RSB-MedNeXt: An attempt at beating the STU-Net through Robust Stem and Bottleneck Design

Cong Thang Pham[1,*], Minh Toan Dinh[1], Thi Thu Thao Tran[2]

[1]The University of Da Nang–University of Science and Technology, Danang, 550000, Vietnam
pcthang@dut.udn.vn, mtoan65@proton.me
[2]The University of Da Nang – University of Economics, Danang, 550000, Vietnam
thaotran@due.udn.vn

Commission II, WG II/8

Keywords: Medical image segmentation, STU-Net, MedNeXt, U-Net.

Abstract

Medical image segmentation is a crucial task that supports clinical diagnosis and treatment planning. This field was revolutionized in both theoretical and practical aspects due to the employment of deep learning, specifically U-Net and its variants. Recently, with the aim of improving scaling and transferable capabilities, which are the drawbacks of U-Net, STU-Net, and other similar works were released. As a result, this led to significant advancements in medical applications practically. However, STU-Net trades efficiency for performance disproportionately, resulting in huge fine-tuning costs to achieve improvement over training from scratch. In this paper, we systematically identify architectural strengths and limitations of STU-Net and MedNeXt that hinder optimal feature learning. Through this analysis, we propose RSB-MedNeXt, a more robust CNN architecture designed to surpass STU-Net while maintaining efficiency. Our architecture introduces two key innovations: (1) a robust stem module with three parallel branches that extract information at multiple scales, (2) a hybrid bottleneck that combines CNN-based feature extraction with self-attention mechanisms to capture both fine-grained details and global context. We integrate our network into the nnU-Net framework and conduct comprehensive experiments on multiple segmentation tasks against STU-Net and MedNeXt. Results demonstrate that RSB-MedNeXt achieves superior performance while requiring fewer computational resources than STU-Net. Through our approach, we hope that the trade-off between performance and efficiency in medical image segmentation can be effectively addressed and offers a promising method in resource-constrained clinical applications.

1. Introduction

Medical image segmentation is a crucial task that aims to accurately segment organs and tumors, thereby supporting doctors in diagnosis and treatment planning via images. Hence, it has long been of high interest to researchers, especially after the emergence of U-Net (Ronneberger et al., 2015, Awais et al., 2025, Zhao et al., 2025), a CNN-based architecture designed specifically for the segmentation of medical image data. U-Net has quickly revolutionized this field and has continually been impactful in other research directions, demonstrated by its citations exceeding 100000 to date. As U-Net’s release, numerous variants have been proposed to improve further performances in various organs and modalities (Zhou et al., 2018, Isensee and Maier-Hein, 2019). However, their inherent disadvantage, which is the need to configure architectures carefully based on specific tasks to achieve state-of-the-art results, has not been ever solved yet (Isensee et al., 2024, Xia et al., 2024, Isensee et al., 2018). As a result, U-Net and its variants are likely to have a lack of generalizability, which is necessary for clinical applications.

Recently, this issue has been mitigated by the introduction of STU-Net (Huang et al., 2023, Zhang and Metaxas, 2024, He et al., 2025, Xia et al., 2024), an architecture with scaling and transfer capabilities, which has resulted in more generalizations. Thanks to its abilities, STU-Net achieved high rankings and even won several medical image segmentation challenges by fine-tuning pre-trained models derived from previous large-scale dataset pre-training. However, it is not without drawbacks, as it trades efficiency, specifically size, and speed, at a disproportionate ratio for impressive performance. This leads to high fine-tuning costs, and hence, there is only a marginal difference between fine-tuning and training from scratch.

Therefore, motivated by this observation, in this paper, we attempt to surpass STU-Net by systematically investigating both STU-Net and MedNeXt (Roy et al., 2023, Liu et al., 2024) - another CNN architecture with competitive performance. From this analysis, we propose a more generalizable and robust network that can surpass both while being more efficient than STU-Net.

Our contributions fold: (1) We analyze the pros and cons of STU-Net and MedNeXt systematically. (2) Based on this in-depth analysis, we propose a new architecture with a robust stem module and enhanced bottleneck. (3) We conduct experiments in nnU-Net to highlight our proposed architecture performance compared to the two aforementioned methods.

