

Lightweight Deep Learning for Processing 3D Medical Images: A Systematic Review

Rahim ullah

University of Glasgow

Hasan Abbas

University of Glasgow

Hazrat Ali

University of Stirling

Syed Waqar Nabi

`syed.nabi@glasgow.ac.uk`

University of Glasgow

Research Article

Keywords: 3D medical imaging, edge deployment, knowledge distillation, lightweight deep learning, medical image segmentation, model compression, pruning, quantization, resource-constrained environments

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10948128/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Lightweight Deep Learning for Processing 3D Medical Images: A Systematic Review

Rahim Ullah¹, Hasan Abbas², Hazrat Ali³, Syed Waqar Nabi^{1*}

^{1*}School of Computing Science, University of Glasgow, Glasgow, Scotland, United Kingdom.

²School of Engineering, University of Glasgow, Glasgow, Scotland, United Kingdom.

³School of Computing, Data, and Mathematical Sciences, University of Stirling, Stirling, Scotland, United Kingdom.

*Corresponding author(s). E-mail(s): syed.nabi@glasgow.ac.uk;
Contributing authors: R.ullah.2@research.gla.ac.uk;
Hasan.Abbas@glasgow.ac.uk; hazrat.ali@live.com;

Abstract

Deep learning for three dimension (3D) modalities such as CT, MRI, and PET offers promise to improve diagnostic accuracy, surgical planning, and visualization of disease progression, as 3D image data provide high resolution volumetric representation. However, 3D deep learning models are computationally heavy and demand powerful hardware that is not always available or affordable, restricting their use in resource-constrained settings such as rural clinics, edge devices, and low- and middle-income countries. Hence, there is a growing interest in building lightweight deep learning models, in general, and for 3D imaging, in particular. This article systematically and critically reviews lightweight deep learning for processing 3D medical images, examining model compression techniques such as knowledge distillation, quantization, pruning and efficient architectural designs, together with the datasets, tools, and deployment environments involved. These techniques substantially reduce parameters, memory, and inference time with modest accuracy loss. We also identify several research gaps in these approaches, such as the accuracy mismatch between lightweight and standard models, difficulty handling small anatomical structures while keeping models compact, limited testing on real resource-constrained hardware, weak integration of clinical information, and the absence of efficiency-aware evaluation metrics. We outline future directions, including robust lightweight designs, purpose-built compact 3D datasets, standardized efficiency metrics, and lightweight multimodal models.

Keywords: 3D medical imaging, edge deployment, knowledge distillation, lightweight deep learning, medical image segmentation, model compression, pruning, quantization, resource-constrained environments.

1 Introduction

Three dimensional (3D) deep learning has substantially advanced medical image analysis, but it comes at a high computational cost. State-of-the-art 3D models depend on high-end GPUs, large annotated datasets, and long training and inference times, which largely confines their deployment to well-equipped hospitals and research centers. Remote and under-resourced settings: rural clinics, mobile health units, and low and middle-income countries often lack such infrastructure. According to the World Economic Forum, nearly five billion people in low and middle-income countries remain excluded from the benefits of medical AI [1]. Combined with high storage demands and long inference times that hinder real-time and portable diagnosis [2], these constraints motivate the central concern of this survey: making advanced 3D medical imaging efficient enough to be deployed in resource-constrained environments.

Medical imaging is a cornerstone of modern diagnostics. Modalities such as X-ray, CT, positron emission tomography (PET), magnetic resonance imaging (MRI) and functional MRI provide rich anatomical and functional information [3], and advances in imaging hardware have improved acquisition speed, spatial resolution, and image quality [4]. Nevertheless, radiological interpretation still relies heavily on human expertise, making it time-consuming, costly, and prone to fatigue and inter-observer variability [5]. As scan volumes continue to grow, automated deep learning based analysis has become essential to assist clinicians and improve both efficiency and accuracy.

While deep learning was first applied to 2D images such as X-rays, the rise of volumetric modalities has driven a transition to 3D deep learning for CT, PET, MRI and fMRI data. 3D architectures such as U-Net and transformer variants exploit spatial and long-range context across all three dimensions, improving the detection and segmentation of complex anatomical structures and pathologies [6]. This capability, however, is computationally expensive: leading 3D models demand multi-gigabyte GPU memory and reach several hundred GFLOPs, as summarised in Table 6, together with large annotated datasets that are difficult to obtain in clinical practice.

The scale of these requirements has broader consequences. Green-AI studies emphasise the high GPU usage, energy consumption, and long training times of large models [7, 8], and Strubell et al. quantified the substantial carbon footprint of training such models on multiple GPUs [9]. This mirrors a broader move toward edge computing in healthcare: portable, point-of-care imaging devices such as handheld ultrasound scanners are increasingly used to deliver diagnosis in remote, mobile, and emergency settings [10]. However, these devices run on low-power processors with limited memory and storage, and cannot support the heavy computational load of full-scale 3D models.

Although several surveys have examined lightweight deep learning and model compression, none focuses specifically on 3D medical imaging under resource constraints.

General overviews of lightweight design [11] and compression methods [12, 13] address 2D or domain-agnostic settings and do not consider the distinct demands of volumetric data, whose memory and computational cost grow rapidly with resolution. Reviews of lightweight convolutional networks [14] concentrate on CNNs and give limited attention to the transformer and hybrid architectures now central to 3D imaging, while more recent surveys remain confined to 2D imaging, whether for semantic segmentation [15] or edge deployment [16]. Table 2 summarises these studies and their limitations. To the best of our knowledge, no existing review consolidates compression and architectural-efficiency strategies specifically for 3D medical imaging in resource-constrained and clinically under-served settings, the gap this review is designed to fill.

This paper presents a systematic review of lightweight deep learning for 3D medical imaging in resource-constrained environments. It examines the motivation for efficient models and the tools and methodologies used to train and deploy them; critically analyses the main compression and acceleration strategies such as pruning, quantization, knowledge distillation, and architectural design; and identifies research gaps and future directions toward energy-efficient, memory-aware, and globally deployable clinical AI. A comprehensive outline of the survey is illustrated in Fig 1.

The remainder of this paper is organized as follows. Section 2 presents the background, including medical imaging modalities, datasets, lightweight deep learning, and resource-constrained environments. Section 3 describes the review methodology. Section 4 presents the core findings, analysing the computational costs of state-of-the-art models and the main efficiency strategies. Section 5 discusses the key challenges and opportunities for lightweight 3D medical imaging, and section 6 outlines future research recommendations. Section 7 presents the threats to validity, and section 8 concludes the paper.

2 Background

Medical imaging is predominantly a volumetric (3D) space. Modalities like MRI, CT, cone-beam CT (CBCT), and PET are inherently three-dimensional, offering far richer anatomical information such as true 3D structure, precise size, and intensity details compared to conventional 2D projections like plain radiography [17], [18].

Deep learning has further strengthened this field, with 3D CNNs, U-Net variants, and transformers achieving strong performance in tasks such as classification and segmentation using volumetric data [19], [17]. Some of the medical imaging applications are shown in Fig. 2, however, these 3D approaches are computationally intensive and memory demanding, limiting their applicability in resource-constrained environments such as rural hospitals, mobile health units, and developing countries where high-end GPUs and large storage are not available [20, 21].

This creates a critical challenge: leveraging the benefits of 3D medical imaging while ensuring accessibility in low-resource settings. Addressing this requires the development of lightweight, efficient, and scalable deep learning solutions.

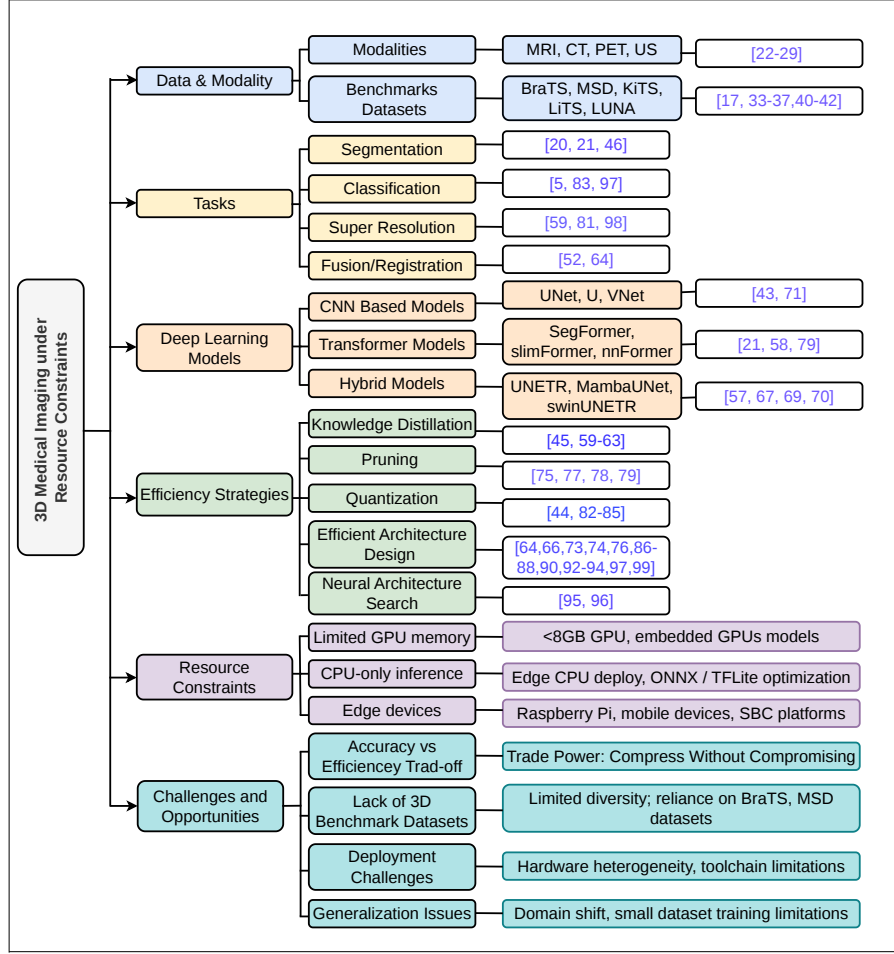


Fig. 1 Comprehensive outline of the survey on 3D medical imaging in resource-constrained environments.

2.1 Why 3D Images

Three-dimensional (3D) medical imaging captures detailed anatomical structures in a continuous volumetric form, providing higher spatial accuracy and detail than conventional 2D imaging [22]. Unlike 2D images, where structures may overlap and obscure spatial relationships, 3D imaging enables precise measurement of distances and accurate estimation of true dimensions such as length, width, and height [23]. Its volumetric nature allows clinicians to view anatomy from multiple angles, isolate structures, and

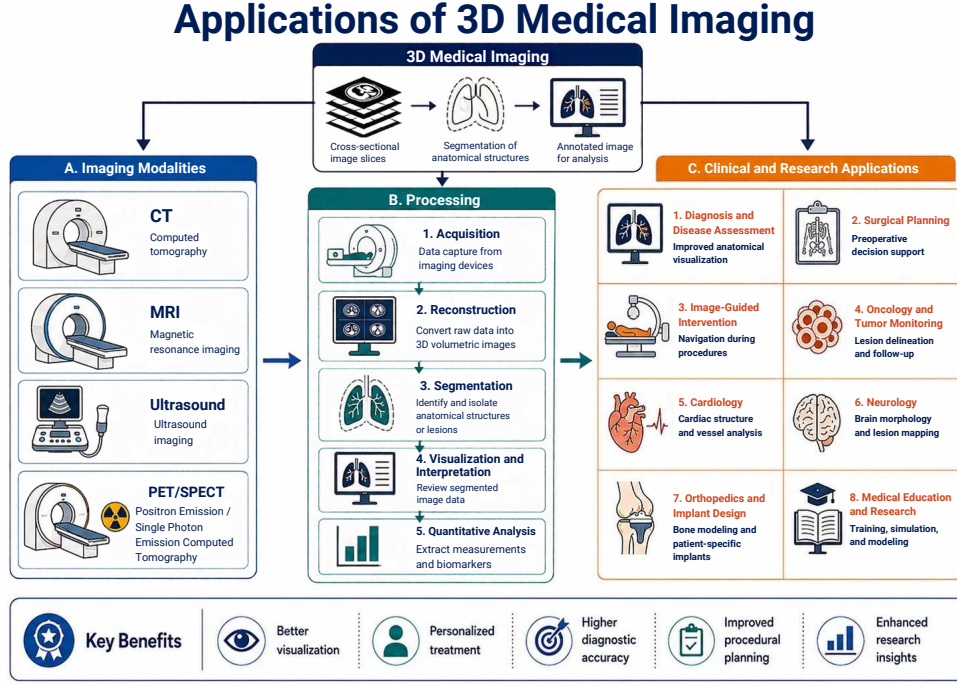


Fig. 2 Key applications of 3D medical imaging. Top row: Overall workflow in 3D medical imaging and segmentation. Panel A: Major 3D imaging modalities. Panel B: Processing and steps. Panel C: Clinical and Research Applications. Bottom row: Key benefits of 3D medical imaging.

perform more accurate quantitative analysis of organs and abnormalities [24]. These capabilities support improved diagnosis, surgical planning, and visualization of disease progression, while also enhancing communication and patient outcomes [23].

Although representing full 3D information on 2D media is challenging, visual comparisons help illustrate the advantages in spatial understanding, as shown in Fig. 3.

2.2 3D Medical Imaging Modalities

Major 3D medical imaging modalities include CT, CBCT, MRI/MRA, ultrasound, angiography, and nuclear imaging techniques such as PET and SPECT, each offering domain-specific advantages [25–28]. CT/CBCT provide high-resolution imaging for bones and dental structures, while MRI/MRA excel in soft-tissue visualization and functional analysis [29]. Hybrid techniques such as PET/CT further enhance diagnostic accuracy by combining anatomical and functional information [30].

