AI-Enhanced Virtual Reality in Medicine: A Comprehensive Survey

Yixuan Wu[1,2,3] , **Kaiyuan Hu**[4] , **Danny Z. Chen**[5] and **Jian Wu**[1,2]
[1]School of Public Health, Zhejiang University, Hangzhou 310058, China
[2]State Key Laboratory of Transvascular Implantation Devices of The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China
[3]Institute of Wenzhou, Zhejiang University, Hangzhou 325036, China
[4]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China
[5]Department of Computer Science and Engineering, University of Notre Dame, USA
wyx_chloe@zju.edu.cn, kaiyuanhu@link.cuhk.edu.cn, dchen@nd.edu, wujian2000@zju.edu.cn

Abstract

With the rapid advance of computer graphics and artificial intelligence technologies, the ways we interact with the world have undergone a transformative shift. Virtual Reality (VR) technology, aided by artificial intelligence (AI), has emerged as a dominant interaction media in multiple application areas, thanks to its advantage of providing users with immersive experiences. Among those applications, medicine is considered one of the most promising areas. In this paper, we present a comprehensive examination of the burgeoning field of AI-enhanced VR applications in medical care and services. By introducing a systematic taxonomy, we meticulously classify the pertinent techniques and applications into three well-defined categories based on different phases of medical diagnosis and treatment: Visualization Enhancement, VR-related Medical Data Processing, and VR-assisted Intervention. This categorization enables a structured exploration of the diverse roles that AI-powered VR plays in the medical domain, providing a framework for a more comprehensive understanding and evaluation of these technologies. To our best knowledge, this work is the first systematic survey of AI-powered VR systems in medical settings, laying a foundation for future research in this interdisciplinary domain.

1 Introduction

The intersection of artificial intelligence (AI) and virtual reality (VR) is an emerging frontier in the realm of medical technologies. Recent advances in these fields have given rise to innovative applications that promise to reshape various aspects of medical care. The integration of AI and VR brings forth a unique confluence of data-driven analytics and immersive experiences, making it a critical area of study and application in contemporary healthcare. AI's contribution to VR in medicine ranges from enhancing diagnostic precision to offering new paradigms in patient care. These technologies are not only redefining existing medical procedures but also paving the way for novel treatment methods. The purpose of this survey is to delve into the technical nuances, practice workflows, and diverse application scenarios of AI-powered VR in medical settings, examining their impacts on the efficiency, accuracy, and effectiveness of healthcare services.

To systematically understand and analyze the role of AI in medical VR, this paper introduces a taxonomy categorizing the applications into three primary categories: Visualization Enhancement, VR-related Medical Data Processing, and VR-assisted Intervention (see Figure 1). This classification allows for a comprehensive overview of the current state of the art of AI in VR applications tailored to medical needs, highlighting each category's unique contributions to healthcare. By categorizing these applications based on their functions and utilities in the spectrum of patient care, we aim to provide an extensive examination that facilitates easier comprehension and identification of existing research gaps.

Visualization Enhancement focuses on the augmentation of a user's visual and sensory perception within a virtual space. This enhancement is crucial in complex medical procedures where understanding intricate anatomical structures and spatial relationships is imperative. We delve into the subcategories of virtual object reconstruction and visual enhancement, both pivotal in improving clinical outcomes and professional training. VR-related Medical Data Processing addresses the advanced capabilities of VR systems to analyze and interpret complex medical data. This category discusses how VR, combined with AI, facilitates in-depth structural and lesion analysis, enhances disease diagnosis, and supports the entire stages of surgical procedures. It underscores the transformation from traditional 2D data interpretation to a more dynamic, 3D analytical approach. VR-assisted Intervention demonstrates the practical application of AI-driven VR in real-time and interactive medical scenarios. This category covers the utilization of VR for direct procedural guidance, intra-operative intervention, and fostering interactive collaborations among medical professionals. It exemplifies the potential of VR to not only assist but also augment medical practice in live environments.

To our best knowledge, this is the first systematic survey of AI-powered VR technologies in medical settings, laying a foundation for future research in this interdisciplinary domain. The scope of this survey extends beyond mere categorization of current techniques and applications. It also explores the

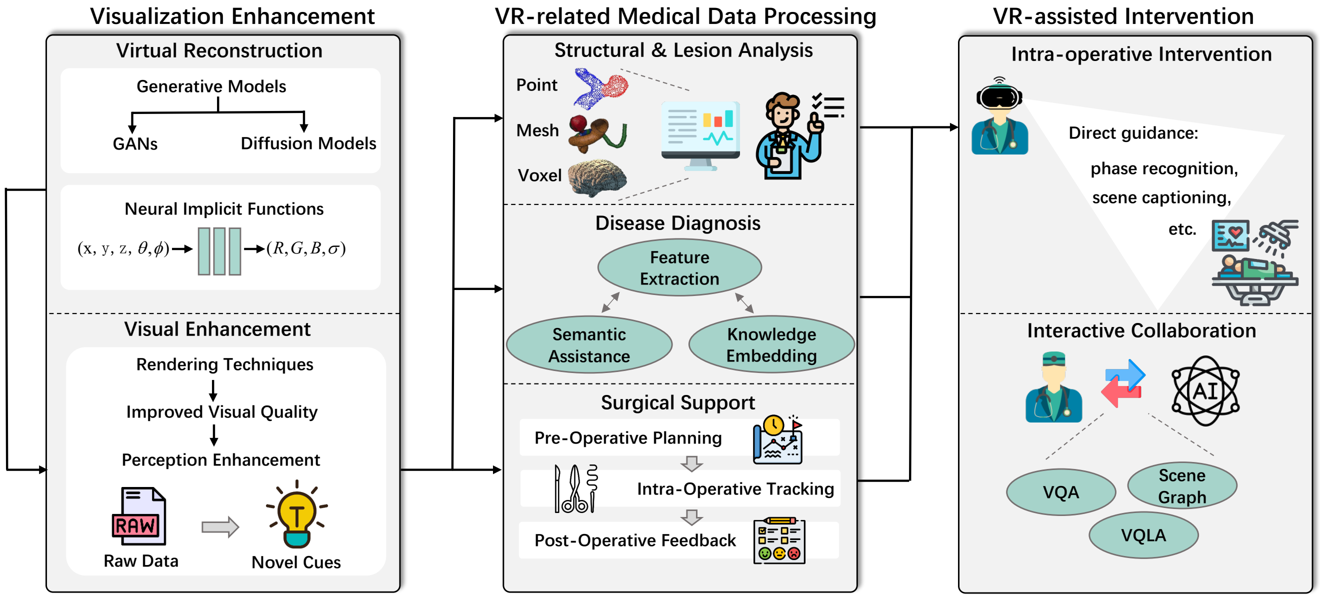


Figure 1: Demonstrating the workflow of AI-enhanced VR-assisted medical services. First, Visualization Enhancement focuses on the augmentation of a user's visual and sensory perception within a virtual space. Next, VR-related Medical Data Processing discusses how VR, combined with AI, facilitates in-depth structural and lesion analysis, enhances disease diagnosis, and supports the entire stages of surgical procedures. Finally, VR-assisted Intervention covers the utilization of VR for direct procedural guidance, intra-operative intervention, and fostering interactive collaborations among medical professionals.

