar models, still face challenges in expressiveness, particularly over extended contexts, where their fixed-size hidden states limit their ability to capture complex and nuanced dependencies.

Recently, TTT (Test-Time Training) [28] have emerged as a new class of sequence modeling layers with linear complexity and an expressive hidden state. TTT treats the traditional fixed hidden state as a machine learning model itself, which can be dynamically updated through self-supervised learning. This dynamic adjustment allows the model to refine its parameters based on test data, providing greater flexibility and expressiveness in capturing intricate long-range dependencies. In comparison to Transformers and Mamba, TTT layers not only maintain efficiency but also offer superior performance in handling long-context sequences.

In this paper, we introduce TTT-UNet, a novel hybrid architecture that incorporates TTT layers within the traditional U-Net framework to address the inherent limitations in modeling long-range dependencies in biomedical image segmentation tasks. The TTT layers dynamically adapt its parameters during test time, allowing it to more effectively capture both localized details and long-range dependencies. Our extensive experiments across various medical imaging datasets demonstrate that TTT-UNet consistently outperforms existing state-of-the-art models. The results highlight the model's effectiveness in handling complex anatomical structures and its robustness in diverse clinical scenarios. Particularly, TTT-UNet has shown significant improvements in biomedical image segmentation tasks, making it a versatile solution for medical image analysis. Our contributions are summarized as follows:

- We introduce TTT-UNet, an enhanced U-Net architecture integrated with TTT layers, which allows the model to perform self-supervised adaptation during test time. This hybrid design effectively tackles the challenge of modeling long-range dependencies and improves the model's generalization capability across diverse data distributions.
- TTT-UNet has been rigorously evaluated on a diverse set of medical imaging datasets, including 3D abdominal organ segmentation in CT and MRI scans, instrument segmentation in endoscopy images, and cell segmentation in microscopy images. The results demonstrate consistent improvements over state-of-the-art models in both 3D and 2D segmentation.

In summary, TTT-UNet represents a significant advancement in biomedical image segmentation, offering a robust and adaptable approach that leverages the strengths of CNNs and TTT layers. This work lays the foundation for future developments in adaptive and context-aware medical image analysis technologies.

2 Related work

2.1 U-Net and variants

CNN-based and Transformer-based models have significantly advanced the field of biomedical image segmentation. U-Net [4], a representative among CNN-based approaches, features a symmetrical encoder-decoder architecture enhanced with skip connections to better preserve details. Various

enhancements [29], such as the self-configuring nnU-Net [30] framework, have been built on this U-shaped design, demonstrating robust performance across a variety of biomedical image segmentation challenges. For Transformer, TransUnet [25] stands out by integrating the Vision Transformer (ViT) [31] for feature extraction in the encoding phase and coupling it with CNN for decoding, demonstrating its capability for processing global information. Swin-UNETR [10] and UNETR [9] blend Transformer architectures with traditional U-Net to enhance 3D imaging analysis. Additionally, Swin-UNet [6] delves into the use of Swin Vision Transformer blocks [32] within a U-Net framework, further expanding the exploration of Transformer technology in medical imaging.

2.2 Hybrid models

SSMs, such as Mamba, have recently gained prominence as a powerful component for developing deep networks, achieving cutting-edge performance in analyzing long-sequence data [33, 34]. In the realm of biomedical image segmentation, U-Mamba [27] presents a novel SSM-CNN hybrid approach, signifying the first application of SSMs in the medical image domain. Further developments include SegMamba [35] and nnMamba [36], which combine SSMs in the encoder with CNNs in the decoder, illustrating the versatility and effectiveness of SSMs in enhancing medical imaging analysis.

