

Self-Attention and State-Space Duality in 3D Medical Segmentation: A Critical Algorithmic Review of Transformers, Mamba Architectures, Modernized ConvNets, and Foundation Models

Abdul Gather S

PhD Scholar

Government Arts College(Autonomous),

Coimbatore, Tamil Nadu

Email: Kadhar094@gmail.com

Soranamageswari M

Associate Professor

Government Arts College (Autonomous),

Coimbatore, Tamil Nadu

Email: swarnakanappan@gmail.com

Cite as: Abdul Gather S, & Soranamageswari M. (2026). Self-Attention and State-Space Duality in 3D Medical Segmentation: A Critical Algorithmic Review of Transformers, Mamba Architectures, Modernized ConvNets, and Foundation Models. Journal of Research and Innovation in Technology, Commerce and Management, Vol. 3(Issue 10), 310001–310009. <https://doi.org/10.5281/zenodo.23153194>

DOI: <https://doi.org/10.5281/zenodo.23153194>

Abstract

Three-dimensional medical image segmentation requires models that can simultaneously preserve fine anatomical boundaries and represent long-range spatial dependencies. CT and MRI volumes contain large numbers of voxels and may exhibit substantial anisotropy, making memory consumption and computational efficiency important architectural considerations. Convolutional neural networks provide strong local spatial priors, whereas Transformer architectures use self-attention to establish global interactions between tokens. However, global attention has quadratic dependence on sequence length. This limitation has motivated hierarchical windowed Transformers, large-kernel convolutional networks,

state-space models (SSMs), and medical foundation-model adaptation [4–15].

This review critically examines sixteen representative works spanning convolutional, Transformer, state-space, foundation-model, generative, and reproducibility paradigms: PG-SAM, limited-interval classifier-free guidance,

Snakemake, 3D U-Net, 3D UX-Net, MedNeXt, Swin UNETR, UNETR, Mamba-UNet, nnU-Net, ConvNeXt, TransUNet, MedSAM, VM-UNet, Mamba-SAM, and the Mamba-2/structured state-space duality framework [1–16]. The comparison focuses on representation construction, receptive-field behavior, computational complexity, volumetric

modeling, information compression, transferability, and failure modes. The evidence does not support a universal winner. Instead, the literature suggests that efficient volumetric segmentation benefits from combining local boundary-sensitive representations with appropriately compressed global context and, where appropriate, pretrained semantic priors. State-space models provide an attractive alternative to explicit pairwise attention, but their practical effectiveness depends on sequence construction, representation capacity, and implementation. The review therefore frames CNNs, Transformers, SSMs, and foundation models as complementary components within a broader architectural design space.

Keywords: 3D Medical Image Segmentation, Vision Transformers, Mamba, ConvNeXt, Foundation Models

1. Introduction

Volumetric medical segmentation differs from conventional two-dimensional image segmentation because the input contains an additional spatial dimension and often exhibits unequal voxel spacing. A volumetric image can be represented as

$$X \in \mathbb{R}^{(H \times W \times D \times C)} \text{ ---- (1)}$$

Where H , W , and D represent the spatial dimensions, and C denotes the number of channels. The segmentation target can be represented as

$$Y \in \{0, 1, \dots, K - 1\}^{(H \times W \times D)} \text{ ---- (2)}$$

The central challenge is to model both local anatomical boundaries and broader anatomical relationships without exceeding available computational resources.

The 3D U-Net established a fundamental encoder-decoder formulation for volumetric segmentation by extending convolutional U-Net operations into three dimensions [4]. Its encoder progressively aggregates contextual information, while skip connections transfer high-resolution features to

the decoder. This design remains influential because it combines spatial locality with multiscale representation.

nnU-Net subsequently demonstrated that segmentation performance is influenced not only by network architecture but also by preprocessing, resampling, patch configuration, normalization, training strategy, and post-processing [10]. This observation is particularly important when comparing newer architectures because numerical improvements cannot automatically be attributed to architectural design alone.

Transformer-based systems introduced an alternative mechanism for contextual modeling. UNETR treats volumetric patches as tokens and uses a Transformer encoder to model relationships among them [8]. TransUNet combines convolutional feature extraction with Transformer-based contextual modeling [12], while Swin UNETR reduces attention complexity by restricting self-attention to local windows and shifting those windows between successive layers [7].

