



Heriot-Watt University
Research Gateway

Is attention all you need in medical image analysis? A review

Citation for published version:

Papanastasiou, G, Dikaios, N, Huang, J, Wang, C & Yang, G 2024, 'Is attention all you need in medical image analysis? A review', *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1398-1411. <https://doi.org/10.1109/JBHI.2023.3348436>

Digital Object Identifier (DOI):

[10.1109/JBHI.2023.3348436](https://doi.org/10.1109/JBHI.2023.3348436)

Link:

[Link to publication record in Heriot-Watt Research Portal](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

IEEE Journal of Biomedical and Health Informatics

Publisher Rights Statement:

© 2023 The Authors.





General rights

Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Is Attention all You Need in Medical Image Analysis? A Review

Giorgos Papanastasiou , Nikolaos Dikaios , Jiahao Huang , Chengjia Wang , and Guang Yang 

Abstract—Medical imaging is a key component in clinical diagnosis, treatment planning and clinical trial design, accounting for almost 90% of all healthcare data. CNNs achieved performance gains in medical image analysis (MIA) over the last years. CNNs can efficiently model local pixel interactions and be trained on small-scale MI data. Despite their important advances, typical CNN have relatively limited capabilities in modelling “global” pixel interactions, which restricts their generalisation ability to understand out-of-distribution data with different “global” information. The recent progress of Artificial Intelligence gave rise to Transformers, which can learn global relationships from data. However, full Transformer models need to be trained on large-scale data and involve tremendous computational complexity. Attention and Transformer compartments (“Transf/Attention”) which can well maintain properties for modelling global relationships, have been proposed as lighter alternatives of full Transformers. Recently, there is an increasing trend to co-pollinate complementary local-global properties from CNN and Transf/Attention architectures, which led to a new era of hybrid models. The past years have witnessed substantial growth in hybrid CNN-Transf/Attention models across diverse MIA problems. In this systematic review, we survey existing hybrid CNN-Transf/Attention models, review and unravel key architectural designs, analyse breakthroughs, and evaluate current and future opportunities as well as challenges. We also introduced an analysis framework on generalisation opportunities of scientific and clinical impact, based on which new data-driven domain generalisation and adaptation methods can be stimulated.

Index Terms—Attention, computed tomography, convolutional neural networks, magnetic resonance imaging,

Manuscript received 24 July 2023; revised 17 November 2023; accepted 21 December 2023. Date of publication 29 December 2023; date of current version 7 March 2024. This work was supported in part by the MIS under Grant 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by EU under the NextGenerationEU Program, in part by ERC IMI under Grant 101005122, in part by H2020 under Grant 952172, in part by MRC under Grant MC/PC/21013, in part by Royal Society under Grant IEC\NSFC\211235, and in part by UKRI Future Leaders Fellowship under Grant MR/V023799/1. (Corresponding author: Giorgos Papanastasiou; Guang Yang.)

Giorgos Papanastasiou is with the Archimedes Unit, Athena Research Centre, 15125 Athens, Greece (e-mail: g.papanastasiou@athenarc.gr).

Nikolaos Dikaios is with the Mathematics Research Centre, Academy of Athens, 10679 Athens, Greece (e-mail: ndikaios@academyofathens.gr).

Jiahao Huang and Guang Yang are with the Bioengineering Department and Imperial-X, Imperial College London, W12 7SL London, U.K. (e-mail: j.huang21@imperial.ac.uk; g.yang@imperial.ac.uk).

Chengjia Wang is with the School of Mathematical and Computer Sciences, Heriot Watt, EH14 4AS Edinburgh, U.K. (e-mail: chengjia.wang@hw.ac.uk).

Digital Object Identifier 10.1109/JBHI.2023.3348436

medical image analysis, positron emission tomography, retinal imaging, transformers.

I. INTRODUCTION

A. Medical Image Analysis and Convolutions

MEDICAL imaging (MI) is a key component in clinical diagnosis, treatment planning, and clinical trial design, accounting for almost 90% of all healthcare data [1], [2]. Medical image analysis (MIA)-derived imaging biomarkers can improve early disease diagnosis, therapy design and treatment response monitoring, beyond visual radiology assessments. MIA is an important component of clinical research, innovation and application [3], [4], [5], [6].

