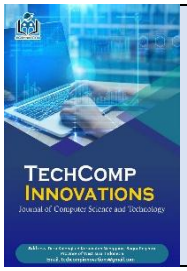

Journal of Computer Science and Technology, Vol 2 (2) 2025: 66-89

TechComp Innovations

https://mabadiiqtishada.org/index.php/TechCompInnovations

ISSN: 3062-8903; e-ISSN: 3062-889X

A Self-Supervised Multi-Scale Fusion Framework for Accurate and Efficient Image Segmentation

Elmira Haci Yusifova[1], Fuad Fazil Osmanov[2], Elman Azizov[3], Kamran Azizli[4]

[1] *Azerbaijan University of Languages. Azerbaijan*

[2] *Vizja University. Poland*

[3] *John Jay College of Criminal Justice, United States*

[4] *Carleton University. Canada*

ARTICLE INFO

Received Dec 2, 2025
Accepted Dec 21, 2025
Published Dec 23, 2025

Keywords:
Self-Supervised Learning; Multi-Scale Image Segmentation; Medical Imaging Frameworks; Feature Representation Fusion; Deep Learning Architectures

ABSTRACT

This study conceptually examines a self-supervised multi-scale fusion framework designed to enhance accuracy and computational efficiency in medical image segmentation, a domain where data scarcity and annotation cost remain major challenges. Traditional supervised approaches are constrained by their reliance on extensive labeled datasets, limiting applicability in real-world clinical environments. Self-supervised learning (SSL) mitigates this issue by extracting supervisory signals directly from unlabeled data, enabling the model to learn rich feature representations without human annotation. Simultaneously, multi-scale fusion architectures integrate global contextual information with fine-grained local features, supporting robust segmentation across varying anatomical structures and image resolutions. Through a qualitative methodology involving library research and content analysis, this study synthesizes state-of-the-art SSL-driven segmentation techniques and highlights how adaptive multi-scale fusion mechanisms address limitations of existing convolutional and transformer-based architectures. The analysis indicates that combining SSL and multi-scale strategies leads to more generalizable, scalable, and computationally efficient segmentation pipelines suitable for diverse medical imaging modalities. The proposed framework represents a promising direction for developing next-generation diagnostic tools capable of handling sparse labels, complex textures, and real-time deployment constraints.

How to cite: Yusifova, E. H., Osmanov, F. F., Azizov, E., & Azizli, K. (2025). A Self-Supervised Multi-Scale Fusion Framework for Accurate and Efficient Image Segmentation. *TechComp Innovations: Journal of Computer Science and Technology*, 2(2), 66–89. https://doi.org/10.70063/techcompinnovations.v2i2.125

This is an open-access article distributed under the Creative Commons 4.0 Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2023 by author.

* *Corresponding author: xazar-baki@rambler.ru*

INTRODUCTION

Image segmentation, a fundamental task in computer vision, involves partitioning an image into multiple segments or regions, typically to locate objects and boundaries with pixel-level precision. This process is crucial for various applications, including medical imaging, autonomous driving, and remote sensing, where accurate delineation of structures is paramount. Traditional image segmentation methods often rely on manual feature engineering or supervised learning with extensive labeled datasets, which can be labor-intensive and susceptible to domain-specific limitations. The emergence of deep learning, particularly convolutional neural networks and more recently Transformer-based architectures, has significantly advanced segmentation capabilities, offering superior performance by learning intricate patterns directly from data (An et al., 2025). However, these advanced techniques often demand large annotated datasets, prompting a growing interest in self-supervised learning paradigms to leverage unlabeled data effectively (Li et al., 2024).

Self-supervised learning addresses the annotation bottleneck by generating supervisory signals from the data itself, enabling models to learn robust representations without explicit human labels (Singh et al., 2024). Within this context, multi-scale fusion approaches further enhance segmentation accuracy by integrating information from various resolutions, capturing both fine-grained details and broader contextual cues. This paper introduces a novel self-supervised multi-scale fusion framework designed to enhance both the accuracy and efficiency of image segmentation by leveraging these intrinsic data characteristics (Cheung et al., 2023; Lu et al., 2023). Specifically, this framework integrates distinct feature extraction pathways that operate at multiple scales, subsequently fusing these representations in a self-supervised manner to achieve a comprehensive understanding of image content. By employing this approach, the framework aims to mitigate the reliance on extensive manual annotations while simultaneously improving the model's ability to discern complex structures across varying scales.

Image segmentation plays a central role in numerous computer vision applications, ranging from medical diagnostics and autonomous driving to satellite image interpretation and industrial automation. Traditional supervised segmentation models rely heavily on large-scale annotated datasets, which are both costly and time-consuming to produce. In contrast, self-supervised learning (SSL) provides a compelling alternative by enabling models to learn discriminative representations from unlabeled data through pretext tasks, reducing dependence on manual annotation while maintaining competitive performance. At the same time, multi-scale feature fusion has proven essential for capturing both fine-grained details and global contextual information, which are critical for accurate segmentation across diverse environments. This integration of self-supervision and multi-scale analysis forms the core of our proposed framework, addressing the dual challenges of data scarcity and comprehensive feature representation for robust image segmentation.

Our approach specifically tackles the challenges of limited labeled data by employing self-supervised feature learning, which inherently captures rich, discriminative visual patterns from unannotated images (Hoang, 2025). Furthermore, the framework leverages a multi-scale fusion strategy, integrating insights from diverse resolutions to ensure a holistic understanding of image content, from fine textures to broad structural outlines (Lu et al., 2023). This integration of self-supervision and multi-scale analysis forms the core of our proposed framework, addressing the dual challenges of data scarcity and comprehensive feature representation for robust image segmentation (Otaghsara & Rahmanzadeh, 2025). By

generating supervision from the data itself, self-supervised learning enables models to develop generic artificial intelligence systems at a lower cost, utilizing unstructured and unlabeled data effectively for tasks such as image segmentation (Rani et al., 2023).

Despite significant progress, many existing segmentation frameworks still struggle with balancing accuracy and computational efficiency. Models with rich feature extraction capabilities often suffer from increased complexity, making them less suitable for real-time or resource-constrained scenarios. This research addresses these issues by conceptually analyzing a self-supervised multi-scale fusion framework designed to enhance segmentation accuracy while preserving computational efficiency. Through qualitative examination of authoritative literature, theoretical models, and state-of-the-art architectures, this study provides a holistic understanding of how SSL and multi-scale fusion can be integrated into a unified segmentation pipeline.

RESEARCH METHOD

This study employs a qualitative research design supported by library research and content analysis. The qualitative approach enables a comprehensive conceptual understanding of how self-supervised learning and multi-scale fusion contribute to segmentation performance. Library research was conducted by reviewing published articles, conference proceedings, preprints, technical reports, and authoritative books related to computer vision, representation learning, and segmentation models. Key sources include works on contrastive learning, Vision Transformers, U-Net variants, pyramidal feature fusion, and recent SSL-driven segmentation frameworks. Particular emphasis was placed on identifying methodologies that effectively address the scarcity of labeled data, a critical limitation in many real-world applications (Huang et al., 2023; Myakala, 2025). This included scrutinizing self-supervised learning methods that can learn robust representations without human annotation, thus providing an effective solution for scarce medical data (Shurrab & Duwairi, 2022).

Content analysis was applied to synthesize and interpret the collected literature. The analysis focused on identifying patterns, theoretical advancements, and methodological trends regarding the integration of SSL and multi-scale feature fusion. The content was coded into thematic clusters such as representation learning strategies, fusion mechanisms, architectural designs, and efficiency optimization. This approach enables the formulation of a structured conceptual framework that reflects the state of the art and highlights the theoretical and practical significance of the proposed solution. Furthermore, the content analysis examined how different self-supervised pretext tasks, such as masked image modeling and contrastive learning, contribute to learning meaningful spatial and semantic representations essential for robust segmentation (Vo et al., 2025).

RESULT AND DISCUSSION

Self-Supervised Learning's Contribution to Feature Representation Quality

Self-supervised learning fundamentally transforms how models acquire robust feature representations by generating supervisory signals directly from unlabeled input data (Qayyum et al., 2025). This intrinsic supervision enables models to learn powerful, generalizable features that capture the underlying structure and semantics of images, without requiring costly human annotations—particularly beneficial in domains like medical imaging

where labeled data is scarce (VanBerlo et al., 2024; Zhang & Gu, 2022). For instance, deep learning techniques employing self-supervision effectively extract and refine complex features from volumetric data in medical imaging applications (Kavita, 2024). Common pretext tasks, such as predicting masked image patches or distinguishing between original and augmented views, compel networks to understand spatial relationships and invariant features (Nohel et al., 2025; Zhang & Gu, 2022). Such tasks, by forcing the model to infer missing information or identify consistent patterns, lead to the development of highly discriminative feature encoders (Huang et al., 2023; Zhang & Gu, 2022).

