Hierarchical Mamba Networks for Multi-Organ Segmentation in Abdominal Medical Imaging

Gerald R. Fields
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
gerald1998@uc.edu

Umesh R. Trivedi
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
umesh.trivedi@unr.edu

Abstract

Abdominal multi-organ segmentation remains a critical task in medical image analysis because quantitative organ delineation supports treatment planning, surgical guidance, and disease monitoring. Although deep learning has produced substantial advances, convolutional approaches tend to emphasize local spatial patterns, while transformer-based methods can capture long-range context at considerable computational cost. Hierarchical Mamba networks represent an emerging alternative that combines selective state space sequence modeling with multiscale representation learning. This paper provides a system-level analysis of hierarchical Mamba networks for multi-organ segmentation in abdominal imaging. Rather than focusing solely on architectural novelty, the discussion addresses structural trade-offs in representation capacity, memory use, scan ordering, and multiscale fusion. It further examines data infrastructure and governance requirements, including annotation quality, demographic bias, privacy protections, and federated learning arrangements. Deployment concerns such as prospective validation, input verification, uncertainty visualization, and integration with clinical workflows are explored. Fairness, regulatory accountability, and sustainability are treated as first-order design constraints rather than post hoc considerations. The paper argues that hierarchical Mamba architectures can support more efficient long-range modeling than attention-based alternatives, but their clinical value depends on robust governance, operational monitoring, and deliberate alignment with diverse patient populations. Future directions include adaptive scan strategies, uncertainty-aware abstention, longitudinal cross-fusion, and resource-aware deployment for lower-resource settings.

Keywords

hierarchical Mamba networks, multi-organ segmentation, abdominal medical imaging, state space models, clinical deployment, data governance, fairness, sustainability.

1. Introduction

Automated multi-organ segmentation in abdominal computed tomography and magnetic resonance imaging has become a foundational problem in medical image analysis because quantitative measurements are increasingly embedded in clinical decision making. Early deep learning approaches established the U-shaped encoder-decoder design with skip connections as a widely adopted framework [1]. Subsequent self-configuring pipelines demonstrated that systematic adaptation of network configuration, preprocessing, and postprocessing could outperform many bespoke segmentation systems [2]. Despite these achievements,

simultaneous segmentation of the liver, spleen, kidneys, pancreas, and other abdominal structures remains difficult because organs vary in shape, neighboring tissues often share similar intensities, and pathologic changes can alter normal boundaries.

One limitation of purely convolutional stacks is that they encode local spatial regularities more naturally than long-range anatomical dependencies. Vision transformer methods [3] and hierarchical windowed transformer designs [4] introduced global attention mechanisms capable of modeling distant context, while hybrid architectures such as UNETR [5] integrated transformer encoders with convolutional decoders for volumetric segmentation. These approaches improved contextual modeling but frequently imposed high memory requirements and computational burdens when applied to high-resolution abdominal images. Attention-based architectures can be especially demanding during training and hyperparameter search, which has practical implications for clinical institutions with limited computational resources.

State space models have recently emerged as a promising alternative for sequence modeling. The selective scan mechanism introduced by Mamba [6] supports long-range dependency modeling while scaling more efficiently than full attention. Early medical adaptations such as Mamba-UNet [7] and SegMamba [8] applied state space principles to medical image segmentation, suggesting that such models can preserve global context without the same quadratic computational growth observed in transformer designs. A cross-fusion Mamba formulation for multi-view medical image classification [9] further demonstrated that complementary representations can be integrated through multiple Mamba-based pathways. These developments motivate the design of hierarchical Mamba networks that operate across several spatial resolutions.

However, architectural performance alone does not determine clinical success. Multi-organ segmentation should be understood as a socio-technical systems problem involving data provenance, institutional governance, deployment infrastructure, human workflow integration, regulatory compliance, and sustainability. The present paper examines hierarchical Mamba networks from this system-level perspective. It emphasizes structural trade-offs that arise when designing multiscale state space models, discusses the governance mechanisms needed for dependable data and model lifecycle management, and considers the fairness and policy implications of deploying these systems in heterogeneous healthcare environments.