More recently, modernized convolutional architectures such as ConvNeXt, 3D UX-Net, and MedNeXt have demonstrated that large receptive fields can be obtained through carefully designed convolutional operations [5,6,11]. State-space approaches provide another alternative. Mamba-based visual architectures process serialized representations through selective state updates rather than explicitly constructing a pairwise attention matrix [9,14]. Foundation models such as MedSAM and PG-SAM introduce pretrained visual representations and semantic priors into medical segmentation pipelines [1,13].

Consequently, the contemporary architectural landscape is better viewed as a continuum rather than a simple competition among CNNs, Transformers, and Mamba. The relevant question is how spatial information should be preserved, compressed, globally exchanged, and dynamically propagated through the segmentation hierarchy.

2. Architectural Paradigms

2.1 Convolutional Paradigm

The 3D U-Net uses volumetric convolutions, downsampling, upsampling, and skip connections to construct a hierarchical representation [4]. A simplified convolution can be written as

$$Y(p) = \sum_{q \in \mathcal{K}} W(q)X(p + q) \text{ ---- (3)}$$

Where $\mathcal{K}$ represents the local convolutional neighborhood.

The principal advantage of convolution is its spatial inductive bias. Neighboring voxels are processed together, making convolution particularly suitable for local structures and boundaries. Its limitation is that long-range relationships must be obtained indirectly through deeper layers, larger kernels, dilation, or downsampling.

nnU-Net addresses a broader methodological issue by automatically configuring important parts of the segmentation pipeline according to characteristics of the training dataset [10]. Its contribution therefore extends beyond a particular network block: it demonstrates that data configuration and optimization can have an effect comparable to architectural modification.

ConvNeXt modernized conventional convolutional design by adopting several principles associated with contemporary Transformer architectures, including large kernels, depthwise convolution, LayerNorm, GELU activation, and redesigned block structures [11]. Although ConvNeXt was developed primarily for natural-image recognition, its design principles influenced subsequent volumetric medical architectures.

3D UX-Net transferred large-kernel depthwise convolution to volumetric segmentation. Its use of kernels beginning at $7 \times 7 \times 7$ was intended to increase the effective receptive field while retaining convolutional inductive biases [5]. The reported experiments showed competitive or superior performance to several Transformer-based baselines on multiple volumetric datasets.

MedNeXt further developed the large-kernel convolutional paradigm for medical segmentation by constructing a fully convolutional encoder-decoder and introducing scalable kernel, depth, and width configurations [6]. Its kernel upsampling strategy was designed to enable larger effective kernels without directly training every large-kernel parameter from scratch.

These studies indicate that convolution should not be equated with purely local processing. Through depth, downsampling, large kernels, and hierarchical feature aggregation, convolutional systems can capture increasingly broad context.

2.2 Transformer Paradigm

UNETR reformulates volumetric segmentation as a sequence-to-sequence problem. The input volume is divided into non-overlapping 3D patches. If the patch size is $P \times P \times P$, the approximate token count is

Number of tokens:

$$N = HWD / P^3 \text{ ---- (4)}$$

Query, Key, and Value matrices:

$$Q = XW_Q \text{ ---- (5)}$$

$$K = XW_K \text{ ---- (6)}$$

$$V = XW_V \text{ ---- (7)}$$

Self-attention:

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T / \sqrt{d_k})V$$

Example:

For a 128^3 volume divided into 4^3 patches:

$$N = 128^3 / 4^3 = 32,768$$

So the corrected paragraph is:

The number of tokens generated by the patch embedding process is

$$N = HWD / P^3$$

The Transformer then operates on these tokens. For self-attention, the Query, Key, and Value matrices are defined as

$$Q = XW_Q$$

$$K = XW_K$$

$$V = XW_V$$

and the self-attention operation is given by

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{d_k})\mathbf{V} \text{ ----- (8)}$$

The matrix $\mathbf{Q}\mathbf{K}^T$ contains pairwise relationships between tokens. Consequently, the computational and memory requirements of global attention grow approximately quadratically with the number of tokens N [8]. For example, a 128^3 volume using 4^3 patches contains

$N = 128^3 / 4^3 = 32,768$ tokens. The corresponding pairwise interaction matrix contains more than one billion entries before considering multiple heads, layers, activations, and other implementation requirements.