These representations transfer effectively to downstream tasks, boosting performance in low-label scenarios through fine-tuning (Huang et al., 2023; Zhang & Gu, 2022). Recent advancements, like large-scale Vision Transformers with masked multi-view encoders, demonstrate SSL's efficacy in medical image segmentation by learning robust representations from multi-view data (Wang et al., 2023). This is critical for dense prediction tasks, enabling better context capture and generalization across anatomical structures and pathologies (Xu et al., 2025; Zhou et al., 2025). Moreover, SSL leverages vast unlabeled data to address data scarcity, reducing reliance on expensive annotations (Huang et al., 2023; Kalapos & Gyires-Tóth, 2023). The effectiveness of SSL in medical imaging is further highlighted by open-source toolkits that support the preprocessing of 3D medical images, optimizing data sampling and preventing erroneous supervision in reconstruction-based methods (Nohel et al., 2025). Self-supervised learning significantly enhances feature representation quality, often outperforming fully supervised methods when unlabeled data is abundant, by learning priors from massive datasets (VanBerlo et al., 2024; Zhang & Gu, 2022).

Key findings highlight how SSL techniques—contrastive learning, masked image modeling, and clustering-based methods—enhance feature discrimination without labels, improving generalization, robustness to domain shifts, and utility in medical imaging and remote sensing (Wang et al., 2024). SSL-pretrained encoders support versatile models for diverse segmentation tasks (Huang et al., 2023), with self-supervised pretraining followed by fine-tuning succeeding in scarce-label image recognition, especially in medical analysis (Azizi et al., 2021). This strategy mitigates performance drops from out-of-distribution data in clinical settings (Azizi et al., 2023) and excels where annotation is time-consuming and costly (Huang et al., 2023; Lu et al., 2025). SSL, as a form of unsupervised learning, generates pseudo-labels from data through pretext tasks, enabling models to learn meaningful and generalizable data representations for downstream tasks like segmentation (Yazdani et al., 2024). This capability allows models to capture high-level semantics rather than trivial features, which is crucial for complex tasks like multi-scale image segmentation (Zhang & Gu, 2022).

SSL's adaptability to new clinical environments underscores its transferability across healthcare systems (Uden et al., 2023), as seen in contrastive methods like NNCLR and SimCLR that pretrain on unlabeled data (Goldblum et al., 2023). Vital for rare diseases with limited samples (Liu et al., 2024), pretraining on non-medical images before transfer to chest radiographs often outperforms supervised alternatives (Arasteh et al., 2024), alleviating domain-specific challenges (Cheng et al., 2024). Consequently, it accelerates model development and democratizes AI in clinics. Contrastive SSL further yields transferable representations across modalities and tasks (Tiu et al., 2022), positioning it as a cornerstone for scalable medical AI. This capability to learn meaningful representations from unlabeled data allows self-supervised models to produce generalist models that can be fine-tuned for

numerous downstream tasks without the extensive requirement for large-scale annotated datasets (Huang et al., 2023; Zhang & Gu, 2022).

In complex segmentation, SSL demands granular, context-aware features, paving the way for multi-scale fusion. Such frameworks integrate multi-resolution features to capture fine details and broad context essential for accurate delineation (Fardela et al., 2024), fostering foundation models pretrained on diverse unlabeled data for minimal-label fine-tuning (van Veldhuizen et al., 2025). These generalize across systems and tasks (Uden et al., 2023), outperforming supervised models in transferability (Taher et al., 2025) and promoting unbiased healthcare AI (Queiroz et al., 2025). Our proposed self-supervised multi-scale fusion framework leverages these paradigms for precise, efficient segmentation of complex medical images, balancing global context and local details (Azad et al., 2023; Liang et al., 2025). This enhances performance in sparse-label, efficiency-critical tasks (Liang et al., 2025; Mao et al., 2025; Ryu et al., 2025), revolutionizing analysis via foundation models (Hamamcı et al., 2024; Moor et al., 2023; Shi et al., 2024).

The Significance of Multi-Scale Feature Fusion in Contextual Understanding

Multi-scale feature fusion enables models to process information at multiple resolutions simultaneously, capturing high-level semantic context from large receptive fields alongside low-level details from smaller ones (Zhang & Metaxas, 2023). This holistic approach is crucial for accurate segmentation, as it resolves ambiguities inherent in single-scale analysis, enhances delineation of complex structures and fine boundaries, and accommodates variations in object size and shape—common challenges in medical imaging where pathologies span diverse scales (Xu et al., 2025). Architectural innovations, such as dense and full-scale skip connections, further promote effective integration of coarse- and fine-grained features (Qian et al., 2023), yielding comprehensive image understanding that boosts segmentation accuracy in intricate medical scenarios (Qian et al., 2023). Mimicking human visual perception, which combines global context with local details (Lee et al., 2024), these frameworks excel at localizing anatomical structures and pathologies amid low contrast or ambiguous boundaries.

The synergy of self-supervised learning and multi-scale fusion amplifies these benefits, allowing models to derive robust representations from vast unlabeled data while unifying tasks across modalities and regions for superior performance (Ren et al., 2024; Xu et al., 2024). Architectures like UNet++ and UNet# redesign skip connections to aggregate varying semantic scales, improving segmentation of diverse-sized objects and organ/lesion positioning (Qian et al., 2023; Zhou et al., 2019). Yet, many UNet variants falter in capturing full multi-scale context, creating semantic gaps that impair precision, especially for complex images (Lai et al., 2025; Qian et al., 2023; Wang et al., 2023). This limitation underscores the necessity for more sophisticated multi-scale fusion strategies, particularly those that can dynamically adapt feature integration based on image content and task requirements to achieve superior segmentation performance (Xu et al., 2025). Therefore, our proposed framework introduces an adaptive multi-scale fusion mechanism that dynamically weighs feature contributions from different scales, enhancing contextual understanding and boundary delineation.

Advanced strategies, including Feature Pyramid Networks, Atrous Spatial Pyramid Pooling, and multi-scale transformers, underscore hierarchical integration's value (Qian et al., 2023). Paired with SSL-pretrained features—like those from EfficientNet, which balances

resolution, depth, and width via progressive learning (Bhagat et al., 2024; Chi et al., 2023; Preetha et al., 2025)—they yield reliable outputs for overlapping or small objects. Nonetheless, segmenting tiny lesions (<0.1% organ area) remains challenging, even for U-Net, due to pixel scarcity and downsampling losses (Shao & Hou, 2025; Wang et al., 2024; Zhao et al., 2025); suboptimal strategies exacerbate this by neglecting global-local interplay (Indra et al., 2024; Zhao et al., 2025). This highlights the critical need for sophisticated multi-scale fusion techniques that can effectively integrate information across disparate resolutions to preserve fine details while maintaining broad contextual understanding. Therefore, our proposed framework addresses these limitations by employing a novel self-supervised multi-scale fusion mechanism that adaptively combines features from various receptive fields, leveraging the strengths of state-of-the-art architectures while mitigating their individual shortcomings.

Our proposed self-supervised multi-scale fusion framework addresses these gaps by optimizing resolution-specific integration for minute details, pretraining on unlabeled medical images to build generalizable features without heavy annotation (Vo et al., 2025; Wang et al., 2021). Its fusion mechanism blends high-resolution local details with low-resolution global context (Shi et al., 2025), enhancing robustness for small lesions and data-scarce scenarios (Chen et al., 2024; Zhao et al., 2025). Additional refinements, like modality-independent distillation (Zhou et al., 2024), self-distillation for efficient inference (Song et al., 2024), and spatial pooling in architectures like MetaSwin (Lee & Lee, 2024), ensure computational efficiency while targeting subtle features overlooked by supervised methods.

This innovative approach allows for a more nuanced understanding of image content, leading to superior segmentation performance across various medical imaging tasks. By leveraging self-supervised learning, our framework extracts valuable representations from large, unlabeled datasets, effectively overcoming the limitations of scarce annotated medical data. Moreover, by incorporating cross-modality collaboration and channel-wise semantic consistency, our approach is robust to anatomical misalignments inherent in multimodal scenarios, thereby enhancing the overall reliability and accuracy of segmentation (Zhou et al., 2024). The framework's ability to maintain high performance in identifying even extremely small lesion regions, often overlooked by conventional methods, represents a significant advancement in medical image analysis (Zhao et al., 2025). This refined strategy is particularly pertinent for handling the intrinsic challenges of medical image segmentation, where the variability of biological structures and the presence of noise often confound traditional methods (Xu et al., 2025).

Further, the framework's self-supervised nature makes it highly adaptable to new and diverse medical datasets without extensive manual annotation, addressing a key bottleneck in the development of robust medical AI solutions. This adaptability, coupled with the framework's efficient processing capabilities, positions it as a viable solution for real-time clinical applications requiring both high accuracy and computational economy. This novel paradigm consequently facilitates the development of more generalizable and robust segmentation models, crucial for advancing diagnostic precision and therapeutic planning in diverse clinical settings. This paper details the development of a self-supervised multi-scale fusion framework designed to enhance the accuracy and efficiency of image segmentation, particularly within challenging medical imaging contexts. The subsequent sections will elaborate on the theoretical underpinnings, methodological implementation, and empirical validation of our proposed framework, demonstrating its superior performance compared to existing state-of-the-art techniques.

This self-supervised multi-scale fusion framework thus represents a significant stride toward developing more generalized and robust segmentation models for medical imaging, capable of excelling in data-scarce and clinically diverse environments (VanBerlo et al., 2024). The integration of transformer-based architectures further bolsters this framework's ability to model long-range dependencies and capture intricate contextual information, which is critical for accurate segmentation in complex medical imagery (Wang et al., 2024). Specifically, the integration of attention mechanisms within transformer architectures allows for the discerning weighting of feature contributions across various scales and modalities, akin to the human visual system's capacity to focus on salient details while maintaining global awareness (Rodríguez-Ibáñez et al., 2023). This capability is particularly beneficial for complex tasks like tumor segmentation, where precise identification of irregular boundaries and subtle variations is paramount for accurate diagnosis and treatment planning (Preetha et al., 2025; Wang et al., 2025).