3 Method

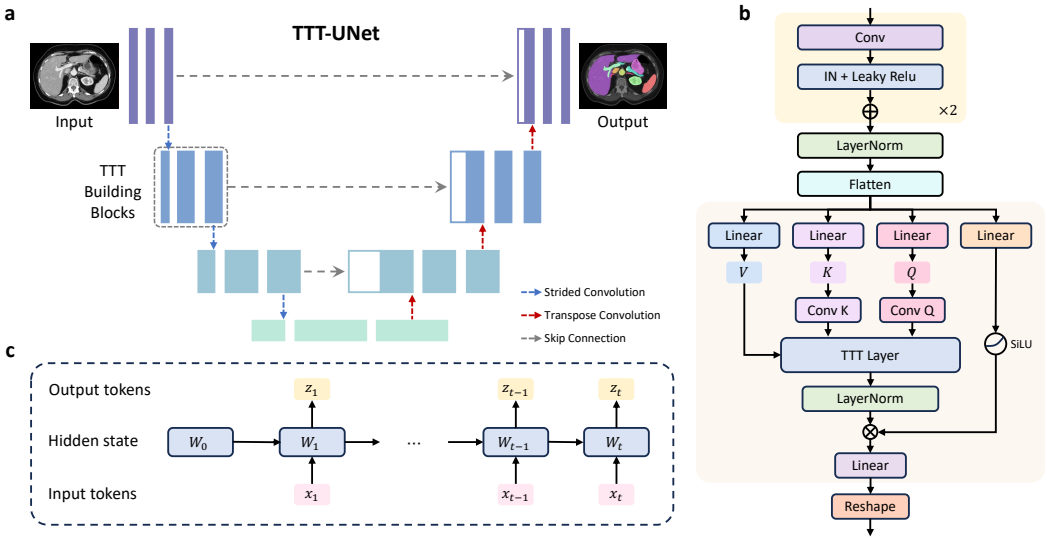

Figure 1: **(a)** The overall framework of TTT-UNet. **(b)** TTT Building Block. **(c)** TTT Layer.

TTT-UNet follows the conventional U-Net structure, designed to effectively capture both local features and long-range dependencies. As shown in Figure 1, TTT-UNet integrates Test-Time Training (TTT) layers into the Mamba blocks within the U-Net network. This integration enables the model to continuously update its parameters based on test data, enhancing its feature extraction capabilities in the encoder and allowing it to adaptively learn long-range dependencies. Subsequently, we introduce the TTT layer and then describe how it is integrated into the Mamba blocks within the U-Net architecture.

3.1 TTT layers

Traditional sequence modeling layers, such as RNNs, compress the context of a sequence $x_1, \ldots, x_t$ into a fixed-size hidden state h_t. For RNNs, the hidden state h_t at time step t is updated based on the current input x_t and the previous hidden state h_{t-1} through linear transformation matrices θ_h and θ_x and a non-linear activation function σ:

$$h_t = \sigma(\theta_h h_{t-1} + \theta_x x_t),$$

where θ_h and θ_x are learned parameters. The output z_t is then generated from the hidden state:

$$z_t = \phi(h_t),$$

where ϕ represents a linear or non-linear transformation.

However, the fixed size of the hidden state limits performance when dealing with long contexts due to its finite capacity to represent contextual information. To address this limitation, a new class of sequence modeling layers, referred to as TTT layers [28] is introduced, where the hidden state is treated as a trainable model and is updated through self-supervised learning.

Specifically, in a TTT layer (**Fig. 1c**) , the hidden state h_t at time step t is treated as a trainable model f with weights W_t, which is updated based on the current input x_t:

$$W_t = W_{t-1} - \eta \nabla \ell(W_{t-1}; x_t)$$

The output token z_t is then generated using trainable model f with weights W_t:

$$z_t = f(x_t; W_t)$$

In the basic naive version, the self-supervised loss ℓ aims to reconstruct the corrupted input $\tilde{x}_t$. This approach is straightforward and focuses on learning to recover the original input from its corrupted version:

$$\ell(W; x_t) = \|f(\tilde{x}_t; W) - x_t\|^2$$

While this naive reconstruction method is effective in certain scenarios, it has inherent limitations in capturing the complex dependencies within the input data, especially in tasks requiring a more nuanced understanding of the input context.