TransUNet reduces this burden by combining convolutional feature extraction with Transformer processing at a lower representation level [12]. The CNN provides local feature extraction, while the Transformer supplies broader contextual interaction.

Swin UNETR addresses the quadratic bottleneck using windowed self-attention [7]. Instead of allowing every token to interact with every other token, attention is calculated inside local 3D windows. Shifted windows then provide communication across neighboring regions in subsequent layers. For a fixed window size M , the attention computation has substantially better scaling than global attention.

The principal trade-off is therefore clear: global attention provides flexible long-range interaction but has high computational cost, whereas windowed attention improves efficiency at the expense of immediate global connectivity.

2.3 State-Space Paradigm

State-space models provide a different mechanism for sequence modeling. A continuous-time linear SSM can be represented as

$$d\mathbf{h}(t)/dt = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t) \text{ ----- (9)}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}\mathbf{x}(t) \text{ ----- (10)}$$

After discretization:

$$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}\mathbf{x}_t \text{ ----- (11)}$$

Where:

$\mathbf{h}(t)$ = hidden state

$\mathbf{x}(t)$ = input

$\mathbf{y}(t)$ = output

$\mathbf{A}$ = state-transition matrix

$\mathbf{B}$ = input matrix

$\mathbf{C}$ = output matrix

$\mathbf{D}$ = direct-feedthrough matrix

$\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ = discretized state-transition and input matrices with an output equation determined by the transformed state.

Mamba introduces input-dependent selective state transitions, allowing the model to dynamically control information propagation [9,14]. Rather than explicitly constructing every pairwise interaction, contextual information is propagated through a latent state.

This provides an important computational advantage for long sequences. However, linear scaling should not be interpreted as an automatic guarantee of lower end-to-end runtime or memory consumption. Actual efficiency depends on sequence length, state dimension, implementation, hardware, parallelization strategy, and surrounding network operations.

Mamba-UNet applies visual Mamba components within a U-Net-like encoder-decoder structure [9]. VM-UNet similarly constructs a U-shaped architecture using visual state-space blocks and directional scanning [14]. These approaches demonstrate the feasibility of replacing explicit global attention with state-space processing in medical segmentation.

The main challenge is that images are spatial rather than naturally sequential. Serialization therefore becomes an architectural decision. A 2D or 3D feature map must be transformed into an ordered sequence, and the selected scan strategy may influence how spatial relationships are represented.

3. Foundation Models and Hybrid Adaptation

MedSAM adapts the Segment Anything framework to medical images through large-scale medical image-mask training [13]. The resulting model uses a prompt-driven segmentation paradigm in which spatial information can be supplied through prompts such as bounding boxes.

The importance of MedSAM is not simply its segmentation architecture. It demonstrates that foundation-model representations can be adapted to medical domains, although differences between natural-image pretraining and medical-image distributions remain significant [13].

PG-SAM introduces medical semantic priors into the SAM framework [1]. The authors use fine-grained medical textual information to improve alignment between semantic descriptions and visual features. Its decoder additionally combines multilevel features and iterative mask refinement. The approach illustrates how foundation-model segmentation can be extended through medical-domain semantic information.

Mamba-SAM represents another hybrid direction by combining SAM-derived representations with visual state-space processing for volumetric segmentation [15]. Its motivation is to exploit pretrained semantic information while reducing the computational burden associated with directly applying conventional attention mechanisms to 3D volumes.

The broader principle is that foundation models need not replace all components of a medical segmentation pipeline. They can instead serve as pretrained feature providers, while lightweight adapters, SSM blocks, convolutional refiners, or cross-attention modules perform domain-specific adaptation.

4. Self-Attention and State-Space Duality

The difference between attention and SSMs is not simply “quadratic versus linear.” It also concerns how contextual information is represented.

In global attention, the interaction matrix

$$A = \text{Softmax}(QK^T / \sqrt{d_k}) \text{ ----- (12)}$$

Here, A represents the attention-weight matrix. explicitly represents relationships among token pairs. This provides high flexibility because every token can directly access other tokens.

An SSM instead propagates information through a state:

$$H_t = f(h_{t-1}, x_t) \text{ ----- (13)}$$

Here, h_t is the current hidden state, h_{t-1} is the previous hidden state, and x_t is the current input.

