Strategies for Deep Learning in Volumetric Medical Imaging: A Survey

João Vitor S. de Oliveira, Danilo F. Vieira, Mateus P. da Silva,
Daniel L. Fernandes, Marcos H. F. Ribeiro, Hugo N. Oliveira
Department of Computer Science, Universidade Federal de Viçosa, Viçosa, Brazil
{ joao.silva.oliveira, danilo.f.vieira, mateus.p.silva, daniel.louzada, marcosh.ribeiro, hugo.n.oliveira }@ufv.br

***Abstract*—Deep learning has transformed medical image analysis. However, building effective models for volumetric data, such as CT, MRI, and PET scans, presents a new set of challenges. These include high computational costs, limited annotated datasets, and the need for architectures that can process volumetric data directly. This survey provides a comprehensive overview of recent advances in deep neural network architectures specifically designed for 3D medical imaging. We analyze the progression of architectural choices along the years, highlighting their innovations and applicability. In addition, we review training strategies such as supervised/self-supervised pretraining and the development of general-purpose 3D foundational models. Also, a comprehensive overview of 3D medical imaging datasets and their associated tasks is presented in the supplementary materials, highlighting the diverse clinical objectives that deep learning models can support across healthcare and diagnostic applications. A dedicated section addresses the emerging field of explainability in volumetric contexts, emphasizing the limitations of adapting 2D explainability-focused tools and the importance of 3D-native explanation frameworks. We conclude by outlining the key trends in the field of 3D medical imaging.**

***Index Terms*—3D medical imaging, Deep learning, Self-supervised learning, Explainable AI, Volumetric segmentation**

I. Introduction

The field of Computer Vision (CV) has evolved significantly in recent decades, driven primarily by the growing number of Machine Learning (ML) studies and, more recently, by the widespread adoption of Deep Learning (DL) [1]. In tasks involving 2D images, DL models have achieved human-level performance across several domains, leveraging large datasets and well-optimized architectures [2]. Naturally, these approaches are being extended to 3D domains, which offer the potential to capture richer and more informative volumetric representations [3, 4]. Three-dimensional, or volumetric, data can be obtained directly from medical contexts, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Positron Emission Tomography (PET) scans [2, 5, 6]. The integration of such data with DL opens up new opportunities for advanced models and systems, but it also introduces a distinct set of limitations and challenges [3, 7].

The challenges associated with volumetric data increase significantly compared to 2D data, as these data add a new spatial dimension in each layer processed by the model, resulting in an even greater need for computational resources to process the convolutions [2, 5, 6]. Beyond computational limitations, there is a scarcity of annotated data [7, 8]. This occurs because the process of annotation can only be performed by specialized professionals, such as physicians and radiologists, leading to a higher financial cost, as a single sample can take hours to label across all its slices [2, 4, 7]. Additionally, there is a limitation in the quantity and modality of available data, since volumetric datasets have few samples and diverse modalities (*e.g.* MRIs, CTs, PETs), resulting in variations in contrast, resolution, noise, and semantic content [3, 8].

This survey aims to review recent advances in the application of DL to 3D medical data, with an emphasis on the inherent challenges of this domain and emerging strategies to overcome these limitations. There are multiple other surveys on ML and volumetric medical imaging describing general aspects [2] of this intersection and focusing on specific tasks as segmentation [3, 7, 9], bounding box detection [10] and COVID-19 detection and classification [4]. This research focuses specifically on Deep Neural Network (DNN) models that operate directly on 3D data representations, excluding approaches based solely on 2D slices or intermediate 2.5D strategies. We expand on these previous surveys by providing a discussion on modern methods for volumetric medical imaging analysis based on DL. We also map the main public datasets and tasks in volumetric medical imaging. At the time of this survey, no existing study provides a comprehensive overview encompassing all the tasks and corresponding datasets in this domain. However, a curated collection of several key datasets and their associated tasks is available on this paper's supplementary materials. By consolidating this overview, we aim to provide a comprehensive and up-to-date review that can serve as a basis for researchers and professionals working at the intersection of medical imaging and ML. Supplementary material, including datasets, links and tables of our literature mapping can be seen in this project's webpage[1].