The European Society of Radiology in coordination with the Radiological Society of North America, has recently provided recommendations for clinically validating MIA techniques [3], [4], [5]. In the era of rapid artificial intelligence (AI) developments and to establish the clinical translation of AI, it is important to review and develop guidelines for innovative AI models.

Since the first “deep” convolutional neural network (CNN) developed by Krizhevsky et al. in 2012 which outperformed the previous state-of-the-art (SOTA, non-deep learning) algorithms on the ImageNet dataset [7], CNNs demonstrated numerous performance gains across all MIA tasks: segmentation, classification, reconstruction, synthesis, denoising, registration, and regression [2]. However, typical CNNs focus on modelling information through small convolutional filter footprints and shared weights, which comes at the cost of introducing local receptive fields thus, limiting their ability to directly model long-range (global) pixel interactions within images. Hence, despite their important advances, CNN-based networks are still focusing on local-scale modelling, with low generic “local-global” modelling capabilities. Their limited ability to model both local and global information from images adds barriers to model generalisability (e.g., across MIA domains or pathology settings) and transfer learning (from one MI modality to another) properties of pure CNN models [2].

B. Hybridisation With Attention Convolutions

First introduced by Bahdanau et al. in 2014, the attention mechanism was initially designed to learn long-range dependencies in natural language processing and improve machine translation [8]. The attention mechanism allows to (soft-)search for a set of positions in a source sentence where the most relevant information is concentrated, encouraging the model to predict a target word based on the context vectors associated with these source positions and all the previous generated target words

[8]. Following attention, the development of the self-attention mechanism in 2016 was designed so that each position (building block) within a self-attention layer (known as query, key and value) can attend to all positions in the output of the previous layer [9], [10], as an additional technique to enhance modelling of long-range dependencies.

The introduction of self-attention and attention mechanisms in the Transformer models made it possible to increase the receptive field and thus, became an efficient solution for modelling long-range dependencies from images [10], [11], [12], with promising results in the field of MIA [13], [14], [15], [16]. The Vision Transformer (ViT) models recruit consecutive multi-head self-attention and attention mechanisms in image patches and have been suggested to even fully replace pure CNN models [10]. The basic concept in ViT is to convert input images to a series of image patches which in turn are transformed into vectors and can be represented as “words” in a normal Transformer. However, as the relationships between an image patch and all other image patches are computed, the computational complexity of the multi-head self-attention modules in ViT becomes quadratic to image size, adding substantial challenges in the setting of analysing high spatial resolution images. Swin Transformers (ST) were designed to overcome these challenges by performing self-attention in non-overlapped image patches [11], [12]. Despite this, ST need to consecutively learn a stack of two successive self-attention blocks with regular and shifted windowing configurations, respectively. This adds computational complexity and limits their applicability in MIA tasks such as segmentation, classification, denoising, reconstruction and registration, where dense predictions at the pixel level and learning representations from high content images are necessary. This is one of the main reasons why full ViT and ST models have been limited to medical image classification and object detection tasks [10], [12], [13], [14], [15], [16], [17].

To reduce the computational complexity and to address both local and global learning in MIA, self-attention and Transformer blocks were incorporated into CNN model architectures (thereafter called as “hybrid” CNN-Transf/Attention models), giving rise to hybrid models. Current evidence shows that by combining local and global modelling capabilities, these hybrid CNN-Transf/Attention models consistently outperform previous SOTA techniques across different MIA tasks [13], [14], [15], [16], [17], [18], [19], [20], [21]. Hybrid models can potentially be also used to improve model interpretability [22], [23], [24], [25].

However, the main drawback of these hybrid models is that they are enormously complex as they have been developed to address particular problems in MIA, which means that their domain generalizability (e.g., from CT to MRI, or from lung to cardiac applications) and transfer learning capabilities can be challenging processes. Given their substantial growth, it is important to methodically assess whether these techniques can generalise across imaging modalities, MIA tasks and clinical applications, or may be over-engineered to specific MIA problems.