Common trade-offs in these procedures include: radiation exposure (CT/CBCT), high cost and long acquisition times (MRI), and limited imaging capability in certain conditions (ultrasound) [31, 32]. Hence, no single modality can be regarded as



Fig. 3 A comparison of 2D and 3D representations of the chest: a conventional 2D X-ray versus a 3D volume rendering of the lungs and airway generated from CT. Both images are reproduced from Wikimedia Commons under their respective licenses.

universally superior, and the choice depends on the specific clinical application and diagnostic requirements.

2.3 3D Medical Imaging Datasets

Although publicly available 3D datasets remain relatively limited, several key benchmarks have significantly advanced the field. The Medical Segmentation Decathlon (MSD) provides multi-organ, multi-modality data for robust segmentation tasks [33], while BraTS is a gold-standard dataset for brain-tumour segmentation using multi-parametric MRI [17]. LIDC-IDRI [34] and LUNA16 [35] are widely used for lung-nodule detection in CT, whereas LiTS [36] and KiTS [37] target liver- and kidney-tumour segmentation, respectively. More recent large-scale benchmarks, such as AMOS [38] for abdominal multi-organ segmentation and TotalSegmentator [39], which covers over 100 anatomical structures in CT, further expand the available 3D data. Other notable datasets include PROMISE12 (prostate), CHAOS (multi-organ), Synapse/BTCV (abdominal CT), and head-and-neck CT datasets [32], [40], [41], [42]. This list is not exhaustive; new 3D datasets continue to be released as the field matures. However, extracting meaningful insights from these datasets requires complex algorithms and high computational resources. With the advancement of deep learning, many applications have been developed to effectively utilize this rich information.

These publicly available datasets are widely used, highly cited, and primarily focused on segmentation tasks, playing a crucial role in advancing deep learning-based 3D medical imaging research. A few commonly used datasets are summarized in Table 1.

Table 1 A few commonly used 3D medical imaging datasets

Dataset	Modality	Target Task	# Cases
MSD [33]	CT, MRI	Multi-organ 3D segmentation	10–500+
BraTS-17-21 [17]	mpMRI	Brain tumor 3D segmentation	200–2000
LIDC-IDRI [34]	CT (Chest)	Nodule detection & segmentation	1018
LUNA16 [35]	CT (Chest)	Nodule detection	888
LiTS [36]	CT (Abdomen)	Liver & tumor 3D segmentation	181
KiTS-19 [37]	CT (Abdomen)	Kidney & tumor 3D segmentation	300
PROMISE12 [40]	MRI (Prostate)	Prostate gland 3D segmentation	50–80
CHAOS [41]	CT, MRI	Multi-organ 3D segmentation	40–60

2.4 Deep Learning in 3D Medical Imaging under Resource Constraints

Several deep learning models, from volumetric CNNs (3D U-Net, V-Net) to self-configuring pipelines like nnU-Net and hybrid transformers & MedSAM3D, achieve good accuracy. However, these models are typically trained on high-performance systems with multi-GB GPUs and large storage. For example, nnU-Net requires around 11 GB GPU memory for stable performance, while lower memory setups risk crashes [43]. Similarly, transformer-based architectures improve global context but at a higher computational cost. As reported in Tables 6 and 7, they involve substantially more parameters and FLOPs (e.g. Swin UNETR at 61.98M parameters and 394.84 G FLOPs) and require on the order of 10–11 hours to train; at inference, Swin UNETR takes about 228.6ms per volume on a GPU and 7.6s on a CPU, roughly 2–2.5× slower than the CNN-based nnU-Net (90.9ms and 3.6s, respectively). Thus, a clear gap exists between performance and resource efficiency.

Recently, efforts have been made to bridge this gap for resource-constrained environments. Post-training quantization (e.g. 8-bit inference) has reduced model size by about 3.9× and latency by 2.7× while maintaining accuracy across models like U-Net, SegResNet, SwinUNETR, and nnU-Net [44]. Knowledge distillation has also been effective; for instance, a 3D U-Net teacher model distilled into a smaller student model improved Dice score from 75% to 84%, while reducing parameters by 75% and inference time by 36% [45]. Efficient architectures such as SegResMamba and MEM-UNet further reduce memory usage to less than half with comparable performance [46], [47].

In short, interest has rapidly increased to develop newer strategies for making 3D models suitable for resource-constrained environments, including memory-efficient designs (e.g. patch-wise inference, 2.5D inputs), model compression techniques (pruning, quantization, distillation), and careful architecture design to reduce parameters while maintaining performance.

2.5 Resource Constrained Environments

In medical imaging, deployment environments are frequently restricted by severe computational and infrastructure constraints. Practical clinical settings, particularly in low and middle income countries (LMICs) often have very limited access to high-end compute, including GPUs. Beyond static computing setups, point-of-care, portable, and edge platforms (e.g. handheld ultrasound units and mobile devices) operate on power constrained ARM processors with limited compute/memory / power and battery capacities [48]. Furthermore, clinical deployment in remote and rural settings is often compounded by unstable power grids and limited network connectivity [49], [10].

These hardware and environmental bottlenecks create a major barrier to adopting computationally intensive 3D deep learning models. Standard 3D networks demand high memory bandwidth and FLOPs, making them unusable for real-time inference on edge devices or standard CPUs. Consequently, there is an urgent need for lightweight and efficient architectures capable of performing key tasks such as 3D segmentation, classification, registration, and anomaly detection directly on low-power hardware [50–52]. Optimizing models for these constrained settings not only enables offline, real-time diagnostic feedback in underserved regions, but also enhances data privacy and lowers energy consumption.

2.6 Related Work

Several surveys have examined lightweight deep learning, but still there is room for exploration; Table 2 summarises their scope and limitations. General overviews of lightweight design and compression by Wang et al. [11] and Liu et al. [12] cover the field broadly but target 2D or domain-agnostic settings and do not address the distinct challenges of 3D volumetric data. Li et al. [13] catalogue compression techniques in a general context without discussing task or modality-specific behaviour, whereas Chen et al. [14] concentrate on lightweight convolutional networks and give limited attention to the transformer and hybrid architectures now common in 3D imaging. The most recent surveys narrow the scope further but remain confined to 2D imaging: Căruntă et al. [15] compare heavy and lightweight models for 2D semantic segmentation only, and Xu et al. [16] review edge deep learning for 2D imaging and general diagnostics. Across these works, three gaps recur: (i) a focus on 2D rather than the computationally demanding 3D volumetric data central to modern diagnostics; (ii) limited attention to deployment in resource constrained clinical settings; and (iii) treatment of individual compression techniques in isolation rather than as a coordinated toolbox. Our review addresses these gaps by consolidating compression and architectural-efficiency strategies specifically for 3D medical imaging in resource-constrained and clinically underserved environments, and by analysing their accuracy, efficiency trade-offs, deployment considerations, and open challenges.

3 Survey Methodology

This survey is carried out using the Systematic Literature Review (SLR) method suggested by Rivera et al. [53].

Table 2 Selected Survey or Review Papers on Lightweight Deep Learning

Reference	Year	Domain	Nature of Study	Contributions	Limitations
Wang et al. [11]	2024	Lightweight deep learning	Overview	Highlighted importance of lightweight deep learning; detailed overview of methods such as distillation, quantization, and pruning.	Did not target 3D domains, which are computationally more challenging.
Liu et al. [12]	2024	Lightweight deep learning	Survey	Comprehensive survey of compression methods including efficient architectures, pruning, quantization, and knowledge distillation for resource-constrained environments.	Focus is general, missing specific 3D data challenges.
Li et al. [13]	2023	Compression Techniques	Survey	Discussed all compression techniques in a general context and further categorized them.	Task-specific issues with image behavior were not discussed.
Chen et al. [14]	2023	Lightweight DCNN	Review	Reviewed lightweight DCNN architectures and model compression; evaluated 50+ representative models, outlining performance, strengths, limitations, and future directions.	Focused on CNN compression; other architectures were not adequately addressed.
Cărunță et al. [15]	2025	Lightweight deep learning for Segmentation	Survey	Surveyed heavy and lightweight models for semantic segmentation; analyzed performance and efficiency; emphasized lightweight trends.	Compared 2D imaging only; 3D segmentation was not covered.
Xu et al. [16]	2025	Edge-Deep Learning	Survey	Reviewed architectures, frameworks, and optimization for on-device intelligence in real-time, resource-constrained scenarios.	Focused on edge devices and 2D imaging only.

Research Scope: While considerable progress has been made on 3D medical imaging using deep learning in resource-constrained environment, there are still many challenges. We aim to investigate prior research in this direction, possible opportunities, and hurdles. The survey aims to address the following research questions.

- **RQ1** What are the computational (CPU, memory) and energy demands of state-of-the-art deep learning models for 3D medical imaging, and why is their critical analysis important?
- **RQ2** Which compression and optimization strategies have been applied to adapt 3D medical imaging models to resource-constrained environments, and how do they impact performance?
- **RQ3** What are the potential opportunities and advantages of deploying models in resource-constrained environments?
- **RQ4** What are the open challenges and future research opportunities in bridging the gap between state-of-the-art 3D medical imaging methods and their adoption in resource-limited environments?

Table 3 Search String Used to Retrieve Relevant Literature

Search Identifier	Search Query String
Search String	("3D medical imaging" OR "volumetric imaging" OR "3D imaging" OR "CT" OR "MRI" OR "CBCT" OR "PET") AND ("deep learning" OR "convolutional neural network" OR "CNN" OR "transformer" OR "neural network") AND ("resource constrained" OR "low-resource" OR "limited-resource" OR "edge device" OR "mobile deployment" OR "lightweight" OR "pruning" OR "quantization" OR "distillation" OR "model compression" OR "computational efficiency")

Following the SLR methodology of [53] and the PRISMA guidelines [54, 55], we first constructed a search string built around our theme, 3D medical imaging in resource-constrained environments using deep learning as shown in Table 3. In addition, the reference lists of the most relevant papers were screened (backward snowballing) to widen coverage.

High-quality and relevant articles were collected from five digital libraries: IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and PubMed. The searches were carried out between August 2025 and March 2026, with the final search conducted on 25th of March; the searches were repeated periodically during this window to capture newly published work. Following the PRISMA framework, the studies were systematically gathered and organized as illustrated in Fig. 4. The inclusion and exclusion criteria for study selection were defined following the protocol established in [53], as summarized in Table 4. The initial search returned 780 records across the five databases. After removing 142 duplicates and 23 records for other reasons, 615 records were screened on title and abstract, of which 312 were excluded. The remaining 303 reports were assessed in full text, and 213 were excluded (not 3D medical imaging, no efficiency focus, not deep learning, or failing quality assessment), yielding 90 studies included in this review.

Fig. 5 presents the analysis of selected papers based on publication trends, task types, and efficiency strategies. The results show a growing research interest, with the highest number of publications in 2025 and 2026. Most studies focus on segmentation, reflecting its clinical importance and complexity. Among efficiency techniques, lightweight architecture design and knowledge distillation are the most commonly used, although some distillation-based studies were excluded as they focused on improving accuracy rather than reducing model complexity.

4 Analysis and Discussion

To evaluate the current state of lightweight 3D deep learning in medical imaging, this section presents the quantitative and qualitative findings from our reviewed literature to systematically address the four defined research questions (RQs). A representative performance summary of lightweight deep learning models in key 3D medical tasks is presented in Table 5, followed by a summary of the state-of-the-art deep learning models with their computation footprints, shown in Table 6. Based on these aggregated results, we analyze the computational bottlenecks of 3D volumetric models,

