Ultrasound Dual-Modality Video and Mamba Architecture in Tumor Diagnosis: A Survey

www.surveyx.cn

Abstract

This survey examines the integration of ultrasound dual-modality video and Mamba architecture in medical imaging, emphasizing their pivotal role in enhancing tumor diagnosis. The dual-modality approach, combining B-mode and contrast-enhanced ultrasound, facilitates improved visualization and characterization of tumors by leveraging complementary imaging strengths. Mamba architecture addresses computational challenges inherent in traditional models, enhancing efficiency and long-range dependency handling, crucial for processing extensive medical imaging sequences. The survey systematically explores the core concepts, including Time-Intensity Curve (TIC) analysis and multimodal feature fusion, which are integral to refining diagnostic precision. Case studies on diverse medical datasets underscore the transformative impact of these technologies, demonstrating significant improvements in diagnostic accuracy across various clinical contexts. Despite the promising advancements, challenges such as data quality, integration into clinical workflows, and computational demands persist. Future research directions focus on enhancing model robustness, exploring lightweight architectures, and addressing ethical and socioeconomic considerations to ensure equitable access to advanced imaging technologies. By fostering innovation and collaboration, these technologies hold the potential to significantly improve diagnostic outcomes and patient care in medical imaging.

www.SurveyX.cn

1 Introduction

1.1 Significance of Advanced Medical Imaging Techniques

Advanced medical imaging techniques have revolutionized tumor diagnosis by enhancing speed and accuracy, ultimately improving patient outcomes [1]. The integration of machine learning has addressed key challenges in diagnostic accuracy and efficiency [2]. Notably, deep learning-based computer-aided systems have significantly advanced breast tumor diagnosis using ultrasound and mammography, highlighting these technologies' critical role in clinical practice [3].

In ultrasound imaging, the limitations of conventional B-mode ultrasound in renal tumor diagnosis necessitate multi-modal approaches for distinguishing benign from malignant tumors [4]. Incorporating human domain knowledge has refined diagnostic capabilities, especially in breast cancer detection [5]. Additionally, advanced imaging techniques in lung ultrasound videos have improved diagnostic precision in pulmonary assessments by facilitating the identification of lung consolidations [6].

The challenges of accurately re-identifying polyps across different colonoscopic views illustrate the necessity of advanced imaging techniques in enhancing tumor diagnosis [7]. Cardiac ultrasound imaging, crucial for heart disease diagnosis, also suffers from variability due to manual processing, underscoring the need for automated and reliable imaging methods [8].

Recent advancements in self-supervised learning methods address the limitations of traditional supervised learning, which often requires extensive labeled datasets [9]. The incorporation of large

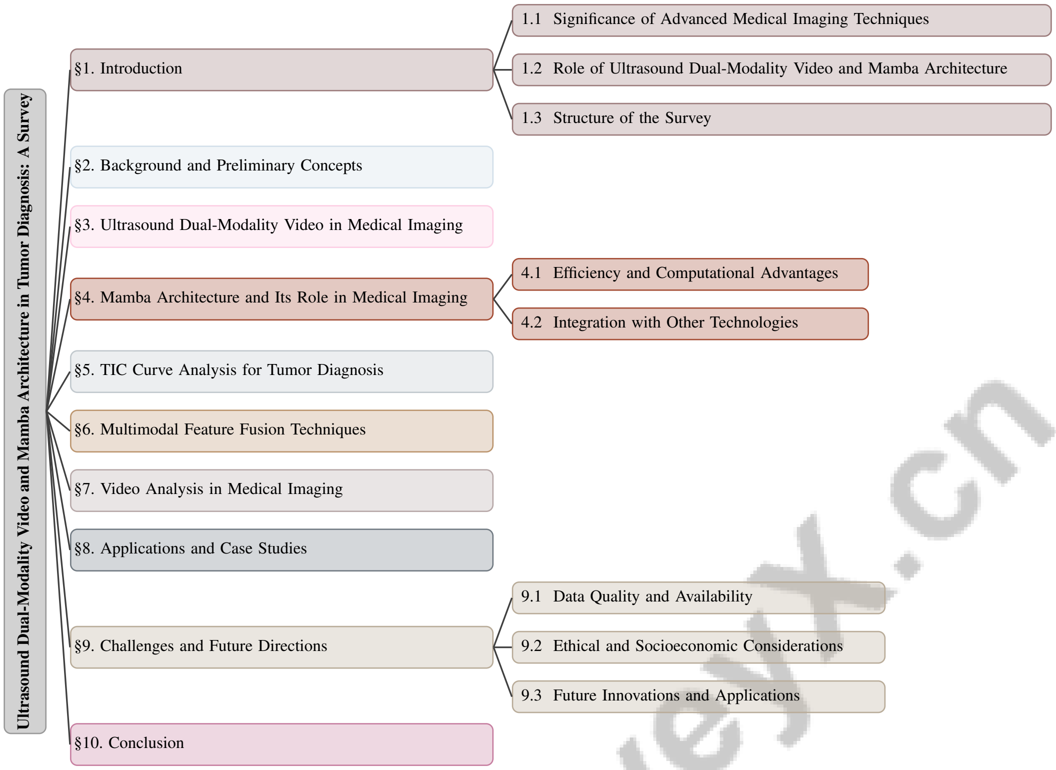

Ultrasound Dual-Modality Video and Mamba Architecture in Tumor Diagnosis: A Survey
§1. Introduction
1.1 Significance of Advanced Medical Imaging Techniques
1.2 Role of Ultrasound Dual-Modality Video and Mamba Architecture
1.3 Structure of the Survey
§2. Background and Preliminary Concepts
§3. Ultrasound Dual-Modality Video in Medical Imaging
§4. Mamba Architecture and Its Role in Medical Imaging
4.1 Efficiency and Computational Advantages
4.2 Integration with Other Technologies
§5. TIC Curve Analysis for Tumor Diagnosis
§6. Multimodal Feature Fusion Techniques
§7. Video Analysis in Medical Imaging
§8. Applications and Case Studies
§9. Challenges and Future Directions
9.1 Data Quality and Availability
9.2 Ethical and Socioeconomic Considerations
9.3 Future Innovations and Applications
§10. Conclusion

Figure 1: chapter structure

language models (LLMs) into medical imaging further exemplifies the transformative impact of deep learning breakthroughs, offering new pathways for enhancing diagnostic accuracy [10]. Collectively, these innovations underscore the pivotal role of advanced medical imaging techniques in improving tumor diagnosis and patient care.

1.2 Role of Ultrasound Dual-Modality Video and Mamba Architecture

The integration of ultrasound dual-modality video and Mamba architecture marks a significant advancement in medical imaging, particularly for enhancing diagnostic accuracy. This dual-modality approach, which combines B-mode and contrast-enhanced ultrasound (CEUS), improves tumor classification and characterization by leveraging the strengths of various imaging modalities, such as renal and thyroid lesions [4]. Additionally, systems that incorporate photoacoustic imaging with ultrasound enable comprehensive tissue assessment and precise lesion monitoring [11].

Mamba architecture addresses computational challenges inherent in traditional deep learning models, such as convolutional neural networks (CNNs) and transformers. Its ability to manage long-range dependencies and enhance computational efficiency makes it particularly suitable for processing extensive sequences in medical imaging [12]. Recent surveys highlight advancements in visual Mamba architectures and their applications across various computer vision tasks, addressing knowledge gaps regarding the model's capabilities and effectiveness [13].

Furthermore, the integration of machine learning techniques, including self-supervised and collaborative learning mechanisms, enhances diagnostic accuracy by leveraging large volumes of unlabeled data and diverse data modalities. This capability is crucial in meeting the growing demand for accurate diagnostic tools amidst a declining number of radiologists [10]. Reynaud et al. demonstrate the potential of transformer architectures in analyzing full temporal sequences of ultrasound videos, facilitating simultaneous regression of Left Ventricular Ejection Fraction (LVEF) and identification of End-Systolic (ES) and End-Diastolic (ED) frames, further exemplifying advancements in diagnostic precision [8].

The recent advancements in medical diagnostics, particularly through ultrasound dual-modality video and the innovative Mamba architecture, represent a transformative shift in the field. The Mamba-

www.SurveyX.cn

UNet model combines U-Net and Mamba's capabilities, enhancing medical image segmentation by effectively capturing both local features and global contextual information, thus addressing the limitations of traditional CNNs and Vision Transformers (ViTs). Additionally, the evolution of Multimodal Large Language Models (MLLMs) in healthcare illustrates their potential to integrate diverse data types—text, images, and audio—into clinical decision support systems, ultimately leading to more comprehensive patient insights and improved diagnostic accuracy. These developments collectively highlight the promising future of integrated technologies in enhancing medical diagnostics and patient care [14, 15, 16]. By improving imaging techniques and computational models, these technologies play a pivotal role in advancing clinical outcomes and the field of medical imaging.

1.3 Structure of the Survey

This survey is systematically structured to provide a comprehensive understanding of the integration of ultrasound dual-modality video and Mamba architecture in tumor diagnosis. It begins with an **Introduction** that emphasizes the significance of advanced medical imaging techniques, detailing their transformative impact on diagnostic accuracy and efficiency. The introduction elaborates on how the integration of ultrasound dual-modality video and Mamba architecture enhances diagnostic capabilities, particularly through the use of both static and dynamic ultrasound images, which leverage complementary information for improved lesion diagnosis. It also discusses Mamba architecture's ability to model complex medical data through advanced techniques like multi-modal feature fusion and self-supervised learning, thereby enhancing diagnostic processes [17, 4, 18, 12].