2. Related Work

CNN-based Architectures Traditional methods of Medical Image Segmentation faced difficulties due to the requirement for manual annotation support and the complexity of medical images (Gotra et al., 2017, Ansari et al., 2022, Li et al., 2015). However, this field was revolutionized when CNN-based networks were applied to it, especially with the release of U-Net

[*] Corresponding author: pcthang@dut.udn.vn

Publisher's note: Copernicus Publications has not received any payments from Russian or Belarusian institutions for this paper.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-219-2025 | © Author(s) 2025. CC BY 4.0 License.

(Ronneberger et al., 2015). It quickly became the most popular baseline for medical image segmentation due to its specific design for this task (Siddique et al., 2021, Azad et al., 2024). Specifically, it consists of an encoder-decoder structure with skip connections, which allows precise localization with limited annotated data. This leads to efficient end-to-end training and strong performance even on small datasets, making it particularly suitable for biomedical applications. After that, many of its variants have been proposed to address specific challenges and enhance U-Net’s performance. The modifications included the introduction of 3D U-Net (Çiçek et al., 2016) and V-Net (Milletari et al., 2016) for volumetric data, replacing 2D convolutions with 3D convolutions to better capture spatial context in CT and MRI scans or nested designs (Huang et al., 2020) such as UNet++ (Zhou et al., 2018) employ dense skip connections and multi-scale feature aggregation to reduce the semantic gap between encoder and decoder. These enhancements further improved segmentation accuracy in challenging scenarios. With the release of nnU-Net (Isensee et al., 2018), a self-adapting framework that automates the configuration of U-Net architectures, preprocessing, training, and inference pipelines for any given dataset, U-Net became not only a top choice for research works but also for practical projects. nnU-Net demonstrated that even standard U-Net architectures can outperform many newly proposed and complex models across diverse benchmarks when it is properly configured and tuned. Specifically, nnU-Net introduced dynamic adaptation of network topology (e.g., 2D, 3D, or cascaded U-Nets), which is automatic adjustment of patch size and pooling operations, and robust data augmentation and normalization strategies. Notably, nnU-Net’s success highlighted that careful pipeline design and rigorous validation are often more impactful than incremental architectural tweaks (Isensee et al., 2018, Isensee et al., 2024). Recent works have also explored scaling U-Net models to larger sizes, both in depth and width, and pre-training them on large, diverse datasets to improve transferability and generalization. For example, STU-Net (Huang et al., 2023) extends nnU-Net by systematically scaling model size up to over a billion parameters and demonstrates that larger models, when trained on sufficiently large datasets, yield consistent performance gains and strong transfer learning capabilities across modalities and tasks, or MedNeXt (Roy et al., 2023), developed from the U-Net architectural framework with individual blocks improved from ConvNeXt (Liu et al., 2022, Woo et al., 2023), achieves competitive performance with STU-Net without pre-training on large datasets.

Transformer-based Architectures To overcome CNN limitations, transformer-based architectures in medical image segmentation were proposed. UNETR (Hatamizadeh et al., 2022) was one of the earliest works that replaced CNNs’ localized receptive fields by processing non-overlapping volumetric patches through a transformer encoder to model global dependencies across entire scans. SwinUNETR (Hatamizadeh et al., 2021) subsequently employed shifted window attention in Swin Transformer (Liu et al., 2021) to reduce computational complexity while maintaining contextual integration. There were also some hybrid approaches like TransUNet (Chen et al., 2021) and CoTr (Xie et al., 2021), which strategically combined transformers with CNNs to benefit from both local feature precision and global context. Their superior ability to capture long-range spatial dependencies, which are critical for segmenting interconnected anatomies or irregular pathologies, led to a significant improvement in performance generally. However, transformers still face significant computational challenges. Their main self-attention mechanisms scale quadratically with input size, causing prolonged inference times and often exceeding GPU memory limits on standard hardware. Recent studies addressed these efficiency constraints through architectural modifications such as interleaved convolution-transformer blocks in nnFormer (Zhou et al., 2022), and state-space-model-based U-Mamba (Ma et al., 2024). Additionally, CNN-based alternatives like MedNeXt (Roy et al., 2023) demonstrated that large-kernel convolutions can approximate transformer-like interactions with lower overhead.