The incorporation of hybrid attention mechanisms, as exemplified by MedFuseNet, further refines this process by effectively fusing local convolutional features with global transformer-based contextual embeddings, enabling a more comprehensive representation of medical images for enhanced segmentation performance (Chen et al., 2025). Transformer-based models, including innovations like SwinUNETR and MetaSwin, have shown remarkable promise in medical image segmentation by effectively capturing both global and local information, even though their training can be resource-intensive (Lee & Lee, 2024; Ranjbarzadeh et al., 2025). The architectural efficiency and scalability of Transformer-based networks, however, facilitate stable training and performance improvements through multi-unit configurations (Rodríguez-Ibáñez et al., 2023). The ability of these models to capture long-range dependencies is particularly advantageous in medical imaging, where contextual information across large anatomical regions can be crucial for accurate segmentation (Li et al., 2023; Rodríguez-Ibáñez et al., 2023). However, a primary trade-off with the self-attention mechanism central to Transformers is its computational expense, especially with large spatial dimensions (Perera et al., 2025).

This is particularly evident in 3D medical image segmentation, where the large input sizes can lead to prohibitively high computational costs for standard Vision Transformer architectures (Perera et al., 2024). Consequently, approaches like the Swin Transformer utilize shifted windowing to enable efficient cross-window information transmission, thereby reducing computational complexity while retaining the ability to model both local and global features (Nemavhola et al., 2025). This window-based multi-scale attention mechanism reduces complexity linearly with the number of tokens, contrasting with the quadratic complexity in traditional Vision Transformers (Kareem et al., 2023). Such advancements are critical for deploying Transformer models in real-world clinical settings where computational resources are often constrained, ensuring that high-resolution medical imaging can be processed efficiently without sacrificing segmentation accuracy. This efficiency is paramount for 3D medical segmentation tasks where the self-attention operation's quadratic complexity historically presents a computational bottleneck (Shaker et al., 2024).

Architectural Considerations for a Unified SSL-Driven Multi-Scale Framework

This section delves into the synergistic integration of self-supervised learning within a multi-scale fusion architecture. It emphasizes how diverse feature representations are learned and combined to optimize segmentation performance across varying object sizes. A primary

objective is to develop a robust feature extraction backbone that captures both fine-grained details and global contextual information—crucial for accurate boundary delineation of anatomical structures and pathologies (Wang et al., 2025). This involves leveraging hierarchical feature maps from a deep neural network, where early layers capture high-resolution local textures and deeper layers encode low-resolution semantic concepts. The self-supervised component employs tasks such as masked autoencoding or contrastive learning on unlabeled medical images to train generalizable feature extractors that reduce reliance on exhaustive manual annotations. Consequently, this pre-training phase enables the model to understand inherent data structures and relationships, significantly boosting performance on downstream segmentation tasks with limited labeled data (Lin et al., 2025; Lu et al., 2025).

The fusion mechanism then strategically combines these multi-scale features—often via attention mechanisms or learned weighting schemes—to emphasize task-relevant information while suppressing noise (He et al., 2025). This integrated approach enhances the model's ability to handle images with varying scales and complexities, thereby improving segmentation accuracy across diverse medical imaging modalities (Zhang et al., 2024). Moreover, the framework's adaptability supports multi-task learning, which mitigates data scarcity by enabling simultaneous training on diverse biomedical imaging datasets and leveraging varied label types and data sources for pretraining image representations (Schäfer et al., 2024). Such versatility yields a more robust model that generalizes across medical imaging challenges and overcomes limitations of specialized datasets. Complementing this, a dense multi-scale attention and depth-supervised network refines feature extraction, sharpening focus on weak features and crucial local details essential for precise medical image segmentation (Fu et al., 2024).

Exemplified by architectures like Attention U-Net and Transformer-based models, this strategy ensures optimal contributions from both global context and fine-grained local features to the final segmentation (Nemavhola et al., 2025; Rodríguez-Ibáñez et al., 2023). In particular, hybrid architectures integrating Vision Transformers with convolutional encoders simultaneously capture global contextual information and localized features—vital for complex tasks like brain tumor segmentation (Ranjbarzadeh et al., 2025). This synergy leverages CNNs' robust local feature extraction alongside Transformers' strength in long-range dependencies, enhancing overall segmentation performance (Chen et al., 2025). Augmented by skip connections, the dual-encoder design facilitates efficient feature transmission across layers, yielding more accurate and robust results in challenging scenarios such as brain tumor identification (Ranjbarzadeh et al., 2025).

Synthesis of prior studies indicates that effective frameworks rely on an encoder-decoder structure: the encoder learns self-supervised representations, while the decoder applies multi-scale fusion modules with components like cross-scale attention, skip connections, and lightweight fusion blocks. SSL integration in the encoder minimizes dependence on annotated datasets, as multi-scale fusion bolsters spatial accuracy and contextual reasoning. Additionally, incorporating instance-adaptive prototype learning addresses intra-class variations by generating segmentation masks via pixel-to-prototype comparison—a promising technique for medical image segmentation (Liang et al., 2024). This refined approach directly confronts data scarcity and the precise delineation of minute structures, key challenges in medical imaging (Ciampi et al., 2025; Manna et al., 2025).

For example, hybrid CNN-ViT models have emerged as powerful tools for medical image segmentation, harnessing CNNs for local features and ViTs for global context (Abıdın

et al., 2024; Andrade-Miranda et al., 2022; Soh & Rajapakse, 2023). The Transformer architecture's efficiency and scalability—originally measuring each term's contribution to a sentence's meaning—extend robustly to image segmentation (Rodríguez-Ibáñez et al., 2023). Its long-range dependency modeling proves especially valuable in medical imaging, where broad contextual cues aid accurate diagnosis and segmentation (Rodríguez-Ibáñez et al., 2023). Models like Swin UNET Transformers thus surpass traditional U-Nets by integrating local dependencies and global context for precise delineation of structures such as brain tumors (Rodríguez-Ibáñez et al., 2023; Saleh & Biswal, 2025).

Within these Transformer-based models, the self-attention mechanism—initially from natural language processing—processes image patches akin to words, capturing complex spatial relationships (Nemavhola et al., 2025; Rodríguez-Ibáñez et al., 2023). Multi-head attention concurrently seizes diverse image relationships, sharpening discernment of intricate patterns (Nemavhola et al., 2025). Coupled with deformable attention, these advances mitigate computational costs in multi-scale processing for architectures like Mask2Former, enabling flexible modeling of non-local dependencies essential for high-accuracy clinical segmentation (Li et al., 2024; Wang et al., 2025; Zeineldin & Mathis-Ullrich, 2024).

Ensemble learning with these architectures further boosts robustness and accuracy by merging predictions from diverse models, offsetting individual limitations (Yao et al., 2023). This synergy improves generalization and reliability, particularly for ambiguous medical cases. Knowledge distillation then transfers insights from complex ensembles to efficient single models, enabling practical clinical deployment with minimal performance loss. Such innovations advance medical image analysis, balancing effectiveness and accessibility. Ongoing refinements via self-supervised learning and efficient training strategies continue to expand performance boundaries across diverse tasks.

These efforts include novel designs optimizing computational efficiency alongside high accuracy, especially for real-time applications demanding rapid processing. Lightweight Transformer alternatives, such as attention and Transformer compartments, reduce computational and memory demands while preserving global relationship modeling (Papanastasiou et al., 2023). Striking a balance between complexity and demands ensures accuracy gains support deployability (Nemavhola et al., 2025). Recent efficient architectures like MobileNet, ShuffleNet, and EfficientNet exemplify this, enabling high-performance deployment on resource-constrained devices (Baareh, 2024). This careful calibration of architectural design with computational constraints is paramount for the practical adoption of advanced segmentation frameworks in clinical settings, particularly for diagnostic tools requiring on-device processing.

Computational Efficiency Challenges and Optimization Strategies

The burgeoning complexity of self-supervised multi-scale fusion frameworks for image segmentation introduces significant computational overhead, demanding novel optimization strategies to ensure their practical applicability in real-world scenarios. This section will delve into the critical challenges posed by the computational and memory demands of these sophisticated models and explore various optimization techniques designed to mitigate them without compromising segmentation accuracy (Nemavhola et al., 2025). For instance, while large models offer superior generalization capabilities, especially in zero-shot or few-shot learning paradigms, their extensive computational requirements necessitate specialized hardware for even basic inference (Rodríguez-Ibáñez et al., 2023). This challenge is

particularly acute for real-time medical imaging applications, where instantaneous processing on devices with limited computational resources is crucial (Jiang et al., 2024). Consequently, researchers are increasingly focusing on efficient architectures and attention mechanisms to achieve a balanced trade-off between accuracy and computational cost, thereby facilitating deployment on mobile and edge devices (Baareh, 2024). This imperative has led to the development of lightweight models that maintain high performance while significantly reducing parameter counts and computational intensity, making them suitable for mobile medical devices (Jiang et al., 2024).