potential future trajectories of AI in VR for medicine, contemplating how ongoing advancements might unfold. The discussion encompasses not only technological aspects but also considers ethical, legal, and practical implications of deploying such advanced systems in sensitive medical environments.

2 Taxonomy

The objective of the taxonomy is to group AI-powered VR applications in medical services with similar goals into the same category, facilitating in-depth investigation in subsequent studies. We classify the techniques and applications into three distinct categories based on diagnosis and treatment procedures: *Visualization Enhancement*, *VR-related Medical Data Processing*, and *VR-assisted Intervention*. A visual presentation of the taxonomy is shown in Figure 2.

2.1 Visualization Enhancement

The rapid advance of AI technologies offers VR platforms the capability of enhancing medical professionals' visual perception during treatment and surgical stages. The techniques of visualization enhancement can be divided into the following two subcategories.

Virtual Reconstruction enables medical professionals to visualize medical data in a more intuitive manner, enhancing their understanding and interpretation of complex anatomical structures. By reconstructing virtual objects (e.g., organs or anatomical models) in a VR environment, professionals can gain a clearer and more comprehensive view of the patient's condition, aiding treatment planning and decision-making.

Visual Enhancement focuses on improving the visual perception of professionals during the training or surgical phase. By immersing the professionals in augmented or virtual surgical scenes, such techniques create an immersive visual perception that enables them to perform treatment procedures with greater precision, resulting in improved treatment outcomes.

2.2 VR-related Medical Data Processing

Enhanced vision in a virtual environment offers an additional analytical capability for processing medical data. By leveraging the immersive and interactive aspect of virtual reality, professionals gain access to a wealth of additional visual information that surpasses the limitations of traditional 2D data usage in different treatment phases, as follows.

Structural and Lesion Analysis. In VR medical contexts, the utilized data formats such as point cloud, mesh, and voxel facilitate AI-powered systems to conduct more comprehensive analysis of anatomical structures and lesion situations, thereby providing additional cognitive information for accurate diagnosis.

Disease Diagnosis. Comprehensive analysis of VR-based medical data serves as an underlying foundation for disease diagnosis through multiple methods, including semantic segmentation, feature extraction, and knowledge embedding.

Surgical Support. Building upon data analysis and diagnosis, surgical support in VR enhances the precision and effectiveness of surgical procedures. Current progress has covered multiple aspects including pre-operative planning, intra-operative recognition and tracking, and post-operative feedback.