3. Methodology

In this section, we begin with a systematic analysis of existing architectures (U-Net, STU-Net, and MedNeXt), identifying their strengths as well as limitations. Based on this analysis, we identify the architectural requirements that need to be achieved, then propose a new architecture through enhancements at multiple granularities: macro design, block design, and micro design. Firstly, for macro design, we determine the overall architectural framework, reusable blocks, and the size of the architecture. Then, for block design, we redesign several blocks to meet the direction of the macro design. Finally, for micro design, we select detailed parameters of the architecture to further optimize its performance. This entire process maintains consistency through our analysis of the selected architectures and the requirements we’ve established, resulting in our novel architecture for medical image segmentation, namely RSB-MedNeXt.

3.1 Analysis of STU-Net and MedNeXt

Overall Architecture The common point of U-Net, STU-Net, and MedNeXt is that they are all based on a symmetric encoder-decoder architecture with skip connections. The strength of this architectural framework is the clear separation of feature compression and decompression phases corresponding to the encoder and decoder, along with the tight connection between these two modules through skip connections at each resolution stage. This helps the model learn features at many different levels, from raw to abstract, while ensuring no information loss during the learning process.

Resolution stage Both Vanilla U-Net, 3D U-Net, and more recently MedNeXt, typically have only 4-5 resolution stages, rarely reaching the seven stages of more complexly configured U-Net versions in the nnU-Net framework. This means the aforementioned architectures lack sufficient depth to learn features at the most abstract level, which is necessary for increasing model performance. However, this depth issue has been fundamentally mitigated in STU-Net, an architecture with a fixed number of 6 resolution stages, allowing the model to have sufficient depth for feature learning while not being excessively deep to cause information loss or gradient vanishing.

Stem Limitations A common disadvantage of U-Net, STU-Net, and MedNeXt is that they were not designed to extract input information sufficiently, resulting in limited feature learning and inaccurate mask prediction in the decompressing phase. While MedNeXt’s stems were too simple, with only a $1 \times 1 \times 1$ convolution, STU-Net’s stems have a better design by using $3 \times 3 \times 3$ and $1 \times 1 \times 1$ convolutions along with a residual connection. However, both blocks lack the ability to extract global-level context information and long-range dependencies, which is done well by transformer-based architectures. Additionally, the lack of diversity in kernel sizes used for feature extraction from the input also causes CNN-based architectures in general and U-Net-based architectures in particular to lack the ability to handle objects of varying sizes. This results in missing important small details or undersegmenting the boundaries of large organs.

Upsampling & Downsampling Limitations U-Net and STU-Net’s downsampling (DS) and upsampling (US) blocks encounter similar limitations as mentioned above, specifically an inadequate number of convolutions and the unemployment of residual connections—an important mechanism for preserving information in deep neural networks. This leads to forgetting details or losing information during up and down sampling processes. As a result, the accuracy of mask prediction is reduced significantly, and even masks do not capture some details, such as tiny tumors or fragments. This issue has been well addressed in MedNeXt’s Upsampling and downsampling blocks, with the core improvement being the enhancement of ConvNeXt’s standard sampling block and the use of residual connections to enable easier gradient flow. These improvements help MedNeXt preserve spatial information at a level comparable to Transformer-based architectures.

Bottleneck Limitations Despite its important role in processing the compressed features and memorizing them the last time before the decompressing phase, the bottleneck is often overlooked when considering improvements in U-Net-based architectures. Similar to stem or DS and US blocks, bottlenecks in U-Net, MedNeXt, and STU-Net only consist of convolution layers, leading to similar disadvantages such as Limited Global Context learning. This has been somewhat addressed by using transformer blocks in transformer-based architectures like SwinUnet and SwinUnetR. However, such pure transformer block bottlenecks face another inherent disadvantage, which is neglecting fine-grained information, a strength of bottlenecks constructed from convolution blocks.

3.2 Macro Design

Having thoroughly analyzed existing architectures, we now present the macro-level design decisions for our proposed model. These foundational choices establish the overall structure upon which our more detailed enhancements are built.

Overall Architecture Due to the advantages mentioned above of the encoder-decoder framework of U-Net-based architectures, our proposed architecture is also based on this architectural framework. This choice provides a proven foundation while allowing us to easily implement architectural improvements and design new modules. By maintaining this established structure, we can focus our innovations on specific components that address the identified limitations while preserving the strengths of the U-Net paradigm (Figure (1)).