In this work, we review the evolution of the hybrid models for in vivo MI: magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), ultrasound, X-rays and retinal imaging. There are numerous recent surveys that describe technical details of CNN models and how these were used to address specific needs in MI [2], [26], [27], [28], [29], as well as some recent survey on ViT in MI [17],

[30], [31], [32], [33]. Differing from previous reviews, we developed a comprehensive systematic review based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for hybrid CNN-Transf/Attention models in MI. We categorised published work on hybrid CNN-Transf/Attention models in MI, analysed key architectural designs and quantitatively as well as qualitatively unravelled the evolution of CNN-Transf/Attention models.

To improve clarity and understanding on these novel techniques, we critically review whether such hybrid models outperform their pure CNN counterparts. We review technical and computational complexities and discuss domain generalization strategies, based on the MI modality, downstream task, and clinical application. We focus on unravelling the importance and potential drawbacks of hybrid models. Finally, we discuss opportunities, challenges (with mitigations, where applicable) and future perspectives of the post-hybrid model era. We consider these review concepts as important pathways towards harmonising and translating these novel techniques into clinically meaningful MIA.

II. METHODS

A. Literature Review Strategy

We performed a systematic review of CNN-Transf/Attention models in MI published between January 1, 2019 and July 1, 2022 using Scopus, Web of Science and Pubmed, based on the PRISMA framework [34]. In our review, we refer to all hybrid models that involve any CNN and Transformer modules-including adaptations of self-attention and attention mechanisms, as hybrid CNN-Transf/Attention models. We only considered MI modalities that involve in vivo body imaging and thus, excluded microscopy and digital pathology slide imaging studies. Therefore, we focused on MRI, CT, PET-CT, ultrasound, retinal imaging and X-rays.

Initial filtering: To broaden the research, we initially mined all publications by searching the following keywords in the abstract, title, and manuscript keywords: (transformer OR self-attention) AND (deep AND learning) OR (convolutional AND neural AND network). This led to 5222 papers (see PRISMA flow in Fig. 1(a)). Subsequently, we focused the search by considering all different combinations of relevant keywords in the abstract, title and keywords of each paper, as follows: (transformer OR self-attention) AND (deep AND learning) OR (convolutional AND neural AND network) AND (medical AND imaging) OR (magnetic AND resonance AND imaging) OR (MRI) OR (computed AND tomography) OR (CT) OR (ultrasound) OR (positron AND emission AND tomography) OR TITLE-ABS-KEY (retin) OR TITLE-ABS-KEY (X-rays) OR TITLE-ABS-KEY (ray). By adding these terms, we removed all irrelevant to MI papers, which led to 656 papers from all three digital libraries. By excluding conference, review and archived (non-peer-reviewed) papers, we then removed all non-journal publications, leaving 352 journal papers for subsequent analysis.

Title and abstract screening: All authors screened titles and abstracts across all 352 journal peer-reviewed papers and removed all irrelevant to the field of study papers, leaving 128 papers for full text review.

Full text screening: Following full paper review, the authors removed 16 journal papers (14 non-relevant to MI or hybrid model studies and 2 papers not written in English). In total,

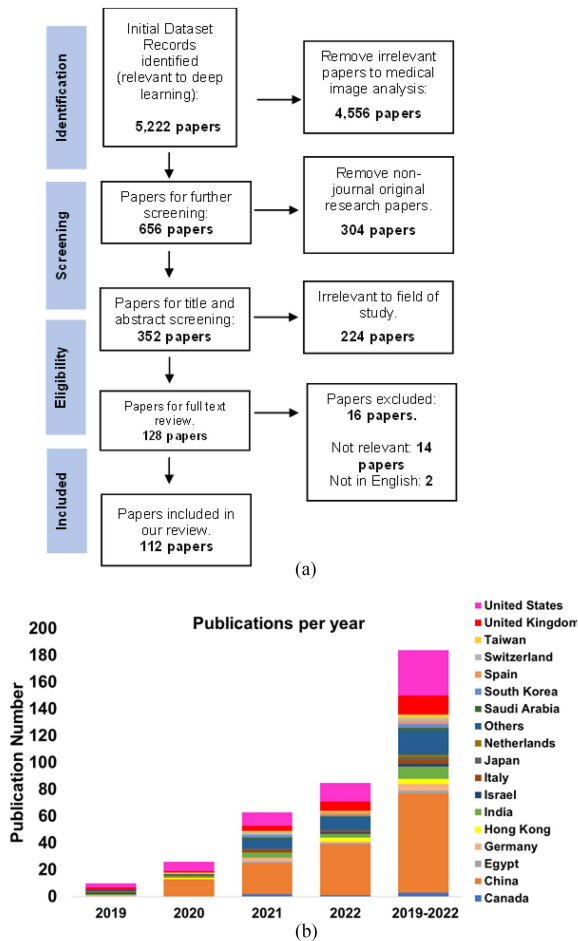


Fig. 1. (a). PRISMA flowchart. The flowchart illustrates inclusion and exclusion of papers at each review stage. (b) Publications per year across the top 18 countries (in terms of publications record). Publications with affiliations from multiple countries have been accumulated on a per country basis.