Following the introduction, **Section 2: Background and Preliminary Concepts** provides an overview of core concepts, including ultrasound dual-modality video, Mamba architecture, TIC curve analysis, and multimodal feature fusion, equipping readers with foundational knowledge for subsequent discussions.

explores the integration of ultrasound video data with complementary imaging techniques, such as photoacoustic imaging, to enhance visualization and analysis in medical diagnostics. This section highlights advancements in real-time imaging capabilities and the development of innovative dual-modality systems, including handheld probes for sentinel lymph node and urinary bladder imaging. It also discusses deep learning models that fuse static images and dynamic videos for improved breast and renal tumor diagnosis, illustrating how these technologies enhance diagnostic accuracy and monitoring of tumor characteristics, ultimately facilitating more effective clinical decision-making [17, 19, 4, 11, 20].

provides an in-depth examination of Mamba architecture, emphasizing its superior computational efficiency and advantages in processing medical images. This section elaborates on how Mamba's design enhances diagnostic accuracy through its ability to model long-range dependencies and integrate seamlessly with existing technologies, including hybrid architectures like Mamba-UNet, which improves medical image segmentation by combining traditional U-Net strengths with Mamba's advanced capabilities. It also discusses various applications of Mamba in medical image analysis, including classification, segmentation, and restoration, showcasing its versatility across different medical domains [15, 21, 16, 12].

In , the survey explores Time-Intensity Curves (TIC) and their critical role in tumor analysis, highlighting complexities in TIC curve interpretation and the potential for integrating deep learning techniques to enhance diagnostic accuracy and efficiency in oncology. This section underscores TIC's significance in improving tumor detection and characterization, contributing to more effective treatment strategies and patient outcomes [22, 23, 24, 25].

delves into the integration of various data sources for holistic tumor analysis. It examines multiple fusion strategies, including input, intermediate, and output fusion, while addressing challenges such as handling noisy, incomplete, and imbalanced multimodal data. The section also highlights recent advancements in deep learning-based multimodal fusion techniques, particularly emerging Transformer-based methods, and discusses their implications for enhancing medical image classification and understanding complex biological interactions in tumor pathology [26, 27, 28].

explores the critical role of video analysis in tumor diagnosis, discussing video segmentation techniques, unique challenges associated with medical video segmentation, and recent advancements leveraging deep learning. These advancements enhance diagnostic accuracy and efficiency by in-

www.SurveyX.cn

tegrating dynamic video data with traditional static imaging methods, ultimately improving tumor identification and characterization in clinical settings [29, 17, 25].

examines real-world implementations of advanced technologies in tumor diagnosis, highlighting various medical datasets and challenges encountered during implementation. The section emphasizes the role of artificial intelligence, particularly deep learning methods, in enhancing cancer subtype detection and gene mutation identification through medical imaging. It also discusses the evolution of multimodal large language models (MLLMs) and their integration of diverse data types, significantly improving clinical decision-making and patient engagement. Furthermore, it reviews the potential of large-scale generative AI applications in radiology to alleviate the shortage of radiologists by streamlining the interpretation process. These case studies illustrate the transformative impact of these technologies on cancer diagnostics and the ongoing need for further validation and research to optimize their effectiveness in clinical settings [30, 14, 31, 32, 23].

analyzes current obstacles in implementing advanced imaging techniques, particularly regarding deep learning and artificial intelligence in medical imaging. It highlights challenges such as the need for large annotated datasets, complexities in data augmentation, and multimodal model integration. The section discusses promising avenues for future research and innovation, including robust data augmentation strategies, improvements in interpretability and uncertainty quantification, and establishing ethical guidelines for generative AI applications in clinical settings. Identifying these critical areas aims to pave the way for advancements that could significantly enhance the efficacy and reliability of medical imaging technologies [14, 31, 32, 25, 33].

The survey culminates in , synthesizing the primary findings and contributions of the research, emphasizing the significant potential impact of integrating ultrasound dual-modality video and Mamba architecture in improving tumor diagnosis. This integration leverages the complementary strengths of dynamic ultrasound videos and static images, enhancing classification accuracy across various tumor types, as evidenced by improved performance metrics in recent studies, including an AUC of 90.0% for breast lesions and 0.901 for predicting central lymph node metastasis in thyroid cancer. These advanced methodologies could greatly enhance clinical workflows and diagnostic precision in oncology [34, 17, 35, 4, 13].The following sections are organized as shown in Figure 1.

www.SurveyX.cn

2 Background and Preliminary Concepts

2.1 Key Concepts Overview

Ultrasound dual-modality video, integrating contrast-enhanced ultrasound (CEUS), enhances visualization of vascular patterns and tissue perfusion, improving diagnostic accuracy, particularly for lung consolidations [6]. Challenges such as speckle noise remain a concern, affecting image quality and interpretation [36].

The Mamba architecture advances medical imaging by modeling long-range interactions and global contextual information, combining state-space modeling with Transformer blocks to enhance computational efficiency and diagnostic accuracy beyond traditional CNNs [8]. This is pivotal for processing extensive sequences in ultrasound videos.

Time-Intensity Curve (TIC) analysis is crucial for tumor diagnosis, offering quantitative insights into tumor vascularity and perfusion. When combined with deep learning, it improves diagnostic precision by addressing limited training data through data augmentation techniques like spatial transformations and noise additions, fostering robust machine learning models for classification and segmentation [32, 25].

Multimodal feature fusion integrates various medical data types, enhancing comprehensive analyses. The Dynamic Multimodal Collaborative Learning (DMCL) method introduces a dynamic fusion strategy that improves model representation and performance in complex scenarios [7]. This aligns with surveys exploring medical image fusion methods and applications [37].

These concepts highlight the transformative potential of integrating advanced imaging techniques and computational models in diagnostics. Leveraging ultrasound dual-modality video, static images, Mamba architecture, TIC analysis, and multimodal feature fusion, significant strides in tumor diagnosis and patient care are achieved. Recent advancements in deep learning-based multimodal fusion, like models combining Vision Transformer (ViT) with radiomics, have shown improved

AI-generated, for reference only.

diagnostic accuracy, with an AUC of 0.901 in predicting central lymph node metastasis in thyroid cancer, underscoring the potential of these approaches to enhance pathology understanding and clinical outcomes [35, 17, 28].

2.2 Innovative Imaging Techniques

Innovative imaging techniques address complexities in medical imaging, such as heterogeneity and distortions, necessitating advanced neural networks that manage diverse features efficiently [38]. Current task-specific models often underperform across modalities, highlighting the need for adaptable solutions [29].

Integrating multimodal feature fusion with knowledge-driven learning enhances training efficiency and classification accuracy, utilizing expert consultations to refine learning processes for complex data [39]. Transformers offer promising potential for fusing and analyzing multimodal medical images, crucial for understanding dependencies [40].

The Multimodal Variational Mixture-of-Experts (MMVM) VAE sets a benchmark for diagnostic performance evaluation from multimodal data, addressing representation learning from weakly-supervised heterogeneous sources [41]. This aligns with self-supervised multimodal contrastive learning, exemplified by ContIG, which aligns representations from unlabeled images and genetic data [42].

Challenges in volume rendering techniques limit their clinical utility for large 3D datasets, necessitating advancements to handle increasing complexity and size [43]. Issues like small sample sizes and heterogeneous data learning emphasize the importance of innovative techniques [26].

Self-supervised Contrastive Video-Speech Representation Learning (SCVSRL) enhances representation learning from ultrasound video and speech audio data through cross-modal contrastive learning [18]. The Segment Anything Model (SAM) shows promise in medical image analysis, though existing variants struggle with accurate segmentation [44].

Microscopic-Mamba effectively models long-range dependencies with linear complexity, capturing local and global features through its architecture [45]. Vande Schaft et al. use maximum-a-posteriori estimation with deep generative priors from MRI to suppress ultrasound speckle noise, enhancing image quality [36]. Collectively, these advancements highlight the transformative potential of novel imaging techniques in improving diagnostic accuracy and efficiency, leading to better patient care and outcomes.

3 Ultrasound Dual-Modality Video in Medical Imaging

Recent advancements in medical imaging have been significantly driven by dual-modality video techniques, enhancing imaging quality and enabling comprehensive analysis of tissue characteristics. This section delves into the improvements in visualization and analysis facilitated by this framework, emphasizing its implications for diagnostic accuracy in medical practice.

3.1 Enhancements in Visualization and Analysis

Combining ultrasound (US) and contrast-enhanced ultrasound (CEUS) in dual-modality video significantly enhances visualization and analysis in medical imaging by leveraging the strengths of both modalities. This approach enables detailed assessments of tissue characteristics and blood flow dynamics. Vakanski et al. improved breast tumor analysis in ultrasound images by using attention blocks with salient maps to focus on tumor regions [5]. Contrastive self-supervised learning, as demonstrated by Chen et al., refines ultrasound video data integration by extracting meaningful spatio-temporal representations, crucial for accurate tumor characterization [6]. Jiao's self-supervised learning approach enhances temporal understanding by correcting reshuffled video frames [46].