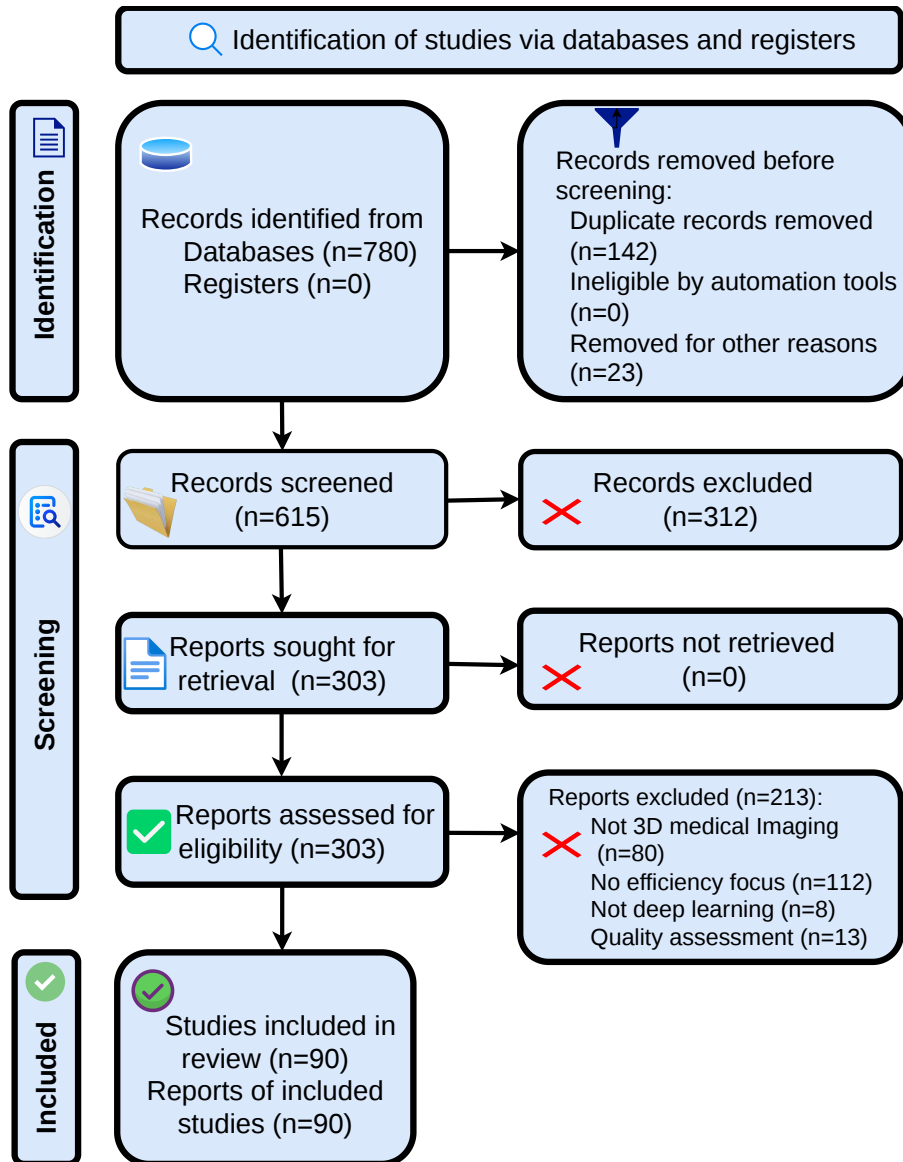
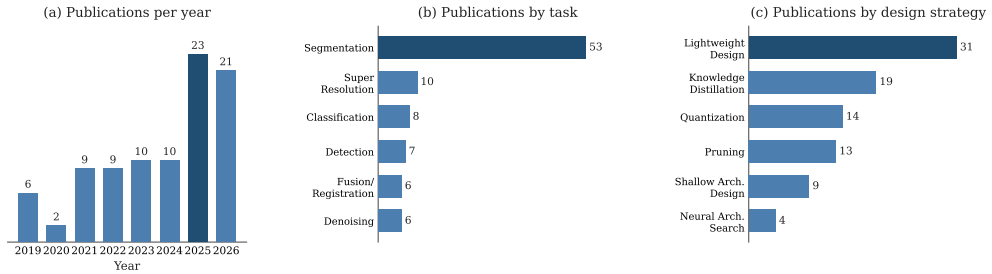


Fig. 4 PRISMA workflow illustrating our methodology to conduct the SLR

evaluate the efficacy of current model compression strategies, and outline critical open challenges and future research directions for resource-constrained clinical deployment.

Table 4 Inclusion and Exclusion Criteria

Category	Inclusion Criteria	Exclusion Criteria
Type of Study	Original research articles and comprehensive reviews	Theses, white papers, book chapters, and short communications
Language	Studies published in English	Non-English papers or duplicate publications
Publication Year	Studies published from 2018 onward	Studies published prior to 2018
Source	Peer-reviewed international journals and major conferences	Publications from unindexed local journals or predatory venues
Intervention	Deep learning architectures and algorithms	Traditional machine learning and non-data-driven statistical methods
Region	No restriction	None
Setting	Deep learning applications in 3D medical imaging	Applications focusing strictly on non-medical domain data
Quality Assessment	Peer-reviewed indexed venues with high citation credibility	Extended abstracts, workshop papers without rigorous peer review, or non-peer-reviewed preprints

**Fig. 5** Distribution of the reviewed studies: (a) number of publications per year, (b) publications by task type, and (c) publications by lightweight design strategy.

4.1 Computational and Energy Costs of 3D deep learning Models

An important step in assessing the resource footprint of 3D deep learning is quantifying the computational and memory demands of heavy architectures. As shown in Table 6, widely used 3D models span a broad range of sizes and costs: parameters range from about 33 M (TransBTS) to 96 M (TransUNet), and FLOPs from roughly 41 G (UNETR) to 554 G (V-Net). Transformer architectures are particularly demanding; Swin UNETR, for example, has 61.98 M parameters and 394.84 G FLOPs. Training these models is correspondingly expensive, on the order of 10 to 11 hours (e.g., about 10 hours for UNETR and 11 hours for nnFormer) on the same machine, and they consume substantial GPU memory and energy resources that are rarely available in constrained settings.

The computing and energy requirements of 3D deep learning depend on input dimensions, resolution, network depth, parameter count, FLOPs, and task complexity [56]. Task type matters a lot: segmentation of large volumes is far more memory-hungry than classification. For a $1 \times 128 \times 128 \times 128$ input, Swin UNETR requires roughly 19.7 GB of GPU memory versus 3.7 GB for the CNN-based nnU-Net (Table 7). Architecture is equally influential, as transformer-based models generally incur more parameters and FLOPs than CNNs and are markedly slower at inference: Swin UNETR takes about 228.6 ms per volume on a GPU and 7.6 s on a CPU, roughly 2 to $2.5\times$ slower than nnU-Net (90.9 ms and 3.6 s, respectively). These figures show that high-resolution volumetric transformers such as Swin UNETR and nnFormer carry a particularly high computational and memory burden [57, 58].

As discussed earlier, this study focuses on 3D medical imaging using deep learning in resource-constrained environments. To this end, we selected 90 high-quality papers that explore strategies for transforming state-of-the-art models into lightweight, resource-efficient versions. Most lightweight models listed in Table 5 are derived from full-scale counterparts through compression techniques or architectural modifications.

In Table 6, we summarize commonly used full-scale models in 3D medical imaging, highlighting their computational and memory requirements, including hardware needs and model complexity in terms of parameters and FLOPs. This serves as a baseline for understanding the resource demands of state-of-the-art architectures. The following section contrasts these with lightweight adaptations that reduce computational and memory burdens for deployment in resource-constrained environments.

To demonstrate a fair comparison of popular CNN and hybrid models, Shaker et al. [68] has conducted a comparison on BraTS dataset under identical experimental setup. An unbiased evaluation of efficiency and performance has been performed using the same input size, pre-processing strategy, and same hardware (Quadro RTX 6000 GPU with 24 GB memory and a 32-core Intel Xeon 4215 CPU). The models are evaluated in terms of No. of parameters, FLOPs, GPU memory, inference time on CPU and GPU (represented as CPU-T and GPU-T) for input size of $1 \times 128 \times 128 \times 128$. The comparison of the said models is depicted in Table 7.

The models mentioned in Table 7 and many recent ones are large-scale architectures that offer high accuracy and strong potential for solving various problems in medical imaging. However, they require substantial computational resources and long execution times, leading to high energy consumption and significant carbon emissions, thereby contributing to environmental degradation [9]. Research shows that high-performance computing systems and data centers are major contributors to global electricity consumption, further intensified by machine learning training and inference processes. Recent studies on the energy and carbon cost of machine learning show that not just training but even inference can produce notable carbon emissions [7, 9], and tools such as the Green Algorithms framework make these footprints quantifiable [72]. These environmental concerns, along with the need for broader and equitable access, highlight the importance of developing computationally efficient and environmentally sustainable models suitable for resource-constrained settings.

A note of caution applies when interpreting Tables 6 and 7. The reported parameters, FLOPs, memory, and inference times are drawn from the respective original

Table 5 Brief Summary of Representative Shortlisted Papers Focusing on Lightweight Models

Reference	Year	Task	Modality	Dataset	Deep Learning Model	Efficiency Strategy
Jianwei Zhao et al. [59]	2025	MRI Super Resolution	MRI	HCP-1200, IXI (public)	Distilled transformer (TDAFD — student / distilled)	KD (distilled transformer)
Ben Loussaief et al. [60]	2025	Segmentation	MRI/CT	Collected	Teacher: DeepLabV3+, PSPNet, FCN Student: ENet, ESPNet, MobileNetV2	KD (adaptive multi-teacher)
Dian Qin [45]	2021	Segmentation	CT	LiTS, KiTS19 (public)	Teacher: RA-UNet Student: ENet, MobileNetV2, ResNet-18	Knowledge Distillation (EMKD), Lightweight networks
Zhao [61]	2023	Segmentation	CT	WORD TCIA+BTCV (public)	Teacher: nnU-Net Student: LCOV-Net	Knowledge Distillation
Yiqing Shen et al. [62]	2024	Segmentation	MRI/CT	AMOS, TotalSegmentor, BraTS (public)	FastSAM3D	Layer-wise progressive distillation
Luís Serrador et al. [63]	2024	Segmentation	CT	VerSe (public)	Teacher: Full U-Net Student: Shallow U-Net	Distillation
Mingwei Wen et al. [64]	2025	Image Fusion	MRI/CT/PET	IXI, RESECT, MMWHS2017, Prostate-MRI, HNSCC (public)	LPM-Net	Shallow architecture
Viraj Aher et al. [65]	2025	Segmentation	CT/MRI/PET /Ultrasound	CVPR BiomedSegFM (public)	MobileSeg3D	Lightweight backbone design using MobileNet
Yan Pang et al. [66]	2024	Segmentation	MRI/CT	BraTS2021, MSD2019 Task-01 Brain Tumor, MSD2021 Task-08 Hepatic Vessel (public)	Slim UNETR	Hierarchical lightweight design

Table 6 Summary of State-of-the-Art Deep Learning Models in Medical Image Segmentation

Model	Dataset	Hardware Used	T/I Time	# Parameters	# FLOPs
UNETR [67]	BTCV, MSD	NVIDIA DGX-1 server	10H/ 12.08S	92.58M	41.19G
UNETR++ [68]	BraTS	A100 40GB GPU	—	92.49M	16.6G
nnFormer [58]	BraTS	NVIDIA 2080 GPU, 11GB VRAM	11H/ 20.25S	37.6M	119.3G
TransBTS [69]	BraTS	8 NVIDIA Titan RTX GPUs (each has 24GB VRAM)	—	32.99M	333G
Swin UNETR [57]	BraTS	8 × NVIDIA V100	—	61.98M	394.84G
TransUNet [70]	BraTS	1 NVIDIA RTX 8000 GPU	—	96.07M	48.3G
VNet [71]	—	NVidia GTX 1080 with 8 GB VRAM	—	45.6M	554.2G

Table 7 COMPARISON OF DEEP LEARNING MODELS ON BRA TS DATASET [68]

Model	#Params (M)	FLOPs (G)	Mem (GB)	GPU-T (ms)	CPU-T (ms)
nnUNet	16.8	410.1	3.7	90.9	3625.5
MedNeXt-M-K3	17.6	248.0	15.3	186.1	19955.4
MedNeXt-M-K5	18.3	308.0	18.8	351.5	64260.8
UNETR	92.5	153.5	3.3	82.5	2145.0
SwinUNETR	62.8	572.4	19.7	228.6	7612.3
CoTr	41.9	668.5	9.8	174.6	4236.2
nnFormer	149.6	421.5	12.6	148.0	5247.5
UNETR++	42.6	70.1	2.4	62.4	1497.7

papers, which differ in dataset, input volume size, patch strategy, precision, and evaluation hardware, and often report only a subset of these metrics. As a result, the values are not directly comparable across studies: a lower FLOP count on one benchmark does not necessarily indicate a more efficient model than a higher count measured under a different setup. We therefore present these figures as indicative of the overall computational burden of state-of-the-art 3D models rather than as a like-for-like ranking. This inconsistency is itself one of the key findings of this review, and it directly motivates the need for a standardized, efficiency-aware evaluation protocol discussed in Section 6.

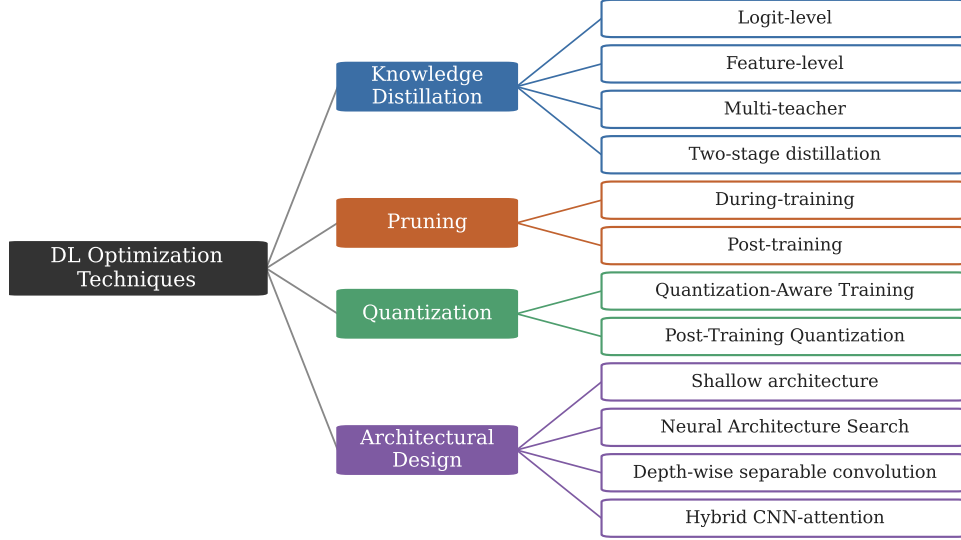


Fig. 6 Taxonomy of deep learning optimization techniques for 3D medical imaging reviewed in this survey, organised into four families: knowledge distillation, pruning, quantization, and architectural design.

The following section contrasts these full-scale models with lightweight adaptations that reduce computational and memory burdens for deployment in resource-constrained environments.

4.2 Efficiency Strategies

In this section we discuss the well-known techniques and methods used to convert a large model into a lightweight model that can be run/deployed in resource-constrained environments. We organise the lightweight techniques identified in this review into four families; (i) knowledge distillation, (ii) pruning, (iii) quantization, and (iv) architectural design, as illustrated in Fig. 6.