112 journal papers (thereafter, referred to as “articles”) were included in our review analysis. See also data extraction for paper content that was reviewed.

B. Data Extraction

During evaluation of article full texts, we considered the following aspects: (1) year of publication; (2) MI modality; (3) CNN backbone; (4) Transf/Attention including all different attention subtypes; (5) MIA task (segmentation, classification, reconstruction, synthesis, denoising, registration, regression); (6) organ or physiology system investigated/ imaged; (7) use of public or private data; (8) data augmentation technique used; implementation details: (9) model optimizer, (10) loss function, (11) metric used to evaluate the results and (12) size of training and testing data. We also considered (13) if computation expense (total number of parameters) was calculated and (14) whether performance was improved against non-hybrid baseline methods.

Furthermore, we assessed the articles in terms of generalisability following 2 objective criteria: whether a CNN-Transf/Attention architecture was a) trained on large datasets,

TABLE I
ALL ARTICLES GROUPED BASED ON THE CLINICAL APPLICATION (ORGAN), MI MODALITY, CNN AND TRANSF/ATTENTION MODEL

Organ	MI	CNN model	Self-Attention	Transformer, ViT, ST (full)	Light Transformer, ViT, ST: Encoder(s), Block(s) or Layer(s)	Other Transf/Attention
Brain	MRI	CNN	[36,60,77,83,98,110,115,137]	[56]	[44,46,53,54,56, 65,66,82,103]	[64]
		UNet	[123]	NA	[61,65,88]	[59, 62]
		GAN	[16,20,79,107,114,115,129]	NA	NA	NA
	CT	Other CNN	[15,21,38,63,94]	[39,47]	[65,70,75]	[59,97]
		CNN	NA	NA	[65,66]	[73]
		UNet	[120,144,136,138]	NA	[65]	NA
		GAN	[16,79,107,136,138]	NA	NA	NA
	Other MI	Other CNN	[107,134]	[41]	NA	NA
		UNet	[136]	NA	[65]	NA
		GAN	[20,118,136]	NA	NA	NA
Lung	CT	Other CNN	[119]	NA	NA	NA
		CNN	[18, 57, 91,99]	NA	[65]	NA
		UNet	[18,95,121,136]	NA	[65]	NA
		GAN	[96,114]	NA	NA	NA
	X-rays	Other CNN	[71,95]	NA	[43,55,85]	[84]
		CNN	NA	NA	NA	[80]
		Other CNN	[92]	[42,49,90]	[43,52,55,93]	NA
	Other MI	CNN	NA	NA	[65]	NA
		UNet	[136]	NA	[65]	NA
		GAN	[118,136]	NA	NA	NA
Multiple organs	MRI	CNN	NA	NA	[65,66]	NA
		UNet	[76]	NA	[65]	NA
		Other CNN	[15,111]	NA	[65,70,109]	NA
		CNN	[113,124]	NA	[65,66]	NA
	CT	UNet	[75,136]	NA	[65]	[127]
		GAN	[124,136]	NA	NA	NA
		Other CNN	[71]	[41]	[65]	[71]
	Other MI	CNN	[35,117]	NA	NA	NA
		UNet	[109]	NA	NA	NA
		GAN	[118]	NA	NA	NA
Retina	All Retinal imaging	CNN	[100,106,133,135]	NA	NA	[105]
	UNet	[45,125,130]	NA	[51]	NA	
	Other CNN	[112]	[41,50]	[40,89]	NA	
Other organ	All MI	CNN	[13,35,68,100,113,117,124, 139]	NA	[37,65,66,69]	NA
		UNet	[74,76,86,108,126,128,131, 136]	NA	[65,67]	[72,102,127]
		GAN	[14,19,87,118,124,132,136]	NA	NA	[19]
		Other CNN	[13,15,48,71,73,96,101, 111,116,122]	[41]	[65,70,81,109]	[13,71,111, 122]