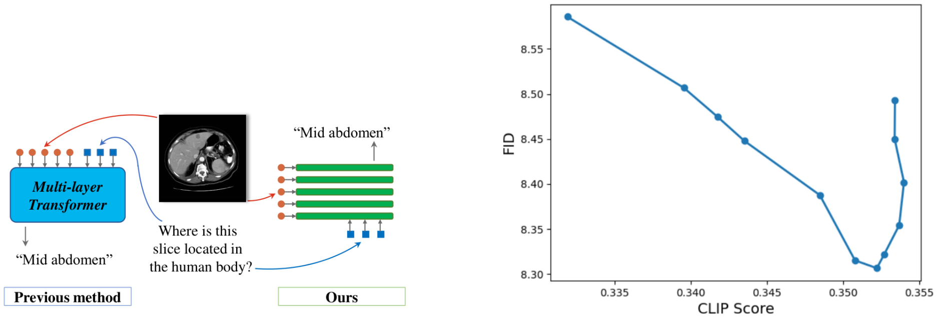
Reynaud et al.'s UVT model encodes ultrasound videos into a lower-dimensional space to extract spatio-temporal features, improving cardiac ultrasound image analysis by facilitating the regression of Left Ventricular Ejection Fraction (LVEF) and frame indices [8]. Vande Schaft et al. enhanced ultrasound image clarity by formulating ultrasound de-speckling as a MAP optimization problem using generative priors from MRI data, reducing noise and improving image quality [36]. The DMCL

AI-generated, for reference only.

www.SurveyX.cn

framework integrates visual features from a ResNet-50 image encoder and textual features from a BERT text encoder, exemplifying the potential of multimodal data integration in improving diagnostic accuracy and clinical outcomes in medical imaging [7].

Advancements in dual-modality video technology, particularly in multi-modal medical imaging, are critical for enhancing visualization and tumor analysis. Research indicates that integrating various imaging modalities, like the Gray Level Co-occurrence Matrix (GLCM) texture analysis in cervical cancer treatment, can improve treatment outcome prediction accuracy. A novel multi-modal ultrasound video fusion network for renal tumor diagnosis merges B-mode and CEUS videos to enhance tumor classification accuracy, promising improved diagnostic precision and paving the way for personalized treatment strategies [22, 14, 4].



(a) Mid Abdomen Localization in CT Images Using a Multi-layer Transformer[47]

(b) The graph shows the relationship between the CLIP score and the FID score for a dataset.[48]

Figure 2: Examples of Enhancements in Visualization and Analysis

As illustrated in Figure 2, the integration of advanced technologies in medical imaging has significantly enhanced visualization and analysis capabilities. The first figure demonstrates a multi-layer transformer for mid-abdomen localization in CT images, showcasing improved precision over traditional methods. The second figure highlights the relationship between the CLIP and FID scores, emphasizing the importance of quantitative metrics in evaluating imaging techniques. Collectively, these advancements illustrate the transformative potential of dual-modality imaging and sophisticated analytical tools in medical diagnostics and treatment planning [47, 48].

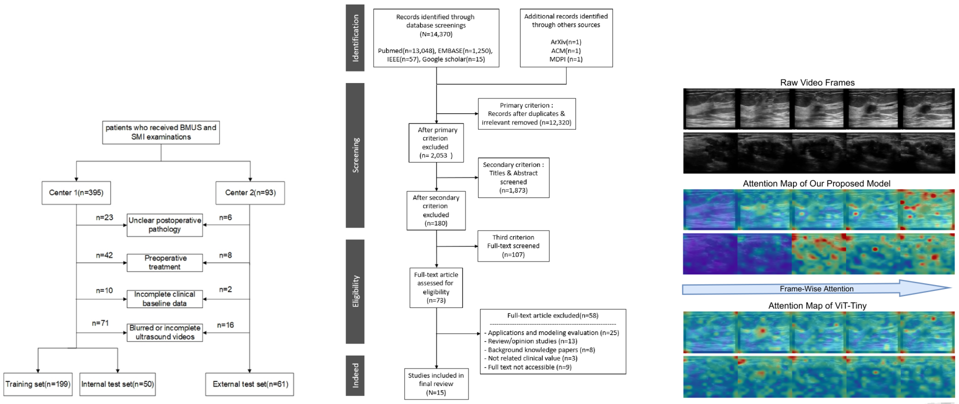
3.2 Applications in Specific Tumor Diagnoses

Dual-modality video in medical imaging has markedly advanced diagnostic capabilities across various tumor types, enhancing accuracy and consistency. The integration of ultrasound with CEUS improves visualization of vascular patterns and tissue perfusion, crucial for diagnosing tumors such as renal and thyroid lesions [4]. In fetal ultrasound examinations, the MMSummary framework automates multimodal summary generation, reducing scanning time and enhancing consistency across operators [34]. The UNICORN Nakagami imaging technique significantly outperforms existing methods in tumor diagnosis and fat fraction estimation, enhancing clinical evaluations [49]. Dual-modality approaches are also effective in breast cancer diagnosis, where B-mode and CEUS imaging allow for comprehensive tissue assessment and precise monitoring of lesion formation [11].

The transformative impact of dual-modality video technologies, such as CEUS and traditional ultrasound, on diagnosing specific tumors, including renal and thyroid cancers, is noteworthy. Advanced imaging techniques leverage complementary information from multiple modalities, enhancing accuracy and efficiency. Studies demonstrate that multi-modal approaches, such as the MUVF-YOLOX network for renal tumors and the Dual-Modality Watershed Fusion Network for thyroid nodules, outperform single-modality methods in classification accuracy and diagnostic performance. The combination of dynamic video and static images in breast lesion diagnosis further illustrates the potential of dual-modality video to improve clinical outcomes and optimize treatment strategies across various cancer types [19, 17, 22, 4]. By leveraging the strengths of multiple imaging modalities, healthcare professionals can achieve more reliable and consistent diagnostic outcomes, ultimately enhancing patient care and treatment planning.

www.Survey

AI-generated, for reference only.


patients who received BMUS and SMI examinations
Center 1(n=395)
Center 2(n=93)
n=23
Unclear postoperative pathology
n=6
n=42
Preoperative treatment
n=8
n=10
Incomplete clinical baseline data
n=2
n=71
Blurred or incomplete ultrasound videos
n=16
Training set(n=199)
Internal test set(n=50)
External test set(n=61)
Identification
Records identified through database screenings (N=14,370)
Pubmed(n=13,048), EMBASE(n=1,250), IEEE(n=57), Google scholar(n=15)
Additional records identified through others sources
ArXiv(n=1)
ACM(n=1)
MDPI (n=1)
Screening
Primary criterion : Records after duplicates & irrelevant removed (n=12,320)
After primary criterion excluded (n= 2,053)
Secondary criterion : Titles & Abstract screened (n=1,873)
After secondary criterion excluded (n=180)
Third criterion Full-text screened (n=107)
Eligibility
Full-text article assessed for eligibility (n=73)
Full-text article excluded(n=58)
- Applications and modeling evaluation (n=25)
- Review/opinion studies (n=13)
- Background knowledge papers (n=8)
- Not related clinical value (n=3)
- Full text not accessible (n=9)
Indeed
Studies included in final review (N=15)
Raw Video Frames
Attention Map of Our Proposed Model
Frame-Wise Attention
Attention Map of ViT-Tiny

(a) Flowchart of Patients Receiving BMUS and SMI Examinations[35]

(b) Flowchart of Study Selection Process[31]

(c) Comparison of Attention Maps for Video Classification with Different Models[17]

Figure 3: Examples of Applications in Specific Tumor Diagnoses

www.SurveyX.cn

As shown in Figure 3, the application of "Ultrasound Dual-Modality Video in Medical Imaging: Applications in Specific Tumor Diagnoses" is illustrated through visual aids that underscore the multifaceted approach in utilizing advanced imaging techniques for tumor diagnosis. The first figure presents a flowchart detailing the patient pathway for BMUS (Brain Magnetic Resonance Imaging) and SMI (Spinal Magnetic Resonance Imaging) examinations, highlighting inclusion criteria and patient distribution. The second figure outlines the study selection process for a comprehensive review, ensuring the inclusion of relevant and high-quality studies. The third figure compares attention maps generated by different models for video classification, illustrating the efficacy of the proposed model in focusing on critical anatomical regions. Collectively, these figures provide a robust framework for understanding the application of ultrasound dual-modality video in enhancing the precision and accuracy of tumor diagnoses in medical imaging [35, 31, 17].

4 Mamba Architecture and Its Role in Medical Imaging

The Mamba architecture marks a significant leap in medical imaging, offering innovative design and computational efficiency vital for addressing the growing demands of generative AI and deep learning techniques in medical diagnostics and clinical practice. These advancements enhance report generation and diagnostic accuracy while navigating challenges related to data diversity and transparency [31, 50, 33, 32]. This section explores how Mamba optimizes medical image processing, improving diagnostic precision across various applications.

4.1 Efficiency and Computational Advantages

Mamba architecture offers substantial computational benefits by efficiently managing long-range dependencies and integrating multimodal features, often achieving accuracy with fewer parameters compared to traditional CNNs and transformers. The Microscopic-Mamba's dual-branch design exemplifies this by enhancing feature extraction and achieving state-of-the-art results on public datasets [45]. Its ability to process long sequences with linear time complexity [12] is crucial for analyzing the extensive temporal data in ultrasound videos, addressing the computational demands of large medical datasets.

The MUVF-YOLOX framework illustrates Mamba's efficiency by integrating multimodal ultrasound video information, thus enhancing diagnostic accuracy [4]. The Mamba-UNet model further showcases improved long-range dependency modeling and feature learning, resulting in superior segmentation performance [15]. Vivim demonstrates enhanced spatiotemporal dependency modeling with lower computational costs compared to transformer-based methods [51], highlighting Mamba's adaptability in integrating varied data modalities.

AI-generated, for reference only.