II. Volumetric Medical Tasks

Organ, Tissue and Tumor Segmentation: Image segmentation is a fundamental task that involves analyzing each voxel of an image to assign labels that define regions or boundaries. It is commonly subdivided into two categories: semantic and instance segmentation. Semantic segmentation assigns a class

The authors would like to thank CAPES for their financial support for this research.

[1]https://github.com/jvoliveira/Volumetric-Medical-Imaging-2025

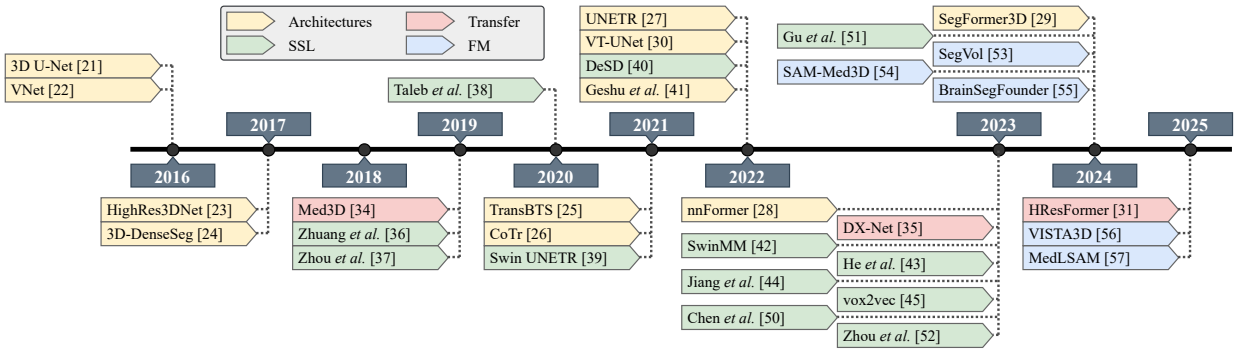

Fig. 1. Timeline for volumetric models, including architectures and Transfer Learning, Self-Supervised Learning and Foundational Model strategies.

to every voxel, generating a class map that distinguishes, *e.g.*, healthy tissue from lesions. Instance segmentation, on the other hand, identifies and separates individual objects within the same class. These tasks can be applied to various fields within medicine, *e.g.*, brain tumor segmentation, organ volumetry, and the detection of internal lesions [2]. Such segmentations are frequently performed on 3D data, such as CT and MRI scans, and support more accurate clinical decisions, from disease diagnosis and monitoring to surgical planning [7].

Object Detection: Object Detection is a CV task focused on identifying and localizing objects of interest within an image, typically by drawing rectangular bounding boxes (bboxes) around them. While it may seem similar to instance segmentation, the two differ in how they delineate objects. Object detection uses bboxes, whereas instance segmentation provides voxel-level outlines. In medical imaging, this ability to accurately locate and identify anatomical or pathological structures is essential [11, 12], serving as a foundation for various diagnostic and interventional procedures.

Diagnosis: The diagnostic task in medical imaging consists of classifying a region by predicting the presence or absence of a disease or condition [2, 4]. In this context, instead of segmenting or locating a particular structure, the goal is to infer a label that represents the pathology or clinical state in question. Unlike segmentation or object detection, classification-based diagnosis does not rely on voxel-wise labels or bboxes. However, it presents greater interpretability challenges, as models must extract discriminative features [13] from potentially small regions within the entire volume, without explicit guidance on where the disease or condition is located. Sec. IV covers the explainability challenges related to such tasks.

Miscelaneous Tasks: Other complementary tasks hold substantial clinical relevance and occur at various stages of the medical workflow. Image registration [6], *e.g.*, is critical in 3D imaging, enabling the alignment of volumes acquired at different time points or through different modalities. This process is particularly important in MRI [14], where accurate alignment is essential for non-invasive comparisons of longitudinal scans. Surgical planning [15] is another key task, where volumetric data are used for surgical navigation or simulation. Finally, longitudinal analysis [16] of sequential scans supports the monitoring of disease progression over time.