4.2.1 Knowledge Distillation

Knowledge Distillation (KD) refers to training a bigger model on the data and transferring its learned knowledge to a much smaller and efficient model [73]. Since its inception, KD has inspired numerous studies and become a key approach for designing lightweight models. The student model learns from both the ground truth labels and the soft labels (logits), as well as the intermediate features or attention maps [74, 75]. As a result, the student retains much of the teacher’s representational power while being smaller and faster, making it suitable for resource-constrained environments such as CPU-only systems or edge devices.

KD has been widely applied in 3D medical imaging to compress heavy architectures such as 3D U-Net, V-Net, and transformer-based models into efficient alternatives.

Common strategies include multi-level distillation, where knowledge is transferred at both logit and feature levels, and two-stage training, where the student is first trained on ground truth and then refined using a pretrained teacher. Representative studies are summarized in Table 8.

From these studies, KD proves to be an effective technique for designing lightweight models by reducing parameters, FLOPs, and inference time, while maintaining competitive accuracy. In some cases, it can even improve performance, as seen with TDAFD [59], due to enhanced feature learning through multi-scale distillation and attention mechanisms. However, performance trade-offs are also observed in some works [45, 60, 61, 63]. Typically, KD achieves a four- to six-fold reduction in parameters and FLOPs, along with two to three fold faster inference. Aggressive reductions in parameters have also been reported such as Luis et al., [63] reported a reduction in the number of parameters from 3.44M to 0.37M and inference time from 14.66s to 1.36s, with only a $\sim 3.5\%$ accuracy drop.

Additionally, performance improvements with KD have been reported by Ben et al., [60], for lightweight architectures such as E-Net, ESPNet, and MobileNetv2. This highlights KD’s dual benefit: compressing models while mitigating or even overcoming performance degradation. It can also be combined with techniques such as pruning, quantization, and shallow design for further efficiency gains.

In summary, KD has emerged as a foundational approach for building lightweight yet high-performing 3D medical imaging models. It enables compact models to inherit structured knowledge from large models without the associated computational burden, achieving significant compression with reliable accuracy. This makes KD particularly valuable for real-time deployment in resource-constrained and developing environments, supporting more accessible and efficient healthcare solutions.

4.2.2 Pruning

Model pruning is a widely used compression strategy that removes redundant parameters while maintaining accuracy within an acceptable limit [76]. It is particularly beneficial for deploying large models in resource-constrained environments and reducing operational costs. Pruning can be broadly categorized into post-training pruning and pruning during training. Post-training pruning eliminates insignificant weights from a pre-trained model based on scoring metrics that estimate parameter importance [77], [78]. For example, Sui et al. [79] proposed a channel-independence criterion that scores each filter’s importance and removes low-scoring filters; such importance-based pruning is directly applicable to volumetric networks. While this approach reduces model size, it often requires fine-tuning to recover performance. In contrast, pruning during training integrates sparsity directly into the learning process. The STAMP algorithm [80] applies channel-wise dropout during U-Net training for 3D medical image segmentation, removing redundant channels while improving inference speed without additional pruning steps.

Researchers typically adopt during-training pruning when simultaneous learning and sparsity optimization are required, as in PSP-Seg [81] and Sauron U-Net [82], while post-training pruning is preferred for rapid deployment with optional fine-tuning. Table 9 summarizes representative studies on pruning in 3D medical imaging.

STAMP and STAMP+ reduce parameters by up to 95% by removing low-activation channels during training, while maintaining accuracy and reducing memory and inference time [80]. Valverde et al. [82] achieved a 93.2% parameter reduction using filter pruning with minimal accuracy loss, though computational gains can sometimes be limited, as reflected by modest FLOPs reduction. APFormer [83] demonstrates the effectiveness of adaptive pruning combined with efficient design, reducing parameters and FLOPs compared to large models while improving accuracy by 1–6%.

Overall, pruning is a key technique for model compression, significantly reducing model size and inference time while maintaining accuracy. Its importance is particularly evident in 3D medical imaging, where computational and memory demands are high, enabling more efficient and practical deployment of AI systems.

4.2.3 Quantization

Quantization is a promising technique to mitigate the challenges of heavy-weight models that require high-performance computing and storage [85]. It works by reducing the precision bit width within a neural network. For example, converting high-precision representations such as 32-bit to lower 8-bit for weights and activations can significantly decrease memory requirements while accelerating processing speed. This makes models suitable for resource-constrained environments and enables faster inference for real-time medical applications.

Quantization is broadly categorized into (a) Quantization Aware Training (QAT) and (b) Post-Training Quantization (PTQ). QAT simulates low-precision computation during training, allowing the model to adapt and maintain accuracy. For instance, DeepSeekV3 [86] achieved accelerated training and reduced GPU requirements using FP8. However, QAT is time-consuming and requires fully labeled datasets, limiting its applicability in medical imaging. In contrast, PTQ quantizes a pre-trained model using a small set of unlabeled data for calibration, making it more practical when retraining is infeasible due to limited resources or data privacy concerns. Table 10 summarizes representative studies in quantization.

A clear trend is that quantization reduces memory usage and inference time while maintaining accuracy comparable to full-precision models. For example, Paschali et al. [87] reduced memory from 63 MB to 2 MB with no loss in Dice score (0.914 $\rightarrow$ 0.914). Similarly, Qu et al. [44] reports 3 $\times$ –10 $\times$ memory reduction across multiple models while preserving accuracy. Memon et al. [88] demonstrated promising performance (Dice = 0.916, accuracy = 99.845%) with significantly reduced memory and inference time, enabling deployment on devices such as Raspberry Pi 4B and smartphones. Other studies [89], [90] also maintain Dice performance while enabling resource-efficient deployment.

Unlike pruning or distillation, quantization does not reduce parameters or FLOPs; instead, it improves efficiency through lower bit-width representations, with gains largely dependent on hardware support. While it effectively reduces memory footprint, power consumption, and latency, it may introduce accuracy degradation, particularly in sensitive 3D segmentation tasks. QAT mitigates this at the cost of higher training complexity, whereas PTQ offers a practical trade-off with minimal calibration data.

Table 8 REPRESENTATIVE PAPERS ON KNOWLEDGE DISTILLATION TECHNIQUES IN 3D MEDICAL IMAGING

Study	Main Model	Lightweight Model	Task	Dataset	Infer. Time (s)	Accuracy / DSC	FLOPs (G)	Params (M)
Zhao et al. [59]	SuperFormer	TDAFD	MRI Super-Res.	HCP-1200, IXI	—	34.54 $\rightarrow$ 34.83 (PSNR)	254.15 $\rightarrow$ 201.06	1.56 $\rightarrow$ 1.21
Qin et al. [45]	RA-UNet	ENet	Segmentation	LiTS, KiTS19	—	0.685 $\rightarrow$ 0.652 (Dice)	24.81 $\rightarrow$ 2.03	22.10 $\rightarrow$ 0.353
Zhao et al. [61]	nnU-Net	LCOV-Net	Segmentation	WORD, TCIA+BTCV	—	85.83 $\rightarrow$ 85.91 (Dice %)	413.42 $\rightarrow$ 82.05	14.98 $\rightarrow$ 2.79
Serrador et al. [63]	U-Net	Shallow U-Net	Segmentation	VerSe	14.66 $\rightarrow$ 1.36	88.22 $\rightarrow$ 84.70 (Dice %)	109.49 $\rightarrow$ 45.07	3.44 $\rightarrow$ 0.37
Loussaief et al. [60]	DeepLabV3+, PSPNet, FCN	ENet, ESPNet, MobileNetV2	Segmentation	Private Dataset	—	0.90 $\rightarrow$ 0.85 (Dice)	96–273 $\rightarrow$ 1.27–19.8	46.7–86.7 $\rightarrow$ 0.18–2.23

Note: Values formatted as “X $\rightarrow$ Y” represent Teacher / Full-scale model $\rightarrow$ Distilled lightweight (student) model. Accuracy/DSC metrics report Dice similarity coefficient unless stated otherwise (e.g., PSNR for super-resolution). A dash (—) denotes unreported metrics in original studies.

Table 9 Summary of Papers on Pruning in 3D Medical Imaging

Reference	Model / Method	GPU Memory	Accuracy / DSC	Inference Time	# FLOPs	# Parameters
Dinsdale et al. [80]	STAMP / STAMP+	Better than baseline	Slightly lower	Faster	—	Reduction up to 85%
Valverde et al. [82]	Sauron	Reduction up to 6.6%	0.9595 $\rightarrow$ 0.9564	—	Reduction up to 93.02%	—
Xian Lin et al. [83]	APFormer	—	83.53 (+1%)	—	3.9 G	2.6 M
Linhao Li et al. [81]	PSP-Seg-S	0.8 GB	84.3 (vs. 85.0)	2 s	—	5.1 M
Hong et al. [84]	SlimFormer-3D	—	0.826	—	149.78 G	69.94 $\rightarrow$ 17.47 M

Note: Values reported as “X $\rightarrow$ Y” denote the baseline / full-scale model $\rightarrow$ the pruned lightweight model.

Overall, quantization is a fundamental technique for efficient deep learning, complementing pruning, distillation, and architectural optimization. It provides a balance between computational efficiency and accuracy, making it essential for deploying 3D medical imaging models in real-world, resource-constrained environments.

4.2.4 Architectural Efficiency Strategies

Lightweight deep learning models can also be designed by optimizing the network architecture itself rather than relying solely on compression techniques. Efficient architectural design can significantly reduce memory consumption and inference latency without substantial loss in accuracy. In this context, several studies employ techniques such as shallow architectures, neural architecture search (NAS), and depth-wise separable convolutions (DSC). Table 11 provides an overview of these approaches and their impact on model size, inference time, and accuracy in 3D medical imaging for resource-constrained environments.

Designing efficient architectures is not only useful for standalone lightweight models but can also complement techniques like knowledge distillation. A lightweight model can be designed first and later enhanced using knowledge from a larger model. Among architectural strategies, depth-wise separable convolutions and channel factorization are particularly effective [91–93] reducing FLOPs by over 50% and significantly lowering parameter counts with minimal accuracy loss. Additional improvements, such as ghost convolutions and hybrid CNN-attention blocks (e.g., GA-UNet and LW-CTrans), further enhance efficiency by reusing features and focusing on salient regions, balancing local and global context with low memory usage. However, these approaches often rely on handcrafted designs, which may limit generalizability.

Another direction is automated optimization through NAS and genetic algorithms, as seen in BiX-NAS, GA-UNet3+, and L-DNASR. These methods explore trade-offs between accuracy and latency, achieving parameter reductions of up to 95% while maintaining competitive performance. However, they introduce significant computational overhead during the search phase. In contrast, manually designed lightweight architectures [78, 86, 94, 95] use compact encoder-decoders and efficient convolutions, showing that thoughtful layer design improves scalability more effectively than simply reducing depth. For example, V-Net Light [96] achieves over 90% parameter reduction while maintaining segmentation accuracy.

Overall, efficient architectural design offers a powerful alternative to compression-based methods, enabling substantial reductions in memory and latency without sacrificing performance. When combined with techniques such as knowledge distillation and quantization, these approaches can further enhance efficiency and generalization for deployment in resource-constrained environments.

4.3 A Comparative Analysis of Efficiency Strategies

In the discussion above in this section, we have discussed various efficiency strategies, including pruning, quantization, knowledge distillation, and architectural optimization, supported by evidence from the literature. While each technique effectively reduces model size, memory footprint, and inference time, an important question is

Table 10 REPRESENTATIVE PAPERS ON QUANTIZATION FOR RESOURCE-LIMITED SETTINGS

Model / Technique	Study	GPU Memory (MB)	Accuracy / DSC	Inference Time (s)
3DQ	Paschali et al. [87]	63.0 $\rightarrow$ 2.0	0.914 $\rightarrow$ 0.915	—
PTQ across 3D Architectures	Qu et al. [44]	U-Net: 23.11 $\rightarrow$ 6.61 TransUNet: 351.85 $\rightarrow$ 91.90 UNesT: 807.30 $\rightarrow$ 25.66 nnU-Net: 26.94 $\rightarrow$ 8.94 SwinUNETR: 61.98 $\rightarrow$ 16.09 SegResNet: 42.40 $\rightarrow$ 17.48 VISTA3D: 65.14 $\rightarrow$ 18.57	U-Net: 0.822 $\rightarrow$ 0.822 TransUNet: 0.816 $\rightarrow$ 0.816 UNesT: 0.702 $\rightarrow$ 0.701 nnU-Net: 0.901 $\rightarrow$ 0.895 SwinUNETR: 0.878 $\rightarrow$ 0.877 SegResNet: 0.882 $\rightarrow$ 0.879 VISTA3D: 0.893 $\rightarrow$ 0.891	U-Net: 2.62 $\rightarrow$ 1.05 TransUNet: 4.09 $\rightarrow$ 1.74 UNesT: 5.50 $\rightarrow$ 2.70 nnU-Net: 2.99 $\rightarrow$ 1.25 SwinUNETR: 9.85 $\rightarrow$ 3.59 SegResNet: 5.14 $\rightarrow$ 2.06 VISTA3D: 4.59 $\rightarrow$ 1.93
NISE (Classification) NIVE (Segmentation)	Memon et al. [88]	NISE: 13.4 $\rightarrow$ 4.9 NIVE: 176.86 $\rightarrow$ 86.24	NISE: 99.84% $\rightarrow$ 99.84% NIVE: 0.9163 $\rightarrow$ 0.9162	Laptop: 1.2074 (NISE) / 1.8241 (NIVE) SBC: 3.7627 (NISE) / 5.2641 (NIVE) Phone: 0.1068 (NISE) / 3.2023 (NIVE) RPi 4B: 3.4700 (NISE) / 6.2100 (NIVE)
Quantized 3D U-Net	Kim et al. [89]	—	0.96 $\rightarrow$ 0.95 (Dice)	—
Integer-arithmetic 3D U-Net	Wang et al. [90]	—	0.888 $\rightarrow$ 0.887	—

Note: Values formatted as “ $X \rightarrow Y$ ” denote Full Precision (FP32) $\rightarrow$ Quantized (e.g., INT8) metric output. A dash (—) indicates metrics not reported in the respective publication.