To keep the information concise, details are prioritised for the top 4 organ areas (brain, lung, multiple organs and retina) in terms of prevalence, the top 2 MI modalities present per organ and the top 3 CNN (ALL) model types per organ. Transf/Attention modules were categorised to: a) Self-Attention, b) Transformer, ViT or ST (full models) and c) Light Transformer, ViT or ST: Encoder(s), Block(s) or Layer(s). All other organs, MI modalities, CNN and Transf/Attention modules are grouped under the term “Other”. Studies occurring in >1 Table cell correspond to model combinations. Missing rows of CNN models corresponds to absence of this technique per MI modality. MI: medical imaging, ViT: Vision Transformer, ST: Swin Transformer.

b) analysed data from heterogeneous MI modalities (e.g., different MRI or CT “sequences”, or MRI and CT, etc.) and/ or multi-modal data (image and text, images and genetics) and/ or was applied to multiple (≥ 2) organ areas (e.g., brain and heart) and/ or multiple (≥ 2) datasets of the same modality and organ. Further, we identify challenges, opportunities and future trends that can be used as suggestions for future work in this field.

III. RESULTS

A. Research Trends

We studied published work on hybrid CNN-Transf/Attention models in MI and observed a consistent increase of these models in 2021 and 2022, against the first 2 years of our observation window: in the period 2019–2022, there were 7, 20, 31 and 54 articles published, respectively.

In Fig. 1(a), we present the PRISMA flow used to search and review articles. In Fig. 1(b), we initially measured the country origin as derived from each affiliation across all articles (all affiliations were considered across publications). Considering the entire review period (2019–2022), the first ten countries in were: China (74 publications), USA (34), U.K. (14), India (9), Germany (5), Hong Kong (4), Canada (3), Taiwan (3), South Korea (3) and Italy (3).

Table I demonstrates all the articles grouped based on the MI modality, CNN backbone, Transf/Attention model and clinical application (organ) [13], [14], [15], [16], [17], [18], [19], [20], [21], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92],

TABLE II

ALL THE ARTICLES GROUPED BASED ON THE DOWNSTREAM MIA TASK, THE NUMBER OF DATA AUGMENTATION TECHNIQUES, OPTIMISER, LOSS FUNCTION AND METRIC USED TO EVALUATE RESULTS

Task	MI	Implementation	Studies
Segmentation	Data augmentation	≤ 3 Combined	[18,58,59,65,70,71,74,76,78,100,106,110,119,120,125-127,131]
		> 3 Combined	[13,51,53,61,67,72,105,113,116,119,122,127]
		NR	[45,57,60,62,69,98,102,103,107,108,111,121,123,130,134]
	Optimiser	Adam	[13,18,45,51,53,58-62,65,70-72,74,78,98,100,105,107,108,110,111,113,119,122,126,127,131,134]
		SGD	[67,76,102,106,121,123,125]
		Other	[57,69,103,116,120,130]
	Loss function	Cross entropy-based	[18,45,51,53,57,74,103,105,106,108,110,111,113,116,120,122,126]
		Combination/Other	[13,59,60-62,65,67,69,72,76,98,100,107,119,121,123,125,131,134]
	Metric	Dice	[45,51,60,67,72,98,102,103,127,131]
		Combination	[13,18,53,57-59,61,62,65,69-71,74,76,78,106-108,110,111,113,119-123,125,126,134]
Classification	Data augmentation	≤ 3 Combined	[35,42-43,84,85,89,96,125,133]
		> 3 Combined	[52,55,93]
		NR	[21,36-40,46,48-50,63,86,87,92,94,95,97]
	Optimiser	Adam	[21,35,37-39,43,46,49,63,84,87,89,92,95,97,133]
		SGD	[36,42,48,52,86,90]
		Other	[40,93]
	Loss function	Cross entropy-based	[21,35-40,42-46,52,55,63,84-86,89,92-95,133]
		Combination/Other	[48,55,87,93]
	Metric	ROC/ AUC/ACC	[21,35-43,46,48,49,52,55,63,84-87,89-97,133]
		Combination	[50-53,55,63,84,89]
Other MIA tasks	Data augmentation	≤ 3 Combined	[13,24,18,55,56,69,70,73,75,76,81,83,85,90,92,98,100,102,103,105,109-112,119,126,127,135,139]
		> 3 Combined	[13,129]
		NR	[15,16,19,20,64,66,68,77,79,80,82,99,101,104,114,115,117,124,128,132,136-138]
	Optimiser	Adam	[13,18,19,14,15,16,20,55,64,66,69,70,73,75-77,79,81-83,85,88,90,92,98-105,109,112,114,115,117-119,124,126-129,135,137-139]
		SGD	[109,135]
		Other	[80,132,136]
	Loss function	Cross entropy-based	[20,73,81,82,135]
		Combination/Other	[13-16,18,19,55,56,64,66,68-70,75-77,79,80,82,83,85,88,90,92,98-105,109,112,114,115,117,118,119,124,126-129,132,136-139]
	Metric	1 metric (e.g., Dice, ROC, PSNR)	[13,18,19,70,73,75,77,85,98-102,104,105,109,112,117,135]
		Combination	[14-16,20,56,66,68,70,79-83,88,105,114,115,117,118,124,127-129,132,136-139]