III. Learning from 3D Data

CV for natural images has advanced in an accelerating pace since the resurgence of Convolutional Neural Networks (CNNs) in the 2010's [1]. Soon after the proposal of traditional CNNs for object recognition, these networks were adapted for segmentation and object detection tasks. Another major development happened when the attention mechanism was proposed [17]. In the early 2020's Vision Transformers (ViTs) [18], CV networks fully based on the attention mechanism were proposed, also being quickly adapted to segmentation and detection.

DL for 2D medical applications followed similar trends to the traditional models for CV. CNNs were repurposed to classify medical images in 2D scenarios [19], while UNets [20] were proposed initially to segment medical images and only afterwards gained popularity in the CV literature for natural images. Detection networks also did not take long to be applied to medical images [12]. RGB-based models were adapted for 2D medical imaging, as this domain is closer to natural images than 3D data, which requires more elaborate strategies.

All of these modern learning strategies have major drawbacks when applied to 3D medical tasks (Sec. II): 1) lack of very large annotated medical datasets; 2) focus on RGB images; 3) architectures hardcoded for 2D images; 4) prohibitive memory costs of adapting the architectures to 3D data. Therefore, distinct customized approaches specifically designed for 3D medical images were proposed to deal with these incompatibilities. Such strategies will be further discussed in Sec. III-A and III-B.

A. 3D Neural Network Architectures

Due to the inherently more computationally expensive nature of volumetric images, Neural Networks for 3D medical data took longer to be developed. The first popular efforts to train DNNs in volumetric radiology were convolutional architectures designed to be more parameter efficient than their 2D counterparts aiming for a smaller memory footprint. 3D UNets [21] and V-Nets [22], for instance, are highly inspired

in 2D UNets [20], both being encoder-decoders with a large amount of skip connections and trainable upsampling via transposed convolutions. Striding for better memory efficiency, these architectures have less encoder/decoder symmetric pairs, considerably less channels per convolutional layer and are designed to receive smaller input volumes.Concurrently to these encoder-decoder architectures, FCN-like architectures were also developed for 3D medical image segmentation, as in HighRes3DNet [23] and 3D-DenseSeg [24]. These networks adapt traditional 2D architectures receiving smaller 3D input volumes in order to decrease memory usage, while still preserving representation capacity.

Despite the large computational requirements of the attention operation [17], hybrid models based on trainable convolutions paired with attention mechanisms were also applied to 3D medical imaging recognition [25, 26]. Multiple convolutional strategies started using the attention mechanism in clever ways, for instance by placing a fully fledged Transformer encoder between a convolutional Encoder-Decoder, similarly to CoTr [26] and TransBTS [25]. Lately, volumetric DNN architectures have been proposed that are based mainly, or even solely, on the attention operation [27–30], incorporating novel designs to mitigate the high computational costs compared to convolutions. Such as, VT-UNet [30], the only full Transformer identified in our review, employs a 3D patch merging scheme on the encoder, instead of the traditional downsampling used in UNet-like architectures, reducing complexity via an information bottleneck. In parallel, other strategies merge Transformer encoders with convolutional decoders, such as in UNETR [27] and SegFormer3D [29]. While the former includes a convolutional UNet-like decoder, the latter is closer to an FCN architecture, using interpolation for upsampling. nnFormer [28] offers a different approach, interleaving self-attention with convolutional blocks for spatial downsampling.

The earlier mentioned segmentation methods are trained using either the Cross-Entropy (CE) loss (*e.g.* 3D U-Net [21] and 3D-DenseSeg [24]) and/or the Dice loss (*e.g.* VNet [22], HighRes3DNet [23] and TransBTS [25]). All of the most recent architectures (*e.g.* CoTr [26], UNETR [27], VT-UNet [30], nnFormer [28], SegFormer3D [29] and HResFormer [31]) use a weighted or non-weighted combination of Dice+CE.