To address these limitations, we follow a more sophisticated self-supervised task that leverages multiple views of the input data. Instead of directly reconstructing the corrupted input, we introduce learnable matrices θ_K and θ_V to project the input into different views. The *training view* $K = \theta_K x_t$ captures the essential information needed for learning, while the *label view* $V = \theta_V x_t$ provides a target for reconstruction:

$$\ell(W; x_t) = \|f(\theta_K x_t; W) - \theta_V x_t\|^2$$

This approach allows the model to selectively focus on the most relevant features of the input, improving its ability to capture long-range dependencies and subtle relationships within the data.

The output token z_t is then generated as follow:

$$z_t = f(\theta_Q x_t; W_t),$$

where f is a function parameterized by W_t, which can be a linear model or a multi-layer perceptron (MLP). Here, the projection θ_Q is used to obtain the test view $Q = \theta_Q x_t$, which introduces additional flexibility by allowing the model to emphasize different aspects of the input data during inference. This approach enables the model to focus on the most informative features in the context of the current test case, thereby enhancing its ability to adapt to new, unseen data.

3.2 TTT-UNet architecture

As shown in **Fig. 1**, the TTT-UNet architecture integrates the traditional U-Net structure with TTT layers, allowing the network to adapt during testing through self-supervised learning dynamically. The architecture is composed of an encoder-decoder structure, where the encoder is enhanced with TTT building blocks to improve adaptability, while the decoder follows the standard U-Net design focused on reconstructing the segmentation map.

Encoder. The encoder in TTT-UNet follows the traditional U-Net design, comprising multiple convolutional layers. These layers are interspersed with TTT building blocks, which are critical components that enable the model to adjust its parameters dynamically during test time. Each layer in the encoder progressively downscales the input image while capturing both local and long-range features essential for segmentation tasks. Including TTT building blocks within the encoder ensures the model can adapt to varying data distributions encountered during testing.

TTT building blocks. As illustrated in **Fig. 1b**, the TTT building blocks are the core components that allow for the test-time adaptability of the model. Initially, the input features pass through two successive Residual blocks [37], each comprising a standard convolutional layer, followed by Instance

Normalization (IN) [38] and Leaky ReLU activation [39]. Subsequently, the features are normalized using Layer Normalization [40], and flattened, making them suitable for linear transformations. Then the flattened features undergo three separate linear transformation branches, obtaining the different features denoted as V, K, and Q respectively. Additional convolutional operations (`Conv` K and `Conv` Q) are applied to the K and Q vectors, allowing the model to focus on specific aspects of the features during test-time training. Meanwhile, the fourth branch performs a linear transformation followed by a SiLU activation function [41], further enriching the feature representations. Then the processed V, K, and Q are fed into the TTT Layer, where self-supervised learning occurs. In this layer, the model dynamically updates its weights based on the self-supervised task applied to the processed V, K, and Q vectors, as detailed in 3.1. The output from the TTT Layer is further normalized using Layer Normalization [40] before being passed on. Finally, this output and the fourth branch output mentioned before are combined via the Hadamard product, followed by a linear transformation and reshaping to fit the required dimensions for subsequent layers in Decoder.

Decoder The decoder in our model maintains the classic U-Net structure, integrating Residual blocks and transposed convolutions to enhance the capture of detailed local features and support resolution recovery. We also incorporate the skip connections in U-Net, ensuring the effective transfer of hierarchical features from the encoder to the decoder. The final output of the decoder is refined through a $1 \times 1 \times 1$ convolutional layer and a Softmax activation, which generates the final segmentation probability map.

Furthermore, we implement two variants of TTT-UNet: TTT-UNet_Bot and TTT-UNet_Enc. In TTT-UNet_Bot, TTT layers are applied solely in the bottleneck, while the rest of the architecture consists of standard Residual blocks. In TTT-UNet_Enc, TTT layers are incorporated throughout the encoder, allowing for a broader integration of self-supervised adaptation.

4 Experiments

4.1 Datasets

To evaluate the performance and scalability of TTT-UNet, we utilize four biomedical image datasets across a variety of segmentation tasks and imaging modalities, including Abdomen CT dataset [42], Abdomen MRI dataset [43], Endoscopy dataset [44] and Microscopy dataset [45]. The basic information of these datasets is summarized in Table 1.

Table 1: Dataset information.