Table 11 SUMMARY OF PAPERS ON ARCHITECTURAL EFFICIENCY STRATEGIES

Model	Optimization Strategy	Base Model	Dataset	Params (M)	FLOPs (G)	Inference Time	Memory (MB)	Accuracy
lightweight 3D U-Net [92]	DSC	U-Net	KiTS	21.33 $\rightarrow$ 16.6	522 $\rightarrow$ 269	–	–	0.80 $\rightarrow$ 0.83 (Dice)
ERV-Net [91]	DSC	–	BraTS 2018	17.3	370.9	–	–	81.81 (Dice %)
LCCNet [93]	DSC	U-Net	ACDC	81.5 $\rightarrow$ 25.2	329.8 $\rightarrow$ 91.0	20.8 $\rightarrow$ 1717.3	–	89.3 $\rightarrow$ 71.0 (Dice %)
GA-UNet [97]	Ghost and Attention U-Net	U-Net	ISIC-2018 etc.	28.95 $\rightarrow$ 2.18	48.33 $\rightarrow$ 4.45	–	–	0.95 $\rightarrow$ 0.96 (Dice)
LPM-Net [64]	Hybrid Lightweight	Densefuse	BraTS 2020 etc.	0.211 $\rightarrow$ 0.022	1159.05 $\rightarrow$ 37.65	200 $\rightarrow$ 90	20.95 $\rightarrow$ 16.56	0.935 $\rightarrow$ 0.927
LW-CTrans [98]	Hybrid and Optimized Attention	H2Former	BraTS2019	33.64 $\rightarrow$ 2.08	24 $\rightarrow$ 31	–	256.76 $\rightarrow$ 15.87	82.43 $\rightarrow$ 83.75 (Dice %)
Slim UNETR [66]	Self-attention decomposition	Swin UNETR	MSD2019	61.99 $\rightarrow$ 1.79	793.92 $\rightarrow$ 12.42	Throughput 20.82 $\rightarrow$ 279.78	Decreased	79.16 $\rightarrow$ 80.27 (Dice %)
AST-Net [99]	lightweight convolutions	nnU-Net	BraTS_2020	30.6 $\rightarrow$ 10.5	263.8 $\rightarrow$ 44.1	–	–	78.7 $\rightarrow$ 77.8 (Dice %)
WIAF [77]	Optimized Attention	Swin UNETR	BraTS2020 and FeTS2022	62.0 $\rightarrow$ 5.23	793.92 $\rightarrow$ 9.75	0.36 $\rightarrow$ 0.28	–	87.3 $\rightarrow$ 85.0 (Dice %)
lightweight 3D CNN-UNet [78]	Lightweight architectural design	U-Net	BraTS 2018	31.05	45	–	–	0.93 (Dice)
PMFSNet [95]	Lightweight Architectural Design	SwinUNETR	Various 3D Datasets	62.19 $\rightarrow$ 0.63	912.35 $\rightarrow$ 15.14	–	–	89.74 $\rightarrow$ 91.3 (Dice %)
BiX-NAS [100]	Neural Architecture Search	U-Net	MoNuSeg and TNBC	8.64 $\rightarrow$ 0.38	65.83 $\rightarrow$ 28.00	–	–	80.7 $\rightarrow$ 82.2 (Dice %)
GA-UNET3+ [101]	Neural Architecture Optimization	Unet 3+	LIDC-IDRI	26.97 $\rightarrow$ 1.02	–	35 ms	–	97.15 $\rightarrow$ 98.97 (Dice %)
PRISM Lite [81]	lightweight network architecture	PRISM	Curated	118 $\rightarrow$ 0.1	557.82 $\rightarrow$ 29.29	10.46 $\rightarrow$ 0.75	–	97.15 $\rightarrow$ 98.00 (Dice %)
PocketNet [102]	Shallow Architecture Design	UNet Full	LiTS, BraTS, NFBS	90.5 $\rightarrow$ 0.8	–	5.89% increase	–	0.922 $\rightarrow$ 0.93 (Dice)
RIRGAN [103]	lightweight GAN architecture	SRGAN	BraTS	30.2 $\rightarrow$ 3.3	–	–	–	PSNR: 25.05 $\rightarrow$ 28.26; SSIM: 0.82 $\rightarrow$ 0.88
Modified UNet [104]	lightweight network architecture	UNet	BraTS	30 $\rightarrow$ 0.49	–	–	–	0.96 (Dice)

Note: Values reported as “X $\rightarrow$ Y” denote the base / full-scale model $\rightarrow$ the lightweight model. A dash (–) indicates a value not reported in the original study.

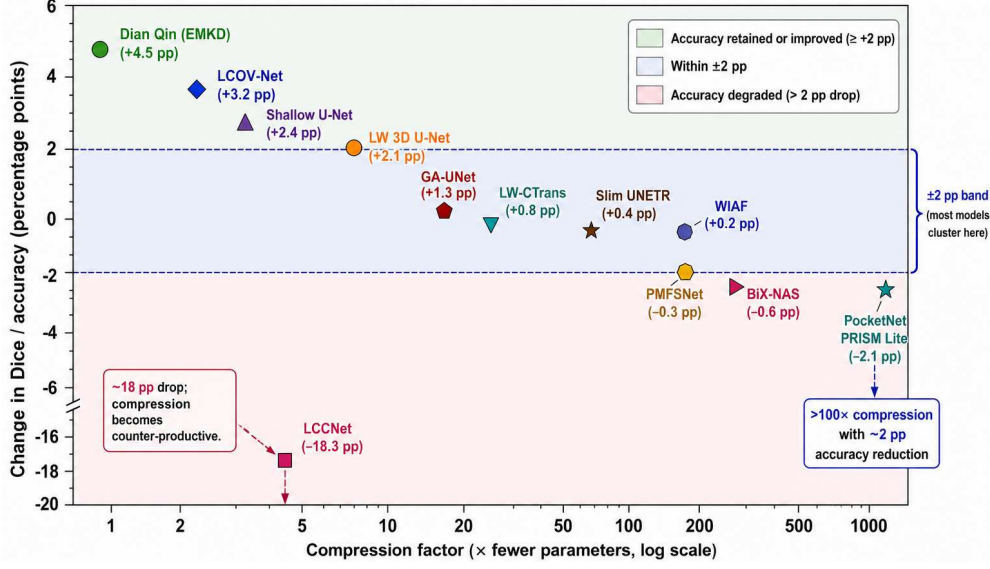


Fig. 7 Accuracy–efficiency trade-off of representative lightweight models.

which technique is most suitable for a given scenario. The answer depends on factors such as computational resources, deployment environment, dataset characteristics, and required accuracy.

For example, pruning is well suited for deploying models on devices with limited memory and computational resources, where a modest loss in performance is acceptable. Quantization, particularly low-bit precision, is effective for accelerating inference on CPUs and edge devices, although it may introduce some performance degradation. When maintaining high accuracy is the primary objective, knowledge distillation is often the preferred approach, as it enables a compact student model to inherit the predictive capability of a high-performing teacher while substantially reducing model complexity. In contrast, architectural optimization techniques, such as depth-wise separable convolutions and lightweight 3D building blocks, are most appropriate when designing models from scratch for resource-constrained environments, as they reduce computational cost while preserving representational capacity. A comparative summary of these techniques is provided in Table 12. Furthermore, as illustrated in Fig. 7, most lightweight models remain within approximately ± 2 percentage points of their teacher baselines while achieving parameter reductions of one to three orders of magnitude. Notably, PocketNet and PRISM Lite exceed $100\times$ compression with only a marginal accuracy reduction of approximately 2 percentage points, whereas LCCNet incurs an accuracy drop of roughly 18 percentage points, illustrating the point at which aggressive compression becomes counter-productive.

In practice, these techniques are complementary rather than mutually exclusive. For instance, a lightweight architecture can be enhanced with knowledge distillation to improve accuracy, followed by pruning or quantization to meet strict deployment

Table 12 COMPARISON OF MODEL OPTIMIZATION TECHNIQUES AND THEIR TRADE-OFFS

Technique	Main Goal	Best Use Case	Accuracy Impact	Speed Improvement	Memory Reduction	Key Limitations
Pruning	Reduce parameters	Memory-constrained devices	Small drop	Moderate improvement	High reduction	May harm accuracy if aggressive
Quantization	Reduce precision	CPU / Edge inference acceleration	Moderate drop	High improvement	High reduction	Accuracy loss at low precision
Knowledge Distillation	Preserve accuracy	High accuracy requirement	Minimal drop	Indirect	Indirect	Requires teacher model
Architectural Optimization	Efficient design	Resource-constrained deployment	Minimal / no drop	High improvement	High reduction	Requires redesign effort

constraints. Therefore, the selection or combination of strategies should be guided by the target environment, the balance between efficiency and accuracy, and task characteristics. By leveraging the strengths of each method, models can be designed to achieve an optimal balance between performance, efficiency, and practical deployment.

4.4 Addressing the Research Questions

Bringing together the evidence from the previous subsections, we now return to the four research questions defined in section 3.

RQ1 (Computational and energy demands). The reviewed state-of-the-art models are heavy on almost every axis. As shown in Tables 6 and 7, parameters range from around 16M for nnU-Net to about 96M for TransUNet, while FLOPs grow from tens to several hundred G, reaching 395–554G for transformer-based designs such as Swin UNETR and V-Net. The rise in inference time is noticeably high on CPU-only devices. For example, for a $1 \times 128 \times 128 \times 128$ input, nnU-Net runs in about 90.9 ms on a GPU but 3625.5 ms on a CPU. Training adds further cost, with nnU-Net needing roughly 11 GB of GPU memory for stable performance [43].

RQ2 (Compression and optimization strategies). Four families of compression and optimization strategies dominate the literature (Fig. 6): knowledge distillation (KD), pruning, quantization, and architectural efficiency design, with lightweight architecture and KD being the most common. Their measured trade-offs are summarised in Tables 8–12. KD usually gives a four- to sixfold reduction in parameters and FLOPs and a two- to threefold faster inference, with a small accuracy change and, in some cases, even a slight gain [59]. Pruning reaches very high parameter reductions, up to 93–95% in STAMP and STAMP+ [80], though it often needs fine-tuning to recover accuracy. Quantization lowers the bit-width (for example 32-bit to 8-bit) and cuts memory by about $3\times$ to $10\times$ with near-lossless accuracy, enabling deployment on devices such as the Raspberry Pi 4B [88]. Architectural design, using depth-wise separable convolutions, NAS, or hybrid CNN-attention blocks, can remove up to 90–95%

of parameters with little or no accuracy loss [95, 100]. These strategies are complementary rather than exclusive, and the right choice depends on whether the target is memory-bound, latency-bound, or accuracy-critical.

RQ3 (Opportunities and advantages). Lightweight 3D models offer faster inference, lower energy use, and deployment on low-cost hardware such as CPU-only laptops, edge processors, and portable ultrasound devices. This extends the benefits of 3D imaging to rural, mobile, and LMIC settings that currently lack high-end GPUs [1, 2, 10]. Beyond access, they support on-device and privacy-preserving inference, offline operation, and real-time tasks, and they align with the goals of Green AI by lowering the carbon footprint of both training and inference [7–9]. The growing availability of public benchmarks and enabling techniques such as synthetic data, self-/semi-supervised learning, and transfer learning further widen these opportunities.

RQ4 (Open challenges and future opportunities). Several challenges remain. First, there is still an accuracy gap between lightweight and full-scale models, especially for small anatomical structures and class-imbalanced tasks. Second, public 3D medical datasets originating from low-resource or portable hardware remain scarce. While high-resolution datasets can be synthetically down-sampled, artificial degradation fails to capture the unique noise profiles, low signal-to-noise ratios, and domain shift characteristic of real edge acquisition environments. Third, most validations are single modality and stay at the prototype stage, with very little real-world clinical testing. Fourth, specialized lightweight hardware and efficient 3D file formats are still lacking. Finally, there is no standard, efficiency-aware evaluation metric that reports accuracy together with parameters, FLOPs, memory, latency, and energy. These gaps directly highlight the open research avenues outlined in Section 6 (Future Work Recommendations).

5 Challenges and Opportunities for Lightweight 3D Medical Imaging

Potential. Deep learning applied to 3D medical imaging has already demonstrated strong clinical value across tasks such as classification, segmentation, detection, and localization, and volumetric AI systems (e.g., RapidAI [105] for stroke imaging) are increasingly integrated into radiology and treatment-planning workflows. By exploiting the full spatial context of CT, MRI, and PET volumes, these methods characterise complex tissues and pathologies more accurately than slice-wise 2D analysis. The open question, therefore, is not whether 3D deep learning is useful, but how to make it efficient and accessible enough for widespread, real-world deployment, which is the focus of this section.

Despite these benefits, widespread adoption of 3D deep learning faces significant challenges, including limited annotated datasets, high computational demands, and concerns about data privacy and security. Training and deploying models on 3D data requires substantial computing power, large memory, and long processing times, restricting their use to well-equipped hospitals and research centers. Cloud computing offers a partial remedy but is constrained by internet dependence, bandwidth limits, and privacy concerns, particularly in low-resource settings. Lightweight deep learning

offers a practical path around these constraints; the remainder of this section examines the key challenges and opportunities in realising efficient, deployable 3D medical imaging.