In addition to segmentation tasks, there are 3D Object Detection (3DOD) DNNs adapted from the traditional two-stage Faster R-CNN [32] and single-stage YOLO [33] architectures. Instead of delivering a voxel-wise classification as in segmentation methods, these models aim to produce a set of volumetric bboxes encoded by 6 points that define the object's location in 3D space. However, there are also a myriad of works that treat the 3DOD task in medical images as a coarse segmentation task [10]. For more information on 3DOD, we refer the readers to the survey conducted by Kern and Mastmeyer [10].

In summary, the evolution of architectures for learning 3D medical images reflects a continuous movement of adaptation and innovation: from the first volumetric versions of CNNs, designed to be lighter and more memory-efficient, through hybrid models that combine convolutions and attention, to recent Transformer architectures. Each advancement has sought to balance representation power and computational cost, highlighting the persistent challenge of fully exploiting the richness of volumetric data without compromising the practical feasibility of training and clinical application.

B. Training Strategies for Volumetric Learning

Another branch of strategies for automated volumetric data analysis using DNNs regards the generalizability of such models for novel domains or even Out-of-Distribution (OOD) data from the same domains. Our research highlights three main branches: 1) supervised pretraining and multi-task learning; 2) Self-Supervised Learning (SSL); and 3) volumetric Foundational Models (FMs). These strategies will be detailed in the following paragraphs.

Regarding supervised knowledge transfer via multi-task learning, the Med3D [34] network architecture employs a multi-task learning scheme that aims to encourage cross-modality transfer learning in volumetric images. This FCN-like strategy uses a relatively standard 3D version of ResNet coupled with spatial pyramid pooling and multiple segmentation heads that focus on distinct tasks trained conjointly. This supervised pretraining strategy yields considerable accuracy gains ranging from 3% to 20% in OOD target tasks. There are also multiple efforts that try to perform supervised knowledge transfer in between 2D and 3D medical data [31, 35].

SSL has also played a crucial role in volumetric image recognition in recent years [36–44], mainly with the advent of large unlabeled 3D datasets (see Sec. II). All variants of SSL can be found in 3D medical imaging, starting in older transformation prediction pretext tasks (*e.g.* solving visual puzzles [36], inpainting [37] and rotation prediction [38]). After major developments in the early 2020's on similarity-based self-supervision that enforced transformation invariance by encouraging latent representations to be similar across distinct views of the same image, such models were quickly ported to 3D data. Both contrastive [41, 45] and non-contrastive [40] approaches were developed for pretraining on unlabeled 3D data by repurposing methods as DINO [46] or SwAV [47].

Due to the inherently expensive nature of such similarity-based SSL models coming from the necessity of larger batch sizes either for negative pairs or training stability, these approaches as vox2vec [45] and Ghesu *et al.* [41] were very expensive to train, requiring computational resources, *e.g.*, GPU memory, that may not be available in most cases. Masked Image Modeling (MIM) [48, 49] approaches based on reconstructing masked patches also were rapidly ported to volumetric data [50, 51]. One major advantage of purely MIM self-supervision is that there is no lower limit on batch size, so these approaches can be trained with very small batches in order to compensate for the higher computational complexity of dealing with volumes instead of images.

There are also hybrid models for 3D SSL, usually pairing either similarity-based or MIM-based self-supervision with simpler transformation prediction pretraining [39, 42–44, 51,

52]. Some of the hybrid approaches take into account regularities of the human anatomy, using anatomical information to guide patch selection when training for similarity [43, 44] or indirectly by enforcing the preservation of geometric shape information in a complementary loss function [51].

FMs for volumetric images are still relatively scarce in medical imaging, however there are some initial efforts in the developments of such highly generalizable models [53–57]. Some customized FM approaches as SegVol [53] and VISTA3D [56] leverage multiple labeled and unlabeled public 3D image sources for training. As these models aim to be generalizable in order to be able to segment most organs or RoIs in the human body, the datasets are very comprehensive in nature, encompassing CT scans, MRIs and PET scans from the head, abdomen, thorax, pelvis and limbs. BrainSegFounder [55] follows a distinct direction, focusing solely on Brain MRIs and specializing in brain structure segmentation.