To keep the information concise, details are prioritised for the top 2 downstream tasks. All other MIA tasks are organised under the "Other MIA tasks". MIA: medical image analysis, NR: not reported, SGD:stochastic gradient descent, ACC: accuracy.

TABLE III

ALL THE ARTICLES GROUPED BASED ON THE DOWNSTREAM MIA TASK, DATA SET (PUBLIC, PRIVATE, BOTH) AND THE DATA SIZE

Task	MI	Implementation	Studies
Segmentation	Data	Public	[13,18,45,51,53,57-62,65,67,70-72,76,78,98,100,102,103,105-108,110,111,103,121,123,125-127]
		Private	[13,69,70,74,98,100,105,110,113,116,119-122,127,130,131,134]
		Both	[13,70,98,100,105,110,113,121,127]
	Data size	≤ 2,000 images	[18,51,53,69,72,74,78,102,103,106,107,110,119,120,131]
		> 2,000 images and ≤ 10,000 images	[13,40,76,105,108,111,113,121,134]
		> 10,000 images	[58,61,62,71,98,100]
	NR	[59,60,67,70,130]	
Classification	Data	Public	[21,35,37,38,42-43,46,48,52,55,63,84,85,89,90-94,97,133]
		Private	[35,36,39,40,49,50,86,87,95,96]
		Both	[35]
	Data size	≤ 2,000 images	[21,35,50,84,85,87,133]
		> 2,000 images and ≤ 10,000 images	[40,43,49]
		> 10,000 images	[39,42,52,55,86,79,90]
	NR	[36,37,38,46,48,63,94-97]	
Other MIA tasks	Data	Public	[13-15,18,19,55,56,64,66,68-70,75-77,80-82,85,90,92,98-100,102-105,109,112,114,115,117,119,124,126,127,129,135]
		Private	[13,24,26,18-20,36,55,64,68-70,73,76,79,82,83,85,88,90,92,98-105,127-119,126-128,132,136-139]
		Both	[13,14,18,19,55,64,68-70,76,82,85,90,92,98-100,102-105,117,119,126,127]
	Data size	≤ 2,000 images	[15,16,20,73,75,77,83,109,112,117,128]
		> 2,000 images and ≤ 10,000 images	[14,66,79,82,88,104,125,135]
		> 10,000 images	[56,80,81,101,114,118,124,129,139]
	NR	[19,74,68,132,136-138]	

To keep the information concise, details are prioritised for the top 2 downstream tasks. All other MIA tasks are organised under the "Other MIA tasks". MIA: medical image analysis, NR: not reported.

[93], [94], [95], [96], [97], [98], [99], [100], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], [124], [125], [126], [127], [128], [129], [130], [131], [132], [133], [134], [135], [136], [137], [138], [139]. Implementation details about the data augmentation technique, optimizer, loss function and the metrics used to evaluate the performance of each hybrid model across studies, are presented in Table II. Table III presents whether public and/ or private were analysed and information about the data size.