Data Availability. The first and most significant challenge is the availability of sufficient *well-annotated* data for training and evaluating models. Although 3D imaging modalities are now routine in healthcare and large volumes of raw volumetric data are generated daily, little of it is usable for research: it is fragmented across institutions, subject to strict privacy regulation, and, most critically, expensive and time-consuming to annotate at the voxel level by experts [39]. Consequently, it is not raw data but *curated, openly available, expertly labelled* 3D datasets that remain scarce, a major barrier to developing and benchmarking deep learning models.

Efforts are under way to narrow this gap. Additional data can be collected and annotated following established technical and ethical procedures, and large population-scale initiatives such as the UK Biobank [106] are building sufficient imaging databases. However, such resources remain comparatively few and are usually access-controlled and curated for high capacity models rather than resource-constrained deployment, so the shortage of openly accessible, well-annotated 3D datasets persists. Where such data does exist, it is valuable not only for direct training but also because knowledge from models trained on it can be transferred or distilled into smaller, deployable networks.

Synthetic Data. To address the challenge of limited datasets, researchers have increasingly turned to synthetic data. Synthetic 3D images, such as CT or MRI volumes generated using AI models include data augmentation, cross-modality translation, and simulation of rare pathologies. A few popular models for synthetic data generation include Generative Adversarial Networks (GANs) [107], Variational Autoencoders (VAEs) [108], and, more recently, diffusion models [109]. Synthetic data serves multiple purposes: augmentation increases the number of training instances, while translation enables cross-modality synthesis, for example, generating CT from MRI [110]. Studies have shown that 3D generative models can produce anatomically coherent and clinically plausible volumes, improving the accuracy and reliability of tasks such as segmentation and classification [111]. However, challenges remain in maintaining inter-slice consistency, preserving clinically significant anatomical details, and avoiding the creation of artificial or non-existent pathologies.

Self-supervised or Semi-supervised Approaches. A major challenge in deep learning for medical imaging is the limited availability of annotated datasets, which are considered the gold standard [76]. Creating and sharing large, high-quality datasets labeled or validated by experts is essential for developing accurate and scalable models. Alternatively, self-supervised or semi-supervised learning allows training on minimally or unlabeled data [93], reducing dependence on extensive annotations and enabling lightweight models with lower computational demands. Combining these approaches with model compression can further enhance efficiency while maintaining performance, helping to overcome data scarcity and computational barriers in 3D medical imaging.

Transfer Learning. Transfer learning can help by enabling a model trained on one task or dataset to be fine-tuned for another task with limited data. However, in 3D medical imaging, pre-trained models are scarce, and training for transfer learning still requires large labeled datasets, which are often unavailable. Recently, 3D variants of

models such as ResNet, UNet, and transformers have been pretrained on MedicalNet [112] and ModelGenesis[113], showing promising results. Yet, these models are not directly suitable for lightweight networks due to their high computational and memory demands, and their weights cannot be easily transferred to smaller architectures.

Specialized Hardware. A major challenge in remote and underdeveloped regions is the limited availability of hardware with sufficient computing power and memory for running deep learning models. Advances in micro and nanotechnology have made more capable edge devices, such as low-power CPUs, AI accelerators, and compact memory modules available, enabling on-device inference for lightweight applications [49, 114]. Nonetheless, there remains a need for task-specific, AI-oriented hardware optimized for 3D medical imaging workloads rather than relying on general-purpose devices [13]. Such targeted hardware, combined with efficient model design, can improve accuracy, reduce inference time, lower costs, and extend the benefits of advanced medical imaging to resource-limited healthcare environments.

Handling Class Imbalance. Another major challenge in deep learning is severe class imbalance. This limitation is particularly pronounced in volumetric medical imaging, where target anatomical structures occupy only a tiny fraction of the overall volume; for example, early-stage lung nodules represent a minuscule portion of the lung volume, and slender structures like the mandibular canal are overwhelmed by the surrounding jawbone [35]. A common solution in large models is a two-stage framework: first localizing or cropping the region of interest (ROI), then applying a second model for the main task, such as classification or segmentation [115]. While this improves focus on minority classes and reduces background interference, it adds computational overhead. Managing class imbalance therefore remains a key challenge in lightweight deep learning, where efficiency and compactness are essential.

Multimodal System. Most existing applications rely solely on medical images, such as X-rays or MRIs, but clinical information like patient history, age, and demographics can also aid accurate diagnosis. Some studies have attempted to integrate clinical data, but improvements were limited due to data scarcity and lack of proper frameworks [116]. Multimodal research is still in its early stages, and developing lightweight 3D models remains challenging. Additionally, most 3D deep learning models operate on image patches due to limited memory [43] or computing power, which neglects the full anatomical context [92]. Efficiently incorporating both clinical and anatomical information remains a key challenge in medical image analysis.

In summary, 3D medical imaging offers enhanced accuracy and spatial understanding but faces challenges such as high computational cost, limited data, and difficulties in model deployment. Recent advances in lightweight architectures, multimodal integration, and resource-efficient training offer promising solutions. Nonetheless, balancing accuracy, efficiency, and real-world applicability remains an open challenge. Future research should aim to develop scalable, memory-efficient 3D models suitable for deployment in resource-constrained healthcare settings.

6 Future Work Recommendations

Despite the significant progress in deep learning and medical imaging, numerous challenges remain, particularly for lightweight deep learning. While various models have been proposed, there is still ample room for improvement to enable clinical integration and more equitable access to medical imaging. Future research should address factors such as algorithm performance, hardware limitations, model efficiency, multi-task learning, and the integration of multimodal clinical information. Building on the challenges identified in section 5, we outline the following key research recommendations.

Specialized Dataset. As noted in section 5, the datasets available today are limited and, more importantly, are built for high-capacity models. A clear direction for future work is therefore to develop compact 3D datasets tailored for lightweight models, featuring lower resolution, fewer slices, optimized memory usage, and appropriate annotation granularity. Such datasets would support the training, benchmarking, and deployment of lightweight deep models in remote or underdeveloped healthcare settings, and standardized compact benchmarks could accelerate research on efficient architectures and edge deployment. Closely related is the question of file formats. Medical images currently use various standards, such as OME-TIFF for multidimensional microscopy [117], but 3D volumes typically demand substantial storage and memory, which is a poor fit for low-resource environments. Developing dedicated lightweight 3D image formats that store volumetric data efficiently while preserving essential diagnostic information is a promising direction for enabling practical deployment on embedded systems.

Handling Class Imbalance. As discussed in section 5, keeping models sensitive to small, underrepresented structures without inflating their size remains unsolved for resource-constrained 3D imaging. Future work should design lightweight architectures and loss formulations that preserve this sensitivity without the overhead of two-stage ROI pipelines or heavy attention modules, for example by co-designing region-focusing components together with compression rather than adding them on top.

Multimodal Systems. Section 5 highlighted that most current systems rely on a single modality and that integrating clinical information remains difficult under resource constraints. The forward-looking challenge is to build lightweight models that can handle multiple modalities at once; for example, a single compact model trained on both CT and MRI that segments accurately regardless of the input modality. Cross-modality knowledge distillation, where a model learns shared representations across modalities, is one practical route and makes this a promising direction for future research.

End-to-End 3D. Most existing medical imaging systems rely on 2D imaging or process 2D slices from 3D volumes. True end-to-end 3D systems that take a full 3D image as input and directly produce a 3D output are still scarce and face challenges such as low accuracy, high storage requirements, long inference times, and single-modality limitations. Developing efficient end-to-end models for volumetric data represents a significant opportunity for future research.

Vision-Language Models. Multimodal AI, which integrates multiple data sources such as text, images, audio, or video, is an emerging field. These models are

often heavyweight due to separate encoders for each modality. A promising direction is the development of lightweight multimodal AI, particularly for integrating 3D imaging data. Vision-Language Models (VLMs), which combine visual and textual information, could be adapted for interactive systems in resource-constrained settings. Challenges remain, especially in creating multimodal datasets with properly aligned textual descriptions.

Generalization of Model. Most existing lightweight models are designed for a single task, such as segmentation or classification, and comparatively few generalise across tasks; PocketNet [102] is one of the few efficient models shown to work across multiple datasets and tasks. A promising direction is to develop unified lightweight models capable of performing several 3D medical imaging tasks such as classification, segmentation, anomaly detection, and registration within a single efficient architecture. By integrating lightweight design strategies such as pruning, quantization, and knowledge distillation, such models could achieve competitive accuracy across these tasks while remaining deployable in resource-constrained environments.

Efficiency-Aware Evaluation Metrics. A recurring gap in existing studies is that lightweight models are often evaluated on accuracy metric, while efficiency is reported inconsistently or not at all, which makes fair comparison across studies difficult. Future work should move toward standardized, efficiency-aware evaluation that reports accuracy together with parameters, FLOPs, memory footprint, inference latency, and energy consumption, ideally combined into a single normalized score such as accuracy-per-MFLOP or Dice-per-megabyte. A shared metric of this kind would let the community compare lightweight 3D models on a common footing and reward genuine efficiency gains rather than accuracy alone.

In conclusion, future research should focus on developing lightweight, efficient, and deployable deep learning models capable of operating in resource-constrained settings. Key directions include creating specialized lightweight 3D datasets, designing compact data formats, addressing class imbalance, enabling cross-modality and multimodal learning, and establishing efficiency-aware evaluation metrics. Integrating efficiency techniques such as pruning, quantization, and knowledge distillation will be crucial to balance performance and efficiency. These advancements will make 3D medical imaging more accessible, scalable, and applicable in real-world clinical and ubiquitous settings.

7 Threats to Validity

This study is subject to several threats to validity, which we sought to mitigate throughout. First, search and selection bias may arise from the choice of databases, keywords, and inclusion criteria, potentially omitting relevant studies. To reduce this, we defined the search string in advance (Table 3), applied a fixed set of inclusion and exclusion criteria (Table 4), searched five major digital libraries, and supplemented the results with backward reference screening (snowballing) to widen coverage. Second, publication bias could limit the evidence base, as studies reporting positive results are more likely to be published than those with negative or inconclusive findings; to partially offset this, we included preprint sources (e.g., arXiv) alongside peer-reviewed venues, although this bias cannot be entirely eliminated. Third, to minimize review

bias during screening and selection, decisions were independently evaluated by a co-author, and disagreements were resolved through discussion. Finally, given the rapid pace of research in deep learning and medical imaging, some recent or emerging models may not be fully captured, and we therefore repeated the searches periodically up to the final search date.

8 Conclusion

This paper presented a systematic review of lightweight deep learning for processing 3D medical images (volumes) in resource-constrained environments. Following the PRISMA methodology, we analysed the reviewed studies and organised the field into four families of efficiency techniques: knowledge distillation, pruning, quantization, and architectural design and examined the datasets, tools, and deployment settings involved. Two findings stand out. First, research on efficient 3D models has grown rapidly, peaking in 2025, with segmentation as the dominant task and lightweight architectural design and knowledge distillation as the most common strategies. Second, and most importantly, lightweight models do not necessarily sacrifice predictive performance: many achieve one to three orders of magnitude reduction in parameters, FLOPs, and inference time while retaining and sometimes improving accuracy, indicating that efficiency and accuracy can be pursued together rather than traded off.

Despite this progress, the field remains at an early stage of clinical translation. The scarcity of well-annotated and compact 3D datasets, weak integration between imaging and clinical information, and limited access to specialised hardware continue to constrain progress. Crucially, evaluations are rarely conducted on real resource-constrained hardware, and reported efficiency gains are measured inconsistently across studies, making fair comparison difficult. Addressing this requires a shift beyond simply reducing model size: future work should adopt standardised, efficiency-aware evaluation that reports accuracy alongside parameters, FLOPs, memory, inference latency, and energy; curate compact 3D datasets and benchmarks tailored to low-resource deployment; and extend lightweight design to cross-modality and multimodal settings. Realising this vision will depend on close collaboration among AI researchers, clinicians, and hardware developers. Ultimately, the goal is not merely smaller models, but reliable, affordable, and deployable 3D medical imaging that operates effectively where computational resources are limited.

Declarations

Funding. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests. The authors have no relevant financial or non-financial interests to disclose.

Ethical Approval. Not applicable. This study is a systematic review of published literature and did not involve human participants or animals.

Consent to Participate. Not applicable.

Consent for Publication. Not applicable.

Data Availability. All data analysed in this review are derived from the published studies cited in the manuscript. The extracted data and search strategy are available from the corresponding author on reasonable request.