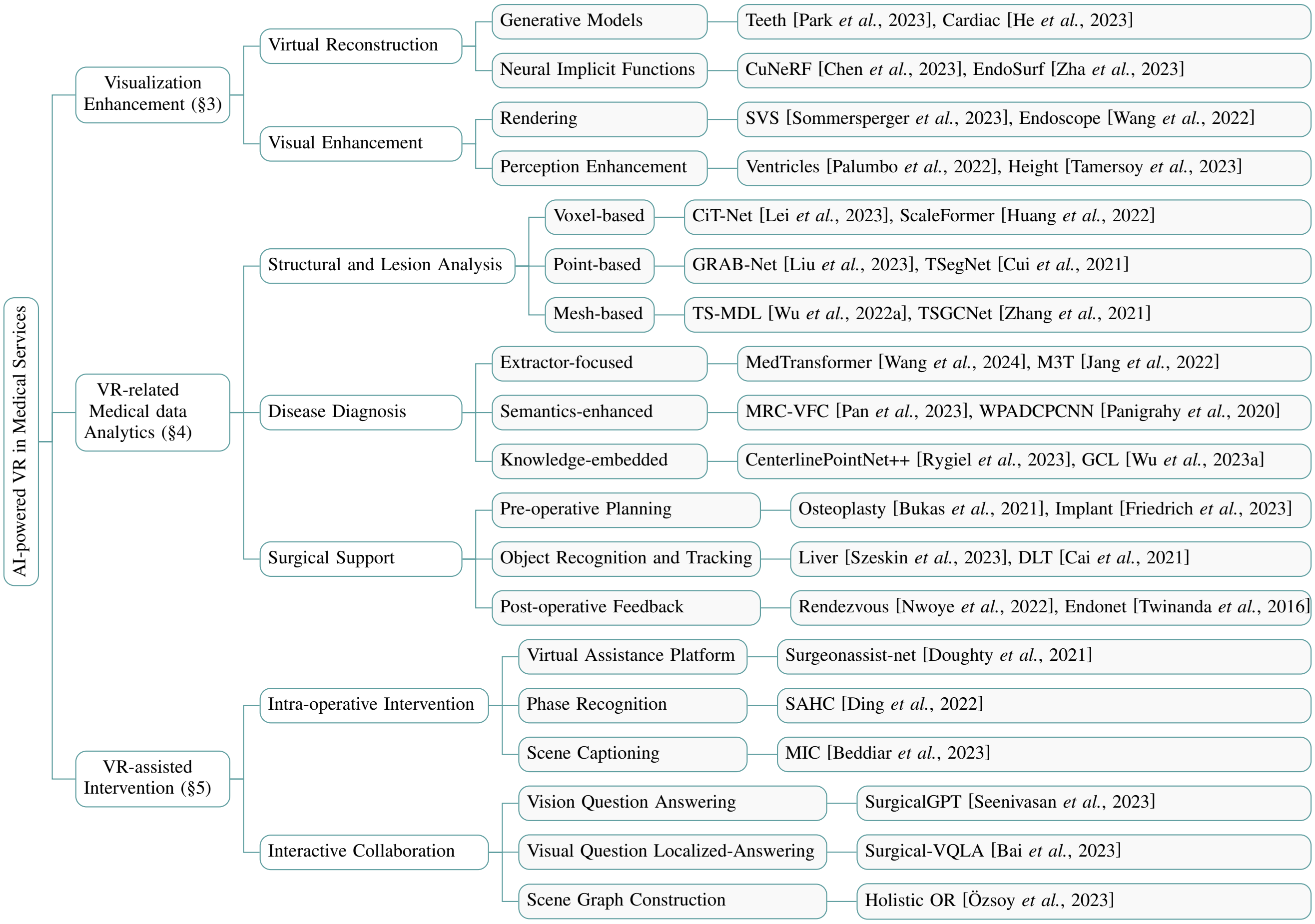

AI-powered VR in Medical Services
Visualization Enhancement (§3)
Virtual Reconstruction
Generative Models
Teeth [Park et al., 2023], Cardiac [He et al., 2023]
Neural Implicit Functions
CuNeRF [Chen et al., 2023], EndoSurf [Zha et al., 2023]
Visual Enhancement
Rendering
SVS [Sommersperger et al., 2023], Endoscope [Wang et al., 2022]
Perception Enhancement
Ventricles [Palumbo et al., 2022], Height [Tamersoy et al., 2023]
VR-related Medical data Analytics (§4)
Structural and Lesion Analysis
Voxel-based
CiT-Net [Lei et al., 2023], ScaleFormer [Huang et al., 2022]
Point-based
GRAB-Net [Liu et al., 2023], TSegNet [Cui et al., 2021]
Mesh-based
TS-MDL [Wu et al., 2022a], TSGCNet [Zhang et al., 2021]
Disease Diagnosis
Extractor-focused
MedTransformer [Wang et al., 2024], M3T [Jang et al., 2022]
Semantics-enhanced
MRC-VFC [Pan et al., 2023], WPADCPCNN [Panigrahy et al., 2020]
Knowledge-embedded
CenterlinePointNet++ [Rygiel et al., 2023], GCL [Wu et al., 2023a]
Surgical Support
Pre-operative Planning
Osteoplasty [Bukas et al., 2021], Implant [Friedrich et al., 2023]
Object Recognition and Tracking
Liver [Szeskin et al., 2023], DLT [Cai et al., 2021]
Post-operative Feedback
Rendezvous [Nwoye et al., 2022], Endonet [Twinanda et al., 2016]
VR-assisted Intervention (§5)
Intra-operative Intervention
Virtual Assistance Platform
Surgeonassist-net [Doughty et al., 2021]
Phase Recognition
SAHC [Ding et al., 2022]
Scene Captioning
MIC [Beddiar et al., 2023]
Interactive Collaboration
Vision Question Answering
SurgicalGPT [Seenivasan et al., 2023]
Visual Question Localized-Answering
Surgical-VQLA [Bai et al., 2023]
Scene Graph Construction
Holistic OR [Özsoy et al., 2023]

Figure 2: A taxonomy of AI-powered VR in medical services with representative examples.

2.3 VR-assisted Intervention

VR-assisted intervention harnesses the power of AI-enhanced visualization and analysis technologies, showcasing its immense potential in augmenting professionals' capabilities during the inspection and surgical phases. This topic explores how VR technologies directly provide guidance in intra-operative interventions and deliver context-aware feedback in an interactive and collaborative manner.

Intra-operative Intervention offers professionals direct guidance and assistance by integrating AI-powered functionalities into VR platforms such as object segmentation, phase recognition, etc. This encompasses the integration of diverse input data formats and platforms.

Interactive Collaboration provides interactive feedback to professionals by leveraging advancements in human-machine collaboration. Techniques such as Vision Question Answering (VQA) and Visual Question Localized-Answering (VQLA) have played a key role in this progress.

3 Visualization Enhancement

Enhanced visualization capability is one of the key advantages of VR technologies, which also shows great potential in the enhancement of medical services. In this section, we cover two key aspects in the application of visualization enhancement: *virtual reconstruction* and *visual enhancement*.

3.1 Virtual Reconstruction

VR technologies offer the capability to provide users with an immersive experience with infinite viewpoints, enabling them to visualize an object freely. This feature shows great potential in medical services, which enables professionals to visualize complex medical data in a more intuitive manner. Traditional medical data acquisition methods like X-ray and ultrasound mostly provide only flattened 2D images which limit the effectiveness of diagnosis. In contrast, AI-powered reconstruction methods provide solutions for reconstructing complete objects with high quality, which can be visualized with VR devices to provide professionals with a better perception of the target objects. Compared with traditional 2D monitoring methods, 3D reconstruction offers notable advantages since it allows users to observe a surgical site from any viewpoint, which dramatically benefits downstream medical applications such as surgeon-centered augmented reality [Nicolau *et al.*, 2011] and virtual reality [Chong *et al.*, 2022]. Multiple learning-based reconstruction methodologies have become prominent, with *neural implicit functions* and *generative models* as the two main approaches.

Generative Models. In practical medical inspection scenarios, due to the limitations of safe radiation exposure, visual occlusion, or anisotropic spatial resolution, the reconstruction of high-quality 3D models in medical scenarios has been challenging. Thanks to the capability to synthesize image details with higher fidelity, generative adversarial networks (GANs) [Goodfellow *et al.*, 2020] have been widely adopted to reconstruct high-quality medical images [Kaplan and Zhu, 2019]. In [Lei *et al.*, 2019], a cycle-consistent generative adversarial network (Cycle GAN) model was proposed to estimate diagnostic quality PET images using low count data. Recently, due to improved sample quality and higher log-likelihood scores compared to GANs, diffusion probabilistic models [Ho *et al.*, 2020] have emerged as a compelling alternative. Such methods have been employed in 3D teeth reconstruction [Park *et al.*, 2023] and cardiac volume reconstruction [He *et al.*, 2023]. Likewise, to tackle the problem of inadequate natural medical data for temporal dynamics analysis and disease progression monitoring, diffusion models have been exploited to generate 4D volumetric data. In [Kim and Ye, 2022], a diffusion deformable model (DDM) was proposed to generate intermediate temporal volumes between source and target volumes.