Dataset	Dimension	#Training Image	#Testing Image	#Targets
Abdomen CT	3D	50 (4794 slices)	50 (10894 slices)	13
Abdomen MRI	3D	60 (5615 slices)	50 (3357 slices)	13
Endoscopy images	2D	1800	1200	7
Microscopy images	2D	1000	101	2

Abdomen CT. The Abdomen CT [42] dataset, from the MICCAI 2022 FLARE challenge, includes the segmentation of 13 abdominal organs from 50 CT scans in both the training and testing sets. The organs segmented include the liver, spleen, pancreas, kidneys, stomach, gallbladder, esophagus, aorta, inferior vena cava, adrenal glands, and duodenum.

Abdomen MRI. The Abdomen MR [43] dataset, from the MICCAI 2022 AMOS Challenge, focuses on the segmentation of the same 13 abdominal organs, using MRI scans. It consists of 60 MRI scans for training and 50 for testing. Additionally, we generate a 2D version of this dataset by converting the 3D abdominal MRI scans into 2D slices. This conversion enables us to evaluate TTT-UNet under the common 2D segmentation setting, which is widely used in practice due to its lower computational requirements. The conversion retains the same 13 organs, ensuring consistent evaluation across both 2D and 3D modalities.

Endoscopy images. From the MICCAI 2017 EndoVis Challenge [44], this dataset focuses on instrument segmentation within endoscopy images, featuring seven distinct instruments, including the large needle driver, prograsp forceps, monopolar curved scissors, cadiere forceps, bipolar forceps,

vessel sealer, and a drop-in ultrasound probe. The dataset is split into 1800 training frames and 1200 testing frames.

Microscopy images. This dataset, from the NeurIPS 2022 Cell Segmentation Challenge [45], is used for cell segmentation in microscopy images, consisting of 1000 training images and 101 testing images. Following U-Mamba [27], we address this as a semantic segmentation task, focusing on cell boundaries and interiors rather than instance segmentation.

4.2 Experimental setup

Table 2: TTT-UNet configurations for each dataset.

Configurations	**Patch size**	**Batch size**	**# Stages**	**# Pooling per axis**
Abdomen CT	(40, 224, 192)	2	6	(3, 3, 5)
3D Abdomen MR	(48, 160, 224)	2	6	(3, 5, 5)
2D Abdomen MR	(320, 320)	30	7	(6, 6)
Endoscopy	(384, 640)	13	7	(6, 6)
Microscopy	(512, 512)	12	8	(7, 7)

The setting of our experiments is the same as that in U-Mamba [27] and nnU-Net [30] to ensure a fair comparison, as shown in Table 2 We adopt an unweighted combination of Dice loss and cross-entropy loss for all datasets and utilize the SGD optimizer with an initial learning rate of 1e-2. The training duration for each dataset is set to 1000 epochs, conducted on a single NVIDIA A100 GPU. Leveraging the self-configuring capabilities from nnU-Net, the number of network blocks adjusts automatically according to the dataset. For evaluation metrics, we employ the Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) to assess performance in abdominal multi-organ segmentation for MR scans, as well as instrument segmentation in Endoscopy images. For the cell segmentation task, we utilize the F1 score to evaluate method performance.

4.3 Baselines and metrics

In our evaluation of TTT-UNet, we compare against two prominent CNN-based segmentation methods: nnU-NET [30] and SegResNet [29]. Additionally, we include a comparison with UNETR [9] and SwinUNETR [10], a Transformer-based network that has gained popularity in biomedical image segmentation tasks. U-Mamba [27], a recent method based on the Mamba model, is also included in our comparison to provide a comprehensive overview of its performance. For each model, we implement their recommended optimizers to ensure consistency in training conditions. To maintain fairness across all comparisons, we apply the default image preprocessing in nnU-NET [30].

4.4 Quantitative segmentation results

Table 3: Results summary of 2D segmentation tasks: organ segmentation in abdomen MRI scans, instruments segmentation in endoscopy images, and cell segmentation in microscopy images.