Another branch of 3D FMs repurpose the now very famous Segment Anything (SAM) model to volumetric imaging [54, 57]. MedLSAM [57] first uses a 3D localization network to find a 3D bbox and pass the 2D slices inside this bbox to the 2D SAM model, leading to potential incongruencies when aggregating the segmented 2D slices. On the other hand, SAM-Med3D [54] is a fully 3D FM also based on SAM, trained from scratch on 131k volumes, the largest training effort yet in volumetric models. FMs usually yield highly generalizable models that can be used for: 1) zero-shot segmentation of OOD data; 2) fine-tuning for more specialized segmentation; or 3) interactive learning via textual or weakly-supervised prompts.

IV. Explainability and Interpretability

The clinical adoption of 3D DL models in medical imaging holds promise, yet it remains constrained by their black-box nature. This lack of transparency underscores the importance of Explainable AI (XAI) as a foundation for building trust and ensuring accountability in clinical settings [58], aligning with the broader vision of Trustworthy AI (TAI) in healthcare [59].

Current XAI efforts are predominantly based on post-hoc attribution techniques, particularly gradient-based saliency methods, which aim to visualize the regions most influential to a model's prediction [60]. Gradient-weighted Class Activation Mapping (Grad-CAM) [61] and Grad-CAM++ [62], are among the most widely adopted methods for 3D data, such as MRI, CT or PET scans, highlighting clinically relevant structures like the hippocampus in Alzheimer's disease classification [63, 64]. Beyond these, a few saliency methods have been designed specifically for 3D data, including Saliency Tubes [65] and Respond-CAM [66]. Other backpropagation-based approaches, such as Layer-wise Relevance Propagation (LRP) [67], generate voxel-wise relevance maps by decomposing model predictions across network layers.

Complementary to these are perturbation-based techniques [68], which modify or occlude parts of the input to assess their impact on the prediction. Although computationally intensive, these methods include solid model-agnostic tools such as Local Interpretable Model-agnostic Explanations (LIME) [69], which approximates the model locally using an interpretable surrogate, and SHapley Additive exPlanations (SHAP) [70], which leverages cooperative game theoretic principles to assign feature importance values with strong theoretical grounding. Moving beyond low-level visual explanations, concept-based methods such as Testing with Concept Activation Vectors (TCAV) [71] enable interpretation in terms of human-understandable clinical concepts, providing a bridge between AI decisions and medical reasoning.

A growing concern, however, lies in the limitations of these post-hoc methods, particularly their potential lack of faithfulness to the model's true decision-making process. This has led to increasing interest in intrinsically interpretable or Self-eXplainable AI (S-XAI) models, which are transparent by design [72]. Despite these advancements, two core challenges persist in the 3D domain: 1) the inadequacy of naively extending 2D explanation techniques to volumetric data, and 2) the absence of standardized benchmarks and quantitative metrics for assessing the quality and reliability of 3D explanations [73]. To the best of our knowledge, SE3D [73] represents the only benchmark specifically designed to evaluate different approaches and techniques for XAI in 3D data. It integrates adaptations of 2D saliency methods alongside intrinsic 3D techniques and systematically assesses their performance across different datasets.

Addressing these gaps represents a critical direction for future research, aiming to develop 3D-native XAI frameworks that can deliver trustworthy and clinically meaningful explanations in high-stakes medical contexts.

V. Final Remarks

In this work we review the literature on DNNs for volumetric medical imaging. We provide an overview on 3D medical imaging tasks and datasets, architectures and other training strategies and literature on explainability and interpretability on volumetric data. We map some major trends in the medical imaging literature, including:

- Datasets with distinct imaging modalities, annotations and functions;
- Distinct architectural design choices that allow models to learn from volumetric data, ranging from convolutional to attention-based;
- Pretraining using labeled/unlabeled data with the aim of enforcing cross-dataset/cross-modality transfer learning;
- The need of development of 3D-native XAI frameworks that leverage the volumetric nature of medical imaging to produce reliable and clinically meaningful explanations.

The domain of DL for volumetric medical imaging is undergoing rapid progress, primarily propelled by innovations in network architectures. Nevertheless, despite notable progress, substantial challenges persist, including the high computational demands, the limited availability of annotated datasets, and the pressing need for inherently three-dimensional explainability techniques capable of fostering clinical trust.