Neural Implicit Functions. In virtual reality, typical methodologies for 3D shape representations encompass point-based [Yang *et al.*, 2020], voxel-based [Lei *et al.*, 2023], and mesh-based [Wu *et al.*, 2022a] techniques. These approaches explicitly utilize individual points, vertices, or faces to delineate 3D structures, which commonly result in substantial data size. Moreover, they necessitate high-quality raw data for precise reconstruction. A notable limitation of these traditional methods, particularly in medical context, is their inability to produce watertight surfaces, leading to significant drawbacks. In contrast, recent advances in neural implicit representation methods have substantially enhanced reconstruction performance. Neural implicit functions represent a scene by learning a mapping from 3D input coordinates to a shape or surface representation using neural networks, typically through a multilayer perceptron (MLP) architecture. These functions capture the scene's characteristics in terms of local densities and associated colors, allowing detailed and flexible modeling of complex shapes and surfaces. In [Chen *et al.*, 2023], an implicit function network was utilized to yield super-resolution medical images at arbitrary scales and free viewpoints in a continuous domain. For deformable tissues that pose a high requirement for watertight reconstruction, EndoSurf [Zha *et al.*, 2023] was proposed, which effectively learns three neural fields to transform 3D points, predict signed distance functions (SDFs), and estimate colors.

3.2 Visual Enhancement

Multiple learning-based methods have been developed to enhance visual perception during the training or surgical phase, improving the skills of surgeons and treatment effect.

Rendering. Recent advances in rendering techniques have significantly enhanced rendering performance, leading to improved visual quality of medical data perception. In Semantic Virtual Shadows (SVS) [Sommersperger *et al.*, 2023], instrument-specific shadows are artificially generated, enabling naturally non-existent but important perceptual cues that are present in microscopic surgeries. On the other hand, in order to deal with the problem of significant topology changes, occlusion, and spare viewpoints, *neural rendering* has been recently developed to address the limitations of traditional 3D rendering techniques. *Neural rendering* involves using neural networks to capture the complex interactions of scene geometry, lighting, and details, enabling the creation of new views from existing scenes. In [Wang *et al.*, 2022], neural rendering is used to enhance 3D modeling from endoscopic videos, even when there is non-rigid deformation and occlusion.

Perception Enhancement. VR devices enable professionals to visualize medical data in a more flexible manner, offering additional cognitive cues during surgical or training procedures. With the aid of AI-powered VR techniques, professionals are able to acquire further information for treatment in virtual environments. For instance, a deep learning-based workflow demonstrated in [Palumbo *et al.*, 2022] supports emergency treatment through automatic segmentation of skin, skull, and ventricles using a 3D U-Net architecture [Çiçek *et al.*, 2016]. The segmented data are then integrated into MR images as 3D point clouds, enhancing the precision of medical interpretations. Besides, a novel method described in [Tamersoy *et al.*, 2023] employs deep learning for estimating a patient's height and weight from depth images alone, using an end-to-end single-value regression approach. Furthermore, Electrocardio Panorama [Chen *et al.*, 2021] incorporates vectorcardiography techniques for enhanced visualization, enabling the synthesis of ECG signals from any viewpoint by modeling an underlying electrocardio field. To provide complementary information for image-guided therapies, *medical image registration* is essential. This process aligns multiple images, volumes, or surfaces within a common coordinate system to identify the common areas. The study in [Chen *et al.*, 2022] introduces a discriminator-free translation network to facilitate the training of the registration network.

4 VR Medical Data Analytics

Enhanced vision in virtual space offers a multitude of visual cues for both human and computer perception, thereby enhancing the analytical capability of medical data. By exploiting the immersive and interactive nature of VR, medical professionals can access a wealth of additional visual information that goes beyond traditional 2D displays.

4.1 Structural and Lesion Analysis

In medical context, the application of VR predominantly utilizes three principal data formats: voxel-based, point-based, and mesh-based. These formats facilitate more sophisticated and nuanced visualization of anatomical structures and disease lesions, which is pivotal for a comprehensive analysis in clinical practice. We provide comparisons and examples of principle data formats in Table 1.

Point-based Format. A point cloud employs a large group of individual data points in space to represent 3D objects. Each point contains spatial coordinates and additional attributes (e.g., color in RGB, transparency). This unique characteristic allows straightforward and direct applications of downstream

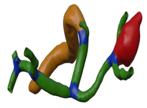
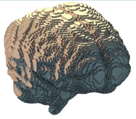
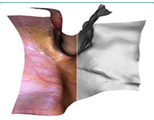
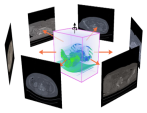
Representation	Size	Visual Quality	Computing Resource	Editability	Visualization
Point Cloud [Yang *et al.*, 2020]	Large	Low	Low	Easy	
Mesh [Wu *et al.*, 2022a]	Medium	Medium	Medium	Medium	
Voxel [Lei *et al.*, 2023]	Medium	Low	Low	Easy	
Implicit Surfaces [Zha *et al.*, 2023]	Medium	Medium	Medium	Hard	
NeRF [Chen *et al.*, 2023]	Medium	Very high	Very high	Hard	

Table 1: Comparisons and visualization of different 3D shape representations.

tasks, including object detection and segmentation. However, the use of point clouds also comes with the drawback of high computational resource consumption due to their large data sizes. In [Liu *et al.*, 2023; Cui *et al.*, 2021], operations are directly conducted on point clouds, without requiring further manipulations of the input point cloud data. IntrA [Yang *et al.*, 2020] and 3DTeethSeg [Ben-Hamadou *et al.*, 2023] are two of the most mainstream point cloud datasets in medical scenarios, and a series of methods in both medical scenarios [Liu *et al.*, 2023; Cui *et al.*, 2021] and general 3D vision [Wu *et al.*, 2019; Li *et al.*, 2018] were implemented on such datasets, showing the transfer-ability of 3D segmentation methods from the general domain to the medical domain. There are also some works [Zhang *et al.*, 2023] that combine different 3D formats to jointly promote the understanding and analysis of anatomical structures. In [Zhang *et al.*, 2023], a point-voxel fusion framework was presented for accurately segmenting the liver into anatomical segments, addressing the challenge of no intensity contrast between adjacent segments by incorporating vessel structure prior.