The sequence history is consequently summarized through a finite-dimensional representation. This can be computationally attractive, but it creates a representational trade-off: the state must retain information that remains useful for future predictions.

The structured state-space duality framework of Dao and Gu provides a deeper connection between SSMs and certain forms of attention [16].

The SSD perspective demonstrates that certain SSM computations and attention-like operations can be represented through structured matrices. Mamba-2 is derived from this framework and provides an efficient refinement of selective SSM computation [16].

Thus, attention and SSMs should not be regarded as completely unrelated mathematical families. They represent different computational factorizations of contextual sequence transformations.

5. Comparative Analysis

Table 1. Comparative Analysis of Representative Architectures for 3D Medical Image Segmentation

Architecture	Paradigm	Main strength	Main limitation
3D U-Net [4]	CNN	Strong local features and skip connections	Limited direct global interaction
nnU-Net [10]	Auto-configured CNN	Dataset-specific pipeline optimization	Architecture remains primarily convolutional
ConvNeXt [11]	Modern CNN	Large-kernel modernized convolution	Not specifically designed as a 3D medical model
3D UX-Net [5]	Large-kernel CNN	Broad receptive field with <u>depthwise</u> convolution	Large kernels increase local computation
MedNeXt [6]	Large-kernel CNN	Scalable volumetric ConvNet	Higher model capacity can increase computation
UNETR [8]	Transformer	Global token interaction	Global attention scales <u>quadratically</u>
TransUNet [12]	CNN + Transformer	Local/global feature combination	Transformer remains computationally demanding
Swin UNETR [7]	Windowed Transformer	More efficient hierarchical attention	Direct global interaction is restricted
Mamba-UNet [9]	Visual SSM	Efficient long-sequence <u>modeling</u>	Spatial serialization remains important
VM-UNet [14]	Visual SSM	Linear-style contextual processing	Resolution and scanning behavior require careful study
MedSAM [13]	Foundation model	Transferable <u>pretrained</u> representation	Medical domain shift and prompt dependence
PG-SAM [1]	Foundation + semantic prior	Medical semantic guidance	Requires multimodal prior construction
Mamba-SAM [15]	Foundation + SSM	Combines <u>pretrained</u> representation and efficient processing	Additional architectural complexity
Mamba-2 / SSD [16]	SSM theory	Connects SSM and attention formulations	Evidence is primarily <u>sequence-modeling</u> rather than medical segmentation

The table should be interpreted qualitatively rather than as a universal ranking. Different studies use different datasets, preprocessing pipelines, training schedules, image resolutions, augmentation methods, and evaluation procedures.

In particular, ConvNeXt's ImageNet performance should not be placed directly beside medical segmentation Dice values. Similarly, Snakemake and diffusion guidance are methodological components rather than segmentation architectures. Their inclusion is useful for discussing reproducible and generative pipelines, but they should not be treated as competing segmentation networks.

6. Failure Modes and Critical Issues

6.1 Quadratic Attention

The principal weakness of global self-attention is the growth of pairwise interactions with token

count. In 3D imaging, increasing resolution can increase token count rapidly, making global attention expensive [7,8].

6.2 Window Locality

Windowed attention improves computational efficiency but restricts direct interaction to tokens within the same window [7]. Shifted windows mitigate this limitation by allowing information to move between neighboring windows over successive layers.

6.3 SSM State Capacity

An SSM does not explicitly preserve all pairwise relationships. Historical information is compressed into a latent state. Very long sequences may therefore create a mismatch between information content and state capacity [16].

6.4 Spatial Serialization

Visual SSMs require feature maps to be scanned or otherwise converted into sequences [9,14]. Different scan directions can expose different spatial relationships. Multi-directional scanning can reduce directional bias but increases computational and architectural complexity.

6.5 Boundary Loss

Large-kernel convolution and frequency-oriented processing can capture broad structures efficiently, but segmentation systems must still preserve high-frequency boundary information. This is particularly important for small lesions, thin anatomical structures, and irregular interfaces.

6.6 Foundation-Model Domain Shift

A foundation model trained predominantly on natural images does not automatically possess medical-domain knowledge. Differences in intensity distributions, acquisition protocols, artifacts, anatomy, and pathology can lead to substantial domain shift [13].

6.7 Benchmark Comparability

Dice, IoU, and HD95 values cannot be compared safely without considering the experimental protocol. Differences in train-test splits, preprocessing, resolution, augmentation, ensemble

methods, and post-processing can produce substantial changes in measured performance.