Methods	Organs in Abdomem MRI		Instruments in Endoscopy		Cells in Microscopy
	DSC	NSD	DSC	NSD	F1
nnU-Net	0.7450±0.1117	0.8153±0.1145	0.6264±0.3024	0.6412±0.3074	0.5383±0.2657
SegResNet	0.7317±0.1379	0.8034±0.1386	0.5820±0.3268	0.5968±0.3303	0.5411±0.2633
UNETR	0.5747±0.1672	0.6309±0.1858	0.5017±0.3201	0.5168±0.3235	0.4357±0.2572
SwinUNETR	0.7028±0.1348	0.7669±0.1442	0.5528±0.3089	0.5683±0.3123	0.3967±0.2621
U-Mamba_Bot	0.7588±0.1051	0.8285±0.1074	0.6540±0.3008	0.6692±0.3050	0.5389±0.2817
U-Mamba_Enc	0.7625±0.1082	0.8327±0.1087	0.6303±0.3067	0.6451±0.3104	0.5607±0.2784
TTT-UNet_Bot	**0.7750±0.1022**	**0.8452±0.1080**	**0.6643±0.3018**	**0.6799±0.3056**	**0.5818±0.2410**
TTT-UNet_Enc	**0.7725±0.1044**	**0.8540±0.1032**	**0.6696±0.3018**	**0.6820±0.3080**	**0.5773±0.2435**

Table 3 presents the results of 2D segmentation tasks, comparing the performance of various models across three datasets: organ segmentation in Abdomen MRI, instrument segmentation in endoscopy images, and cell segmentation in microscopy images. For the organ segmentation task in Abdomen MRI, TTT-UNet models significantly outperformed other methods. The TTT-UNet_Bot variant

achieved the highest DSC of 0.7750±0.1022 and NSD of 0.8452±0.1080, while TTT-UNet_Enc closely followed with a DSC of 0.7725±0.1044 and an NSD of 0.8540±0.1032. These results suggest that the TTT-UNet's ability to adapt its parameters at test time provides a considerable advantage in accurately segmenting organs, where anatomical variability is common. This adaptability likely contributes to the model's superior performance, enabling it to better generalize across different patient scans and handle the complexities of MRI data.

In the instrument segmentation task for endoscopy images, TTT-UNet variants again demonstrated superior performance, with TTT-UNet_Bot achieving a DSC of 0.6643±0.3018 and NSD of 0.6799±0.3056. The TTT-UNet_Enc model further improved these metrics, with a DSC of 0.6696±0.3018 and NSD of 0.6820±0.3080. These results indicate that the TTT layers within the model effectively capture the fine details of surgical instruments, which are often challenging to segment due to their small size and variability in appearance. The test-time adaptation provided by TTT layers helps in refining the segmentation boundaries, making the model more precise in instrument delineation.

For the cell segmentation task in microscopy images, the TTT-UNet models once again outperformed their counterparts. The TTT-UNet_Bot achieved the highest F1 score of 0.5818±0.2410, followed by TTT-UNet_Enc with an F1 score of 0.5773±0.2435. The superior performance in this task highlights the robustness of TTT-UNet in handling high variability and noise in microscopy data. The ability to dynamically adjust to new test samples allows the model to focus on relevant features, thereby improving segmentation accuracy even in challenging scenarios like cell segmentation.

Table 4: Results summary of 3D organ segmentation on abdomen CT and MRI datasets.

Methods	Organs in Abdomen CT		Organs in Abdomen MRI	
	DSC	NSD	DSC	NSD
nnU-Net	0.8615±0.0790	0.8972±0.0824	0.8309±0.0769	0.8996±0.0729
SegResNet	0.7927±0.1162	0.8257±0.1194	0.8146±0.0959	0.8841±0.0917
UNETR	0.6824±0.1506	0.7004±0.1577	0.6867±0.1488	0.7440±0.1627
SwinUNETR	0.7594±0.1095	0.7663±0.1190	0.7565±0.1394	0.8218±0.1409
U-Mamba_Bot	0.8683±0.0808	**0.9049±0.0821**	0.8453±0.0673	0.9121±0.0634
U-Mamba_Enc	0.8638±0.0908	0.8980±0.0921	0.8501±0.0732	0.9171±0.0689
TTT-UNet_Bot	**0.8709±0.1011**	0.8995±0.0721	**0.8677±0.0482**	**0.9247±0.0631**