Mamba's real-time capabilities in dual-modality applications enhance contrast and depth, making it suitable for numerous clinical applications [11]. Its miniaturization and compatibility with other imaging modalities, such as MRI, facilitate comprehensive patient assessments [20]. The DiM framework emphasizes scalability and computational efficiency, enabling high-resolution media generation with reduced resources [52]. Additionally, self-supervised learning approaches minimize the need for extensive human annotation across diverse medical contexts [46].

Attention mechanisms, as discussed in [5], enhance Mamba architectures by focusing processing on significant image regions, improving efficiency and reducing computational overhead. MambaClinix exemplifies these gains, achieving accuracy across multiple datasets and establishing itself as a promising tool for clinical segmentation tasks [53]. Unsupervised learning approaches leverage existing MRI data to enhance computational efficiency, reducing the need for extensive ultrasound data acquisition [36]. Advancements in modeling long-range dependencies and computational efficiency in visual tasks are directly applicable to the Mamba architecture [13].

As illustrated in Figure 4, the hierarchical structure of Mamba architecture applications in medical imaging is categorized into efficient feature integration, real-time capabilities, and advanced learning techniques. Each category highlights specific models and frameworks that demonstrate the transformative potential of Mamba in enhancing diagnostic accuracy and computational efficiency in medical diagnostics. MambaVision demonstrates these advantages by outperforming existing Mamba and Transformer-based models in image classification and downstream tasks, alongside its efficient processing capabilities [54]. Collectively, these advancements underscore Mamba's transformative potential in enhancing the efficiency and effectiveness of medical diagnostics.

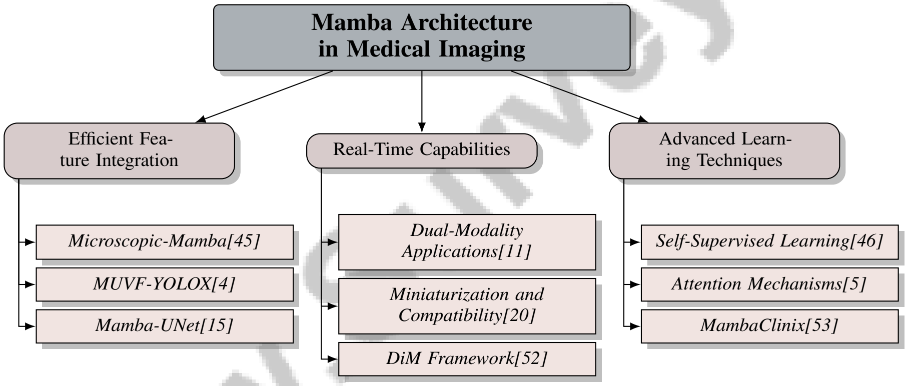

Mamba Architecture in Medical Imaging
Efficient Feature Integration
Real-Time Capabilities
Advanced Learning Techniques
Microscopic-Mamba[45]
MUVF-YOLOX[4]
Mamba-UNet[15]
Dual-Modality Applications[11]
Miniaturization and Compatibility[20]
DiM Framework[52]
Self-Supervised Learning[46]
Attention Mechanisms[5]
MambaClinix[53]

Figure 4: This figure shows the hierarchical structure of Mamba architecture applications in medical imaging, categorized into efficient feature integration, real-time capabilities, and advanced learning techniques. Each category highlights specific models and frameworks that demonstrate the transformative potential of Mamba in enhancing diagnostic accuracy and computational efficiency in medical diagnostics.

4.2 Integration with Other Technologies

The integration of Mamba architecture with other technologies significantly enhances diagnostic accuracy and efficiency in medical imaging. MambaVision, a hybrid model combining Mamba's efficient processing with Transformers' self-attention mechanisms, captures both local and global features in medical images, improving image classification and downstream tasks [54]. This demonstrates the potential of leveraging Mamba alongside Transformer models to meet the computational demands of high-dimensional visual data [12].

MambaClinix further illustrates this integration through a hierarchical gated convolutional network, enhancing local and global feature extraction in medical images [53]. This synergy improves segmentation and diagnostic accuracy across various imaging tasks, showcasing Mamba's adaptability to complex challenges.

Wei et al.'s method employs separate branches for spatial and temporal information extraction, utilizing advanced aggregation and rolling techniques to enhance representation [55]. This aligns

AI-generated, for reference only.

with hierarchical modeling principles, demonstrating Mamba's capacity for effective integration with advanced computational techniques.

Bansal et al.'s survey organizes existing Mamba methods into stages such as optimizations, scanning techniques, and applications across various medical domains, highlighting the evolution and performance of these architectures [12]. This comprehensive overview underscores the versatility and transformative potential of Mamba architecture in advancing medical imaging technologies.

These integrations highlight Mamba architecture's transformative potential in enhancing medical imaging. By combining Mamba with state-of-the-art models and methodologies, researchers can improve the precision and efficiency of diagnostic processes, leading to better patient outcomes and more informed clinical decision-making. The fusion of diverse data sources through Mamba's architecture, such as electronic health records and medical imaging, significantly advances medical image analysis capabilities, ultimately transforming patient care in healthcare settings [14, 21, 12, 56].

5 TIC Curve Analysis for Tumor Diagnosis

Analyzing Time-Intensity Curves (TIC) is crucial for tumor diagnosis, offering insights into tumor hemodynamics through dynamic contrast-enhanced imaging. This section delves into TICs' role in improving diagnostic accuracy, particularly in distinguishing malignant from benign lesions, by integrating advanced computational techniques and addressing inherent challenges in TIC analysis. Figure 5 illustrates the hierarchical structure of TIC analysis for tumor diagnosis, highlighting the concept and importance of TICs, the challenges faced in their analysis, and the integration with deep learning techniques to enhance diagnostic accuracy and clinical applicability. This visual representation not only complements the textual discussion but also emphasizes the multifaceted approach required for effective TIC analysis in clinical settings.

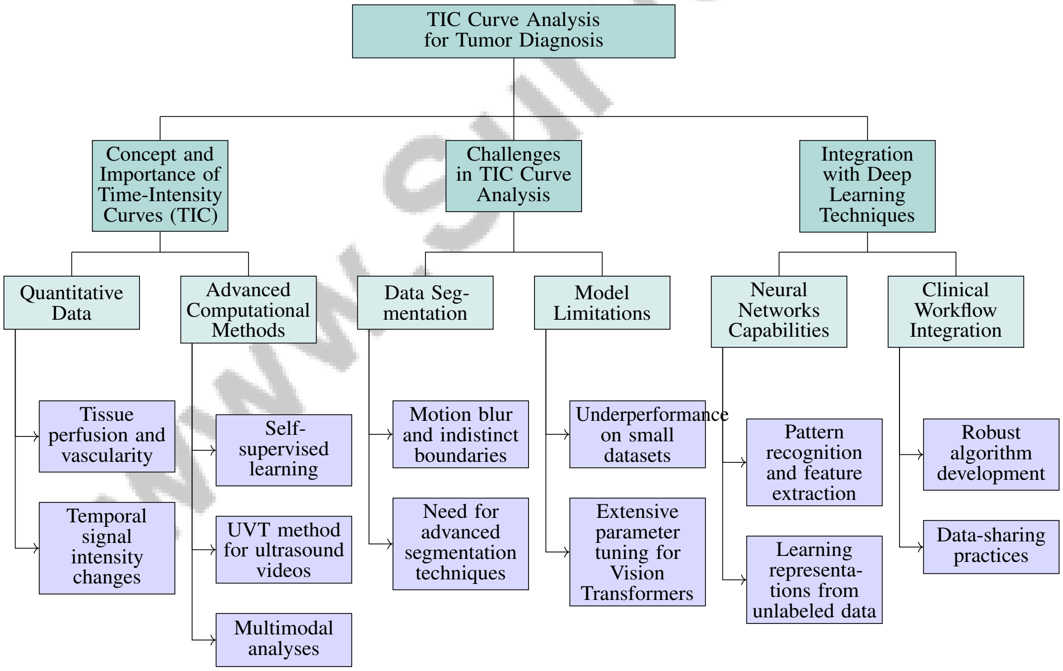

TIC Curve Analysis for Tumor Diagnosis
Concept and Importance of Time-Intensity Curves (TIC)
Challenges in TIC Curve Analysis
Integration with Deep Learning Techniques
Quantitative Data
Advanced Computational Methods
Data Segmentation
Model Limitations
Neural Networks Capabilities
Clinical Workflow Integration
Tissue perfusion and vascularity
Temporal signal intensity changes
Self-supervised learning
UVT method for ultrasound videos
Multimodal analyses
Motion blur and indistinct boundaries
Need for advanced segmentation techniques
Underperformance on small datasets
Extensive parameter tuning for Vision Transformers
Pattern recognition and feature extraction
Learning representations from unlabeled data
Robust algorithm development
Data-sharing practices

Figure 5: This figure illustrates the hierarchical structure of Time-Intensity Curve (TIC) analysis for tumor diagnosis, highlighting the concept and importance of TICs, the challenges faced in their analysis, and the integration with deep learning techniques to enhance diagnostic accuracy and clinical applicability.

5.1 Concept and Importance of Time-Intensity Curves (TIC)

Time-Intensity Curves (TIC) provide vital quantitative data on tissue perfusion and vascularity, essential for tumor analysis in dynamic contrast-enhanced imaging. They capture temporal signal

www.surveyx.cn

AI-generated, for reference only.

intensity changes post-contrast agent administration, aiding in differentiating malignant from benign lesions through vascular pattern analysis [6, 5]. TICs are particularly valuable in overcoming challenges like operator dependency and sensitivity in dense breast tissue, offering objective tumor assessments [3]. Their role in holistic evaluation underscores their necessity for nuanced anatomical and disease recognition [10].