2. Hierarchical Mamba Networks and Structural Trade-offs

Hierarchical Mamba networks can be understood as multiscale sequence models that apply selective state space operations across multiple spatial resolutions. General-purpose visual Mamba architectures [10] introduced bidirectional scanning and hierarchical patch merging, enabling the network to retain local detail while aggregating global semantic context. The selective state space mechanism [6] determines how information is propagated along the sequence, allowing the model to emphasize salient boundaries and suppress irrelevant background regions. In abdominal imaging, this selectivity can be particularly useful because organs such as the pancreas may have ambiguous margins that require contextual evidence from surrounding structures.

Medical adaptations have demonstrated that state space operations can be integrated with U-shaped decoders [7] and volumetric scanning strategies [8]. The cross-fusion Mamba classifier [9] suggests that multiple scan pathways with distinct ordering rules can be fused to improve representational completeness. For multi-organ segmentation, cross-fusion can be extended across hierarchical levels, where fine-grained representations resolve small ducts

and vascular structures while coarse representations stabilize larger organs such as the liver and kidneys. Hybrid designs that embed transformer-based attention blocks, such as TransUNet [11], provide useful comparisons because they also combine global context with convolutional localization, but their attention complexity and memory profile differ substantially from state space alternatives.

Structural decisions in hierarchical Mamba networks have direct clinical consequences. The number of hierarchical stages influences how well the model can separate organs with different scales, but deeper hierarchies also increase the number of latent feature maps that must be stored and synchronized during training. The internal state dimension controls selective memory capacity, yet larger states increase memory use and can reduce throughput. Bidirectional scanning improves robustness to ordering but approximately doubles the amount of state space computation unless shared parameters or efficient fusion mechanisms are used. Such trade-offs are not merely technical; they influence how easily a model can be validated, inspected, retrained, and deployed within hospital infrastructure.

For abdominal imaging, the heterogeneity of organ scales is especially challenging. The pancreas may occupy a small region with low contrast, while the liver and spleen present large but variable structures. Hierarchical Mamba networks can allocate several scales to different organs, but system architects must decide whether to use shared weights across scales or specialized branches. Shared weights reduce parameter counts and may lower overfitting risk, whereas specialized branches permit scale-specific inductive biases. This choice also affects fairness across patient subgroups because rare anatomical variants and small pathologies may be underrepresented in training distributions. A model optimized for average performance may fail precisely for the cases that require the most clinical attention.

Another structural trade-off concerns the ordering of image regions when flattening three-dimensional volumes into sequences. Simple raster scanning can produce artificial discontinuities at slice boundaries, while anatomically aware ordering may require prior detection or segmentation. Hierarchical methods can combine multiple orderings through cross-fusion [9], but this added complexity may reduce interpretability. Clinicians often need to understand why a boundary was placed, and recursive scan paths can make attribution more difficult. Consequently, architectural choices should be paired with explainability strategies that are tailored to the clinical task and the intended user.

The design of skip connections and feature fusion also affects system behavior. U-shaped networks transmit fine spatial information from encoder to decoder, but state space pathways may transform this information in ways that are difficult to audit. Hierarchical Mamba networks can include gated fusion blocks that selectively combine coarse and fine features, but these blocks must be validated for stability across different imaging protocols. A segmentation system that works well on portal venous phase computed tomography may exhibit unexpected behavior on non-contrast studies if the fusion mechanisms rely on contrast-dependent intensity patterns. System-level design should therefore include protocol metadata and acquisition-aware normalization as explicit inputs to the segmentation pipeline.

3. Data Infrastructure and Governance

Robust multi-organ segmentation depends heavily on data quality and provenance. A broad survey of deep learning in medical image analysis documented both rapid progress and persistent challenges related to limited annotations and heterogeneous acquisition protocols [12]. Preparing medical imaging data for machine learning requires careful attention to

curation, labeling, harmonization, and institutional review before model training begins [13]. In multi-organ segmentation, consistent anatomical labeling across multiple raters is especially difficult because small organs and vascular structures can be interpreted differently. Without rigorous annotation standards, even a well-designed hierarchical Mamba network will learn from noisy boundaries and propagate these errors into clinical measurements.