Table 4 presents the results of 3D organ segmentation tasks on the Abdomen CT and Abdomen MRI datasets, comparing the performance of several state-of-the-art models. For the Abdomen CT dataset, TTT-UNet_Bot achieves the highest DSC score of 0.8709±0.1011, slightly outperforming U-Mamba_Bot (0.8683±0.0808) and U-Mamba_Enc (0.8638±0.0908). The small variance in TTT-UNet_Bot's performance suggests that the model is not only accurate but also consistent across different test samples. Additionally, the NSD score of 0.8995±0.0721 for TTT-UNet_Bot further supports its ability to preserve organ shapes and boundaries effectively.

For the Abdomen MRI dataset, TTT-UNet_Bot again demonstrates superior performance, achieving a DSC of 0.8677±0.0482 and an NSD of 0.9247±0.0631, outperforming all other models. The small variance in the DSC and NSD scores highlights the robustness of TTT-UNet_Bot, suggesting that it generalizes well across different MRI samples, even in the presence of anatomical variability and challenging contrasts.

Overall, TTT-UNet_Bot consistently outperforms other models in both CT and MRI segmentation tasks, not only achieving higher mean performance scores but also demonstrating lower variance, which indicates stable and reliable segmentation results. The test-time adaptation enabled by TTT layers plays a key role in this enhanced performance, allowing the model to effectively handle the complexities of 3D biomedical image segmentation.

4.5 Qualitative segmentation results

As shown in **Figure 2**, the segmentation results on the Abdomen MRI dataset reveal the effectiveness of TTT-UNet in handling complex anatomical structures. The comparison between the ground truth labels and the TTT-UNet predictions indicates a strong alignment, particularly in regions with significant anatomical variability. This suggests that TTT-UNet's ability to adapt its parameters

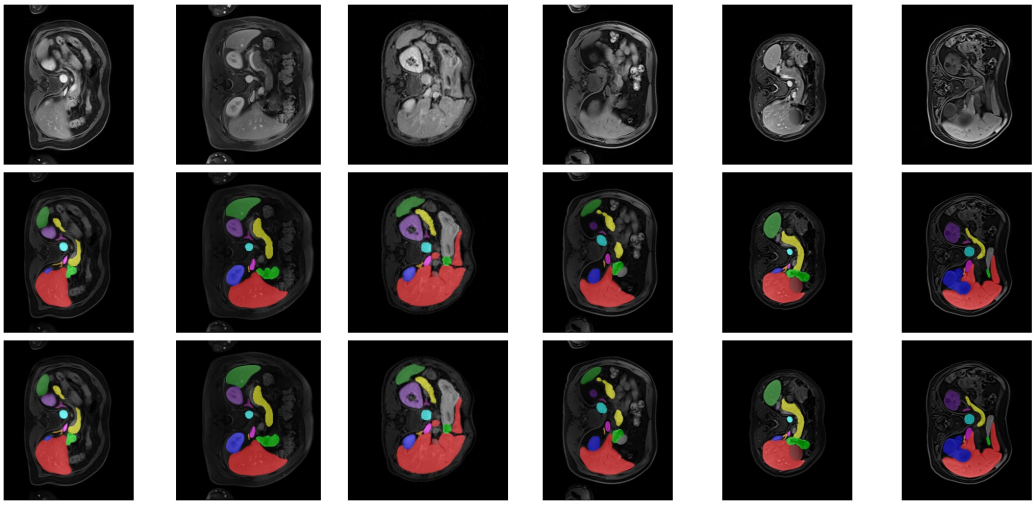

Figure 2: The visualization results of TTT-UNet on Abdomen MRI datasets. The first row shows the original images, the middle row shows the ground truth, and the bottom row shows the TTT-UNet predictions.