Future works of this literature review include a benchmark of 3D medical imaging models, as the ones presented in

Sec. III-A and III-B. Our team intends to improve the current work with a thorough investigation of which architectures and supervised/unsupervised pretraining strategies work best for OOD learning in volumetric medical data.

REFERENCES

[1] A. Krizhevsky *et al.*, "ImageNet Classification with Deep Convolutional Neural Networks," *NeurIPS*, vol. 25, 2012.

[2] S. P. Singh *et al.*, "3D Deep Learning on Medical Images: A Review," *Sensors*, vol. 20, no. 18, p. 5097, 2020.

[3] S. Niyas *et al.*, "Medical Image Segmentation with 3D Convolutional Neural Networks: A Survey," *Neurocomputing*, vol. 493, pp. 397–413, 2022.

[4] I. S. Ahmad *et al.*, "Deep Learning Models for CT Image Classification: A Comprehensive Literature Review," *QIMS*, vol. 15, no. 1, p. 962, 2024.

[5] M. Saraei *et al.*, "Deep Learning-Based Medical Object Detection: A Survey," *IEEE Access*, 2025.

[6] J. Chen *et al.*, "A Survey on Deep Learning in Medical Image Registration: New Technologies, Uncertainty, Evaluation Metrics, and Beyond," *MIA*, vol. 100, p. 103385, 2025.

[7] R. Nambiar *et al.*, "Harnessing Medical Imaging Applications: Survey on 3D Deep Learning Models for Image Segmentation," in *CCIS*, pp. 1–6, 2024.

[8] R. Wang *et al.*, "Medical Image Segmentation Using Deep Learning: A Survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, 2022.

[9] Y. He *et al.*, "Deep Learning Based 3D Segmentation: A Survey," *arXiv preprint arXiv:2103.05423*, 2021.

[10] D. Kern *et al.*, "3D Bounding Box Detection in Volumetric Medical Image Data: A Systematic Literature Review," in *ICIEA*, pp. 509–516, 2021.

[11] H. Kasban *et al.*, "A Comparative Study of Medical Imaging Techniques," *Int. J. Intell. Syst.*, vol. 4, no. 2, pp. 37–58, 2015.

[12] M. A. Al-masni *et al.*, "Detection and Classification of the Breast Abnormalities in Digital Mammograms via Regional Convolutional Neural Network," in *EMBC*, pp. 1230–1233, 2017.

[13] L. Cai *et al.*, "A Review of the Application of Deep Learning in Medical Image Classification and Segmentation," *Ann. Transl. Med.*, vol. 8, no. 11, p. 713, 2020.

[14] T. Huang *et al.*, "3D Deformable Convolution for Medical Image Registration," in *PRICAI*, pp. 179–191, 2024.

[15] X. Chen *et al.*, "Artificial Intelligence Driven 3D Reconstruction for Enhanced Lung Surgery Planning," *Nat. Commun.*, vol. 16, no. 1, p. 4086, 2025.

[16] K. E. Link *et al.*, "Longitudinal Deep Neural Networks for Assessing Metastatic Brain Cancer on a Large Open Benchmark," *Nat. Commun.*, vol. 15, no. 1, p. 8170, 2024.

[17] A. Vaswani *et al.*, "Attention Is All You Need," *NeurIPS*, vol. 30, 2017.

[18] A. Dosovitskiy *et al.*, "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint*, 2020.

[19] Q. Li *et al.*, "Medical Image Classification with Convolutional Neural Network," in *ICARCV*, pp. 844–848, 2014.

[20] O. Ronneberger *et al.*, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, pp. 234–241, 2015.

[21] Ö. Çiçek *et al.*, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *MICCAI*, pp. 424–432, 2016.

[22] F. Milletari *et al.*, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *3DV*, pp. 565–571, 2016.

[23] W. Li *et al.*, "On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task," in *IPMI*, pp. 348–360, 2017.

[24] T. D. Bui *et al.*, "3D Densely Convolutional Networks for Volumetric Segmentation," *arXiv preprint*, 2017.