Mesh-based Format. A polygon mesh consists of vertices, edges, and faces that define the shape of a polyhedral object. The 3D mesh format is a collection of meshes that represent the spatial surface, color, and texture of the object. Compared with point clouds, the mesh is suitable for representing complex geometry by a combination of smooth surfaces with a smaller data size. However, the complicated composition of the mesh introduces higher processing costs. Mesh-based methods [Wu *et al.*, 2022a; Zhang *et al.*, 2021; Lian *et al.*, 2020] take triangle meshes as input and produce corresponding labels for each mesh. In [Lian *et al.*, 2020], cascaded graph-constrained learning modules were introduced for extracting multi-scale local contextual features for automated tooth labeling on raw dental meshes. In [Zhang *et al.*, 2021], a two-stream graph convolutional network was proposed to learn multi-view geometric information from different geometric attributes. In [Wu *et al.*, 2022a], a two-stage framework was designed for concurrent segmentation of meshes and regression of anatomical landmarks.

Voxel-based Format. Compared with pixels in 2D images, voxels are the most analogous representation in 3D space with a smaller domain gap. Therefore, many 2D methods for medical data processing, by taking into account the additional spatial dimension, can be adapted to encode 3D features and capture the relationships in three-dimensional space. In [Lei *et al.*, 2023; Huang *et al.*, 2022; Dong *et al.*, 2022], voxel-based methods operate on volumetric data represented as a grid of 3D pixels, known as voxels. CiT-Net [Lei *et al.*, 2023] was designed as a hybrid architecture of 3D convolutional neural networks (CNNs) hand in hand with 3D Vision Transformers for volumetric medical image segmentation. Furthermore, 2D imaging [Salari *et al.*, 2023], projections [Wu *et al.*, 2022b], and depth maps [Dima *et al.*, 2023] are frequently employed to facilitate the comprehension and interpretation of voxel-based data. In [Dima *et al.*, 2023], 2D labels were mapped to the 3D space using depth information to bridge the gap between 3D supervision and 2D supervision for 3D arterial segmentation. In [Salari *et al.*, 2023], real-time 2D imaging was leveraged to assist high-precision 3D voxel-based visual landmark detection.

4.2 Disease Diagnosis

Recently, based on enhanced visual data, various methods for disease diagnosis have been developed, which can be grouped into three lines according to their design philosophies.

Extractor-focused Methods. This line of methods [Wang *et al.*, 2024; Wu *et al.*, 2023b; Shao *et al.*, 2022; Jang *et al.*, 2022] focuses on designing efficient and effective local feature extractors. Early work was usually based on using MLPs [Liu *et al.*, 2018], 3D CNNs [Çiçek *et al.*, 2016], and GNNs [Zhang *et al.*, 2021]. More recently, inspired by the success of Transformers in the general vision domain, Transformer-based backbones [Wang *et al.*, 2024; Wu *et al.*, 2023b; Shao *et al.*, 2022; Jang *et al.*, 2022] have gradually unified feature extraction for various 3D medical enhanced visual data. For example, D-former [Wu *et al.*, 2023b] introduced a dilated attention to effectively reduce computation for long-dependent feature extraction process of Transformer.

Semantics-enhanced Methods. Semantics-based methods [Pan *et al.*, 2023; Zhao *et al.*, 2023; Xia *et al.*, 2019] are concerned with integrating various intrinsic information sources to enhance the comprehension of the underlying semantics associated with features and target objects. This can involve leveraging multi-view consistency [Pan *et al.*, 2023], multi-task consistency [Zhao *et al.*, 2023], and multi-modality consistency [Xia *et al.*, 2019] to improve the accuracy of diagnostic algorithms. In [Pan *et al.*, 2023], a multi-view relation-aware consistency and virtual feature compensation (MRC-VFC) framework was proposed for long-tailed medical image classification.

Knowledge-embedded Methods. Knowledge-embedded methods [Rygiel *et al.*, 2023; Wu *et al.*, 2023a] aim to incorporate external knowledge and priors to enable networks to conduct diagnosis in a more informative and effective manner. Based on geometric priors, a geometry-based architecture was proposed to utilize implicit geometry embedding for directly estimating hemodynamic features [Rygiel *et al.*, 2023]. Meta labels [Wu *et al.*, 2023a], representing specific attribute information of medical images, serve as valuable cues for networks to enhance the identification and classification of pathological findings. Such incorporation of domain-specific knowledge not only enhances the interpretability of AI-driven diagnostics but also increases the reliability of systems in clinical settings.

4.3 Surgical Support

An important application scenario of AI-powered VR is in surgical support, where it can offer surgeons more precise planning, real-time tracking, and reliable feedback throughout the surgical procedures.

Pre-operative Planning. Pre-operative planning is a crucial phase in surgery, involving detailed patient-specific analysis and strategic procedure mapping to ensure optimal surgical outcomes and minimize risks. VR-enhanced planning in surgery primarily serves two key purposes: (1) facilitating the arrangement of surgical procedures [Bukas *et al.*, 2021], and (2) enabling the simulation of surgical outcomes [Friedrich *et al.*, 2023]. In [Bukas *et al.*, 2021], a personalized automatic high-level framework was presented for planning osteoplasty procedures. In [Friedrich *et al.*, 2023], a novel approach was proposed for implant generation based on a combination of 3D point cloud diffusion models and voxelization networks.

Real-time Object Recognition and Tracking. Real-time object recognition and tracking play a crucial role in surgical procedures, ensuring immediate and accurate identification of surgical instruments and anatomical structures. Based on the categories of the targets, existing works can be broadly classified into two types: lesion-centric and device-centric. Lesion-centric works [Szeskin *et al.*, 2023; Cai *et al.*, 2021] involve accurate measurement and monitoring of pathological changes during surgical procedures. In [Szeskin *et al.*, 2023], a fully automatic end-to-end pipeline was presented for liver lesion change analysis in consecutive (prior and current) abdominal CECT scans of oncology patients. Highlighting unusual lesion labels and lesion change patterns helps radiologists identify possibly overlooked or faintly visible lesions. In contrast, device-centric approaches [Gsaxner *et al.*, 2021] focus on real-time localization and control of surgical instruments. In [Gsaxner *et al.*, 2021], 6-DOF tracking was novelly achieved by purely utilizing on-board stereo cameras of a HoloLens2 to track the same retro-reflective marker spheres used by current optical navigation systems.