Resolution Stage Our analysis revealed that STU-Net, with the number of resolution stages fixed at 6, has demonstrated superior performance as well as better generalization, transferability, and scalability compared to U-Net and MedNeXt. Therefore, to serve similar purposes in the future and maintain optimal information flow between abstraction levels, the number of resolution stages in our architecture is also fixed at 6. This design choice balances the need for sufficient depth to capture abstract features while avoiding excessive depth that could lead to information loss or gradient vanishing problems. Besides, similar to STU-Net Large, each RSB-MedNeXt’s resolution stage includes N MedNeXt block, with the value of N = 2.

3.3 Block Design

Basic block Based on our analysis above, our requirement for the basic block is robustness in learning features as well as preserving information. As shown in Figure (2), MedNeXt block consists of a Depthwise (DW) convolution layer with kernel size $k \times k \times k$ (k=1), along with group normalization (GNorm), resulting in C output channels. Next, the expansion layer, which includes a convolution layer with Gaussian Error Linear Unit (GELU) activation, is placed, followed by a convolution layer with $1 \times 1 \times 1$ kernel as the compression layer. This layer is enhanced by a skip connection. Thus, the MedNeXt block itself, which is an improvement from the ConvNeXt block, is already sufficient to meet these requirements. Therefore, this MedNeXt block will be reused as the basic block in our architecture to inherit its powerful capabilities.

Upsampling & Downsampling Blocks Similar to the requirements for the basic block, for DS and US blocks, our goal is the ability to maintain features during the sampling process. For this reason, MedNeXt DS and US blocks, which are customized from the MedNeXt block to serve sampling, effectively address the identified limitations above in STU-Net’s approach by incorporating adequate convolutions and residual connections. Thus, they are sufficient to ease these aforementioned issues and can also be reused in our architecture.

Robust Stem Module Based on our analysis, our goal for the stem is a new module with the ability to extract input information more powerfully at different levels using kernels of diverse sizes. Therefore, a more robust stem module is proposed with three parallel branches to extract information in multiple levels, which includes one branch with DW $7 \times 7 \times 7$ and $5 \times 5 \times 5$ convolutions for global-level context, another branch with two stacked DW $3 \times 3 \times 3$ convolutions for important information, and the remaining branch with a DW $1 \times 1 \times 1$ convolution for local-level detail. Residual connections are employed to concatenate extracted information before feeding to the latter $1 \times 1 \times 1$ convolution for a combination. Stride with the value of 1 is applied to all convolution layers in this module (Figure (3)).

Hybrid Bottleneck Based on our analysis, our goal for the bottleneck is similar to that for the stem, which is not only to learn fine-grained features well but also to capture global information. Therefore, a new bottleneck is proposed consisting of 2 parallel branches along with a residual connection (Figure (4)). In the first branch, MedNeXt’s bottleneck is utilized because it leverages the MedNeXt block, which is superior to other basic residual convolution blocks. In the other branch, a self-attention block is employed to learn global relationships within the compressed features. As a result, our hybrid bottleneck can benefit from the strengths of two distinct blocks, which are merged by concatenate operation to serve the decoder’s decompression process.

4. Experiments

4.1 Datasets

Our model is evaluated on diverse datasets encompassing both organ and tumor segmentation tasks to demonstrate its comprehensive performance. These datasets include ATLAS (Quinton