[25] W. Wang *et al.*, "TransBTS: Multimodal Brain Tumor Segmentation Using Transformer," in *MICCAI*, pp. 109–119, 2021.

[26] Y. Xie *et al.*, "CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation," in *MICCAI*, pp. 171–180, 2021.

[27] A. Hatamizadeh *et al.*, "UNETR: Transformers for 3D Medical Image Segmentation," in *WACV*, pp. 574–584, 2022.

[28] H.-Y. Zhou *et al.*, "nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer," *IEEE TIP*, vol. 32, pp. 4036–4045, 2023.

[29] S. Perera *et al.*, "SegFormer3D: An Efficient Transformer for 3D Medical Image Segmentation," in *CVPR*, pp. 4981–4988, 2024.

[30] H. Peiris *et al.*, "A Robust Volumetric Transformer for Accurate 3D Tumor Segmentation," in *MICCAI*, pp. 162–172, 2022.

[31] S. Ren and X. Li, "HResFormer: Hybrid Residual Transformer for Volumetric Medical Image Segmentation," *Trans. Neural Netw. Learn.*, 2025.

[32] K. Kaluva *et al.*, "An Automated Workflow for Lung Nodule Follow-Up Recommendation Using Deep Learning," in *ICIAR*, pp. 369–377, 2020.

[33] J. Sobek *et al.*, "MedYOLO: A Medical Image Object Detection Framework," *JIIM*, vol. 37, no. 6, pp. 3208–3216, 2024.

[34] S. Chen *et al.*, "Med3D: Transfer Learning for 3D Medical Image Analysis," *arXiv preprint*, 2019.

[35] H. Messaoudi *et al.*, "Cross-Dimensional Transfer Learning in Medical Image Segmentation with Deep Learning," *MIA*, vol. 88, p. 102868, 2023.

[36] X. Zhuang *et al.*, "Self-supervised Feature Learning for 3D Medical Images by Playing a Rubik's Cube," in *MICCAI*, pp. 420–428, 2019.

[37] Z. Zhou *et al.*, "Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis," in *MICCAI*, pp. 384–393, Springer, 2019.

[38] A. Taleb *et al.*, "3D Self-Supervised Methods for Medical Imaging," *NeurIPS*, vol. 33, pp. 18158–18172, 2020.

[39] Y. Tang *et al.*, "Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis," in *CVPR*, pp. 20730–20740, 2022.

[40] Y. Ye *et al.*, "DeSD: Self-Supervised Learning with Deep Self-Distillation for 3D Medical Image Segmentation," in *MICCAI*, pp. 545–555, 2022.

[41] F. C. Ghesu *et al.*, "Contrastive Self-Supervised Learning from 100 Million Medical Images with Optional Supervision," *JMI*, vol. 9, no. 6, p. 064503, 2022.

[42] Y. Wang *et al.*, "SwinMM: Masked Multi-view with Swin Transformers for 3D Medical Image Segmentation," in *MICCAI*, pp. 486–496, 2023.

[43] Y. He *et al.*, "Geometric Visual Similarity Learning in 3D Medical Image Self-supervised Pre-training," in *CVPR*, pp. 9538–9547, 2023.

[44] Y. Jiang *et al.*, "Anatomical Invariance Modeling and Semantic Alignment for Self-supervised Learning in 3D Medical Image Analysis," in *ICCV*, pp. 15859–15869, 2023.

[45] M. Goncharov *et al.*, "vox2vec: A Framework for Self-supervised Contrastive Learning of Voxel-level Representations in Medical Images," in *MICCAI*, pp. 605–614, 2023.

[46] M. Caron *et al.*, "Emerging Properties in Self-Supervised Vision Transformers," in *ICCV*, pp. 9650–9660, 2021.

[47] M. Caron *et al.*, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," *NeurIPS*, vol. 33, pp. 9912–9924, 2020.

[48] K. He *et al.*, "Masked Autoencoders are Scalable Vision Learners," in *CVPR*, pp. 16000–16009, 2022.

[49] Z. Xie *et al.*, "SimMIM: A Simple Framework for Masked Image Modeling," in *CVPR*, pp. 9653–9663, 2022.