Real-time segmentation requires models that balance complexity with speed. Literature suggests that lightweight architectures—such as MobileNet-based encoders, edge-aware fusion modules, and pruning-aware designs—are essential for efficient deployment. SSL also contributes to efficiency by reducing the need for extensive supervised fine-tuning. Compression techniques and quantization further improve inference speed without degrading accuracy significantly. For example, quantization-aware training pipelines have been successfully applied to models like MedSAM to enable efficient deployment on platforms with limited resources, such as laptops, thereby addressing the trade-off between accuracy and processing speed (Lu et al., 2025). Moreover, strategies like knowledge distillation transfer the learned representations from complex, high-performing models to simpler, more efficient ones, further enhancing their suitability for deployment in resource-constrained environments (He et al., 2024; Rodríguez-Ibáñez et al., 2023). Furthermore, novel lightweight architectures, like MobileNet, utilize depthwise separable convolution to drastically reduce parameters and computational complexity, which is crucial for efficient deployment on mobile terminals and embedded devices despite a minor accuracy trade-off (Chi et al., 2023).

Additionally, models like MobileNetV2 and MobileNetV3 further refined these principles by introducing inverted residuals and attention mechanisms, respectively, to optimize parameter efficiency while sustaining performance (Chi et al., 2023). Similarly, architectural innovations such as those found in ResNet-50 and DenseNet-121 also prioritize efficient deep network training and enhanced information flow through residual blocks and dense connectivity, respectively, demonstrating diverse approaches to achieving computational efficiency within complex models (Ryando et al., 2025). The EfficientNet series further exemplifies this trend by uniformly scaling network depth, width, and resolution using a compound coefficient, demonstrating state-of-the-art accuracy with significantly fewer parameters and FLOPs compared to prior models (Chi et al., 2023). These advances are critical for embedded medical devices where computational and memory constraints are severe, often necessitating a departure from high-resolution inputs to achieve real-time throughput (Ostrowski & Shafique, 2024). Specifically, EfficientNetV2 introduced an improved progressive learning method that dynamically adjusts training based on image size, enhancing both speed and accuracy through the Fused-MBConv structure, which replaces depthwise convolution with standard 3x3 convolution (Chi et al., 2023).

This strategy effectively combines the benefits of efficient computation with robust feature extraction, crucial for real-time medical image analysis on resource-constrained platforms (Nemavhola et al., 2025). Such architectural optimizations are vital for pushing the boundaries of what is achievable in edge AI, enabling the deployment of sophisticated AI systems directly onto portable medical equipment (Xu et al., 2025). LeNet-5, an early CNN model, demonstrated the potential of such architectures for complex pattern recognition but

faced hardware limitations (Chi et al., 2023). Subsequent advancements, such as the adoption of deeper networks like VGGNet, ResNet, and Inception, have shown that increasing network depth can significantly improve performance by capturing more complex patterns, albeit with increased computational demands (Baareh, 2024). Despite these successes, the computational burden associated with these deeper architectures necessitates exploration into novel techniques for maintaining accuracy while simultaneously reducing operational costs, especially in clinical settings where real-time processing is paramount (Islam et al., 2024; Manzari et al., 2025).

This has led to a focus on architectures like GoogLeNet, which employs an inception module that allows for multi-scale feature extraction through varying filter sizes within a single block, enhancing efficiency and accuracy (Kavita, 2024). This multi-scale approach within GoogLeNet enables the network to capture intricate patterns at different resolutions, a critical aspect for robust image segmentation that is further refined in modern architectures like EfficientNetV2 and ResNet-RS (Haripriya et al., 2025). Other notable architectures like AlexNet and VGG-16/19, while foundational in their contributions to deep learning, are characterized by their deeper and more parameter-heavy structures, making them less suitable for real-time edge device deployment compared to more optimized designs like Inception-V3 (Kavita, 2024). These developments underscore a continuous push towards architectural innovations that facilitate high performance with reduced computational footprints, enabling the practical application of advanced segmentation techniques in resource-limited environments. The challenge intensifies when considering gigapixel medical images, where the sheer volume of data combined with real-time inference demands necessitates sophisticated pruning strategies and hardware-aware optimizations to overcome deployment limitations (Sreeram & Kainz, 2023).

This often involves integrating specialized hardware accelerators, such as GPUs or TPUs, which are engineered to handle the parallel processing demands of deep neural networks efficiently. However, for widespread adoption in resource-constrained medical settings, the reliance on such high-end hardware is often impractical, necessitating a paradigm shift toward intrinsically efficient model designs (Iqbal et al., 2023). Therefore, the development of self-supervised multi-scale fusion frameworks represents a crucial step towards achieving accurate and efficient image segmentation without requiring extensive manual annotation or high computational resources. These frameworks leverage unlabeled data to learn robust feature representations, thereby mitigating the dependency on labor-intensive data labeling processes and enhancing generalizability across diverse datasets (Jiang et al., 2024). This approach can significantly reduce the cost and time associated with preparing large annotated datasets, which is often a major bottleneck in medical image analysis.

This is particularly vital in medical imaging, where data scarcity due to privacy concerns and the need for expert annotation significantly hinders the application of supervised learning techniques (Zhang et al., 2024). Self-supervised learning, by contrast, exploits the inherent structure within the data itself to generate supervisory signals, enabling models to learn powerful representations without explicit human labels (Bougourzi & Hadid, 2025). This capability is particularly beneficial for tasks like image segmentation, where pixel-level annotations are exceptionally time-consuming and require highly specialized knowledge (Singh et al., 2024). Furthermore, self-supervised methods can overcome the limitations of domain shift, a common challenge in medical imaging where models trained on one dataset

often perform poorly on others due to variations in imaging protocols, equipment, and patient populations (Mao et al., 2025). The integration of multi-scale fusion within self-supervised frameworks can further enhance robustness by enabling models to learn hierarchical representations from various contextual perspectives, thereby improving segmentation accuracy across diverse anatomical structures and pathological conditions.

Theoretical Advantages and Practical Implications

Self-supervised learning offers a compelling solution to the annotation bottleneck prevalent in medical imaging by enabling models to learn from vast amounts of unlabeled data, thus drastically reducing the need for costly and time-consuming manual annotations (Huang et al., 2023; Wang et al., 2021). This approach is particularly critical in medical imaging analysis, where the scarcity of high-quality annotated datasets significantly impedes the advancement and application of machine learning models (Huang et al., 2023; Shurrab & Duwairi, 2022). By formulating pretext tasks, self-supervised methods can encourage networks to capture high-level semantics from unlabeled data, thereby boosting performance on downstream tasks with limited annotated examples (Huang et al., 2023; Zhang & Gu, 2022). This inherent capability allows for the creation of generalist models that can be fine-tuned for a multitude of downstream tasks, circumventing the need for extensive, task-specific labeled datasets (Huang et al., 2023). This approach is particularly advantageous in medical imaging, where ethical considerations and the expertise required for accurate labeling make large-scale annotation prohibitively expensive and often impractical (Qu et al., 2022).

For instance, self-supervised pretraining has demonstrated significant improvements in 2D medical image segmentation tasks, especially in scenarios with scarce labeled data, by leveraging unlabeled examples to learn robust feature representations (Kalapos & Gyires-Tóth, 2023). Moreover, the application of self-supervised learning extends beyond segmentation to other diagnostic tasks, such as classification and detection, across various modalities like X-ray, CT, MRI, and ultrasound, proving its versatility and impact (VanBerlo et al., 2024; Zhang & Gu, 2022). Specifically, multi-encoder architectures with self-supervised pretraining have shown exceptional performance in medical image segmentation, even outperforming transformer models by achieving high Dice Similarity Coefficients on diverse MRI modalities (Otaghsara & Rahmanzadeh, 2025). This demonstrates the ability of self-supervised methods to effectively generalize across different imaging types and tasks, mitigating the need for modality-specific labeled datasets (Wang et al., 2024; Zhang & Gu, 2022).

These approaches underscore the potential of self-supervised learning to bridge the gap between abundant unlabeled medical imaging data and the performance requirements of clinical applications, thereby facilitating more widespread adoption of AI in healthcare (Azizi et al., 2021; Huang et al., 2023). Furthermore, by reducing the dependency on extensive labeled datasets, these frameworks facilitate the development of more robust and generalizable models capable of performing effectively across varied clinical settings and patient populations, overcoming common issues like domain shift and data heterogeneity (Azizi et al., 2023). The integration of self-supervised learning with multi-scale fusion further strengthens these capabilities, as models can extract both local and global contextual information from unlabeled images, leading to more comprehensive and accurate feature learning (Arasteh et al., 2023). This synergy allows for the capture of richer, more informative

representations that are crucial for precise segmentation in complex anatomical structures and pathologies (van Veldhuizen et al., 2025).

The proposed conceptual framework offers several theoretical advantages: improved generalization, reduced annotation dependency, enhanced detail preservation, and computational efficiency. Practically, such a framework can be applied to medical image diagnosis, autonomous driving perception systems, industrial defect detection, and environmental monitoring. Integrating SSL and multi-scale fusion creates a robust and adaptable segmentation pipeline suitable for real-world implementation. This approach is particularly critical for whole-heart segmentation from CT and MRI scans, where modality-specific biases and the need for extensive labeled datasets present significant challenges (Qayyum et al., 2025). Specifically, advancements in self-supervised pre-training using Swin Transformers have demonstrated remarkable generalizability across various medical imaging modalities, outperforming single-modality models, especially on out-of-distribution tasks (Talasila et al., 2024). This hierarchical design, leveraging shifted windowing methods and multi-head self-attention, allows for efficient cross-window information transmission and captures both local and global features, crucial for complex image analysis (Nemavhola et al., 2025).