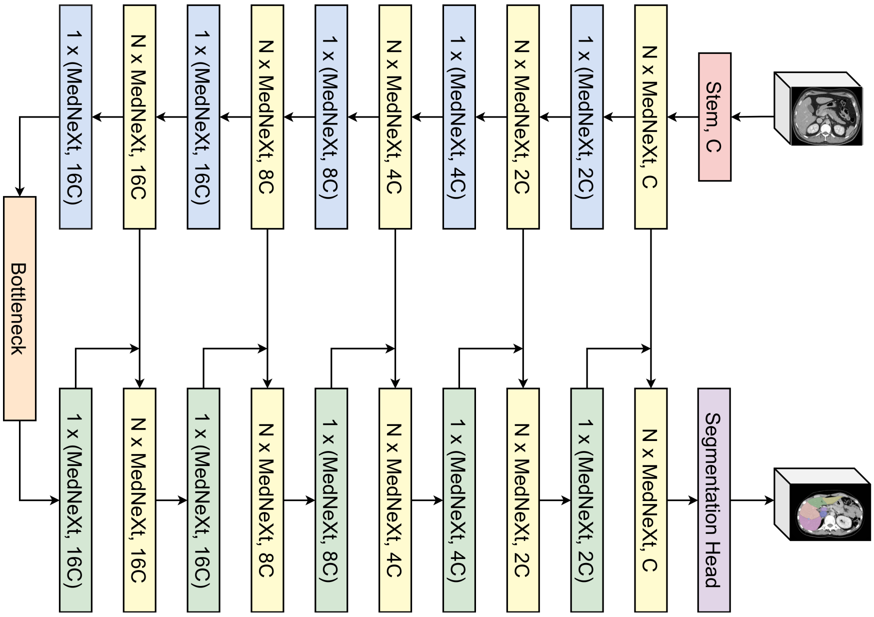

Figure 1. Overall architecture showing the integration of our enhanced components across the six resolution stages, with the encoder pathway (left), decoder pathway (right), and skip connections between corresponding levels.

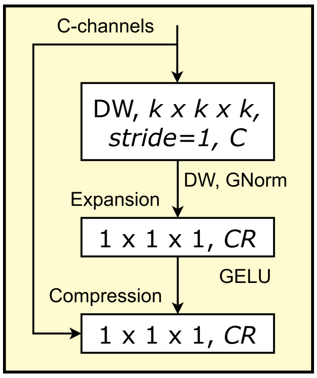

Figure 2. MedNeXt block.

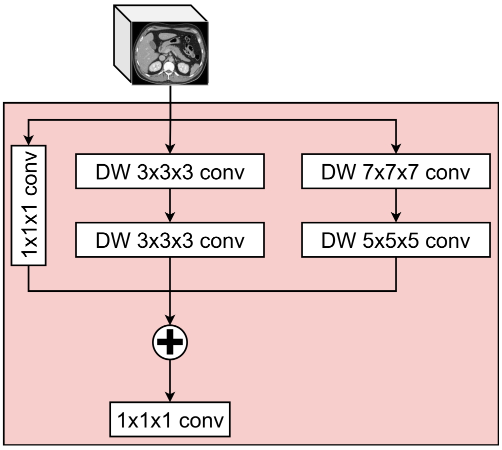

Figure 3. Robust stem with three parallel branches for multi-scale feature extraction.

et al., 2023), AMOS22 (Ji et al., 2022), and KiTs2023 (Heller et al., 2023).

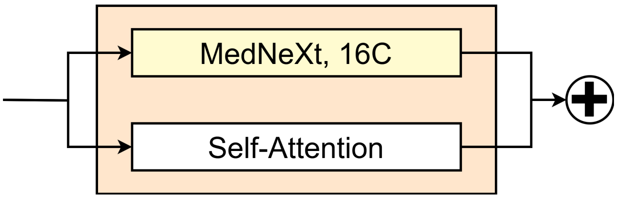

Figure 4. Hybrid bottleneck with self-attention mechanism.

4.2 Experimental Setup

Training Strategy To validate the efficiency of RSB-MedNeXt to other U-Net-based architectures in challenging tasks like tumor segmentation, specifically STU-Net and MedNeXt, nnU-Net is leveraged as our standardized framework. Our models are trained from scratch in 200 epochs with an initial learning rate of 1×10^{-2} for all selected datasets. Additionally, other hyperparameters will be configured by default by nnU-Net to ensure fairness. Only for ATLAS, MedNeXt are trained with the same training recipe of RSB-MedNeXt, and STU-Net are finetuned in 100 epochs.

Evaluation strategy Regarding evaluation strategy, K-fold cross-validation with the common value K = 5 is employed to ensure fairness, especially with small-scale datasets like medical images. This scheme helps maintain consistent experimental conditions across different architectures and facilitates further comparisons with one of the largest benchmarks conducted in the study of Isensee et al. (Isensee et al., 2024), as well as results from other works. Therefore, the metrics from this benchmark are also reused to evaluate our model's performance, specifically the average and standard deviation Dice Similarity Coefficients (DSC) are the primary and secondary metrics, respectively.

4.3 Experimental Results

RSB-MedNeXt consistently outperforms previous state-of-the-art models across diverse datasets. Our experimental res-

Network	Average Dice Similarity Coefficients (%)↑		
	ATLAS	AMOS22	KiTs2023
STU-Net L	83.36	89.34	85.84
MedNeXt L k3	84.70	89.62	88.25
MedNeXt L k5	85.04	89.73	87.74
RSB-MedNeXt	**85.81**	89.69	**88.94**

Table 1. Performance comparison of RSB-MedNeXt against STU-Net Large and MedNeXt Large across datasets. ↑: Higher is better. **Bold** indicates the best, and underline indicates the second best.