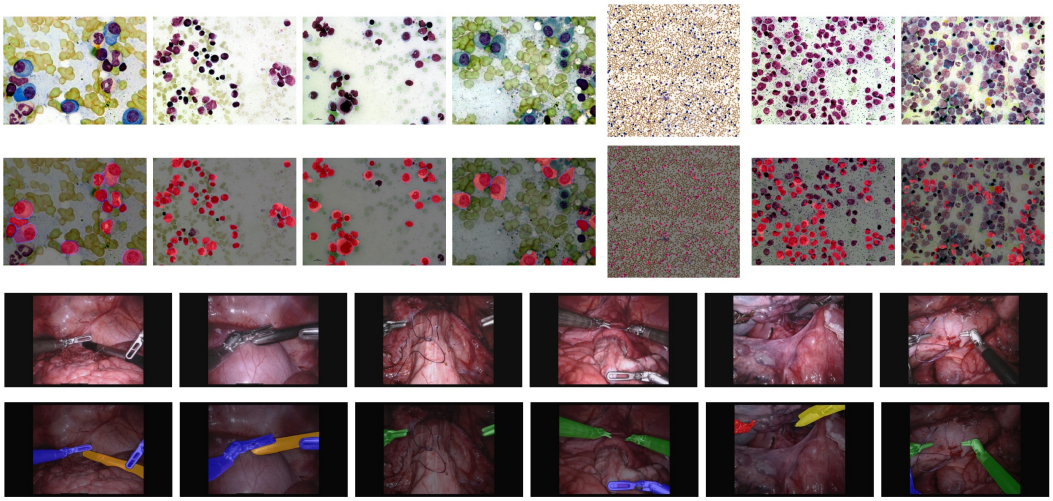

Figure 3: Visualization results of TTT-UNet on Microscopy and Endoscopy datasets. The first and second rows show the original images and TTT-UNet predictions on the Microscopy dataset, respectively. The third and fourth rows show the original images and TTT-UNet predictions on the Endoscopy dataset.

during test time enhances its precision in segmenting intricate and variable abdominal organs, making it a robust tool in scenarios where consistent accuracy is crucial.

Figure 3 provides further insights through the segmentation results on the Endoscopy and Microscopy datasets. In the Endoscopy dataset, TTT-UNet successfully delineates the surgical instruments, which are challenging due to their small size and diverse appearances. This capability underlines the model's strength in capturing fine details and its adaptability to various shapes and textures. Similarly, in the Microscopy dataset, TTT-UNet demonstrates its robustness by accurately segmenting cell boundaries and interiors, even amidst high variability and noise levels. The model's performance in these diverse settings highlights its versatility and reliability across different medical imaging modalities.

The visual evidence presented in Figures 2 and 3 aligns with the quantitative improvements reported in Table 3. These results underscore TTT-UNet's consistent ability to deliver high-quality segmentations, affirming its potential as a state-of-the-art approach in medical image analysis.

5 Discussion and conclusion

The experimental results across multiple biomedical image segmentation tasks consistently demonstrate that TTT-UNet provides a significant improvement over state-of-the-art methods. The key factor driving this improvement is the integration of TTT layers, which allow the model to adapt to the characteristics of each test image dynamically. This capability leads to enhanced generalization, especially in tasks involving diverse and complex imaging modalities, such as 3D abdomen CT, abdomen MRI, endoscopy, and microscopy datasets.

Furthermore, TTT-UNet's superior performance in capturing long-range dependencies and handling high anatomical variability positions it as a robust tool for clinical applications. In the case of both large-scale anatomical structures and smaller, intricate features, TTT-UNet has demonstrated the ability to adapt and deliver accurate segmentation results. This versatility is particularly crucial in clinical scenarios where precision and adaptability are essential for effective diagnosis and treatment.

One of the primary advantages of TTT-UNet lies in its capacity to dynamically adjust model parameters during the test phase, which significantly enhances segmentation accuracy. Additionally, the lower variance in performance across different datasets emphasizes the model's robustness and consistency. However, it is important to acknowledge that the computational cost associated with test-time training could be a limitation for real-time applications. Future work should focus on optimizing the TTT layers to minimize computational overhead without compromising performance.