References

- [1] World Economic Forum: AI in Healthcare Risks Could Exclude 5 Billion People: Here's What We Can Do About It. [Online]. Available: <https://www.weforum.org/stories/2025/10/ai-in-healthcare-risks-could-exclude-5-billion-people-here-s-what-we-can-do-about-it/>. Accessed: Dec. 3, 2025 (2025)
- [2] Ciecierski-Holmes, T., Singh, R., Axt, M., Brenner, S., Barteit, S.: Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: A systematic scoping review. *npj Digital Medicine* **5**(1), 162 (2022) <https://doi.org/10.1038/s41746-022-00688-9>
- [3] Doi, K.: Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics* **31**(4-5), 198–211 (2007) <https://doi.org/10.1016/j.compmedimag.2007.02.002>
- [4] Thibault, J.-B., Sauer, K.D., Bouman, C.A., Hsieh, J.: A three-dimensional statistical approach to improved image quality for multislice helical ct. *Medical Physics* **34**(11), 4526–4544 (2007) <https://doi.org/10.1118/1.2789499>
- [5] Ker, J., Singh, S.P., Bai, Y., Rao, J., Lim, T., Wang, L.: Image thresholding improves 3d convolutional neural network diagnosis of different acute brain hemorrhages on computed tomography scans. *Sensors* **19**(9), 2167 (2019) <https://doi.org/10.3390/s19092167>
- [6] Singh, S.P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., Gulyás, B.: 3d deep learning on medical images: A review. *Sensors* **20**(18), 5097 (2020) <https://doi.org/10.3390/s20185097>
- [7] Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green ai. *Communications of the ACM* **63**(12), 54–63 (2020) <https://doi.org/10.1145/3381831>
- [8] IEEE Spectrum: Energy-Efficient Green AI Strategies. [Online]. Available: <https://spectrum.ieee.org/energyefficient-green-ai-strategies>. Accessed: Dec. 3, 2025 (2025)
- [9] Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in nlp. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, pp. 3645–3650 (2019). <https://doi.org/10.18653/v1/P19-1355>

- [10] Nelson, B.P., Melnick, E.R., Li, J.: Portable ultrasound for remote environments, part i: feasibility of field deployment. *J. Emerg. Med.* **40**(2), 190–197 (2011) <https://doi.org/10.1016/j.jemermed.2009.09.006>
- [11] Wang, C.-H., Huang, K.-Y., Yao, Y., Chen, J.-C., Shuai, H.-H., Cheng, W.-H.: Lightweight deep learning: an overview. *IEEE Consum. Electron. Mag.* **13**(4), 51–64 (2024) <https://doi.org/10.1109/MCE.2022.3181759>
- [12] Liu, H.-I., Galindo, M., Xie, H., Wong, L.-K., Shuai, H.-H., Li, Y.-H., Cheng, W.-H.: Lightweight deep learning for resource-constrained environments: a survey. *ACM Comput. Surv.* **56**(10), 1–42 (2024) <https://doi.org/10.1145/3657282>
- [13] Li, Z., Li, H., Meng, L.: Model compression for deep neural networks: a survey. *Computers* **12**(3), 60 (2023) <https://doi.org/10.3390/computers12030060>
- [14] Chen, F., Li, S., Han, J., Ren, F., Yang, Z.: Review of lightweight deep convolutional neural networks. *Arch. Comput. Methods Eng.* **31**(4), 1915–1937 (2024) <https://doi.org/10.1007/s11831-023-10032-z>
- [15] Cărunță, C., Cărunță, A., Popa, C.-A.: Heavy and lightweight deep learning models for semantic segmentation: a survey. *IEEE Access* **13**, 17745–17765 (2025) <https://doi.org/10.1109/ACCESS.2025.3529812>
- [16] Xu, Y., Khan, T.M., Song, Y., Meijering, E.: Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. *Artif. Intell. Rev.* **58**(3), 93 (2025) <https://doi.org/10.1007/s10462-024-11033-5>
- [17] Menze, B.H., *et al.*: The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging* **34**(10), 1993–2024 (2015) <https://doi.org/10.1109/TMI.2014.2377694>
- [18] Litjens, G., *et al.*: A survey on deep learning in medical image analysis. *Medical Image Analysis* **42**, 60–88 (2017) <https://doi.org/10.1016/j.media.2017.07.005>
- [19] IoT Analytics: Number of Connected IoT Devices. [Online]. Available: <https://iot-analytics.com/number-connected-iot-devices/>. Accessed: Dec. 3, 2025 (2025)
- [20] Xia, Q., Chen, H., Liu, X., *et al.*: A comprehensive review of deep learning for medical image segmentation. *Neurocomputing* **613**, 128740 (2025) <https://doi.org/10.1016/j.neucom.2024.128740>
- [21] Perera, S., Navard, P., Yilmaz, A.: Segformer3d: an efficient transformer for 3d medical image segmentation. In: 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, pp. 4981–4988 (2024). <https://doi.org/10.1109/CVPRW63382.2024.00503>

- [22] Zhou, L., Fan, M., Hansen, C., Johnson, C.R., Weiskopf, D.: A review of three-dimensional medical image visualization. *Health Data Science* **2022**, 9840519 (2022) <https://doi.org/10.34133/2022/9840519>
- [23] OpenMedScience: Emergence of 3D Imaging. [Online]. Available: <https://openmedscience.com/a-dive-into-the-world-of-3d-medical-imaging/>. Accessed: Dec. 3, 2025
- [24] Lev-Ari, A.: Pharma Intelligence. [Online]. Available: <https://pharmaceuticalintelligence.com/category/cardiac-and-cardiovascular-surgical-procedures/>. Accessed: Dec. 3, 2025
- [25] Cline, H.E., *et al.*: 3d surface rendered mr images of the brain and its vasculature. *J. Comput. Assist. Tomogr.* **15**(2), 344–351 (1991) <https://doi.org/10.1097/00004728-199103000-00035>
- [26] Alder, M.E., Deahl, S.T., Matteson, S.R.: Clinical usefulness of two-dimensional reformatted and three-dimensionally rendered computerized tomographic images. *J. Oral Maxillofac. Surg.* **53**(4), 375–386 (1995) [https://doi.org/10.1016/0278-2391\(95\)90707-6](https://doi.org/10.1016/0278-2391(95)90707-6)
- [27] Werner, H.: 3d reconstruction in fetal medicine. *Ultrasound Med. Biol.* **45**, 44–45 (2019) <https://doi.org/10.1016/j.ultrasmedbio.2019.07.557>
- [28] Bajura, M., Fuchs, H., Ohbuchi, R.: Merging virtual objects with the real world: Seeing ultrasound imagery within the patient. *ACM SIGGRAPH Comput. Graph.* **26**(2), 203–210 (1992) <https://doi.org/10.1145/142920.134061>
- [29] Sakas, G.: Trends in medical imaging: From 2d to 3d. *Comput. Graph.* **26**(4), 577–587 (2002) [https://doi.org/10.1016/S0097-8493\(02\)00103-6](https://doi.org/10.1016/S0097-8493(02)00103-6)
- [30] Israel, O., Mariani, S., Giammarile, F.: Two decades of spect/ct – the coming of age of a technology: An updated review of literature evidence. *Eur. J. Nucl. Med. Mol. Imaging* **46**(10), 1990–2012 (2019) <https://doi.org/10.1007/s00259-019-04404-6>
- [31] Frush, D.P., Callahan, M.J., Coley, B.D., Nadel, H.R., Guillerman, R.P.: Comparison of the different imaging modalities used to image pediatric oncology patients: A cog diagnostic imaging committee/spr oncology committee white paper. *Pediatr. Blood Cancer* **70**(S4), 30298 (2023) <https://doi.org/10.1002/pbc.30298>
- [32] Ronkainen, A.-P., Al-Gburi, A., Liimatainen, T., Matikka, H.: A dose-neutral image quality comparison of different cbct and ct systems using paranasal sinus imaging protocols and phantoms. *Eur. Arch. Otorhinolaryngol.* **279**(9), 4407–4414 (2022) <https://doi.org/10.1007/s00405-022-07271-4>

- [33] Antonelli, M., Reinke, C., Bakas, A., *et al.*: The medical segmentation decathlon. Nat. Commun. **13**(1), 4128 (2022) <https://doi.org/10.1038/s41467-022-30695-9>
- [34] Armato, S.G., McLennan, G., Bidaut, L., *et al.*: The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. Med. Phys. **38**(2), 915–931 (2011) <https://doi.org/10.1118/1.3528204>
- [35] Setio, A.A.A., Ciompi, F., Litjens, G., *et al.*: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge. Med. Image Anal. **42**, 1–13 (2017) <https://doi.org/10.1016/j.media.2017.06.015>
- [36] Bilic, P., Christ, P., Vorontsov, E., *et al.*: The liver tumor segmentation benchmark (lits). Med. Image Anal. **84**, 102680 (2023) <https://doi.org/10.1016/j.media.2022.102680>
- [37] Heller, N., Isensee, M., Maier-Hein, K., *et al.*: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv (2019) <https://doi.org/10.48550/ARXIV.1904.00445>
- [38] Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., Luo, P.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2022)
- [39] Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), 230024 (2023) <https://doi.org/10.1148/ryai.230024>
- [40] Litjens, G., Kooi, T., Bejnordi, B.E., *et al.*: Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. Med. Image Anal. **18**(2), 359–373 (2014) <https://doi.org/10.1016/j.media.2013.12.002>
- [41] Kavur, A.E., Depeursinge, E.C., Sanchez, C.I., *et al.*: Chaos challenge – combined (ct-mr) healthy abdominal organ segmentation. Med. Image Anal. **69**, 101950 (2021) <https://doi.org/10.1016/j.media.2020.101950>
- [42] Raudaschl, P.F., Zaffino, C., Sharp, M., *et al.*: Evaluation of segmentation methods on head and neck ct: Auto-segmentation challenge 2015. Med. Phys. **44**(5), 2020–2036 (2017) <https://doi.org/10.1002/mp.12197>
- [43] Isensee, F., *et al.*: nnU-Net: Self-Adapting Framework for U-Net-Based Medical Image Segmentation. arXiv:1809.10486 (2018). <https://doi.org/10.48550/arXiv.1809.10486>

- [44] Qu, C., Chen, Y., Zhang, W., *et al.*: Post-training quantization for 3d medical image segmentation: A practical study on real inference engines. arXiv (2025) <https://doi.org/10.48550/ARXIV.2501.17343>
- [45] Qin, D., *et al.*: Efficient medical image segmentation based on knowledge distillation. IEEE Trans. Med. Imaging **40**(12), 3820–3831 (2021) <https://doi.org/10.1109/TMI.2021.3098703>
- [46] Das, B.K., Singh, A., Islam, S., Zhao, G., Maier, A.: Segresmamba: An efficient architecture for 3d medical image segmentation. arXiv (2025) <https://doi.org/10.48550/ARXIV.2503.07766>
- [47] Lin, C.-C., Ruan, S.-J., Wang, Y.-J.: Mem-unet: Morphology-enhanced 3d mamba unet for esophagus segmentation. IEEE Journal of Biomedical and Health Informatics (2026)
- [48] IEEE Spectrum: Edge-AI. <https://spectrum.ieee.org/edge-ai>. Accessed: Dec. 3, 2025
- [49] Banbury, C., *et al.*: Micronets: neural network architectures for deploying tinymml applications on commodity microcontrollers. arXiv (2020) <https://doi.org/10.48550/ARXIV.2010.11267>
- [50] Li, Q., Huang, X., Zhang, Y., Fang, B., Hong, D., Chen, J.: Empowering 2d neural network for 3d medical image segmentation via neighborhood information fusion. Pattern Recognition, 113035 (2026)
- [51] Xu, K., Guo, X., Pan, B., Zhang, Y., Liu, Y., Zhang, X., Zeng, X., Liu, Y.: Litefanet: A lightweight unet-based fusion-attention segmentation network for 2d and 3d medical images. Biomedical Signal Processing and Control **112**, 108632 (2026)
- [52] Jian, S., Wei, G., Tian, Y., Ai, Y.: Swinmamba-match: An efficient unsupervised 3d medical image registration network based on swinmamba. Biomedical Signal Processing and Control **116**, 109611 (2026)
- [53] Carrera-Rivera, A., Ochoa, W., Larrinaga, F., Lasa, G.: How-to conduct a systematic literature review: a quick guide for computer science research. MethodsX **9**, 101895 (2022) <https://doi.org/10.1016/j.mex.2022.101895>
- [54] Ullah, N., Khan, S.U., Niazi, M., Esposito, M., Khan, A.A., Nasir, J.A.: Solutions to cybersecurity challenges in secure vehicle-to-vehicle communications: a multivocal literature review. Inf. Softw. Technol. **179**, 107639 (2025) <https://doi.org/10.1016/j.infsof.2024.107639>
- [55] Inamullah, I., Razzak, I., Jameel, S.: The eye as a window to systemic health: A survey of retinal imaging from classical techniques to oculomics. The Journal

- [56] DeepSeek-AI: DeepSeek-V3 technical report. arXiv (2024). <https://doi.org/10.48550/ARXIV.2412.19437>
- [57] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021). Lect. Notes Comput. Sci., vol. 12962, pp. 272–284. Cham, Switzerland (2022). https://doi.org/10.1007/978-3-031-08999-2_22
- [58] Zhou, H.-Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE Trans. Image Process. **32**, 4036–4045 (2023) <https://doi.org/10.1109/TIP.2023.3293771>
- [59] Zhao, J., Hong, T., Qi, H., Zhou, Z., Wang, H.: A lightweight 3d distillation volumetric transformer for 3d mri super-resolution. IEEE J. Biomed. Health Inform. **29**(7), 5083–5094 (2025) <https://doi.org/10.1109/JBHI.2025.3555603>
- [60] Loussaief, E.B., Rashwan, H.A., Ayad, M., Khalid, A., Puig, D.: Adaptive weighted multi-teacher distillation for efficient medical imaging segmentation with limited data. Knowl.-Based Syst. **315**, 113196 (2025) <https://doi.org/10.1016/j.knosys.2025.113196>
- [61] Zhao, Q., *et al.*: Efficient multi-organ segmentation from 3d abdominal ct images with lightweight network and knowledge distillation. IEEE Trans. Med. Imaging **42**(9), 2513–2523 (2023) <https://doi.org/10.1109/TMI.2023.3262680>
- [62] Shen, Y., *et al.*: Fastsam3d: an efficient segment anything model for 3d volumetric medical images. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lect. Notes Comput. Sci., vol. 15012, pp. 542–552. Cham, Switzerland (2024). https://doi.org/10.1007/978-3-031-72390-2_51
- [63] Serrador, L., Villani, F.P., Moccia, S., Santos, C.P.: Knowledge distillation on individual vertebrae segmentation exploiting 3d u-net. Comput. Med. Imaging Graph. **113**, 102350 (2024) <https://doi.org/10.1016/j.compmedimag.2024.102350>
- [64] Wen, M., Zhang, X.: Lpm-net: lightweight pixel-level modeling network based on cnn and mamba for 3d medical image fusion. Inf. Fusion **123**, 103306 (2025) <https://doi.org/10.1016/j.inffus.2025.103306>
- [65] Viraj, A.: MobileSeg3D: a lightweight framework for multi-modality 3D medical image segmentation. CVPR 2025 Workshop MedSegFM Submission (2025)