7. Generative Augmentation and Reproducibility

Generative augmentation can complement architectural improvements by increasing the diversity of training samples, particularly when annotated medical datasets are limited. Kynkäänniemi et al. demonstrated that applying guidance within a limited interval can improve sample and distribution quality in diffusion models [2]. In medical segmentation research, such techniques may therefore be explored as a supporting strategy for increasing training-data diversity.

However, synthetic data should not automatically be considered equivalent to real clinical data. Generated samples require quality control, distributional validation, and careful assessment of anatomical plausibility and pathological fidelity. Augmentation should therefore be treated as a supporting methodology rather than a substitute for representative clinical datasets [2].

Reproducibility also depends on systematic management of preprocessing, training, evaluation, and data-generation workflows. Snakemake provides a workflow-management framework for organizing computational analysis pipelines and supporting reproducible data-analysis processes [3]. A reproducible segmentation study should additionally report dataset partitions, preprocessing operations, augmentation settings, model configurations, random seeds, evaluation metrics, and hardware/software environments [3].

8. Architectural Design Principles

Five design principles emerge from the reviewed literature.

First, preserve local boundary information. Convolutional layers and high-resolution skip pathways remain valuable even when global context is modeled by Transformers or SSMs [4,12].
Second, compress before expensive global processing. Patch aggregation and hierarchical

feature extraction can reduce token count before attention or state-space processing [7,8,12].

Third, match representation capacity to sequence complexity. Extremely long sequences should not automatically be passed into a recurrent state representation without considering state dimension and information compression [16].

Fourth, use foundation models selectively. A pretrained model can provide transferable semantic features while medical-specific modules adapt those representations to volumetric data [1,13,15].

Fifth, evaluate both region overlap and boundary accuracy. Dice should be complemented by HD95 or other surface-based metrics, particularly when small or thin anatomical structures are clinically important.

9. Future Research Directions

Future 3D segmentation systems should investigate genuinely volumetric state-space processing rather than treating 3D data as independent 2D slices. Three-dimensional scanning strategies, tri-plane representations, and topology-aware sequence construction may reduce the limitations introduced by serialization [9,15].

Hybrid CNN-SSM-Transformer systems also deserve systematic investigation. A possible architecture could use a convolutional stem for high-frequency boundary information, compressed latent representations for global contextual modeling, an SSM for efficient long-range propagation, and a lightweight attention module for selected cross-region interactions.

Foundation models could further incorporate multimodal medical information. CT or MRI features could be combined with radiological text, clinical variables, anatomical descriptions, or other structured information to produce more clinically informed segmentation representations.

State-capacity-aware tokenization is another promising direction. Instead of fixing the token count independently of the input volume, an adaptive system could determine an appropriate

latent representation according to spatial complexity, anatomical content, and available computational resources.

Future benchmark studies should report Dice, IoU, HD95, sensitivity to small structures, inference time, GPU memory, parameter count, FLOPs, robustness to resolution changes, and sensitivity to anisotropic voxel spacing. Such reporting would provide a more meaningful comparison than a single Dice score.

10. Conclusion

The evolution of 3D medical segmentation has moved from predominantly local convolution toward increasingly global and computationally efficient representations. The 3D U-Net established the volumetric encoder-decoder foundation [4], while nnU-Net demonstrated the importance of automated dataset-specific configuration [10]. ConvNeXt, 3D UX-Net, and MedNeXt subsequently showed that modernized convolution and large kernels can provide broad contextual representations without explicit global attention [5,6,11].

UNETR and TransUNet demonstrated the potential of Transformer-based global modeling [8,12], while Swin UNETR reduced its computational burden through hierarchical shifted-window attention [7]. Mamba-based approaches introduced state-space processing as another mechanism for long-range representation [9,14]. Their computational advantages are promising for large sequences, but their practical behavior depends on sequence construction, state capacity, and implementation.

Medical foundation models introduce an additional source of transferable knowledge. MedSAM demonstrates the potential of adapting SAM to medical segmentation [13], while PG-SAM and Mamba-SAM illustrate increasingly sophisticated combinations of semantic priors, pretrained representations, and efficient volumetric processing [1,15].