Furthermore, the robustness of these models is significantly enhanced by multi-axis consistency frameworks, which generate reliable pseudo-labels, thereby improving the overall semi-supervised learning process (Lu et al., 2023). These sophisticated self-supervised methodologies, including contrastive learning and multi-view consistency, excel at extracting robust feature representations from unlabeled data, thereby mitigating the need for extensive human annotation in medical imaging tasks (Gomariz et al., 2024; Sun et al., 2024). This reduction in annotation dependency is crucial for democratizing AI in healthcare, allowing smaller institutions or researchers with limited resources to develop and deploy highly accurate segmentation models. Moreover, the integration of multi-scale fusion allows these self-supervised models to simultaneously capture fine-grained local details and broader contextual information, which is paramount for segmenting intricate anatomical structures and subtle pathological changes (Lu et al., 2023; Xu et al., 2025). By leveraging self-supervised learning, models can learn robust representations from unlabeled data, thereby significantly improving segmentation accuracy and generalizability, particularly when combined with multi-scale fusion techniques (Qayyum et al., 2024).

This strategy is particularly beneficial for segmentation tasks involving irregularly shaped organs or complex pathologies where distinct feature scales are critical for accurate delineation (Zheng et al., 2025). The development of foundation models, such as PRISM, pretrained on large-scale multi-sequence MRI datasets, further exemplifies the potential for generalizable representations that adapt robustly to diverse clinical applications, overcoming heterogeneity among MRI sequences (Qiu et al., 2025). Indeed, these models, particularly those leveraging Vision-language Foundation Models, show promise in integrating computer vision and natural language processing to address complex tasks like disease classification and cross-modal retrieval, thereby reducing reliance on labor-intensive annotations (Ryu et al., 2025; Xu et al., 2025). This capability is further augmented by self-supervised learning methods that can generate pseudo-labels for unlabeled data, effectively expanding the training pool without manual intervention (Zhou et al., 2021).

This approach allows for the creation of more robust and adaptable models that can generalize effectively across different datasets and clinical scenarios, significantly enhancing

the utility of AI in medical imaging. This paradigm shift towards self-supervision offers a potent solution to the challenges of data scarcity and annotation cost in medical imaging, paving the way for more autonomous and efficient diagnostic tools (Singh et al., 2024; Yazdani et al., 2024). Moreover, the inherent versatility of these models allows them to be fine-tuned for various downstream tasks, such as disease classification and prognosis, with minimal labeled data, thereby broadening their clinical utility (van Veldhuizen et al., 2025; Zhang & Metaxas, 2023). The development of generalist foundation models, particularly in 3D imaging, further supports this trend by integrating comprehensive natural language annotations into the learning process, eliminating the need for explicit manual image annotations and enhancing adaptability across various tasks (Hamamcı et al., 2024).

This approach significantly reduces the time and effort required for model deployment in new clinical applications, thereby accelerating the translation of AI research into practical healthcare solutions. The evolution of these foundational models, trained on extensive and diverse datasets using self-supervision, marks a significant advancement, moving beyond task-specific AI techniques to models capable of learning generalized representations applicable across various medical imaging modalities (Azad et al., 2023; Bian et al., 2025). Such models, often utilizing Large Vision Language Models, process diverse data types and learn from large-scale unlabeled datasets or natural pairs of different modalities, significantly contributing to the development of more robust and versatile AI systems in healthcare (Huang et al., 2024). This systematic analysis of papers published between 2012 and 2024, screening 1,144 papers from medical and AI domains, highlights the rapid progression and potential of these multimodal frameworks in clinical applications (Huang et al., 2024). This comprehensive review underpins the transformative potential of self-supervised multi-scale fusion frameworks, emphasizing their capacity to address critical challenges in medical image segmentation, particularly data scarcity and annotation burden.

CONCLUSION

This study concludes that integrating self-supervised learning with multi-scale fusion offers a powerful paradigm shift in medical image segmentation. By enabling models to learn meaningful patterns from unlabeled data while simultaneously leveraging multi-resolution features, this framework overcomes key limitations of traditional supervised methods, including annotation dependency, domain sensitivity, and inefficient representation learning. The conceptual synthesis demonstrates that such architectures enhance segmentation accuracy, improve generalization across imaging modalities, and reduce computational burdens, making them suitable for real-time and resource-constrained clinical environments. Consequently, self-supervised multi-scale fusion represents a viable foundational approach for future diagnostic and decision-support systems in healthcare.

LIMITATIONS AND FUTURE RESEARCH RECOMMENDATIONS

This research is restricted to qualitative conceptual analysis and does not provide empirical validation through benchmark datasets or comparative performance metrics. It also focuses predominantly on medical imaging use cases, limiting its generalizability to broader computer vision applications. Moreover, the study does not evaluate hardware-dependent implementation constraints or deployment complexities associated with transformer-driven segmentation models.

Future research should empirically evaluate the proposed framework using large-scale, multimodal clinical datasets and real-world testing environments. Experimental comparisons with state-of-the-art architectures, including hybrid CNN–Vision Transformer models, will be essential to quantify performance gains. Further exploration into lightweight, hardware-aware model optimization, federated training for privacy-preserving data sharing, and foundation models capable of cross-domain segmentation can extend the applicability and robustness of self-supervised multi-scale frameworks within clinical and industrial ecosystems.

REFERENCES

Abıdın, Z. U., Naqvi, R. A., Haider, A., Kim, H. S., Jeong, D., & Lee, S. W. (2024). Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: a prospective survey [Review of Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: a prospective survey]. *Frontiers in Bioengineering and Biotechnology*, 12, 1392807. Frontiers Media. https://doi.org/10.3389/fbioe.2024.1392807

An, T., Huang, W., Xu, D., He, Q., Hu, J., & Lou, Y. (2025). *A Deep Learning Framework for Boundary-Aware Semantic Segmentation*. https://doi.org/10.48550/ARXIV.2503.22050

Andrade-Miranda, G., Jaouen, V., Bourbonne, V., Lucia, F., Visvikis, D., & Conze, P.-H. (2022). Pure Versus Hybrid Transformers For Multi-Modal Brain Tumor Segmentation: A Comparative Study. *2022 IEEE International Conference on Image Processing (ICIP)*, 1336. https://doi.org/10.1109/icip46576.2022.9897658

Arasteh, S. T., Misera, L., Kather, J. N., Truhn, D., & Nebelung, S. (2023). Enhancing Network Initialization for Medical AI Models Using Large-Scale, Unlabeled Natural Images. *arXiv*. https://doi.org/10.48550/ARXIV.2308.07688

Arasteh, S. T., Misera, L., Kather, J. N., Truhn, D., & Nebelung, S. (2024). Enhancing diagnostic deep learning via self-supervised pretraining on large-scale, unlabeled non-medical images. *European Radiology Experimental*, 8(1). https://doi.org/10.1186/s41747-023-00411-3

Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., & Merhof, D. (2023). Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.18689

Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomašev, N., Mitrović, J., Strachan, P. H., Mahdavi, S. S., Wulczyn, E., Babenko, B., Walker, M., Loh, A., Chen, P.-H. C., Liu, Y., Bavishi, P., McKinney, S. M., … Natarajan, V. (2023). Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6), 756. https://doi.org/10.1038/s41551-023-01049-7

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., & Norouzi, M. (2021). Big Self-Supervised Models Advance Medical Image Classification. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3458. https://doi.org/10.1109/iccv48922.2021.00346

Baareh, A. K. (2024). Facial Recognition and Discovery Using Convolution Deep Learning Neural Network. *Journal of Computer Science*, 20(11), 1559. https://doi.org/10.3844/jcssp.2024.1559.1568

Bhagat, D., Vakil, A., Gupta, R. K., & Kumar, A. (2024). Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN). *Procedia Computer Science*, 235, 2079. https://doi.org/10.1016/j.procs.2024.04.197

Bian, Y., Li, J., Ye, C., Jia, X., & Yang, Q. (2025). Artificial intelligence in medical imaging: From task-specific models to large-scale foundation models [Review of Artificial intelligence in medical imaging: From task-specific models to large-scale foundation models]. *Chinese Medical Journal*. Lippincott Williams & Wilkins. https://doi.org/10.1097/cm9.0000000000003489

Bougourzi, F., & Hadid, A. (2025). Recent Advances in Medical Imaging Segmentation: A Survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2505.09274

Chen, J., Chen, R., Wang, W., Cheng, J., Zhang, L., & Chen, L. (2024). TinyU-Net: Lighter Yet Better U-Net with Cascaded Multi-receptive Fields. In *Lecture notes in computer science* (p. 626). Springer Science+Business Media. https://doi.org/10.1007/978-3-031-72114-4_60

Chen, R., He, S., Xie, J., Wang, T., Xu, Y., Fang, J., Zhao, X., Zhang, S., Wang, G., Lu, H., & Yang, Z. (2025). MedFuseNet: fusing local and global deep feature representations with hybrid attention mechanisms for medical image segmentation. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-89096-9

Cheng, S., Starosolski, Z. A., & Subramanian, D. (2024). *Self-Supervised Learning for Building Robust Pediatric Chest X-ray Classification Models*. https://doi.org/10.48550/ARXIV.2409.00231

Cheung, Y. S. J., Chen, X., Yang, L., & Zhao, H. (2023). A Lightweight Clustering Framework for Unsupervised Semantic Segmentation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2311.18628