Advanced computational methods, such as self-supervised learning, enhance TIC analysis by extracting intricate patterns from unlabeled data, thus increasing tumor characterization robustness [46]. The UVT method exemplifies TICs' utility in identifying key frames and estimating cardiac function from ultrasound videos, underscoring their significance in comprehensive tumor assessments [8]. TICs are also integral to multimodal analyses, facilitating real-time monitoring of therapeutic interventions and improving treatment precision [53]. Collectively, TICs offer critical insights into vascular characteristics, advancing diagnostic accuracy and patient care.

5.2 Challenges in TIC Curve Analysis

TIC analysis faces challenges like motion blur and indistinct boundaries due to the amorphous nature of relevant classes in medical imaging, complicating data segmentation and interpretation [38]. This necessitates advanced segmentation techniques for diverse tumor characteristics. The reliance on large datasets for effective TIC analysis, particularly with transformer networks, poses another challenge, as these models often underperform on small datasets [40]. Robust data augmentation and synthesis methods are needed to enhance TIC applicability across clinical scenarios.

Segmentation accuracy is critical, as models like the Segment Anything Model (SAM) often require substantial manual corrections for reliable results [44]. Improved automated segmentation techniques are essential to reduce manual intervention and enhance TIC analysis efficiency. The complexity of overlapping regions with multiple level-set functions presents additional obstacles, potentially leading to excessive distinct regions that do not correspond to desired outcomes [57]. Advanced computational models are needed to accurately delineate complex tumor boundaries, enhancing TIC analysis precision.

Extensive parameter tuning and computational resources required for optimizing Vision Transformers (ViTs) in medical image segmentation often exceed practical limits, posing significant challenges for TIC analysis [58]. Addressing these demands is crucial for enhancing TIC analysis scalability across diverse medical imaging contexts. The integration of AI and large-scale generative models can address existing limitations in tumor diagnosis, improving accuracy and efficiency in medical imaging. AI, particularly deep learning algorithms, aids in cancer detection and characterization, facilitating effective treatment strategies, though further validation is essential for clinical workflows [31, 23].

5.3 Integration with Deep Learning Techniques

Integrating deep learning with TIC analysis enhances tumor diagnosis by utilizing neural networks' pattern recognition and feature extraction capabilities. Deep learning models, especially those employing self-supervised learning, can learn representations from unlabeled data, as demonstrated by the ContIG model, which aligns medical images and genetic modalities using contrastive loss [42]. This is crucial for TIC analysis, where large labeled datasets are scarce, enhancing diagnostic model robustness and accuracy.

Incorporating machine learning into clinical workflows facilitates robust algorithm development and improves data-sharing practices, enhancing TIC analysis applicability in real-world settings [2]. Deep learning also addresses data heterogeneity and variability, leading to consistent diagnostic outcomes. Future research directions, such as those proposed by Mo et al., suggest enhancing deep learning models like the DiM framework to handle diverse video types and generate dynamic scenes [52]. These advancements are applicable to TIC analysis, where modeling complex temporal dynamics and interactions between imaging modalities is essential for accurate tumor characterization.

Overall, integrating deep learning with TIC analysis offers promising avenues for improving tumor diagnosis precision and reliability. By leveraging neural networks' capabilities to learn intricate patterns from large datasets, researchers can enhance the interpretability and clinical applicability of TIC analysis. This advancement improves diagnostic accuracy and treatment planning, contributing to

AI-generated, for reference only.

www.SurveyX.cn

better patient care and outcomes, particularly in complex medical imaging tasks like digital pathology and various imaging modalities such as CT and MRI [33, 25].

6 Multimodal Feature Fusion Techniques

6.1 Fusion Strategies and Taxonomies

Method Name	Fusion Strategies	Applicable Scenarios	Integration Techniques
M4oE[59]	Attention-based Fusion	Medical Imaging	M4oe Framework
MMD[60]	Dynamic Fusion Strategy	Medical Diagnosis	Sparse Feature Coding
3DMResNet[61]	Feature-data Fusion	Vmat Plan QA	3dmresnet
DMCL[7]	Dynamic Multimodal Fusion	Polyp Re-identification	Deep Multimodal Collaborative

Table 1: Overview of multimodal fusion strategies and their applications in medical imaging. The table lists various methods, the fusion strategies they employ, the scenarios in which they are applicable, and the integration techniques used. This comprehensive tabulation highlights the diversity and specificity of approaches in enhancing diagnostic accuracy and model robustness.

Multimodal feature fusion is crucial in medical imaging, enhancing diagnostic precision by integrating diverse data sources. The M4oE framework exemplifies a mixture of experts approach, optimizing representation learning across various imaging modalities, thereby bolstering the robustness of medical image segmentation models [59]. Current methodologies are categorized into strategies such as input fusion, single-level fusion, hierarchical fusion, attention-based fusion, and output fusion [28]. Input fusion combines raw data from different modalities, while single-level fusion integrates features at specific network layers. Hierarchical fusion employs multiple integration layers for nuanced feature combinations. Attention-based fusion selectively highlights relevant features, enhancing model integration [60]. Output fusion synthesizes information from multiple models or modalities, improving prediction accuracy and diagnostic outcomes. Table 1 provides a detailed overview of the different multimodal fusion strategies utilized in medical imaging, illustrating the methods, applicable scenarios, and integration techniques that enhance diagnostic precision.

The feature-data fusion (FDF) approach merges imaging and non-imaging modalities, enhancing prediction accuracy in diagnostics [61]. In disease diagnosis, various machine learning algorithms and fusion strategies are employed to improve model performance [56]. The dynamic fusion strategy in the Multimodal Dynamics method models feature-level and modality-level variations, further enhancing data integration [60]. The DMCL framework combines visual and textual representations, demonstrating the potential of multimodal approaches to bolster diagnostic accuracy in medical imaging [7]. Collectively, these strategies illustrate the transformative potential of multimodal feature fusion in advancing medical imaging and patient care.

www.SurveyX.cn

6.2 Challenges in Multimodal Feature Fusion

Multimodal feature fusion in medical imaging faces challenges that hinder effective data integration. A primary challenge is the variability and heterogeneity of medical data, encompassing diverse imaging modalities, electronic health records, and clinical notes [7]. This diversity necessitates sophisticated fusion strategies for harmonizing disparate data types for reliable diagnostic outcomes. Another significant challenge is the alignment of multimodal data, requiring precise synchronization to ensure accurate fusion [60]. Differences in data acquisition techniques, resolutions, and temporal dynamics can complicate this process, leading to misalignment and information loss.

The computational complexity of processing large volumes of multimodal data presents another obstacle, often demanding substantial resources and advanced algorithms [59]. This complexity can limit the scalability and applicability of multimodal fusion techniques in clinical settings, where real-time analysis is crucial. Developing robust and generalizable models that effectively fuse multimodal data across diverse patient populations remains challenging, as the dynamic nature of medical data requires models that can adapt to new information without compromising performance [61]. Lastly, the interpretability of fused data outputs is crucial, as clinicians must understand and trust the insights generated by multimodal models. Ensuring transparency and explainability is essential for the acceptance of these models in clinical practice [56].

Addressing these challenges is vital for enhancing the integration of diverse medical data sources, significantly improving diagnostic accuracy and patient care outcomes. By overcoming issues related to data quality, modality incompleteness, and appropriate fusion technique selection, researchers can leverage deep learning and artificial intelligence to create robust multimodal systems that provide comprehensive insights into patient health [56, 27, 28].

6.3 Innovations and Advanced Techniques

Recent innovations in feature fusion have significantly advanced medical imaging, enhancing the integration of diverse data sources for improved diagnostic accuracy and efficiency. The Auxiliary Online Learning (AuxOL) approach, which integrates a small auxiliary model with the Segment Anything Model (SAM), exemplifies this trend by utilizing expert annotations for real-time segmentation accuracy improvement [44]. This adaptive learning strategy highlights the potential of enhancing multimodal feature fusion in dynamic clinical environments.

Attention-based fusion strategies have also made substantial contributions, allowing multimodal models to prioritize pertinent features from various data types—such as text, images, and audio—thereby improving the integration of complex biomedical information. This capability facilitates comprehensive insights into patient health and enhances the interpretability of diagnostic outcomes, ultimately supporting clinical decision-making and patient engagement [14, 31, 30]. Hierarchical modeling principles further enhance the extraction of local and global features from multimodal data, improving the robustness and adaptability of fusion models.

Hybrid models, combining the strengths of different architectures like Mamba and Transformer models, have shown significant improvements in processing efficiency and diagnostic accuracy. MambaVision models, for instance, integrate Mamba architectures' efficiency with Transformers' self-attention mechanisms, optimizing feature fusion in medical imaging and improving performance metrics such as Top-1 accuracy and image throughput on benchmark datasets like ImageNet-1K, MS COCO, and ADE20K [54, 62].

Dynamic fusion strategies that model variations at both feature and modality levels represent a significant milestone in multimodal classification, enhancing the reliability of heterogeneous data integration in safety-critical applications like medical diagnosis. These strategies dynamically assess the informativeness of different modalities and features, allowing for a more accurate and robust fusion process [60, 56, 27, 28]. By accommodating the dynamic nature of medical data, these approaches enhance the adaptability and performance of fusion models across diverse clinical applications.