Post-operative Feedback. Post-operative feedback is essential in evaluating the success of surgical procedures, guiding patient recovery, and informing future improvements in surgical practice and patient care, where motion analysis and outcome assessment are two key technical aspects. Motion analysis [Nwoye *et al.*, 2022; Twinanda *et al.*, 2016] refers to precisely detecting fine-grained actions of surgeons, with a particular focus on their hand movement. Rendezvous (RDV) [Nwoye *et al.*, 2022] was proposed to recognize surgical action triplets (i.e., <instrument, verb, target>) in endoscopic videos by leveraging attention at the spatial and semantic levels, respectively. POV-Surgery dataset [Wang *et al.*, 2023] is a large-scale, synthetic, egocentric dataset focusing on pose estimation of hands with different surgical gloves and orthopedic surgical instruments. Outcome assessment [Jamzad *et al.*, 2023] involves systematic evaluation of the results and impact of a treatment or intervention, measuring its effectiveness, safety, and patient satisfaction.

5 VR-assisted Intervention

Based on combination of AI-powered visualization and analysis technologies, VR-assisted intervention technologies have significantly enhanced professionals' capabilities during the inspection and surgical phases. This section explores two key applications of VR-assisted intervention: *Intra-operative Intervention* and *Interactive Collaboration*.

5.1 Intra-operative Intervention

Intra-operative intervention focuses on leveraging VR technologies to enhance professionals' abilities during surgery. VR platforms attached with AI algorithms offer additional guidance/assistance during the complex intervention phases. Leveraging the advantages of augmented vision, a virtual assistance platform based on commodity optical see-through head-mounted displays (OST-HMDs) was implemented [Doughty *et al.*, 2021], which provides users with action-and-workflow-driven assistance with near-real-time performance on the Microsoft HoloLens2 OST-HMD. To achieve real-time performance, CNN for spatial feature extraction and RNN for temporal relationship learning were introduced to form the core of the system's analytical capabilities. A similar attempt [Palumbo *et al.*, 2022] was also implemented on the same platform (Microsoft HoloLens2) in the emergency treatment scenario, where 3D U-Net [Çiçek *et al.*, 2016] architecture is exploited for skill, face skin, and ventricle segmentation. These works provide valuable insights into developing real-time surgical guidance systems based on VR platforms. Furthermore, regarding surgical scene understanding and reasoning, which serve as a backbone of VR-assisted intervention, phase recognition [Ding *et al.*, 2022] and scene captioning [Beddiar *et al.*, 2023] techniques play a crucial role, aiding surgeons to obtain direct responses and comprehensive understanding of the current state of surgery. This task encompasses real-time processing and integration of diverse data streams, including visual inputs from the surgical field, patient vitals, and historical medical data.

5.2 Interactive Collaboration

Recent advances in human-machine collaboration have significantly reduced human effort in field operations, thereby fostering widespread implementation of AI-aided interactive collaboration. At the forefront of these advances is Interactive Query. In [Seenivasan *et al.*, 2023], a Vision Question Answering (VQA) system was proposed specifically for medical scenarios, providing a robust and reliable surgical visual question answering solution that can respond to questions by inferring from context-enriched surgical scenes. Beyond that, in [Bai *et al.*, 2023], Visual Question Localized-Answering (VQLA) was further developed that offers more guidance by highlighting the specific areas in the images related to the question and answer. This allows a better understanding of complex medical diagnoses and surgical scenes. Notably, Scene Graph [Holm *et al.*, 2023; Özsoy *et al.*, 2023] is a more holistic, semantically meaningful, and human-readable way to represent surgical videos while encoding all anatomical structures, tools, and their interactions.

Datasets play a critical role in this domain. For instance, in [Sharghi *et al.*, 2020], a dataset was created to capture different robot-assisted interventions, focusing on the phase recognition of the OR scene. The MVOR dataset [Srivastav *et al.*, 2018] contains synchronized multi-view frames from real surgeries with human pose annotations. This dataset facilitates research on human motion analysis and understanding during surgical procedures. Additionally, the 4D-OR dataset [Özsoy *et al.*, 2022] is the first publicly available 4D surgical semantic scene graphs dataset, contributing to advancements in surgical planning and decision-making.

6 Challenges and Future Prospects

6.1 Challenges

Data and Integration Limitations. A primary challenge in AI-driven VR for medical applications lies in the quality and availability of data. High-quality, diverse datasets are essential for training effective AI models. Yet, such data currently are still scarce and fragmented. Additionally, integrating the advanced AI-VR technologies with existing healthcare systems and workflows is a complex task, requiring both technical compatibility and operational adjustments.

Ethical and Legal Issues. Ethical issues, particularly concerning with patient privacy, data security, and informed consent, are crucial. AI-VR systems must comply with stringent healthcare regulations to ensure patient confidentiality and safety. Furthermore, the issue of liability in the cases of AI-VR related errors remains unresolved, complicating the legal landscape for healthcare providers, technology developers, and governments.

Technological and User Acceptance. Achieving realism and accuracy in VR simulations is vital for effective medical training and treatment, yet it remains technically challenging. The 'black box' nature of some AI systems can impede user trust and acceptance, as medical professionals often need to understand the decision-making processes behind AI recommendations. Additionally, designing user-friendly interfaces for a diverse range of users and overcoming technophobia among healthcare professionals and patients is critical for the widespread adoption of AI-driven VR systems.

6.2 Future Prospects

Immersive Medicine through Technological Advancements. Future advancements in AI algorithms enable VR technologies to provide more accurate, efficient solutions in immersive medical services. For example, the integration of Natural Language Processing (NLP) approaches within VR environments will facilitate an enhanced understanding of patient speech and text records. Besides, advanced interactive sonification technologies [Matinfar *et al.*, 2023] can also lead to more comprehensive and detailed diagnostics, thereby enhancing the accuracy and effectiveness of diagnoses and treatments.

Tailored Therapeutic Interventions with AI-Driven Analytics. Utilization of AI algorithms in analyzing patient interactions within VR scenarios enables the customization of therapeutic approaches and options, especially in mental healthcare settings. Furthermore, the integration of biofeedback into VR environments, where artificial intelligence (AI) adapts experiences according to real-time physiological data, signifies a remarkable advancement in tailoring patient care and treatment experience.

AI-Enhanced Real-Time Clinical Analytics. AI-powered VR has the potential to transform telemedicine and remote medical training, especially in underserved areas. By integrating real-time analytics, AI algorithms can offer immediate diagnostic support, improving decision-making in both clinical and non-clinical settings. Additionally, these technologies can enhance healthcare accessibility, quality, and equity by delivering high-quality medical services remotely.