- [66] Pang, Y., Liang, J., Huang, T., Chen, H., Li, Y., Li, D., Huang, L., Wang, Q.: Slim unetr: Scale hybrid transformers to efficient 3d medical image segmentation under limited computational resources. *IEEE Trans. Med. Imaging* **43**(3), 994–1005 (2024) <https://doi.org/10.1109/TMI.2023.3326188>
- [67] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: 2022 IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp. 574–584 (2022). <https://doi.org/10.1109/WACV51458.2022.00181>
- [68] Shaker, A.M., Maaz, M., Rasheed, H., Khan, S., Yang, M.-H., Khan, F.S.: Unetr++: Delving into efficient and accurate 3d medical image segmentation. *IEEE Trans. Med. Imaging* **43**(9), 3377–3390 (2024) <https://doi.org/10.1109/TMI.2024.3398728>
- [69] Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. *Lect. Notes Comput. Sci.*, vol. 12901, pp. 109–119. Cham, Switzerland (2021). https://doi.org/10.1007/978-3-030-87193-2_11
- [70] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* (2021) <https://doi.org/10.48550/ARXIV.2102.04306>
- [71] Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth Int. Conf. on 3D Vision (3DV), Stanford, CA, USA, pp. 565–571 (2016). <https://doi.org/10.1109/3DV.2016.79>
- [72] Lannelongue, L., Grealey, J., Inouye, M.: Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science* **8**(12), 2100707 (2021) <https://doi.org/10.1002/advs.202100707>
- [73] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *arXiv* (2015) <https://doi.org/10.48550/arXiv.1503.02531>
- [74] Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (ICLR) (2015)
- [75] Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (ICLR) (2017)

- [76] Zhao, Q., Wang, H., Wang, G.: Lcov-net: a lightweight neural network for covid-19 pneumonia lesion segmentation from 3d ct images. In: 2021 IEEE 18th Int. Symp. Biomed. Imaging (ISBI), Nice, France, pp. 42–45 (2021). <https://doi.org/10.1109/ISBI48211.2021.9434023>
- [77] Zhong, Y., Wang, S., Miao, Y., Zhang, T., Li, H.: Lightweight brain tumor segmentation through wavelet-guided iterative axial factorization attention. *Brain Sci.* **15**(6), 613 (2025) <https://doi.org/10.3390/brainsci15060613>
- [78] Chetty, G., Yamin, M., White, M.: A low-resource 3d u-net based deep learning model for medical image analysis. *Int. J. Inf. Technol.* **14**(1), 95–103 (2022) <https://doi.org/10.1007/s41870-021-00850-4>
- [79] Sui, Y., Yin, M., Xie, Y., Phan, H., Zonouz, S.A., Yuan, B.: CHIP: Channel independence-based pruning for compact neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 24604–24616 (2021)
- [80] Dinsdale, N.K., Jenkinson, M., Namburete, A.I.L.: Stamp: Simultaneous training and model pruning for low data regimes in medical image segmentation. *Med. Image Anal.* **81**, 102583 (2022) <https://doi.org/10.1016/j.media.2022.102583>
- [81] Li, H., *et al.*: Prism lite: a lightweight model for interactive 3d placenta segmentation in ultrasound. In: Colliot, O., Mitra, J. (eds.) *Medical Imaging 2025: Image Processing*, p. 12. SPIE, San Diego, USA (2025). <https://doi.org/10.1117/12.3047410>
- [82] Valverde, J.M., Shatillo, A., Tohka, J.: Sauron u-net: simple automated redundancy elimination in medical image segmentation via filter pruning. *Neurocomputing* **594**, 127817 (2024) <https://doi.org/10.1016/j.neucom.2024.127817>
- [83] Lin, X., Yu, L., Cheng, K.-T., Yan, Z.: The lighter the better: rethinking transformers in medical image segmentation through adaptive pruning. *IEEE Trans. Med. Imaging* **42**(8), 2325–2337 (2023) <https://doi.org/10.1109/TMI.2023.3247814>
- [84] Hong, Y., Zhang, L., Ye, X., Mo, J.: Slimformer-3d: a layer-adaptive lightweight transformer for efficient 3d medical image segmentation. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., Park, J. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Lect. Notes Comput. Sci.*, vol. 15961, pp. 537–546. Cham, Switzerland (2026). https://doi.org/10.1007/978-3-032-04937-7_51
- [85] Nuechterlein, N., Mehta, S.: 3d-espnet with pyramidal refinement for volumetric brain tumor image segmentation. In: Crimi, A., Bakas, S., Kuijff, H., Keyvan, F., Reyes, M., Walsum, T.V. (eds.) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Lect. Notes Comput. Sci.*, vol. 11384, pp. 245–253.

- Cham, Switzerland (2019). https://doi.org/10.1007/978-3-030-11726-9_22
- [86] Kim, Y.B., Le, T.V., Lee, J.Y.: Lightweight brain mr image super-resolution using 3d convolution. *Multimed. Tools Appl.* **83**(3), 8785–8795 (2024) <https://doi.org/10.1007/s11042-023-15969-8>
 - [87] Paschali, M., Gasperini, S., Roy, A.G., Fang, M.Y.-S., Navab, N.: 3dq: compact quantized neural networks for volumetric whole-brain segmentation. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. *Lect. Notes Comput. Sci.*, vol. 11766, pp. 438–446. Cham, Switzerland (2019). https://doi.org/10.1007/978-3-030-32248-9_49
 - [88] Memon, K., *et al.*: Edge computing for ai-based brain mri applications: a critical evaluation of real-time classification and segmentation. *Sensors* **24**(21), 7091 (2024) <https://doi.org/10.3390/s24217091>
 - [89] Kim, S., Cha, M., Yi, C., Nguyen, X.T.: Towards eight-bit quantization for 3d u-net medical image segmentation via roi-based calibration and background-aware shift. In: *2024 IEEE Int. Conf. Consumer Electron.-Asia (ICCE-Asia)*, Da Nang, Vietnam, pp. 1–3 (2024). <https://doi.org/10.1109/ICCE-Asia63397.2024.10774036>
 - [90] Wang, W., Lin, B.: Optimizing 3d u-net-based brain tumor segmentation with integer-arithmetic deep learning accelerators. *ACM J. Emerg. Technol. Comput. Syst.* **18**(2), 1–16 (2022) <https://doi.org/10.1145/3495210>
 - [91] Zhou, X., Li, X., Hu, K., Zhang, Y., Chen, Z., Gao, X.: Erv-net: an efficient 3d residual neural network for brain tumor segmentation. *Expert Syst. Appl.* **170**, 114566 (2021) <https://doi.org/10.1016/j.eswa.2021.114566>
 - [92] Li, Z., Pan, J., Wu, H., Wen, Z., Qin, J.: Memory-efficient automatic kidney and tumor segmentation based on non-local context guided 3d u-net. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racocanu, D., Joskowicz, L. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. *Lect. Notes Comput. Sci.*, vol. 12264, pp. 197–206. Cham, Switzerland (2020). https://doi.org/10.1007/978-3-030-59719-1_20
 - [93] Song, L., Yi, J., Peng, J.: Lcc-net: a lightweight cross-consistency network for semi-supervised cardiac mr image segmentation. *Comput. Math. Methods Med.* **2021**, 1–9 (2021) <https://doi.org/10.1155/2021/9960199>
 - [94] Shahid, G.E.S., *et al.*: Liu-net: lightweight inception u-net for efficient brain tumor segmentation from multimodal 3d mri images. *PeerJ Comput. Sci.* **11**, 2787 (2025) <https://doi.org/10.7717/peerj-cs.2787>

- [95] Zhong, J., *et al.*: Pmfsnet: polarized multi-scale feature self-attention network for lightweight medical image segmentation. *Comput. Methods Programs Biomed.* **261**, 108611 (2025) <https://doi.org/10.1016/j.cmpb.2025.108611>
- [96] Yaniv, O., Portnoy, O., Talmon, A., Kiryati, N., Konen, E., Mayer, A.: V-net light – parameter-efficient 3-d convolutional neural network for prostate mri segmentation. In: 2020 IEEE 17th Int. Symp. Biomed. Imaging (ISBI), Iowa City, IA, USA, pp. 442–445 (2020). <https://doi.org/10.1109/ISBI45749.2020.9098643>
- [97] Pang, B., Chen, L., Tao, Q., Wang, E., Yu, Y.: Ga-unet: a lightweight ghost and attention u-net for medical image segmentation. *J. Imaging Inform. Med.* **37**(4), 1874–1888 (2024) <https://doi.org/10.1007/s10278-024-01070-5>
- [98] Kuang, H., *et al.*: Lw-ctrans: a lightweight hybrid network of cnn and transformer for 3d medical image segmentation. *Med. Image Anal.* **102**, 103545 (2025) <https://doi.org/10.1016/j.media.2025.103545>
- [99] Wang, P., Liu, S., Peng, J.: Ast-net: lightweight hybrid transformer for multi-modal brain tumor segmentation. In: 2022 26th Int. Conf. Pattern Recognit. (ICPR), Montreal, QC, Canada, pp. 4623–4629 (2022). <https://doi.org/10.1109/ICPR56361.2022.9956705>
- [100] Wang, X., *et al.*: Bix-nas: searching efficient bi-directional architecture for medical image segmentation. In: Bruijne, M.D., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. *Lect. Notes Comput. Sci.*, vol. 12901, pp. 229–238. Cham, Switzerland (2021). https://doi.org/10.1007/978-3-030-87193-2_22
- [101] Said, Y., *et al.*: Ai-driven genetic algorithm-optimized lung segmentation for precision in early lung cancer diagnosis. *Sci. Rep.* **15**(1), 23058 (2025) <https://doi.org/10.1038/s41598-025-08116-w>
- [102] Celaya, A., Actor, J.A., Muthusivarajan, R., Gates, E., Chung, C., Schellinghout, D., Riviere, B., Fuentes, D.: Pocketnet: A smaller neural network for medical image analysis. *IEEE Trans. Med. Imaging* **42**(4), 1172–1184 (2023) <https://doi.org/10.1109/TMI.2022.3224873>
- [103] Yu, M., *et al.*: Rirgan: an end-to-end lightweight multi-task learning method for brain mri super-resolution and denoising. *Comput. Biol. Med.* **167**, 107632 (2023) <https://doi.org/10.1016/j.compbiomed.2023.107632>
- [104] Ali, O., Ali, H., Shah, S.A.A., Shahzad, A.: Implementation of a modified u-net for medical image segmentation on edge devices. *IEEE Trans. Circuits Syst. II: Express Briefs* **69**(11), 4593–4597 (2022) <https://doi.org/10.1109/TCSII.2022.3181132>

- [105] RapidAI evaluation study authors: Rapidai compared with human readers of acute stroke imaging for detection of intracranial vessel occlusion. *Stroke: Vascular and Interventional Neurology* (2024) <https://doi.org/10.1161/SVIN.123.001145>
- [106] Littlejohns, T.J., Holliday, J., Gibson, L.M., *et al.*: The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications* **11**, 2624 (2020) <https://doi.org/10.1038/s41467-020-15948-9>
- [107] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2014)
- [108] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: *International Conference on Learning Representations (ICLR)* (2014)
- [109] Khader, F., *et al.*: Denoising diffusion probabilistic models for 3d medical image generation. *Sci. Rep.* **13**(1), 7303 (2023) <https://doi.org/10.1038/s41598-023-34341-2>
- [110] Rafiq, M., Ali, H., Mujtaba, G., Shah, Z., Azmat, S.: Cross modality medical image synthesis for improving liver segmentation. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* **13**(1), 2476702 (2025) <https://doi.org/10.1080/21681163.2025.2476702>
- [111] Bortsova, G., *et al.*: Adversarial attack vulnerability of medical image analysis systems: unexplored factors. *Med. Image Anal.* **73**, 102141 (2021) <https://doi.org/10.1016/j.media.2021.102141>
- [112] Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. *arXiv* (2019) <https://doi.org/10.48550/arXiv.1904.00625>
- [113] Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. *Med. Image Anal.* **67**, 101840 (2021) <https://doi.org/10.1016/j.media.2020.101840>
- [114] Maier, O., Menze, C.M., Gablentz, J., *et al.*: Isles 2015 – a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Med. Image Anal.* **35**, 250–269 (2017) <https://doi.org/10.1016/j.media.2016.07.009>
- [115] Roth, H.R., Oda, H., Zhou, X., Shimizu, N., Yang, Y., Hayashi, Y., Oda, M., Fujiwara, M., Misawa, K., Mori, K.: An application of cascaded 3d fully convolutional networks for medical image segmentation. *Comput. Med. Imaging Graph.* **66**, 90–99 (2018) <https://doi.org/10.1016/j.compmedimag.2018.03.001>
- [116] Kooi, T., *et al.*: Large-scale deep learning for computer-aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017) <https://doi.org/10.1016/j.media.2016.07.009>

1016/j.media.2016.07.007

- [117] Moore, J., *et al.*: Ome-ngff: a next-generation file format for expanding bioimaging data-access strategies. Nat. Methods **18**(12), 1496–1498 (2021) <https://doi.org/10.1038/s41592-021-01326-w>