The advancements in feature fusion techniques in medical imaging underscore their transformative potential, enhancing diagnostic accuracy and patient care through the integration of diverse data sources and sophisticated computational models. Recent developments in deep learning-based multimodal fusion methods have emerged as effective tools for improving medical image classification by leveraging complementary information from various imaging modalities. This integration addresses challenges such as incomplete data management and optimal network architecture selection, promising significant growth in the field as research progresses, particularly with the potential of Transformer-based approaches [37, 28].

7 Video Analysis in Medical Imaging

7.1 Role of Video Segmentation in Tumor Diagnosis

Video segmentation plays a crucial role in tumor diagnosis by enabling precise delineation of tumor boundaries and extraction of relevant features from medical imaging data. This precision is essential for accurately identifying tumor regions and assessing their characteristics, which are vital for effective diagnosis and treatment planning. The use of advanced techniques, particularly deep learning models, has significantly improved the accuracy and efficiency of tumor analysis by automating boundary identification and reducing reliance on manual intervention [7].

Integrating segmentation with multimodal imaging data enhances the capture of complex anatomical structures and tumor characteristics, thereby improving diagnostic precision. Attention-based models facilitate focused analysis of relevant regions within video data, allowing for detailed assessments of

AI-generated, for reference only.

www.SurveyX.cn

tumor morphology and vascular patterns [5]. This approach effectively addresses challenges posed by variability and heterogeneity in tumor presentations across patients and imaging modalities.

Segmentation techniques in dynamic contrast-enhanced imaging enable real-time monitoring of tumor perfusion and vascularity, offering insights into tumor behavior and treatment response. This capability is crucial for tailoring therapeutic interventions and evaluating their efficacy over time [6]. Developing robust segmentation models that adapt to the dynamic nature of medical video data is essential for enhancing clinical utility.

Video segmentation significantly enhances tumor diagnosis by improving accuracy and efficiency in analyzing medical imaging data. By facilitating precise tumor boundary delineation and the extraction of clinically relevant features, advanced segmentation techniques inform personalized treatment planning and improve patient outcomes. These techniques streamline large-scale dataset analysis and support deep learning models that learn from diverse medical image data, advancing diagnostic tools and disease progression monitoring [2, 32, 25, 44, 29].

7.2 Challenges in Medical Video Segmentation

Medical video segmentation faces numerous challenges due to the complexities inherent in medical imaging data. Variability in image quality and resolution can lead to inconsistent segmentation outcomes, complicating tumor boundary identification [7]. Noise and artifacts, especially in ultrasound video data, obscure critical anatomical details, further complicating the segmentation process [36].

The dynamic nature of medical video data, characterized by continuous changes in anatomical structures and tumor morphology, requires advanced segmentation models capable of adapting to these changes. Ensuring accurate and consistent tumor delineation across different frames is crucial, particularly in real-time clinical settings that demand efficient algorithms for rapid and reliable segmentation without sacrificing accuracy [6].

The heterogeneity of tumor presentations across patients and imaging modalities adds complexity to segmentation efforts. Tumors exhibit diverse shapes, sizes, and contrast levels, necessitating robust models that generalize across various clinical scenarios and imaging conditions [5]. Limited availability of annotated medical video datasets further restricts the training and validation of segmentation models, hindering their development and clinical deployment.

Integrating multimodal imaging data, which combines information from different techniques, presents challenges in aligning and fusing these diverse data sources for accurate segmentation [7]. Developing sophisticated fusion strategies to harmonize disparate modalities and enhance segmentation interpretability is critical for advancing medical video segmentation.

Addressing these challenges is essential for improving the accuracy and efficiency of medical video segmentation, vital for clinical applications such as disease diagnosis, treatment planning, and monitoring. Enhancing segmentation techniques through universal models like MedSAM, leveraging large-scale datasets and advanced deep learning methods, can significantly advance medical imaging capabilities. This progress will facilitate precise and timely diagnostics and support personalized treatment plans, ultimately leading to improved patient outcomes [32, 29, 25].

7.3 Advancements in Video Analysis Technologies

Recent advancements in video analysis technologies have significantly enhanced medical imaging capabilities, particularly in tumor diagnosis. The development of advanced neural network architectures, including those utilizing self-supervised learning, has improved the extraction of meaningful spatio-temporal representations from ultrasound video data, enhancing the accuracy and reliability of diagnostic models [6]. These models can learn from unlabeled data, addressing challenges posed by limited annotated datasets and enabling robust analysis of complex medical video data.

Incorporating attention mechanisms into video analysis frameworks has further improved the interpretability and diagnostic utility of medical imaging technologies. By focusing on salient regions within video data, attention-based models enhance relevant feature extraction, allowing for detailed assessments of tumor characteristics and vascular patterns [5]. This approach helps overcome the variability and heterogeneity inherent in medical video data, ensuring consistent and reliable diagnostic outcomes.

www.SurveyX.cn

AI-generated, for reference only.

Innovations in multimodal feature fusion have also advanced video analysis technologies. By integrating diverse data sources, such as electronic health records and various imaging modalities, these fusion strategies enhance the precision and accuracy of diagnostic models [7]. Dynamic fusion strategies that model both feature-level and modality-level variations further enhance the adaptability and performance of video analysis technologies across a wide range of clinical applications.

The application of deep learning techniques in video analysis has facilitated real-time processing and analysis of medical video data, enabling rapid and accurate tumor boundary delineation and treatment response assessment [46]. These advancements underscore the transformative potential of video analysis technologies in enhancing the diagnostic and therapeutic capabilities of medical imaging, ultimately improving patient care and clinical outcomes.

8 Applications and Case Studies

8.1 Case Studies on Diverse Medical Datasets

The exploration of advanced imaging techniques and multimodal feature fusion through case studies on diverse medical datasets highlights their significant impact on tumor diagnosis. The surge in publications from 2016 to 2023 underscores the emphasis on multimodal medical classification, especially in brain studies, which integrate varied data sources to boost diagnostic accuracy and predictive capabilities [28]. Studies on datasets like BRCA, LGG, ROSMAP, and KIPAN demonstrate the clinical efficacy of these techniques, employing modalities such as mRNA expression, DNA methylation, and miRNA expression. This integration enriches the understanding of complex disease mechanisms, thus enhancing diagnostic model accuracy and clinical relevance [60].

Multimodal feature fusion's success illustrates the benefits of combining imaging and non-imaging modalities, improving predictive accuracy and offering a comprehensive view of complex data scenarios. This approach effectively addresses issues related to low-quality data, including noise and incompleteness, paving the way for future research in robust multimodal analysis [61, 27, 28]. By leveraging diverse data sources, researchers can develop models that enhance diagnostic precision and reliability, ultimately advancing patient care and treatment outcomes.

Collectively, these case studies highlight significant advancements enabled by technologies such as ultrasound dual-modality video, Mamba architecture, and multimodal feature fusion in medical imaging. They showcase the effectiveness of deep learning and large-scale generative AI across various clinical applications, including digital pathology, chest, brain, cardiovascular, and abdominal imaging, offering critical insights into their potential to enhance tumor diagnosis, improve imaging accuracy, and address challenges like the shortage of radiologists and the need for robust data analysis techniques [31, 23, 33, 32].

8.2 Real-World Implementation Challenges

The adoption of advanced imaging technologies, such as ultrasound dual-modality video and Mamba architecture, in clinical settings faces substantial challenges. Variability in data quality and availability significantly affects the performance and reliability of diagnostic models [28]. The heterogeneity of medical data necessitates robust data harmonization strategies to ensure consistent diagnostic outcomes across diverse healthcare environments.

Integrating these technologies into clinical workflows often requires significant modifications, making the process resource-intensive and time-consuming, thus posing barriers to seamless adoption [60]. Furthermore, the need for specialized training to operate and interpret outputs from these imaging systems complicates implementation, potentially limiting accessibility in resource-constrained settings.

The computational demands of processing large volumes of multimodal data also present a challenge. High-performance computing infrastructure and advanced algorithms are essential for effective data management, which can be prohibitive in low-resource environments [59]. This underscores the necessity for scalable and efficient computational models compatible with existing healthcare system constraints.

Ethical and regulatory considerations related to data privacy and security must also be addressed for responsible clinical use. Integrating diverse data sources, including electronic health records

www.SurveyX.cn

and imaging data, requires robust data governance frameworks to protect patient confidentiality and ensure compliance with regulatory standards [56].

These challenges highlight the need for ongoing innovation and collaboration among researchers, clinicians, and policymakers. By overcoming these barriers, stakeholders can unlock the full potential of advanced imaging technologies, including large-scale generative AI and deep learning methods, which promise to enhance tumor diagnosis and improve patient care outcomes. A collaborative approach is crucial to leveraging multimodal data integration, such as combining electronic health records with imaging data, facilitating more accurate diagnostics and personalized treatment strategies, thereby transforming clinical practice in oncology [31, 23, 33, 56].

9 Challenges and Future Directions

9.1 Data Quality and Availability

Developing diagnostic models in medical imaging is significantly challenged by data quality and availability. High-quality saliency maps are crucial, as inferior maps can compromise model performance, leading to inaccurate segmentations and affecting overall data quality [5]. The acquisition of expert annotations for ultrasound videos is resource-intensive, complicating the collection of adequately labeled data [6]. This dependency on extensive labeled datasets highlights the importance of data quality in achieving precise diagnostic results.

To address data quality issues, Kumar et al. propose methods that enhance segmentation accuracy and robustness while reducing the dependency on large labeled datasets [58]. However, challenges persist with video data quality, as noise or artifacts in ultrasound images can affect learned representations in self-supervised learning frameworks [46]. The quality of textual descriptions is also vital, particularly in the DMCL method, where inaccuracies can impair model effectiveness [7]. Additionally, high processing times for iterative MAP inference limit the practical use of advanced imaging techniques in real-time applications, further constraining data availability [36].