Chi, J., On, C. K., Zhang, H., & Chai, S. S. (2023). A Review of Deep Convolutional Neural Networks in Mobile Face Recognition [Review of A Review of Deep Convolutional Neural Networks in Mobile Face Recognition]. *International Journal of Interactive Mobile Technologies*, 17(23), 4. kassel university press. https://doi.org/10.3991/ijim.v17i23.40867

Ciampi, L., Lagani, G., Amato, G., & Falchi, F. (2025). *Semi-Supervised Biomedical Image Segmentation via Diffusion Models and Teacher-Student Co-Training*. https://doi.org/10.48550/ARXIV.2504.01547

Fardela, R., Milvita, D., Almuhayar, M., Mardiansyah, D., Rasyada, L. A., & Hakim, L. (2024). Classification of Thoracic X-Ray Images of COVID-19 Patients Using the Convolutional Neutral Network (CNN) Method. *Journal of Computer Science*, 20(4), 357. https://doi.org/10.3844/jcssp.2024.357.364

Fu, Z., Li, J., Chen, Z., & Ren, L. (2024). DmADs-Net: dense multiscale attention and depth-supervised network for medical image segmentation. *International Journal of Machine Learning and Cybernetics*. https://doi.org/10.1007/s13042-024-02248-7

Goldblum, M., Souri, H., Ni, R., Shu, M., Prabhu, V., Somepalli, G., Chattopadhyay, P., Ibrahim, M., Bardes, A., Hoffman, J., Chellappa, R., Wilson, A. G., & Goldstein, T. (2023). Battle of the Backbones: A Large-Scale Comparison of Pretrained Models across Computer Vision Tasks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.19909

Gomariz, Á., Kikuchi, Y., Li, Y. Y., Albrecht, T., Maunz, A., Ferrara, D., Lu, H., & Göksel, O. (2024). Joint semi-supervised and contrastive learning enables zero-shot domain-

adaptation and multi-domain segmentation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.05336

Hamamcı, İ. E., Er, S., Almas, F., Simsek, A. G., Esirgün, S. N., Doğan, İ., Dasdelen, M. F., Durugol, O. F., Wittmann, B., Amiranashvili, T., Simsar, E., Simsar, M., Erdemir, E. B., Alanbay, A., Sekuboyina, A., Lafci, B., Blüthgen, C., Özdemir, M. K., & Menze, B. (2024). Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography. *Research Square*. https://doi.org/10.21203/rs.3.rs-5271327/v1

Haripriya, R., Khare, N., & Pandey, M. (2025). Privacy-preserving federated learning for collaborative medical data mining in multi-institutional settings. *Scientific Reports*, 15(1), 12482. https://doi.org/10.1038/s41598-025-97565-4

He, Y., Bi, Y., Li, L., Pun, C.-M., Jiao, W., & Jin, Z. (2025). *Mutual Evidential Deep Learning for Medical Image Segmentation*. https://doi.org/10.48550/ARXIV.2505.12418

He, Y., Yu, H., Liu, X., Yang, Z., Sun, W., Anwar, S., & Mian, A. (2024). *Deep Learning Based 3d Segmentation: A Survey*. https://doi.org/10.2139/ssrn.4824679

Hoang, C. M. (2025). Unsupervised Instance Segmentation with Superpixels. *Pattern Recognition*, 172, 112402. https://doi.org/10.1016/j.patcog.2025.112402

Huang, S.-C., Jensen, M., Yeung, S., Lungren, M. P., Poon, H., & Chaudhari, A. (2024). Multimodal Foundation Models for Medical Imaging - A Systematic Review and Implementation Guidelines [*Review of Multimodal Foundation Models for Medical Imaging - A Systematic Review and Implementation Guidelines]. medRxiv (Cold Spring Harbor Laboratory)*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2024.10.23.24316003

Huang, S.-C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., & Chaudhari, A. (2023). Self-supervised learning for medical image classification: a systematic review and implementation guidelines [*Review of Self-supervised learning for medical image classification: a systematic review and implementation guidelines]. Npj Digital Medicine*, 6(1). Nature Portfolio. https://doi.org/10.1038/s41746-023-00811-0

Indra, Z., Jusman, Y., Elfizar, E., Salambue, R., Kurniawan, R., & Melia, T. (2024). *Computer-Assisted Disease Diagnosis Application for Malaria Early Diagnosis Based on Modified CNN Algor*.pdf.

Iqbal, S., Khan, T., Alhussein, M., Naqvi, S. S., Usman, M., & Aurangzeb, K. (2023). LDMRes-Net: Enabling Efficient Medical Image Segmentation on IoT and Edge Platforms. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.06145

Islam, M. S., Rahman, Md. A., Mazumder, Md. T. R., & Shourov, S. H. (2024). Comparative Analysis of Neural Network Architectures for Medical Image Classification: Evaluating Performance Across Diverse Models. *American Journal of Advanced Technology and Engineering Solutions*, 4(1), 1. https://doi.org/10.63125/feed1x52

Jiang, J., Wang, M., Tian, H., Cheng, L., & Liu, Y. (2024). LV-UNet: A Lightweight and Vanilla Model for Medical Image Segmentation. *2021 IEEE International Conference on Bioinformatics and Biomedicine*, 4240. https://doi.org/10.1109/bibm62325.2024.10822465

Kalapos, A., & Gyires-Tóth, B. (2023). Self-supervised Pretraining for 2D Medical Image Segmentation. In *Lecture notes in computer science* (p. 472). Springer Science+Business Media. https://doi.org/10.1007/978-3-031-25082-8_31

Kareem, D. N. A., Fiaz, M., Novershtern, N., Hanna, J. H., & Cholakkal, H. (2023). Improving 3D Medical Image Segmentation at Boundary Regions using Local Self-attention and

Global Volume Mixing. *IEEE Transactions on Artificial Intelligence*, 5(6), 3233. https://doi.org/10.1109/tai.2023.3346833

Kavita, R. S. C. (2024). Innovative Integration of Convolutional Neural Networks for Enhanced Face Recognition., 20, 1941. https://doi.org/10.52783/jes.1740

Lai, M., Chen, J., Huang, Y., Wang, X., Xu, N., Zhou, S., Zhu, X., Wu, Y., Yang, B., Chen, G., & Li, J. (2025). LESS-Net: a lightweight network for epistaxis image segmentation using similarity-based contrastive learning. *Frontiers in Physiology*, 16. https://doi.org/10.3389/fphys.2025.1644589

Lee, H. H., 池谷裕二, Zhao, T., Xu, Y., Yang, J., Usuyama, N., Wong, C., Mu, W., Landman, B. A., Huo, Y., Santamaría-Pang, A., & Poon, H. (2024). Foundation Models for Biomedical Image Segmentation: A Survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.07654

Lee, S., & Lee, M. (2024). MetaSwin: a unified meta vision transformer model for medical image segmentation. *PeerJ Computer Science*, 10. https://doi.org/10.7717/peerj-cs.1762

Li, B., Mao, J., Sun, Z., Li, C., Zhao, Q., & Tanaka, T. (2024). MAProtoNet: A Multi-scale Attentive Interpretable Prototypical Part Network for 3D Magnetic Resonance Imaging Brain Tumor Classification. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.08917

Li, G. Y., Chen, J., Jang, S., Gong, K., & Li, Q. (2023). SwinCross: Cross-modal Swin transformer for head-and-neck tumor segmentation in PET/CT images. *Medical Physics*, 51(3), 2096. https://doi.org/10.1002/mp.16703

Li, X., Ding, H., Yuan, H., Zhang, W., Pang, J., Cheng, G., Chen, K., Liu, Z., & Loy, C. C. (2024). Transformer-Based Visual Segmentation: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 10138. https://doi.org/10.1109/tpami.2024.3434373

Liang, G., Zhou, Q., Chen, J., Wang, Z., & Yao, C. (2024). *Advancing Medical Image Segmentation via Self-supervised Instance-adaptive Prototype Learning*. 1056. https://doi.org/10.24963/ijcai.2024/117

Liang, P., Pu, B., Huang, H., Li, Y., Wang, H., Ma, W., & Chang, Q. (2025a*). Vision Foundation Models in Medical Image Analysis: Advances and Challenges*. https://doi.org/10.48550/ARXIV.2502.14584

Liang, P., Pu, B., Huang, H., Li, Y., Wang, H., Ma, W., & Chang, Q. (2025b). Vision Foundation Models in Medical Image Analysis: Advances and Challenges. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.14584

Lin, Q., Liu, G.-H., Li, Z., Li, Y., Jiang, Y., & Wu, X. (2025). *Multimodal Medical Image Classification via Synergistic Learning Pre-training*. https://doi.org/10.48550/ARXIV.2509.17492

Liu, K., Ling, S., & Liu, S. (2024). Semi-Supervised Medical Image Classification with Pseudo Labels Using Coalition Similarity Training. *Mathematics*, 12(10), 1537. https://doi.org/10.3390/math12101537

Lu, H., Fu, Y., Zhang, F., & Zhang, L. (2025). Efficient Quantization-Aware Training on Segment Anything Model in Medical Images and Its Deployment. In *Lecture notes in computer science* (p. 137). Springer Science+Business Media. https://doi.org/10.1007/978-3-031-81854-7_9

Lu, Y., Fan, Z., & Xu, M. (2023). Multi-dimensional Fusion and Consistency for Semi-supervised Medical Image Segmentation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.06618