Acknowledgments

This research was partially supported by National Natural Science Foundation of China under grants No. 62176231, No. 62106218, No. 82202984, No. 92259202 and No. 62132017, Zhejiang Key R&D Program of China under grant No. 2023C03053.

Contribution Statement

Yixuan Wu and Kaiyuan Hu contributed equally to this work. Jian Wu and Danny Z. Chen are the corresponding authors.

References

[Bai *et al.*, 2023] Long Bai, Mobarakol Islam, et al. Surgical-VQLA: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. In *ICRA*, 2023.

[Beddiar *et al.*, 2023] Djamila-Romaissa Beddiar, Mourad Oussalah, et al. Automatic captioning for medical imaging (MIC): A rapid review of literature. *Artificial Intelligence Review*, 2023.

[Ben-Hamadou *et al.*, 2023] Achraf Ben-Hamadou, Oussama Smaoui, et al. 3DTeethSeg'22: 3D teeth scan segmentation and labeling challenge. *arXiv*, 2023.

[Bukas *et al.*, 2021] Christina Bukas, Bailiang Jian, et al. Patient-specific virtual spine straightening and vertebra inpainting: an automatic framework for osteoplasty planning. In *MICCAI*, 2021.

[Cai *et al.*, 2021] Jinzheng Cai, Youbao Tang, et al. Deep lesion tracker: Monitoring lesions in 4D longitudinal imaging studies. In *CVPR*, 2021.

[Chen *et al.*, 2021] Jintai Chen, Xiangshang Zheng, et al. Electrocardio panorama: synthesizing new ecg views with self-supervision. 2021.

[Chen *et al.*, 2022] Zekang Chen, Jia Wei, et al. Unsupervised multi-modal medical image registration via discriminator-free image-to-image translation. In *IJCAI*, 2022.

[Chen *et al.*, 2023] Zixuan Chen, Lingxiao Yang, et al. CuNeRF: Cube-based neural radiance field for zero-shot medical image arbitrary-scale super resolution. In *ICCV*, 2023.

[Chong *et al.*, 2022] Nannan Chong, Yazhong Si, et al. Virtual reality application for laparoscope in clinical surgery based on siamese network and census transformation. In *MICCAI*, 2022.

[Çiçek *et al.*, 2016] Özgün Çiçek, Ahmed Abdulkadir, et al. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *MICCAI*, 2016.

[Cui *et al.*, 2021] Zhiming Cui, Changjian Li, et al. TSegNet: An efficient and accurate tooth segmentation network on 3D dental model. *MIA*, 2021.

[Dima *et al.*, 2023] Alina F Dima, Veronika A Zimmer, et al. 3D arterial segmentation via single 2D projections and depth supervision in contrast-enhanced CT images. In *MICCAI*, 2023.

[Ding *et al.*, 2022] Xinpeng Ding, Xiaomeng Li, et al. Exploring segment-level semantics for online phase recognition from surgical videos. *TMI*, 2022.

[Dong *et al.*, 2022] Zhangfu Dong, Yuting He, et al. MNet: Rethinking 2D/3D networks for anisotropic medical image segmentation. In *IJCAI*, 2022.

[Doughty *et al.*, 2021] Mitchell Doughty, Karan Singh, et al. SurgeonAssist-Net: Towards context-aware head-mounted display-based augmented reality for surgical guidance. In *MICCAI*, 2021.

[Friedrich *et al.*, 2023] Paul Friedrich, Julia Wolleb, et al. Point cloud diffusion models for automatic implant generation. In *MICCAI*, 2023.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, et al. Generative adversarial networks. *Commun. ACM*, 2020.

[Gsaxner *et al.*, 2021] Christina Gsaxner, Jianning Li, et al. Inside-out instrument tracking for surgical navigation in augmented reality. In *VRST*, 2021.

[He *et al.*, 2023] Xiaoxiao He, Chaowei Tan, et al. DMCVR: Morphology-guided diffusion model for 3D cardiac volume reconstruction. In *MICCAI*, 2023.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, et al. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[Holm *et al.*, 2023] Felix Holm, Ghazal Ghazaei, et al. Dynamic scene graph representation for surgical video. In *ICCV*, 2023.

[Huang *et al.*, 2022] Huimin Huang, Shiao Xie, et al. ScaleFormer: Revisiting the Transformer-based backbones from a scale-wise perspective for medical image segmentation. In *IJCAI*, 2022.

[Jamzad *et al.*, 2023] Amoon Jamzad, Fahimeh Fooladgar, et al. Bridging ex-vivo training and intra-operative deployment for surgical margin assessment with evidential graph Transformer. In *MICCAI*, 2023.

[Jang *et al.*, 2022] Jinseong Jang, Dosik Hwang, et al. M3T: Three-dimensional medical image classifier using multi-plane and multi-slice Transformer. In *CVPR*, 2022.

[Kaplan and Zhu, 2019] Sydney Kaplan and Yang-Ming Zhu. Full-dose PET image estimation from low-dose PET image using deep learning: A pilot study. *Dig. imaging*, 2019.

[Kim and Ye, 2022] Boah Kim and Jong Chul Ye. Diffusion deformable model for 4D temporal medical image generation. In *MICCAI*, 2022.

[Lei *et al.*, 2019] Yang Lei, Xue Dong, et al. Whole-body PET estimation from low count statistics using cycle-consistent generative adversarial networks. *Physics in Medicine & Biology*, 2019.

[Lei *et al.*, 2023] Tao Lei, Rui Sun, et al. CiT-Net: Convolutional neural networks hand in hand with vision Transformers for medical image segmentation. In *IJCAI*, 2023.

[Li *et al.*, 2018] Yangyan Li, Rui Bu, et al. PointCNN: Convolution on X-transformed points. *NeurIPS*, 2018.

[Lian *et al.*, 2020] Chunfeng Lian, Li Wang, et al. Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3D intraoral scanners. *TMI*, 2020.

[Liu *et al.*, 2018] Bingbin Liu, Michelle Guo, et al. 3D point cloud-based visual prediction of ICU mobility care activities. In *MLHC*, 2018.

[Liu *et al.*, 2023] Yifan Liu, Wuyang Li, et al. GRAB-Net: Graph-based boundary-aware network for medical point cloud segmentation. *TMI*, 2023.

[Matinfar *et al.*, 2023] Sasan Matinfar, Mehrdad Salehi, et al. From tissue to sound: Model-based sonification of medical imaging. In *MICCAI*, 2023.