The principal conclusion is therefore not that CNNs, Transformers, or SSMs universally outperform one

another. Instead, successful 3D segmentation depends on matching the representation mechanism to the spatial scale, information density, anatomical structure, and computational constraints of the problem.

The most promising direction is consequently a hybrid architecture in which convolution preserves local precision, compressed representations support global reasoning, state-space mechanisms provide efficient long-range propagation, selective attention handles important cross-region interactions, and foundation models contribute transferable semantic knowledge. Such a design provides a principled route toward scalable, memory-efficient, anatomically precise, and clinically robust volumetric segmentation.

References

- [1] Y. Zhong, Z. Luo, C. Liu, F. Tang, Z. Peng, M. Hu, Y. Hu, J. Su, Z. Ge, and I. Razzak, "PG-SAM: Prior-Guided SAM with Medical for Multi-organ Segmentation," *arXiv preprint arXiv:2503.18227*, 2025. doi: [10.48550/arXiv.2503.18227](https://doi.org/10.48550/arXiv.2503.18227).
- [2] T. Kynkäänniemi, M. Aittala, T. Karras, S. Laine, and T. Aila, "Applying Guidance in a Limited Interval Improves Sample and Distribution Quality in Diffusion Models," in *Advances in Neural Information Processing Systems*, vol. 37, 2024. doi: [10.52202/079017-3892](https://doi.org/10.52202/079017-3892).
- [3] F. Mölder *et al.*, "Sustainable Data Analysis with Snakemake," *F1000Research*, vol. 10, p. 33, 2021. doi: [10.12688/f1000research.29032.3](https://doi.org/10.12688/f1000research.29032.3).
- [4] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016*, 2016, pp. 424–432. doi: [10.1007/978-3-319-46723-8_49](https://doi.org/10.1007/978-3-319-46723-8_49).
- [5] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman, "3D UX-Net: A Large Kernel Volumetric ConvNet Modernizing Hierarchical Transformer for Medical Image Segmentation," in *Proc. 11th International Conference on Learning Representations (ICLR)*, 2023. doi: [10.48550/arXiv.2209.15076](https://doi.org/10.48550/arXiv.2209.15076).

- [6] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger, and K. Maier-Hein, "MedNeXt: Transformer-Driven Scaling of ConvNets for Medical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2023*, 2023, pp. 405–415. doi: **10.1007/978-3-031-43901-8_39**.
- [7] A. Hatamizadeh *et al.*, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in *BrainLesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 2022. doi: **10.1007/978-3-031-08999-2_22**.
- [8] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D Medical Image Segmentation," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 574–584. doi: **10.1109/WACV51458.2022.00181**.
- [9] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, and L. Li, "Mamba-UNet: UNet-Like Pure Visual Mamba for Medical Image Segmentation," arXiv preprint arXiv:2402.05079, 2024. doi: 10.48550/arXiv.2402.05079.
- [10] F. Isensee *et al.*, "nnU-Net: Self-Adapting Framework for U-Net-Based Medical Image Segmentation," in *Bildverarbeitung für die Medizin 2019*, 2019, pp. 22–29. doi: **10.1007/978-3-658-25326-4_7**.
- [11] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11976–11986. doi: **10.1109/CVPR52688.2022.01167**.
- [12] J. Chen *et al.*, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv preprint arXiv:2102.04306, 2021. doi: **10.48550/arXiv.2102.04306**.
- [13] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment Anything in Medical Images," *Nature Communications*, vol. 15, Art. no. 654, 2024. doi: **10.1038/s41467-024-44824-z**.
- [14] J. Ruan, J. Li, and S. Xiang, "VM-UNet: Vision Mamba UNet for Medical Image Segmentation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2025. doi: **10.1145/3767748**.
- [15] M. G. Shahraki, M. Rezaeian, and M. Ghasemzadeh, "A Hybrid Mamba-SAM Architecture for Efficient 3D Medical Image Segmentation," arXiv preprint arXiv:2602.00650, 2026. Doi: 10.48550/arXiv.2602.00650.
- [16] T. Dao and A. Gu, "Transformers are SSMS: Generalized Models and Efficient Algorithms Through Structured State Space Duality," in *Proc. 41st International Conference on Machine Learning (ICML)*, PMLR, vol. 235, pp. 10041–10071, 2024. doi: **10.48550/arXiv.2405.21060**.