Future research must prioritize clinical validation of model predictions and enhancements to existing architectures to improve performance and data utilization [8]. Addressing these challenges is essential for advancing medical imaging reliability in clinical practice.

9.2 Ethical and Socioeconomic Considerations

The integration of advanced imaging technologies, including ultrasound dual-modality video and Mamba architecture, into medical diagnostics raises ethical and socioeconomic concerns. Ethical issues primarily revolve around patient privacy and data security, especially with the use of machine learning models and large language models (MLLMs) alongside sensitive medical data [14]. Ensuring robust data protection and compliance with regulatory standards like GDPR and HIPAA is essential to maintain public trust.

Socioeconomic factors significantly influence the adoption of advanced imaging techniques, affecting access to resources, availability of trained professionals, and integration into clinical practice. The shortage of radiologists has increased the demand for AI solutions to enhance efficiency and diagnostic accuracy. Challenges in acquiring diverse training datasets and the costs of implementing cutting-edge technologies underscore the importance of socioeconomic considerations in deploying advanced imaging modalities [31, 32, 25, 29, 33]. The high costs of acquiring and maintaining state-of-the-art imaging equipment and the necessary computational infrastructure for processing multimodal data create barriers to implementation, particularly in low-resource settings, potentially exacerbating healthcare disparities.

Integrating these technologies into clinical practice requires specialized training and expertise, which can be resource-intensive and limit skilled personnel availability in certain regions. Addressing these socioeconomic challenges is crucial for equitable deployment of advanced imaging technologies across diverse healthcare environments. Collaboration among healthcare providers, policymakers, and industry stakeholders can foster strategies that enhance ethical practices and socioeconomic sustainability while leveraging advancements in AI and multimodal data integration. This approach is vital for improving patient care and clinical outcomes, as demonstrated by the increasing effectiveness of multimodal fusion techniques in diagnosing and predicting diseases. Implementing human-centered

www.SurveyX.cn

AI-generated, for reference only.

design principles in developing machine learning models can ensure these technologies meet user needs and are clinically relevant, ultimately leading to more transparent and effective healthcare solutions [14, 50, 56].

9.3 Future Innovations and Applications

The future of medical imaging is poised for significant advancements through innovative techniques and novel applications. A promising direction involves exploring the scalability of FST-Mamba, with potential improvements via pretraining techniques to enhance adaptability across diverse imaging tasks [55]. This approach aims to refine model performance, ensuring robustness and effectiveness in various clinical settings.

Enhancing model robustness and exploring transfer learning techniques are crucial for future innovations in medical imaging, as suggested by Jiménez-Gaona et al. [3]. These strategies are instrumental in developing models that generalize across different imaging modalities and patient populations, improving diagnostic accuracy and reliability.

Future research could focus on automating feedback mechanisms or improving auxiliary model architectures to enhance performance in scenarios with limited expert input [44]. This direction is essential for reducing reliance on extensive manual annotations and facilitating the scalability of advanced imaging techniques in resource-constrained environments.

The exploration of lightweight Mamba architectures for real-time applications and enhancing multimodal learning capabilities are critical areas for future research [12]. These efforts aim to address challenges related to training with limited labeled data and improve diagnostic model efficiency in dynamic clinical settings.

Additionally, the robustness of the UNICORN method could be further improved in diverse clinical settings, with research exploring its applicability to other types of tissue characterization [49]. This exploration will contribute to refining the precision and reliability of diagnostic models across various medical contexts.

Future research will also focus on integrating deep learning techniques for automated feature extraction and enhancing segmentation frameworks to address more complex imaging problems [57]. This integration is crucial for advancing medical diagnostics and improving patient care outcomes.

Further refinements to augmentation techniques and the applicability of self-supervised learning methods to other medical imaging modalities represent another promising avenue for future innovations [6]. These advancements will enhance the robustness and adaptability of diagnostic models, ensuring effectiveness across a wide range of clinical applications.

Refining adaptive configuration strategies and exploring the application of MambaClinix in other medical imaging tasks will be a focus of future research [53]. This exploration will contribute to developing more versatile and efficient diagnostic models capable of addressing diverse imaging challenges.

Exploring different architectures of large language models (LLMs) and optimizing them for specific medical imaging tasks will be crucial for balancing model size and performance [58]. This direction will facilitate the development of efficient and scalable models capable of delivering accurate diagnostic outcomes in real-world clinical settings.

Finally, scaling visual Mamba models and exploring their application in dense prediction tasks, alongside integrating advanced techniques for improved performance, will be critical for future research [13]. These efforts will pave the way for significant advancements in medical imaging, offering promising avenues for enhancing diagnostic accuracy, efficiency, and patient care. By fostering collaboration among researchers, clinicians, and policymakers, the field can continue to evolve, ultimately improving patient outcomes and advancing the state of medical diagnostics.

10 Conclusion

The integration of ultrasound dual-modality video with Mamba architecture represents a pivotal advancement in medical imaging, significantly enhancing diagnostic accuracy and efficiency in tumor diagnosis. This survey highlights the transformative potential of multimodal feature fusion,

www.SurveyX.cn

which integrates diverse data types to improve diagnostic outcomes. The MUVF-YOLOX framework exemplifies this potential by outperforming single-modal and other state-of-the-art methods in renal tumor diagnosis, thereby supporting more informed clinical decision-making.

Advanced imaging techniques, such as the MVCC method, demonstrate substantial promise in breast cancer diagnosis from ultrasound videos by addressing challenges related to limited annotations and complex lesion patterns. Furthermore, dual-modality endoscopes that combine high-resolution imaging for photoacoustic tomography and white light videoscopy illustrate the clinical applicability of these technologies in laparoscopic surgery.

The significance of self-supervised learning frameworks in medical imaging is underscored, enhancing performance on downstream tasks while reducing dependence on labor-intensive manual annotation. The DMCL framework further exemplifies the impact of integrating multimodal representations, achieving state-of-the-art performance in polyp re-identification tasks.

Key findings underscore the potential of PAT/US imaging in improving diagnostic accuracy and patient outcomes, with substantial promise for future clinical applications. The effectiveness of multimodal fusion in enhancing clinical outcomes is noted, alongside the necessity for more comprehensive datasets to advance this research area. Additionally, the method proposed by Vande Schaft et al. demonstrates significant improvements in denoising ultrasound images, thus enhancing image quality.

Collectively, these findings underscore the critical role of advanced imaging techniques, such as ultrasound dual-modality video and Mamba architecture, in enhancing tumor diagnosis. By refining imaging techniques and computational models, these technologies advance clinical outcomes and the field of medical imaging. Continued innovation and collaboration among researchers, clinicians, and policymakers are essential to address existing challenges and fully realize the potential of these technologies in improving patient care.

www.SurveyX.cn

AI-generated, for reference only.

References

[1] Rafael Orozco, Mathias Louboutin, Ali Siahkoohi, Gabrio Rizzuti, Tristan van Leeuwen, and Felix Herrmann. Amortized normalizing flows for transcranial ultrasound with uncertainty quantification, 2023.

[2] Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. Machine learning for medical imaging. *radiographics*, 37(2):505–515, 2017.

[3] Yuliana Jiménez-Gaona, María José Rodríguez-Álvarez, and Vasudevan Lakshminarayanan. Deep learning based computer-aided systems for breast cancer imaging : A critical review, 2020.

[4] Junyu Li, Han Huang, Dong Ni, Wufeng Xue, Dongmei Zhu, and Jun Cheng. Muvf-yolox: A multi-modal ultrasound video fusion network for renal tumor diagnosis, 2023.

[5] Aleksandar Vakanski, Min Xian, and Phoebe Freer. Attention enriched deep learning model for breast tumor segmentation in ultrasound images, 2020.

[6] Li Chen, Jonathan Rubin, Jiahong Ouyang, Naveen Balaraju, Shubham Patil, Courosh Mehanian, Sourabh Kulhare, Rachel Millin, Kenton W Gregory, Cynthia R Gregory, Meihua Zhu, David O Kessler, Laurie Malia, Almaz Dessie, Joni Rabiner, Di Coneybeare, Bo Shopsin, Andrew Hersh, Cristian Madar, Jeffrey Shupp, Laura S Johnson, Jacob Avila, Kristin Dwyer, Peter Weimersheimer, Balasundar Raju, Jochen Kruecker, and Alvin Chen. Contrastive self-supervised learning for spatio-temporal analysis of lung ultrasound videos, 2023.

[7] Suncheng Xiang, Jincheng Li, Zhengjie Zhang, Shilun Cai, Jiale Guan, and Dahong Qian. Deep multimodal collaborative learning for polyp re-identification, 2024.

[8] Hadrien Reynaud, Athanasios Vlontzos, Benjamin Hou, Arian Beqiri, Paul Leeson, and Bernhard Kainz. Ultrasound video transformers for cardiac ejection fraction estimation, 2021.

[9] Zelong Liu, Andrew Tieu, Nikhil Patel, Georgios Soultanidis, Louisa Deyer, Ying Wang, Sean Huver, Alexander Zhou, Yunhao Mei, Zahi A. Fayad, Timothy Deyer, and Xueyan Mei. Vis-mae: An efficient self-supervised learning approach on medical image segmentation and classification, 2025.