Lu, Y., Wu, Y., Kateb, R., & Chaddad, A. (2025). *SEMI-Supervised Medical Image Segmentation via Dual Networks*. 1. https://doi.org/10.1109/isbi60581.2025.10980890

Manna, S., Das, S., Ghosh, S., & Bhattacharya, S. (2025). *Federated Self-Supervised Learning for One-Shot Cross-Modal and Cross-Imaging Technique Segmentation*. https://doi.org/10.48550/ARXIV.2503.23507

Manzari, O. N., Asgariandehkordi, H., Koleilat, T., Xiao, Y., & Rivaz, H. (2025). *Medical Image Classification with KAN-Integrated Transformers and Dilated Neighborhood Attention*. https://doi.org/10.48550/ARXIV.2502.13693

Mao, Y., Li, H., Lai, Y., Papanastasiou, G., Qi, P., Yang, Y., & Wang, C. (2025). *Semi-Supervised Medical Image Segmentation via Knowledge Mining from Large Models*. https://doi.org/10.48550/ARXIV.2503.06816

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence *[Review of Foundation models for generalist medical artificial intelligence]. Nature*, 616(7956), 259. Nature Portfolio. https://doi.org/10.1038/s41586-023-05881-4

Myakala, P. K. (2025). *Scaling AI with Limited Labeled Data: A Self-Supervised Learning Approach*. 2(1), 26. https://doi.org/10.62762/tetai.2025.607708

Nemavhola, A., Chibaya, C., & Viriri, S. (2025). A Systematic Review of CNN Architectures, Databases, Performance Metrics, and Applications in Face *Recognition [Review of A Systematic Review of CNN Architectures, Databases, Performance Metrics, and Applications in Face Recognition]. Information*, 16(2), 107. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/info16020107

Nohel, M., Ulrich, C., Suprijadi, J., Wald, T., & Maier-Hein, K. (2025). A Unified Framework for Foreground and Anonymization Area Segmentation in CT and MRI Data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.04361

Ostrowski, E., & Shafique, M. (2024). *Embedded Deployment of Semantic Segmentation in Medicine through Low-Resolution Inputs*. https://doi.org/10.48550/ARXIV.2403.05340

Otaghsara, S. S. T., & Rahmanzadeh, R. (2025). *Multi-encoder nnU-Net outperforms transformer models with self-supervised pretraining*. https://doi.org/10.48550/ARXIV.2504.03474

Papanastasiou, G., Δικαίος, N., Huang, J., Wang, C., & Yang, G. (2023). Is Attention all You Need in Medical Image Analysis? A Review [Review of Is Attention all You Need in Medical Image Analysis? A Review]. *IEEE Journal of Biomedical and Health Informatics*, 28(3), 1398. Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/jbhi.2023.3348436

Perera, S., Erzurumlu, Y., Gulati, D., & Yılmaz, A. (2025). MobileUNETR: A Lightweight End-to-End Hybrid Vision Transformer For Efficient Medical Image Segmentation. In *Lecture notes in computer science* (p. 281). Springer Science+Business Media. https://doi.org/10.1007/978-3-031-91721-9_18

Perera, S., Navard, P., & Yılmaz, A. (2024). SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 4981. https://doi.org/10.1109/cvprw63382.2024.00503

Preetha, R., M, J. P. P., & Nisha, J. S. (2025). Brain tumor segmentation using multi-scale attention U-Net with EfficientNetB4 encoder for enhanced MRI analysis. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-94267-9

Qayyum, A., Mazher, M., Ugurlu, D., Lemus, J. A. S., Rodero, C., & Niederer, S. A. (2025). *Foundation Model for Whole-Heart Segmentation: Leveraging Student-Teacher Learning in Multi-Modal Medical Imaging*. https://doi.org/10.48550/ARXIV.2503.19005

Qayyum, A., Xu, H., Halliday, B. P., Rodero, C., Lanyon, C. W., Wilkinson, R. D., & Niederer, S. (2024). Transforming Heart Chamber Imaging: Self-Supervised Learning for Whole Heart Reconstruction and Segmentation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.06643

Qian, L., Wen, C., Li, Y., Hu, Z., Zhou, X., Xia, X., & Kim, S.-H. (2023). Multi-scale context UNet-like network with redesigned skip connections for medical image segmentation. *Computer Methods and Programs in Biomedicine*, 243, 107885. https://doi.org/10.1016/j.cmpb.2023.107885

Qiu, Z., Wang, X., Xie, Z., Zhou, J., Wang, Y., Yang, L., Jiang, X., Bae, J., Son, M. H., Ye, Q., Chen, D., Zhang, R., Li, T., Mahboobani, N. R., Vardhanabhuti, V., Duan, X., Zhao, Y., & Chen, H. (2025). *Large-scale Multi-sequence Pretraining for Generalizable MRI Analysis in Versatile Clinical Applications*. https://doi.org/10.48550/ARXIV.2508.07165

Qu, L., Liu, S., Liu, X., Wang, M., & Song, Z. (2022). Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis [*Review of Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis]. Physics in Medicine and Biology*, 67(20). IOP Publishing. https://doi.org/10.1088/1361-6560/ac910a

Queiroz, D., Carlos, A., Anjos, A., & Berton, L. (2025). *Fair Foundation Models for Medical Image Analysis: Challenges and Perspectives*. https://doi.org/10.48550/ARXIV.2502.16841

Rani, V., Nabi, S. T., Kumar, M., Mittal, A., & Kumar, K. (2023). Self-supervised Learning: A Succinct Review [Review of Self-supervised Learning: A Succinct Review]. *Archives of Computational Methods in Engineering*, 30(4), 2761. Springer Science+Business Media. https://doi.org/10.1007/s11831-023-09884-2

Ranjbarzadeh, R., Anari, S., Crane, M., & Bendechache, M. (2025). A Hybrid UNet and Vision Transformer Architecture with Multi-scale Fusion for Brain Tumor Segmentation. In *Lecture notes in electrical engineering* (p. 484). Springer Science+Business Media. https://doi.org/10.1007/978-981-96-3863-5_44

Ren, S., Huang, X., Li, X., Xiao, J., Mei, J., Wang, Z., Yuille, A., & Zhou, Y. (2024). Medical Vision Generalist: Unifying Medical Imaging Tasks in Context. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.05565

Rodríguez-Ibáñez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P.-M. (2023). A review on sentiment analysis from social media platforms [Review of A review on sentiment analysis from social media platforms]. *Expert Systems with Applications*, 223, 119862. Elsevier BV. https://doi.org/10.1016/j.eswa.2023.119862

Ryando, C., Sigit, R., Setiawardhana, S., & Dewantara, B. S. B. (2025). Face Recognition for Logging in Using Deep Learning for Liveness Detection on Healthcare Kiosks. JOIV

International Journal on Informatics Visualization, 9(1), 295. https://doi.org/10.62527/joiv.9.1.2759

Ryu, J.-W., Kang, H., Chu, Y., & Yang, S. (2025). Vision-language foundation models for medical imaging: a review of current practices and innovations [Review of Vision-language foundation models for medical imaging: a review of current practices and innovations]. *Biomedical Engineering Letters*, 15(5), 809. Springer Science+Business Media. https://doi.org/10.1007/s13534-025-00484-6

Saleh, M. A., & Biswal, B. B. (2025). From U-Net to Swin-Unet Transformers: The Next-Generation Advances in Brain Tumor Segmentation with Deep Learning. *Journal of Biomedical Science and Engineering*, 18(8), 328. https://doi.org/10.4236/jbise.2025.188024

Schäfer, R., Nicke, T., Höfener, H., Lange, A., Merhof, D., Feuerhake, F., Schulz, V., Lotz, J., & Kießling, F. (2024). Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nature Computational Science*, 4(7), 495. https://doi.org/10.1038/s43588-024-00662-z

Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M., & Khan, F. S. (2024). UNETR++: Delving Into Efficient and Accurate 3D Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 43(9), 3377. https://doi.org/10.1109/tmi.2024.3398728

Shao, H., & Hou, Q. (2025). *MedSeg-R: Medical Image Segmentation with Clinical Reasonin*g. https://doi.org/10.48550/ARXIV.2506.18669

Shi, C., Rezai, R., Yang, J., Dou, Q., & Li, X. (2024). *A Survey on Trustworthiness in Foundation Models for Medical Image Analysis*. https://doi.org/10.48550/ARXIV.2407.15851

Shi, K., Li, Y., Ho, B., Wang, J., & Guo, K. (2025). Universal Lesion Segmentation Challenge 2023: A Comparative Research of Different Algorithms. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.10608

Shurrab, S., & Duwairi, R. (2022). Self-supervised learning methods and applications in medical imaging analysis: a survey. *PeerJ Computer Science*, 8. https://doi.org/10.7717/peerj-cs.1045

Singh, P., Chukkapalli, R., Chaudhari, S., Chen, L., Chen, M., Pan, J., Smuda, C., & Cirrone, J. (2024). Shifting to machine supervision: annotation-efficient semi and self-supervised learning for automatic medical image segmentation and classification. *Scientific Reports*, 14(1). https://doi.org/10.1038/s41598-024-61822-9

Soh, W. K., & Rajapakse, J. C. (2023). Hybrid UNet transformer architecture for ischemic stoke segmentation with MRI and CT datasets. *Frontiers in Neuroscience*, 17. https://doi.org/10.3389/fnins.2023.1298514