B. Experimental Settings and Key Architectural Designs

1) Medical Imaging Modality Recruited: We reviewed the publication record of the MI modality used per year (Fig. 2(a)). Most of the studies involved MRI (50 studies), followed by CT (42), retinal imaging (14), X-rays (12), ultrasound (7) and PET-CT (5). Although MRI was less frequent than CT the first 2 years of our observation time frame, it outnumbered CT in the last two years (2021 and 2022).

2) CNN Model Used: In Fig. 2(b), we demonstrate all CNN backbone models used across studies. Standard CNN architectures have been implemented in most of the studies (40 articles),

followed by UNet (30), GAN (14), ResNet (14), DenseNet (7), None (i.e., no CNN backbone-only Transformer model used) (7), fully connected networks (FCN) (6) and VGG (3).

3) Transformer and Attention Mechanisms: Fig. 2(c) illustrates the evolution of Transf/Attention models recruited per year. It is obvious that self-attention mechanisms have been most widely used (64 studies out of 112 in total), followed by Transformer (22 studies), ViT (9 studies), channel- and spatial-attention (6), ST (4), attention (2) and other (11).

It is known that to exploit the performance capabilities of full Transformer models, a combination of large data and supercomputer facilities are necessary [10], [11], [12]. In our review, there were numerous studies that either analysed relatively small (i.e., <2000 images) data (~27%), and/ or private data alone (~29%) and/ or did not report the data size (~21%). Details are presented in Table III. Furthermore, computational resources were not reported in most studies, with only 29 out of 83 reporting the number of model parameters and/ or time for training. Of note, only 8 out of 112 studies (~7%) described use of full original Transformer, ViT or ST models, with the rest ~93% involving: self-attention, channel- and spatial-attention, attention and simplified and light Transformer versions including transformer blocks (of stacked layers), layers, or encoders (Table I).

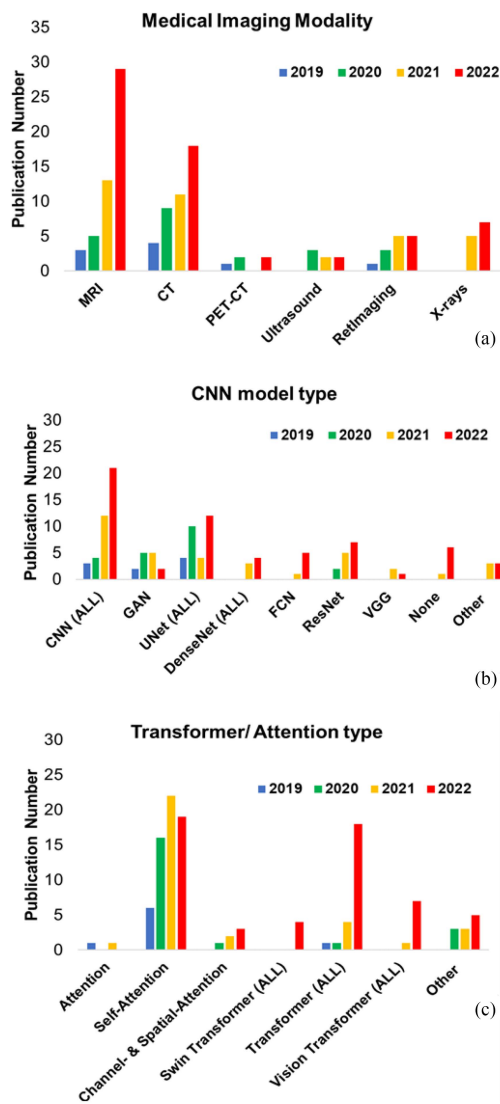


Fig. 2. Publication record over time for the Medical Imaging modality (a), the CNN model type (b), and Transf/Attention architectures (c). In b), the CNN (ALL) term describes all standard CNN models captured across studies: CNN encoder-decoder (E-D), CNN layers, CNN decoder only and CNN (E-D) (3D). GAN describes either GAN or CycleGAN models. UNet (ALL) and DenseNet (ALL) represent 2D and 3D model variants. The term “Other” includes all other models identified across studies (a total of 6 articles): EfficientNet, multi-linear perceptron (MLP), Deep Belief Network and long short-term memory (LSTM). In c), the Transformer, ViT and ST (ALL) include all model adaptations identified across studies: Transformer (full), Transformer encoder(s) only, Transformer blocks or layer(s), ViT (full), ViT encoder(s) only, ViT blocks, ST (full model), ST blocks, ST layers, respectively. The term “Other” includes all distinct architectural adaptations of Transf/Attention mechanisms extracted across studies: DistilGPT2 (1), channel self-attention (3), criss-cross self-attention (2), cross-attention (2), cross spatial-attention (1), multi-head self-attention (1) and spatial self-attention (1).