Network	Standard Deviation of Dice Coefficients (%)		
	ATLAS	AMOS22	KiTs2023
STU-Net L	2.6	0.45	2.1
MedNeXt L k3	2.1	0.43	0.94
MedNeXt L k5	2.0	0.43	1.2
RSB-MedNeXt	1.77	0.41	0.98

Table 2. Standard deviation of Dice Similarity Coefficients across different architectures and datasets.

ults in Table (1) demonstrate the robustness of RSB-MedNeXt, which achieves higher average DSC scores across most of the datasets compared to STU-Net Large (L), MedNeXt Large with k3 and k5 kernel versions. RSB-MedNeXt has significant improvements when compared to STU-Net L, specifically with substantial gains of +2.45% on ATLAS (Figure (5)) and +3.10% on KiTS2023 datasets. These improvements suggest that our architecture effectively addresses the limitations identified in previous models, especially in feature extraction at multiple levels.

Performance gains of RSB-MedNeXt are stable across popular benchmarks. Table (2) shows the evidence that RSB-MedNeXt exhibits lower standard deviations (1.77%, 0.41%, 0.98%) compared to STU-Net L (2.6%, 0.45%, 2.1%) across all three datasets. This indicates that our model not only achieves higher average performance but also provides more reliable results. Specifically, our model without pretraining on large-scale datasets is still more robust than STU-Net across different anatomical structures and imaging conditions, which is a critical factor for clinical applications.

The robust stem design significantly improves feature extraction capabilities. RSB-MedNeXt's performance in mean and std of DSC scores are superior compared to other networks, especially in two challeging tumor segmentation tasks ATLAS and KiTs2023, with the gains of 2.45% and 3.1% over STU-Net L respectively. Despite the large number of resolution stages, our network are still robust and generalizable thanks to the stem module. This demonstrates that the parallel branch structure in our stem module enables efficiently simultaneous extraction of multi-scale features, addressing a key limitation in previous architectures that relied primarily on fixed kernel sizes.

Hybrid bottleneck architecture effectively balances global context capture and fine-grained detail preservation. Similar to the stem, our bottleneck also contributes to the efficiency of RSB-MedNeXt, specifically in highest average DSC score in ATLAS and KiTs2023 (85.81% and 88.94%). For AMOS22, despite marginal gap compared to MedNeXt L k5 (89.73%), which is the best results, our network still significantly outperforms STU-Net L.

5. Conclusion

In this work, we attempt to beat STU-Net by carefully investigating it and other state-of-the-art architectures. Based on our analysis, we propose a new architecture with strategic designs that can outperform STU-Net and be more efficient than it, specifically RSB-MedNeXt. Our method's performance is demonstrated on several challenging tasks against other strong baselines. We hope that RSB-MedNeXt can serve as a top choice for medical image segmentation in clinical practice, offering both improved accuracy and reduced computational requirements.

References

Ansari, M. Y., Abdalla, A., Ansari, M. Y., Ansari, M. I., Malluhi, B., Mohanty, S., Mishra, S., Singh, S. S., Abinahed, J., Al-Ansari, A. et al., 2022. Practical utility of liver segmentation methods in clinical surgeries and interventions. *BMC medical imaging*, 22(1), 97.

Awais, M., Naseer, M., Khan, S., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M.-H., Khan, F. S., 2025. Foundation Models Defining a New Era in Vision: a Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4), 2245-2264.

Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., Merhof, D., 2024. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 10076-10095.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv:2102.04306*, 1–13.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, Springer, 424–432.

Gotra, A., Sivakumaran, L., Chartrand, G., Vu, K.-N., Vandenbroucke-Menu, F., Kauffmann, C., Kadoury, S., Gallix, B., de Guise, J. A., Tang, A., 2017. Liver segmentation: indications, techniques and future directions. *Insights into imaging*, 8, 377–392.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., Xu, D., 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *International MICCAI brainlesion workshop*, Springer, 272–284.