Song, Z., Zhao, Y., Li, X., Fei, M., Zhao, X., Liu, M., Chen, C., Yeh, C., Wang, Q., Zheng, G., Ai, S., & Zhang, L. (2024). REHRSeg: Unleashing the Power of Self-Supervised Super-Resolution for Resource-Efficient 3D MRI Segmentation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.10097

Sreeram, S., & Kainz, B. (2023). Sculpting Efficiency: Pruning Medical Imaging Models for On-Device Inference. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.05090

Sun, K., Xue, S. D., Sun, F., Sun, H., Luo, Y., Wang, L., Wang, S., Guo, N., Liu, A., Zhao, T., Xinzhou, W., Yang, L., Jin, S., Yan, J., & Dong, J. (2024). Medical Multimodal Foundation Models in Clinical Diagnosis and Treatment: Applications, Challenges, and Future Directions. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.02621

Taher, M. R. H., Haghighi, F., Gotway, M. B., & Liang, J. (2025). Large-scale benchmarking and boosting transfer learning for medical image analysis. *Medical Image Analysis*, 103487. https://doi.org/10.1016/j.media.2025.103487

Talasila, A., Maity, M., & Priyakumar, U. D. (2024). Self-Supervised Modality-Agnostic Pre-Training of Swin Transformers. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.12781

Tiu, E., Talius, E., Patel, P. R., Langlotz, C. P., Ng, A. Y., & Rajpurkar, P. (2022). Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12), 1399. https://doi.org/10.1038/s41551-022-00936-9

Uden, C. V., Irvin, J., Huang, M., Dean, N. C., Carr, J., Ng, A. Y., & Langlotz, C. P. (2023). How to Train Your CheXDragon: Training Chest X-Ray Models for Transfer to Novel Tasks and Healthcare Systems. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.08017

van Veldhuizen, V., Botha, V., Lu, C., Cesur, M. E., Lipman, K. G., de Jong, E. D., Horlings, H., Sanchez, C. I., Snoek, C. G. M., Wessels, L., Mann, R., Marcus, E., & Teuwen, J. (2025). *Foundation Models in Medical Imaging -- A Review and Outlook*. https://doi.org/10.48550/ARXIV.2506.09095

VanBerlo, B., Hoey, J., & Wong, A. (2024). A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound [Review of A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound]. *BMC Medical Imaging*, 24(1). BioMed Central. https://doi.org/10.1186/s12880-024-01253-0n

Vo, H. Q., Yuan, P., Yin, Z., Wong, K. K., Ezeana, C. F., Ly, S. T., Wong, S. T. C., & Nguyen, H. V. (2025). *MIRAM: Masked Image Reconstruction Across Multiple Scales for Breast Lesion Risk Prediction*. https://doi.org/10.48550/ARXIV.2503.07157

Wang, C., Chen, Z., Yang, J., & Kim, J. (2025). Improving Lesion Segmentation in Medical Images by Global and Regional Feature Compensation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.08675

Wang, G., Li, Y., Chen, W., Ding, M., Cheah, W. P., Qu, R., Ren, J., & Shen, L. (2024). {S^3-Mamba}: Small-Size-Sensitive Mamba for Lesion Segmentation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.14546

Wang, H., Cao, P., Yang, J., & Zaïane, O. R. (2023). MCA-UNet: multi-scale cross co-attentional U-Net for automatic medical image segmentation. *Health Information Science and Systems*, 11(1). https://doi.org/10.1007/s13755-022-00209-4

Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., Liu, X., Chen, J., Zhou, H., Ayed, I. B., & Zheng, H. (2021). Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, 12(1). https://doi.org/10.1038/s41467-021-26216-9

Wang, Y., Li, Z., Mei, J., Wei, Z., L, L., Wang, C., Sang, S., Yuille, A., Xie, C., & Zhou, Y. (2023). SwinMM: Masked Multi-view with Swin Transformers for 3D Medical Image Segmentation. In *Lecture notes in computer science* (p. 486). Springer Science+Business Media. https://doi.org/10.1007/978-3-031-43898-1_47

Wang, Y., Ni, H., Zhou, J., Liu, L., Lin, J., Yin, M., Gao, J., Zhu, S., Yin, Q., Zhu, J., & Li, R. (2024). A Semi-Supervised Learning Framework for Classifying Colorectal Neoplasia

Based on the NICE Classification. *Journal of Imaging Informatics in Medicine*, 37(5), 2342. https://doi.org/10.1007/s10278-024-01123-9

Wang, Y., Wang, L., Huang, Z., Zhang, Y., & Han, Y. (2024). Multimodal Medical Image Fusion Method based on the Swin Transformer and Self-supervised Contrast Learning. *Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering)*, 18(7), 819. https://doi.org/10.2174/0123520965311272240709053422

Wang, Y., Xu, J., Guan, Y., Ahmad, F., Mahmood, T., & Rehman, A. (2025). MSegNet: A Multi-View Coupled Cross-Modal Attention Model for Enhanced MRI Brain Tumor Segmentation. *International Journal of Computational Intelligence Systems*, 18(1). https://doi.org/10.1007/s44196-025-00787-7

Xu, G., Udupa, J. K., Luo, J., Zhao, S., Yu, Y., Raymond, S. B., Peng, H., Ning, L., Rathi, Y., Liu, W., & Zhang, Y. (2025). Is the medical image segmentation problem solved? A survey of current developments and future directions. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2508.20139

Xu, Y., Khan, T., Song, Y., & Meijering, E. (2025). Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. *Artificial Intelligence Review*, 58(3). https://doi.org/10.1007/s10462-024-11033-5

Xu, Z., Chen, D., & Gong, W. (2024). UMSSNet: A Unified Multi-Scale SegmentationNetwork for Heterogeneous Medical Images. *Research Square*. https://doi.org/10.21203/rs.3.rs-4001737/v1

Yao, L., Zhang, Z., & Bağcı, U. (2023). Ensemble Learning with Residual Transformer for Brain Tumor Segmentation. *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1. https://doi.org/10.1109/isbi53787.2023.10230404

Yazdani, E., Karamzadeh-Ziarati, N., Cheshmi, S. S., Sadeghi, M., Geramifar, P., Vosoughi, H., Jahromi, M. K., & Kheradpisheh, S. R. (2024). Automated segmentation of lesions and organs at risk on [68Ga]Ga-PSMA-11 PET/CT images using self-supervised learning with Swin UNETR. *Cancer Imaging*, 24(1). https://doi.org/10.1186/s40644-024-00675-x

Zeineldin, R. A., & Mathis-Ullrich, F. (2024). *Unified HT-CNNs Architecture: Transfer Learning for Segmenting Diverse Brain Tumors in MRI from Gliomas to Pediatric Tumors*. https://doi.org/10.48550/ARXIV.2412.08240

Zhang, C., Deng, X.-Y., Ling, S. H., Zhang, C., Deng, X.-Y., & Ling, S. H. (2024). Next-Gen Medical Imaging: U-Net Evolution and the Rise of Transformers [Review of Next-Gen Medical Imaging: U-Net Evolution and the Rise of Transformers]. *Sensors*, 24(14), 4668. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/s24144668

Zhang, C., & Gu, Y. (2022). Dive into Self-Supervised Learning for Medical Image Analysis: Data, Models and Tasks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2209.12157

Zhang, S., & Metaxas, D. (2023). On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, 91, 102996. https://doi.org/10.1016/j.media.2023.102996

Zhao, L., Wang, T., Chen, Y., Zhang, X., Tang, H., Lin, F., Li, C., Li, Q., Tan, T., Kang, D., & Tong, T. (2025). A novel framework for segmentation of small targets in medical images. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-94437-9

Zhao, Y., Wang, C., Hao, Y., Li, L., & Liao, T. (2025). *DyGLNet: Hybrid Global-Local Feature Fusion with Dynamic Upsampling for Medical Image Segmentation*. https://doi.org/10.48550/ARXIV.2509.12763

Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., Naseem, U., & Gu, Y. (2025). Robust and explainable framework to address data scarcity in diagnostic imaging. *Computers in Biology and Medicine*, 197, 111052. https://doi.org/10.1016/j.compbiomed.2025.111052

Zheng, H., Chen, Y., Mu, N., & Li, X. (2025). *Multi-Disease-Aware Training Strategy for Cardiac MR Image Segmentation*. https://doi.org/10.48550/ARXIV.2503.17896

Zhou, P., Zhang, L., & Li, W. (2025). *Boosting Semi-Supervised Medical Image Segmentation via Masked Image Consistency and Discrepancy Learning*. https://doi.org/10.48550/ARXIV.2503.14013

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Ginneken, B. van, Madabhushi, A., Prince, J. L., Rueckert, D., Summers, R. M., Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Ginneken, B. van, Madabhushi, A., Prince, J. L., Rueckert, D., & Summers, R. M. (2021). A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *Proceedings of the IEEE*, 109(5), 820. https://doi.org/10.1109/jproc.2021.3054390

Zhou, X., Sun, Y., Deng, M., Chu, C. W., & Dou, Q. (2024). Robust Semi-supervised Multimodal Medical Image Segmentation via Cross Modality Collaboration. In *Lecture notes in computer science* (p. 57). Springer Science+Business Media. https://doi.org/10.1007/978-3-031-72378-0_6

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(6), 1856. https://doi.org/10.1109/tmi.2019.2959609