Data governance frameworks must address consent, de-identification, and geographic differences in privacy law. Large abdominal datasets often combine images from multiple scanners and populations, and the aggregation of such data can magnify existing biases if underlying demographics and acquisition parameters are not documented. Data cascades in high-stakes artificial intelligence demonstrate how upstream data quality problems propagate into downstream model failures [14]. A segmentation model may appear accurate on internal validation data while failing on patients whose imaging protocols, body habitus, or organ appearances differ from the training majority. This risk is especially relevant in abdominal imaging because clinical protocols vary widely across institutions.

Fairness considerations begin at the data level. If training sets overrepresent certain scanner vendors, disease prevalences, or demographic groups, performance can vary substantially across subgroups [15]. Health equity researchers argue that such disparities should be treated as system failures rather than external limitations [16]. Federated learning [17] offers one governance pathway by enabling multiple institutions to contribute model updates without sharing raw images, but it also introduces coordination burdens, audit requirements, and risks of indirect information leakage. Hierarchical Mamba networks may be well suited to federated settings because their sequence operations can be configured for relatively efficient computation, yet this advantage depends on the specific communication and security infrastructure available across sites.

An effective data infrastructure should include versioned annotation guidelines, periodic inter-rater agreement audits, and explicit documentation of acquisition parameters. Multi-organ segmentation also requires quality assurance for label boundaries, because small shifts in organ contours can affect volumetry and radiation dose planning. Governance committees should establish criteria for when a model may be retrained, how new sites are onboarded, and which data elements are stored for audit. These decisions should involve radiologists, technologists, data stewards, and patient representatives rather than being left solely to machine learning engineers.

4. Training, Deployment and Operational Robustness

Operational robustness in clinical artificial intelligence requires more than high segmentation accuracy on retrospective data. Responsible machine learning for health care emphasizes prospective validation, explicit uncertainty communication, and monitoring for errors that could harm patients [18]. For hierarchical Mamba networks, training should include realistic augmentation strategies that simulate imaging artifacts, contrast phase differences, and patient motion. Multi-organ segmentation models can learn spurious correlations if training data are not representative, and these correlations may remain hidden until deployment. The training process should therefore be designed as part of a broader quality management system rather than as an isolated computational experiment.

Deployment scenarios range from cloud-based processing of large cohorts to embedded inference on hospital workstations. Hierarchical Mamba networks can reduce peak memory relative to transformer-based alternatives because their sequence modeling avoids storing full

attention matrices. However, the selective scan mechanism may require optimized kernels and careful memory alignment to maintain production throughput. The system architecture must include failover pathways, input validation, and versioned model serving so that segmentation results remain reproducible. If a scanner produces a series with unusual spacing or contrast, the system should flag it instead of silently producing uncertain contours.

Human factors are equally important. Radiologists and surgeons will not adopt an automated system if it creates excessive false positives or requires unexplained manual corrections. Segmentation output should be embedded within existing picture archiving and communication systems, with interactive overlay and confidence visualization. Regulatory frameworks increasingly expect medical software to provide a valid clinical claim and to define the intended use population [19]. For abdominal segmentation, intended use may vary across oncologic staging, transplant planning, or emergency trauma workflows, and each use requires distinct validation. A hierarchical Mamba model that is suitable for volumetric reporting may not be appropriate for real-time intraoperative guidance.

Continuous monitoring is necessary because imaging protocols, scanner upgrades, and patient demographics change over time. A deployed model may experience performance drift that is not immediately visible from aggregate statistics. Institutions should define thresholds for alerting and evaluation cycles, and they should maintain datasets that reflect current operational conditions. Hierarchical Mamba networks can support such monitoring by encoding uncertainty at multiple scales, but systems must be designed to log these outputs without compromising patient privacy. The operational infrastructure should also include mechanisms for user feedback, because clinicians who edit contours are producing valuable signals that can improve model safety and local adaptation.