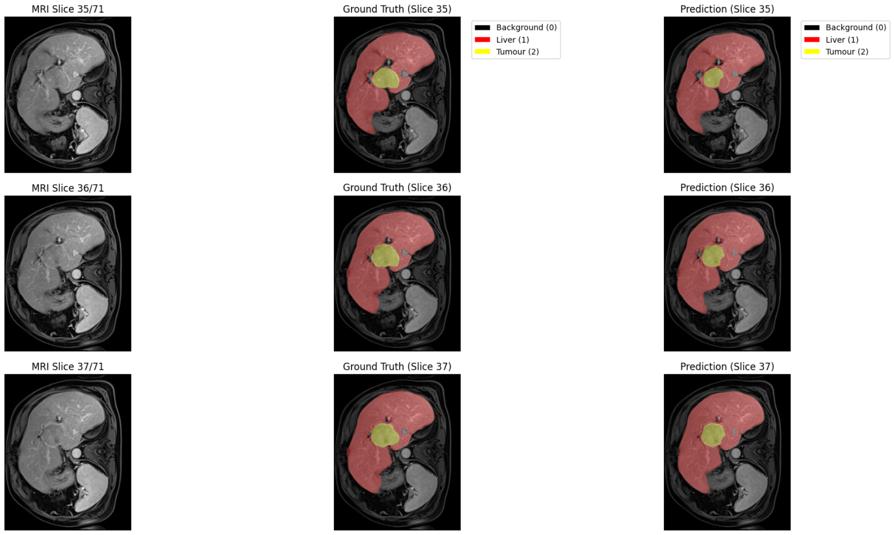

Figure 5. An visualized example of a patient in ATLAS dataset

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., Xu, D., 2022. Unetr: Transformers for 3d medical image segmentation. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 574–584.

He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., Chen, H., 2025. Foundation model for advancing healthcare: challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, 18, 172-191.

Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D. et al., 2023. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv:2307.01984*, 1–34.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Wu, J., 2020. Unet 3+: A full-scale connected unet for medical image segmentation. *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 1055–1059.

Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S. et al., 2023. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv:2304.06716*, 1–15.

Isensee, F., Maier-Hein, K. H., 2019. An attempt at beating the 3D U-Net. *arXiv:1908.02182*, 1–8.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S. et al., 2018. NNU-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv:1809.10486*, 1–11.

Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P. F., 2024. Nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 488–498.

Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X. et al., 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35, 36722–36732.

Li, G., Chen, X., Shi, F., Zhu, W., Tian, J., Xiang, D., 2015. Automatic liver segmentation based on shape constraints and deformable graph cut in CT images. *IEEE Transactions on Image Processing*, 24(12), 5315–5329.

Liu, B., Li, B., Chen, Y., Sreeram, V., Li, S., 2024. Fast-MedNeXt: Accelerating the MedNeXt Architecture to Improve Brain Tumour Segmentation Efficiency. *International Journal of Imaging Systems and Technology*, 34(6), e23196.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Ma, J., Li, F., Wang, B., 2024. U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv:2401.04722*, 1–17.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 fourth international conference on 3D vision (3DV)*, IEEE, 565–571.

Quinton, F., Popoff, R., Presles, B., Leclerc, S., Meriaudeau, F., Nodari, G., Lopez, O., Pellegrinelli, J., Chevallier, O., Ginhac, D. et al., 2023. A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data*, 8(5), 1-9.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 234–241.

Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P. F., Maier-Hein, K. H., 2023. Mednext: transformer-driven scaling of convnets for medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 405–415.

Siddique, N., Sidike, P., Elkin, C., Devabhaktuni, V., 2021. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access*, 9, 82031-82057.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., Xie, S., 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16133–16142.

Xia, Q., Zheng, H., Zou, H., Luo, D., Tang, H., Li, L., Jiang, B., 2024. A Comprehensive Review of Deep Learning for Medical Image Segmentation. *Neurocomputing*, 128740.

Xie, Y., Zhang, J., Shen, C., Xia, Y., 2021. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, Springer, 171–180.

Zhang, S., Metaxas, D., 2024. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis*, 91, 102996.

Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H. H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J. et al., 2025. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1), 166–176.

Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y., 2022. nnformer: Interleaved transformer for volumetric segmentation. *arXiv:2307.01984*, 1-10.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. *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*, Springer, 3–11.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-219-2025 | © Author(s) 2025. CC BY 4.0 License.