[50] Z. Chen *et al.*, "Masked Image Modeling Advances 3D Medical Image Analysis," in *WACV*, pp. 1970–1980, 2023.

[51] P. Gu *et al.*, "Self Pre-Training with Topology- and Spatiality-aware Masked Autoencoders for 3D Medical Image Segmentation," *arXiv preprint*, 2024.

[52] L. Zhou *et al.*, "Self Pre-Training with Masked Autoencoders for Medical Image Classification and Segmentation," in *ISBI*, pp. 1–6, 2023.

[53] Y. Du *et al.*, "SegVol: Universal and Interactive Volumetric Medical Image Segmentation," *NeurIPS*, vol. 37, pp. 110746–110783, 2024.

[54] H. Wang *et al.*, "SAM-Med3D: Towards General-purpose Segmentation Models for Volumetric Medical Images," in *ECCV*, pp. 51–67, 2024.

[55] J. Cox *et al.*, "BrainSegFounder: Towards 3D Foundation Models for Neuroimage Segmentation," *MIA*, vol. 97, p. 103301, 2024.

[56] Y. He *et al.*, "VISTA3D: A Unified Segmentation Foundation Model For 3D Medical Imaging," in *CVPR*, pp. 20863–20873, 2025.

[57] W. Lei *et al.*, "MedLSAM: Localize and Segment Anything Model for 3D CT Images," *MIA*, vol. 99, p. 103370, 2025.

[58] D. Bhati *et al.*, "A Survey on Explainable Artificial Intelligence (XAI) Techniques for Visualizing Deep Learning Models in Medical Imaging," *J. Imaging*, vol. 10, no. 10, p. 239, 2024.

[59] N. Hasani *et al.*, "Trustworthy Artificial Intelligence in Medical Imaging," *PET Clin.*, vol. 17, no. 1, p. 1, 2022.

[60] K. Borys *et al.*, "Explainable AI in Medical Imaging: An Overview for Clinical Practitioners - Beyond Saliency-based XAI Approaches," *Eur. J. Radiol.*, vol. 162, p. 110787, 2023.

[61] R. R. Selvaraju *et al.*, "Grad-CAM: Why Did You Say That?," *arXiv preprint*, 2016.

[62] A. Chattopadhay *et al.*, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in *WACV*, pp. 839–847, 2018.

[63] C. Yang *et al.*, "Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification," in *AMIA*, p. 1571, 2018.

[64] T. Mahmud *et al.*, "An Explainable AI Paradigm for Alzheimer's Diagnosis Using Deep Transfer Learning," *Diagnostics*, vol. 14, no. 3, p. 345, 2024.

[65] A. Stergiou, G. Kapidis, G. Kalliatakis, C. Chrysoulas, R. Veltkamp, and R. Poppe, "Saliency Tubes: Visual Explanations for Spatio-Temporal Convolutions," in *ICIP*, pp. 1830–1834, IEEE, 2019.

[66] G. Zhao, B. Zhou, K. Wang, R. Jiang, and M. Xu, "Respond-CAM: Analyzing Deep Models for 3d Imaging Data by Visualizations," in *MICCAI*, pp. 485–492, Springer, 2018.

[67] S. Bach *et al.*, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS One*, vol. 10, no. 7, p. e0130140, 2015.

[68] M. Ivanovs *et al.*, "Perturbation-based Methods for Explaining Deep Neural Networks: A Survey," *Pattern Recog. Lett.*, vol. 150, pp. 228–234, 2021.

[69] M. T. Ribeiro *et al.*, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *SIGKDD*, pp. 1135–1144, 2016.

[70] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *NeurIPS*, vol. 30, 2017.

[71] B. Kim *et al.*, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in *ICML*, pp. 2668–2677, 2018.

[72] J. Hou *et al.*, "Self-eXplainable AI for Medical Image Analysis: A Survey and New Outlooks," *arXiv preprint*, 2024.

[73] M. Wiśniewski *et al.*, "SE3D: A Framework For Saliency Method Evaluation In 3D Imaging," in *ICIP*, pp. 89–95, 2024.