5. Fairness, Generalization and Policy Implications

Broader policy questions concern whether automated segmentation systems widen or narrow disparities in access to high-quality imaging interpretation. Topol [20] argues that artificial intelligence can humanize medicine if it is used to augment clinical attention rather than replace it, but this requires careful attention to which patient groups benefit. Multi-organ segmentation can reduce workload and improve consistency, yet it may perform less well for rare anatomic variants or populations underrepresented in training data. These disparities must be measured and reported in ways that are meaningful to clinicians and regulators.

Generalization across scanner manufacturers and imaging protocols remains a central challenge. A model trained on portal venous phase computed tomography may fail on non-contrast or arterial phase images, and this limitation should be documented rather than hidden. Regulatory policy in multiple jurisdictions has moved toward lifecycle accountability for medical artificial intelligence, including postmarket surveillance and risk-based classification [19]. Hierarchical Mamba networks do not inherently escape these requirements. Their architectural novelty may even raise regulatory questions about explainability, failure analysis, and the stability of selective state space mechanisms under distribution shift.

Fairness audits should examine model performance across sex, age, body size, and disease subtype. Existing bias and fairness frameworks [15] provide structured methods for measuring disparities, but translating such metrics to segmentation tasks requires organ-specific definitions of clinically meaningful differences. Health equity guidance [16] emphasizes that fairness is not simply statistical parity; it is the alignment of model behavior with patient needs across diverse populations. For abdominal segmentation, clinically

meaningful errors include missed small lesions, incorrect organ boundaries that affect radiation fields, and inconsistent labeling between nearby structures. A small numerical improvement in average Dice performance may be less important than reducing failures for underserved groups.

Policy implications also extend to data sharing and intellectual property. Collaborative consortia must balancethe benefits of data pooling against the risks of privacy loss and competitive misuse. Federated learning [17] can reduce raw data sharing, but it introduces new forms of coordination overhead and does not automatically prevent indirect leakage or inequitable participation. Policy makers should consider whether reference datasets used for regulatory validation are themselves representative, because a model may receive approval based on a narrow population and then fail in broader clinical use. The segmentation system should therefore be treated as a socio-technical intervention whose benefits and harms are distributed unevenly across patients, providers, and institutions.

6. Sustainability and Infrastructure Economics

The computational cost of training and serving large segmentation models has direct implications for sustainability and equitable access. Deep learning research has repeatedly demonstrated that accuracy gains can come at the cost of substantial energy use and carbon emissions [21]. Hierarchical Mamba networks may reduce attention-related memory burdens, but they still require careful optimization of training schedules, mixed precision, and distributed communication. The infrastructure economics of medical artificial intelligence are not neutral: institutions with access to high-performance computing can iterate faster and adapt models to local populations, while smaller centers may depend on external cloud services that introduce data governance complexities.

A comparative infrastructure analysis should consider not only peak memory and training time but also the environmental footprint of hyperparameter tuning and repeated evaluation. The computational limits of deep learning warn that continued scaling is not guaranteed to yield proportional clinical benefit [22]. For abdominal segmentation, where many organs are already well segmented by modern methods, additional model complexity may primarily improve rare cases. System designers should therefore ask whether a hierarchical Mamba model is the most resource-appropriate choice for a given clinical task, or whether a lighter convolutional network with strong preprocessing could suffice. This question matters for long-term maintainability and for the ability of lower-resource settings to participate in model development and validation.

Sustainability also includes human and institutional sustainability. A model that demands frequent retraining or specialized hardware may divert resources from other health system priorities. The governance mechanisms discussed earlier should include explicit budget assessments for energy, compute, and expert time. By treating sustainability as part of the design space rather than an afterthought, hierarchical Mamba networks can be deployed in ways that align with institutional capacity and broader environmental responsibilities.