[10] Zhengliang Liu, Hanqi Jiang, Tianyang Zhong, Zihao Wu, Chong Ma, Yiwei Li, Xiaowei Yu, Yutong Zhang, Yi Pan, Peng Shu, Yanjun Lyu, Lu Zhang, Junjie Yao, Peixin Dong, Chao Cao, Zhenxiang Xiao, Jiaqi Wang, Huan Zhao, Shaochen Xu, Yaonai Wei, Jingyuan Chen, Haixing Dai, Peilong Wang, Hao He, Zewei Wang, Xinyu Wang, Xu Zhang, Lin Zhao, Yiheng Liu, Kai Zhang, Liheng Yan, Lichao Sun, Jun Liu, Ning Qiang, Bao Ge, Xiaoyan Cai, Shijie Zhao, Xintao Hu, Yixuan Yuan, Gang Li, Shu Zhang, Xin Zhang, Xi Jiang, Tuo Zhang, Dinggang Shen, Quanzheng Li, Wei Liu, Xiang Li, Dajiang Zhu, and Tianming Liu. Holistic evaluation of gpt-4v for biomedical imaging, 2023.

[11] Kathyayini Sivasubramanian. *Dual modality ultrasound-photoacoustic clinical imaging system and its applications*. PhD thesis, 2018.

[12] Shubhi Bansal, Sreekanth Madisetty, Mohammad Zia Ur Rehman, Chandravardhan Singh Raghaw, Gaurav Duggal, Nagendra Kumar, et al. A comprehensive survey of mamba architectures for medical image analysis: Classification, segmentation, restoration and beyond. *arXiv preprint arXiv:2410.02362*, 2024.

[13] Rui Xu, Shu Yang, Yihui Wang, Yu Cai, Bo Du, and Hao Chen. Visual mamba: A survey and new outlooks. *arXiv preprint arXiv:2404.18861*, 2024.

[14] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[15] Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation, 2024.

www.Survey.cn

AI-generated, for reference only.

[16] Feng Wang, Jiahao Wang, Sucheng Ren, Guoyizhe Wei, Jieru Mei, Wei Shao, Yuyin Zhou, Alan Yuille, and Cihang Xie. Mamba-r: Vision mamba also needs registers. *arXiv preprint arXiv:2405.14858*, 2024.

[17] Yunwen Huang, Hongyu Hu, Ying Zhu, and Yi Xu. Breast lesion diagnosis using static images and dynamic video, 2023.

[18] Jianbo Jiao, Yifan Cai, Mohammad Alsharid, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Self-supervised contrastive video-speech representation learning for ultrasound, 2020.

[19] Rui Li, Jingliang Ruan, and Yao Lu. Dual-modality watershed fusion network for thyroid nodule classification of dual-view ceus video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 220–229. Springer, 2024.

[20] Shaoyan Zhang, Edward Z Zhang, Paul C Beard, Adrien E Desjardins, and Richard J Colchester. Dual-modality fibre optic probe for simultaneous ablation and ultrasound imaging. *Communications engineering*, 1(1):20, 2022.

[21] Haohao Qu, Liangbo Ning, Rui An, Wenqi Fan, Tyler Derr, Hui Liu, Xin Xu, and Qing Li. A survey of mamba. *arXiv preprint arXiv:2408.01129*, 2024.

[22] Haotian Feng, Emi Yoshida, and Ke Sheng. Multi-modality and temporal analysis of cervical cancer treatment response, 2024.

[23] Mario Coccia. Artificial intelligence technology in oncology: a new technological paradigm, 2019.

[24] Yan Tian, Zhaocheng Xu, Yujun Ma, Weiping Ding, Ruili Wang, Zhihong Gao, Guohua Cheng, Linyang He, and Xuran Zhao. Survey on deep learning in multimodal medical imaging for cancer detection, 2023.

[25] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017.

[26] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022.

[27] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*, 2024.

[28] Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification, 2024.

[29] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[30] Shentong Mo and Paul Pu Liang. Multimed: Massively multimodal and multitask medical understanding, 2024.

[31] Inwoo Seo, Eunkyoung Bae, Joo-Young Jeon, Young-Sang Yoon, and Jiho Cha. The era of foundation models in medical imaging is approaching : A scoping review of the clinical value of large-scale generative ai applications in radiology, 2024.

[32] Manuel Cossio. Augmenting medical imaging: A comprehensive catalogue of 65 techniques for enhanced data analysis, 2023.

[33] S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises, 2021.

www.SurveyX.cn

AI-generated, for reference only.

[34] Xiaoqing Guo, Qianhui Men, and J. Alison Noble. Mmsummary: Multimodal summary generation for fetal ultrasound video, 2024.

[35] Peng-Fei Zhu, Xiao-Feng Zhang, Yu-Xiang Mao, Pu Zhou, Jian-Jun Lin, Long Shi, Xin-Wu Cui, and Ying He. Predicting central lymph node metastasis in papillary thyroid carcinoma using a fusion model of vision transformer and traditional radiomics based on dynamic dual-modality ultrasound. 2024.

[36] Vincent van de Schaft and Ruud J. G. van Sloun. Ultrasound speckle suppression and denoising using mri-derived normalizing flow priors, 2021.

[37] A. P. James and B. V. Dasarathy. Medical image fusion: A survey of the state of the art, 2013.

[38] Negin Ghamsarian, Sebastian Wolf, Martin Zinkernagel, Klaus Schoeffmann, and Raphael Sznitman. Deeppyramid+: Medical image segmentation using pyramid view fusion and deformable pyramid reception, 2023.

[39] Danilo Avola, Luigi Cinque, Alessio Fagioli, Sebastiano Filetti, Giorgio Grani, and Emanuele Rodolà. Multimodal feature fusion and knowledge-driven learning via experts consult for thyroid nodule classification, 2021.

[40] Yin Dai and Yifan Gao. Transmed: Transformers advance multi-modal medical image classification, 2021.

[41] Andrea Agostini, Daphné Chopard, Yang Meng, Norbert Fortin, Babak Shahbaba, Stephan Mandt, Thomas M. Sutter, and Julia E. Vogt. Weakly-supervised multimodal learning on mimic-cxr, 2024.

[42] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics, 2021.

[43] Wenhui Zhang. Medical volume reconstruction techniques, 2018.

[44] Tianyu Huang, Tao Zhou, Weidi Xie, Shuo Wang, Qi Dou, and Yizhe Zhang. Improving segment anything on the fly: Auxiliary online learning and adaptive fusion for medical image segmentation, 2024.

[45] Shun Zou, Zhuo Zhang, Yi Zou, and Guangwei Gao. Microscopic-mamba: Revealing the secrets of microscopic images with just 4m parameters, 2024.

[46] Jianbo Jiao, Richard Droste, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Self-supervised representation learning for ultrasound video, 2020.

[47] Shanshan Song, Jiangyun Li, Jing Wang, Yuanxiu Cai, and Wenkai Dong. Mf2-mvqa: A multi-stage feature fusion method for medical visual question answering, 2022.

[48] Walter H. L. Pinaya, Mark S. Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F. da Costa, Ashay Patel, Hyungjin Chung, Can Zhao, Wei Peng, Zelong Liu, Xueyan Mei, Oeslle Lucena, Jong Chul Ye, Sotirios A. Tsaftaris, Prerna Dogra, Andrew Feng, Marc Modat, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Generative ai for medical imaging: extending the monai framework, 2023.

[49] Kwanyoung Kim, Jaa-Yeon Lee, and Jong Chul Ye. Unicorn: Ultrasound nakagami imaging via score matching and adaptation, 2024.

[50] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. Explainable medical imaging ai needs human-centered design: Guidelines and evidence from a systematic review, 2022.

[51] Yijun Yang, Zhaohu Xing, Lequan Yu, Chunwang Huang, Huazhu Fu, and Lei Zhu. Vivim: a video vision mamba for medical video segmentation, 2024.

[52] Shentong Mo and Yapeng Tian. Scaling diffusion mamba with bidirectional ssms for efficient image and video generation, 2024.

www.SurveyX.cn

AI-generated, for reference only.

[53] Chenyuan Bian, Nan Xia, Xia Yang, Feifei Wang, Fengjiao Wang, Bin Wei, and Qian Dong. Mambaclinix: Hierarchical gated convolution and mamba-based u-net for enhanced 3d medical image segmentation, 2024.

[54] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone, 2024.

[55] Yuxiang Wei, Anees Abrol, Reihaneh Hassanzadeh, and Vince Calhoun. Hierarchical spatio-temporal state-space modeling for fmri analysis, 2024.

[56] Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. Artificial intelligence-based methods for fusion of electronic health records and imaging data, 2022.

[57] Nadja Gruber, Johannes Schwab, Sebastien Court, Elke Gizewski, and Markus Haltmeier. Lifting-based variational multiclass segmentation algorithm: design, convergence analysis, and implementation with applications in medical imaging, 2023.

[58] Gurucharan Marthi Krishna Kumar, Aman Chadha, Janine Mendola, and Amir Shmuel. Medvisionllama: Leveraging pre-trained large language model layers to enhance medical image segmentation, 2024.

[59] Yufeng Jiang and Yiqing Shen. M^4oe: A foundation model for medical multimodal image segmentation with mixture of experts, 2024.

[60] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20707–20717, 2022.

[61] Ting Hu, Lizhang Xie, Lei Zhang, Guangjun Li, and Zhang Yi. Deep multimodal neural network based on data-feature fusion for patient-specific quality assurance. *International journal of neural systems*, 32(01):2150055, 2022.

[62] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*, 2024.

www.SurveyX.cn

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

www.SurveyX.cn

AI-generated, for reference only.