In conclusion, TTT-UNet represents a significant advancement in biomedical image segmentation by offering a flexible and adaptive solution. Its ability to consistently outperform other models in both 2D and 3D segmentation tasks reinforces its potential as a state-of-the-art model for medical image analysis. As the model evolves, further optimization of test-time adaptation strategies and integration with large-scale datasets will pave the way for broader clinical adoption and deployment.

6 Acknowledgements

This work is partially supported by the National Science Foundation grants (MRI-2215789, IIS-2319451, CRII-2246067, ATD-2427915, POSE-2346158), National Institutes of Health grants (R21EY034179), and Lehigh's grants under CORE (001250) and FRG (S00011497).

References

[1] Imran Qureshi, Junhua Yan, Qaisar Abbas, Kashif Shaheed, Awais Bin Riaz, Abdul Wahid, Muhammad Waseem Jan Khan, and Piotr Szczuko. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90:316–352, 2023.

[2] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.

[3] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[5] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[6] 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. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[7] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[8] 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.

[9] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[10] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

[11] Zhiling Yan, Weixiang Sun, Rong Zhou, Zhengqing Yuan, Kai Zhang, Yiwei Li, Tianming Liu, Quanzheng Li, Xiang Li, Lifang He, et al. Biomedical sam 2: Segment anything in biomedical images and videos. *arXiv preprint arXiv:2408.03286*, 2024.

[12] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13, 2024.

[13] Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310, 2024.

[14] Weixiang Sun, Xiaocao You, Ruizhe Zheng, Zhengqing Yuan, Xiang Li, Lifang He, Quanzheng Li, and Lichao Sun. Bora: Biomedical generalist video generation model. *arXiv preprint arXiv:2407.08944*, 2024.

[15] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[16] Ruizhe Zheng and Lichao Sun. Llm4brain: Training a large language model for brain video understanding. *arXiv preprint arXiv:2409.17987*, 2024.

[17] Rong Zhou, Houliang Zhou, Brian Y Chen, Li Shen, Yu Zhang, and Lifang He. Attentive deep canonical correlation analysis for diagnosing alzheimer's disease using multimodal imaging genetics. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 681–691. Springer, 2023.

[18] Rong Zhou, Houliang Zhou, Li Shen, Brian Y Chen, Yu Zhang, and Lifang He. Integrating multimodal contrastive learning and cross-modal attention for alzheimer's disease prediction in brain imaging genetics. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1806–1811. IEEE, 2023.

[19] Xiyue Wang, Sen Yang, Yuqi Fang, Yunpeng Wei, Minghui Wang, Jing Zhang, and Xiao Han. Sk-unet: An improved u-net model with selective kernel for the segmentation of lge cardiac mr images. *IEEE Sensors Journal*, 21(10):11643–11653, 2021.

[20] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[21] SM Kamrul Hasan and Cristian A Linte. U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 7205–7211. IEEE, 2019.

[22] Sirojbek Safarov and Taeg Keun Whangbo. A-denseunet: Adaptive densely connected unet for polyp segmentation in colonoscopy images with atrous convolution. *Sensors*, 21(4):1441, 2021.

[23] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. 3d transunet: Advancing medical image segmentation through vision transformers. *arXiv preprint arXiv:2310.07781*, 2023.

[24] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1655–1663, 2021.

[25] 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.

[26] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[27] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[28] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

[29] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 311–320. Springer, 2019.

[30] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[31] 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.

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[33] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models. In *International Conference on Machine Learning*, pages 7616–7633. PMLR, 2022.

[34] Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

[35] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560*, 2024.

[36] Haifan Gong, Luoyao Kang, Yitao Wang, Xiang Wan, and Haofeng Li. nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model. *arXiv preprint arXiv:2402.03526*, 2024.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[38] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[39] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.

[40] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[41] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[42] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023.

[43] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35:36722–36732, 2022.

[44] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019.

[45] Jun Ma, Ronald Xie, Shamini Ayyadhury, Cheng Ge, Anubha Gupta, Ritu Gupta, Song Gu, Yao Zhang, Gihun Lee, Joonkee Kim, et al. The multi-modality cell segmentation challenge: Towards universal solutions. *arXiv preprint arXiv:2308.05864*, 2023.