7. Future Directions

Future research should move beyond isolated architectural benchmarks and toward longitudinal system evaluation in heterogeneous clinical environments. One promising direction is adaptive scan ordering, in which the hierarchical Mamba network learns to reorder image regions based on preliminary segmentation confidence, thereby focusing selective state space capacity on ambiguous boundaries. Such adaptivity could improve

segmentation of small organs and rare variants, but it requires careful validation to ensure that adaptive strategies do not introduce protocol-dependent failures. Another direction is uncertainty-aware abstention, where the model quantifies agreement across different scan paths and hierarchical scales. If cross-fusion confidence is low, the system could flag the case for manual review or request additional imaging. This behavior aligns with responsible machine learning principles that emphasize explicit uncertainty communication in health care [18].

A further opportunity lies in longitudinal abdominal imaging. Patients may undergo multiple scans over time, and changes in organ size, shape, and appearance can provide important diagnostic signals. Hierarchical Mamba networks could be extended to process temporal sequences of volumetric images, but this requires careful handling of registration variability and missing time points. Cross-fusion mechanisms similar to those used in multi-view classification [9] may be adapted to combine current and prior scans without relying on perfect alignment. Longitudinal systems also raise new governance questions, because model predictions could influence future clinical decisions and create feedback loops that are difficult to audit.

Generative data augmentation and synthetic imaging present additional avenues. If governed appropriately, synthetic abdominal images can improve representation of rare anatomical variants and scanner conditions, but they also raise concerns about data authenticity and model collapse. Any future deployment should include prospective clinical studies that measure workflow impact, clinician trust, and patient outcomes rather than relying solely on geometric overlap metrics. Policy engagement should run parallel to technical development, with regulators, patient advocates, and clinical societies jointly defining acceptable performance thresholds for automated abdominal segmentation.

8. Conclusion

Hierarchical Mamba networks offer a promising system architecture for multi-organ segmentation in abdominal medical imaging because they combine multiscale representation learning with efficient long-range sequence modeling. The selective state space mechanism can capture contextual anatomical relationships without the prohibitive memory demands of full attention, and cross-fusion strategies can integrate multiple scan paths to improve representational completeness. However, this paper has argued that technical performance should not be evaluated in isolation. The structural trade-offs of hierarchy, state size, scan ordering, and fusion must be weighed against clinical workflows, institutional governance, and patient-centered outcomes. Data infrastructure and annotation standards are decisive, and fairness cannot be achieved without deliberate attention to demographic, protocol, and anatomical diversity. Deployment requires prospective validation, uncertainty communication, continuous monitoring, and regulatory alignment. Sustainability considerations further demand that computational efficiency be understood as part of responsible health system design. By treating segmentation as a socio-technical system, hierarchical Mamba networks can move beyond benchmark performance toward dependable, equitable, and resource-aware support for abdominal imaging practice.

References

1. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234–241). Springer.

2. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203–211.

3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

4. 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. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012–10022).

5. 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. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 574–584).

6. Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

7. Wang, Z., Zheng, J.-Q., Zhang, Y., Cui, G., & Li, L. (2024). Mamba-UNet: UNet-like pure visual Mamba for medical image segmentation. arXiv preprint arXiv:2402.05079.

8. Xing, Z., Ye, T., Yang, Y., Liu, G., & Zhu, L. (2024). SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024 (pp. 578–588). Springer.

9. Zheng, X., Chen, X., Gong, S., Griffin, X., & Slabaugh, G. (2025, September). Xfmamba: Cross-fusion mamba for multi-view medical image classification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 672-682). Cham: Springer Nature Switzerland.

10. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. In Proceedings of the 41st International Conference on Machine Learning.

11. 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 preprint arXiv:2102.04306.

12. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.

13. Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing medical imaging data for machine learning. Radiology, 295(1), 4–15.

14. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–15).

15. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

16. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872.

17. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, Article 119.

18. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337–1340.

19. Muehlematter, U. J., Daniore, P., & Vokinger, K. N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe. The Lancet Digital Health, 3(3), e195–e203.

20. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

21. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).

22. Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. arXiv preprint arXiv:2007.05558.