[Nicolau *et al.*, 2011] Stéphane Nicolau, Luc Soler, et al. Augmented reality in laparoscopic surgical oncology. *Surgical Oncology*, 2011.

[Nwoye *et al.*, 2022] Chinedu Innocent Nwoye, Tong Yu, et al. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *MIA*, 2022.

[Özsoy *et al.*, 2022] Ege Özsoy, Evin Pınar Örnek, et al. 4D-OR: Semantic scene graphs for or domain modeling. In *MICCAI*, 2022.

[Özsoy *et al.*, 2023] Ege Özsoy, Tobias Czempiel, et al. Holistic OR domain modeling: A semantic scene graph approach. *International Journal of Computer Assisted Radiology and Surgery*, 2023.

[Palumbo *et al.*, 2022] Maria Chiara Palumbo, Simone Saitta, et al. Mixed reality and deep learning for external ventricular drainage placement: A fast and automatic workflow for emergency treatments. In *MICCAI*, 2022.

[Pan *et al.*, 2023] Li Pan, Yupei Zhang, et al. Combat long-tails in medical classification with relation-aware consistency and virtual features compensation. In *MICCAI*, 2023.

[Panigrahy *et al.*, 2020] Chinmaya Panigrahy, Ayan Seal, et al. MRI and SPECT image fusion using a weighted parameter adaptive dual channel PCNN. *IEEE SPL*, 2020.

[Park *et al.*, 2023] Sihwa Park, Seongjun Kim, et al. 3D teeth reconstruction from panoramic radiographs using neural implicit functions. In *MICCAI*, 2023.

[Rygiel *et al.*, 2023] Patryk Rygiel, Paweł Płuszka, et al. CenterlinePointNet++: A new point cloud based architecture for coronary artery pressure drop and vFFR estimation. In *MICCAI*, 2023.

[Salari *et al.*, 2023] Soorena Salari, Amirhossein Rasoulian, et al. Towards multi-modal anatomical landmark detection for ultrasound-guided brain tumor resection with contrastive learning. In *MICCAI*, 2023.

[Seenivasan *et al.*, 2023] Lalithkumar Seenivasan, Mobarakol Islam, et al. SurgicalGPT: End-to-end language-vision GPT for visual question answering in surgery. *arXiv*, 2023.

[Shao *et al.*, 2022] Di Shao, Xuequan Lu, et al. 3D intracranial aneurysm classification and segmentation via unsupervised dual-branch learning. *JBHI*, 2022.

[Sharghi *et al.*, 2020] Aidean Sharghi, Helene Haugerud, et al. Automatic operating room surgical activity recognition for robot-assisted surgery. In *MICCAI*, 2020.

[Sommersperger *et al.*, 2023] Michael Sommersperger, Shervin Dehghani, et al. Semantic virtual shadows (SVS) for improved perception in 4D OCT guided surgery. In *MICCAI*, 2023.

[Srivastav *et al.*, 2018] Vinkle Srivastav, Thibaut Issenhuth, et al. MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation. *arXiv*, 2018.

[Szeskin *et al.*, 2023] Adi Szeskin, Shalom Rochman, et al. Liver lesion changes analysis in longitudinal CECT scans by simultaneous deep learning voxel classification with SimU-Net. *MIA*, 2023.

[Tamersoy *et al.*, 2023] Birgi Tamersoy, Felix Alexandru Pîrvan, et al. Accurate and robust patient height and weight estimation in clinical imaging using a depth camera. In *MICCAI*, 2023.

[Twinanda *et al.*, 2016] Andru P Twinanda, Sherif Shehata, et al. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *TMI*, 2016.

[Wang *et al.*, 2022] Yuehao Wang, Yonghao Long, et al. Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. In *MICCAI*, 2022.

[Wang *et al.*, 2023] Rui Wang, Sophokles Ktistakis, et al. POV-Surgery: A dataset for egocentric hand and tool pose estimation during surgical activities. In *MICCAI*, 2023.

[Wang *et al.*, 2024] Yifeng Wang, Ke Chen, et al. MedTransformer: Accurate AD diagnosis for 3D MRI images through 2D Vision Transformers. *arXiv*, 2024.

[Wu *et al.*, 2019] Wenxuan Wu, Zhongang Qi, et al. PointConv: Deep convolutional networks on 3D point clouds. In *CVPR*, 2019.

[Wu *et al.*, 2022a] Tai-Hsien Wu, Chunfeng Lian, et al. Two-stage mesh deep learning for automated tooth segmentation and landmark localization on 3D intraoral scans. *TMI*, 2022.

[Wu *et al.*, 2022b] Yixuan Wu, Bo Zheng, et al. Self-learning and one-shot learning based single-slice annotation for 3D medical image segmentation. In *MICCAI*, 2022.

[Wu *et al.*, 2023a] Yixuan Wu, Jintai Chen, et al. GCL: Gradient-guided contrastive learning for medical image segmentation with multi-perspective meta labels. In *ACM Multimedia*, 2023.

[Wu *et al.*, 2023b] Yixuan Wu, Kuanlun Liao, et al. D-former: A U-shaped dilated transformer for 3D medical image segmentation. *Neural Computing and Applications*, 2023.

[Xia *et al.*, 2019] Kai-jian Xia, Hong-sheng Yin, et al. A novel improved deep convolutional neural network model for medical image fusion. *Cluster Computing*, 2019.

[Yang *et al.*, 2020] Xi Yang, Ding Xia, et al. IntrA: 3D intracranial aneurysm dataset for deep learning. In *CVPR*, 2020.

[Zha *et al.*, 2023] Ruyi Zha, Xuelian Cheng, et al. EndoSurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In *MICCAI*, 2023.

[Zhang *et al.*, 2021] Lingming Zhang, Yue Zhao, et al. TS-GCNet: Discriminative geometric feature learning with two-stream graph convolutional network for 3D dental model segmentation. In *CVPR*, 2021.

[Zhang *et al.*, 2023] Xukun Zhang, Yang Liu, et al. Anatomical-aware point-voxel network for Couinaud segmentation in liver CT. In *MICCAI*, 2023.

[Zhao *et al.*, 2023] Yan Zhao, Xiuying Wang, et al. Multi-task deep learning for medical image computing and analysis: A review. *Computers in Biology and Medicine*, 2023.