4) *Medical Image Analysis (Downstream) Task:* Further, we extracted all MIA (downstream) tasks across studies (Fig. 3(a)). Most of the studies aimed to solve segmentation tasks (43 studies), followed by classification (35), reconstruction (14), synthesis (10), denoising (7), localisation (5), regression (4), registration (3) and radiology report generation (2).

5) *Organs Analysed:* We reviewed the organ under investigation across all studies (Fig. 3(b)). Most of the 112 studies

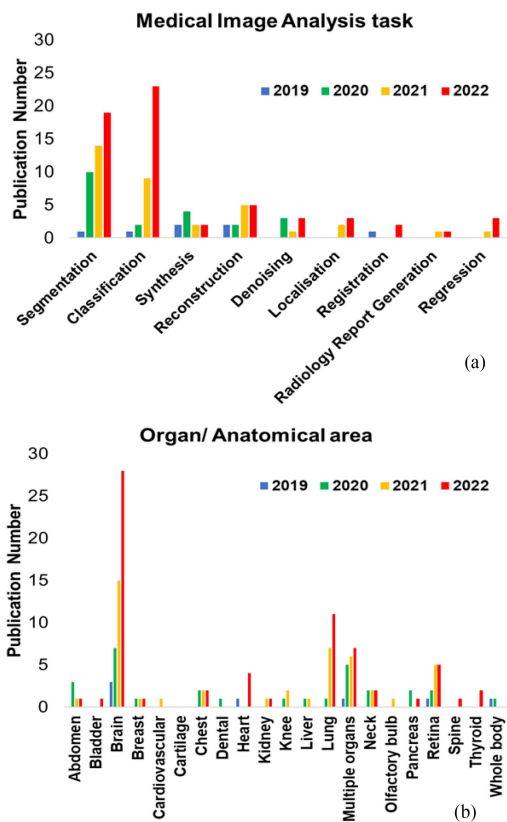


Fig. 3. Publication record over time for the medical image analysis task (a), and the organ/anatomical area (clinical application) under investigation (b), respectively. The term “whole body” (Table I, Fig. 3(b)) by Xue et al. [118] and Dong et al. [136] describes simultaneous imaging covering the entire body using a single PET-CT session, a dedicated full body technique that has been recently developed [2]. The term “multiple organs” describes all studies that imaged more than one organ in the same imaging setting [15], [19], [35], [41], [65], [66], [76], [70], [71], [76], [100], [108], [109], [111], [113], [117], [124], [127], including whole body studies [118], [136].

analysed medical images from the brain (53 studies), followed by lung (20), multiple organs (20), retina (13), chest (6), neck (6), abdomen (5), heart (5), breast (3) knee (3) and pancreas (3). All other organs were examined in equal to or less than 2 studies (Fig. 3(b)). Studies on the top 3 most frequently analysed organs (brain, lung, multiple organs) were constantly increasing each year (Fig. 3(b)).

6) *Transformers and Medical Imaging:* We reviewed which Transf/Attention mechanisms were implemented per MI modality (Fig. 4(a)). Self-attention was mostly recruited in CT (23 studies) and MRI imaging (22), followed by ultrasound (7), retinal imaging (6), X-rays (2) and PET-CT (2).

Transformers were the most frequent choice in MRI (16), followed by CT (4), X-rays (3) and retinal imaging (1). ViT was mostly applied to X-rays and retinal imaging (3 studies each), followed by CT (2). Channel- and spatial-attention mechanisms were used in MRI, CT, and X-rays (2 studies each). ST was only used in MRI (4 studies).

7) *CNN and Transf/Attention Combinations:* In Fig. 4(b), we show that the incorporation of self-attention mechanisms was the dominant choice distributed across all CNN model types. Transformers were the second most frequent type and was used

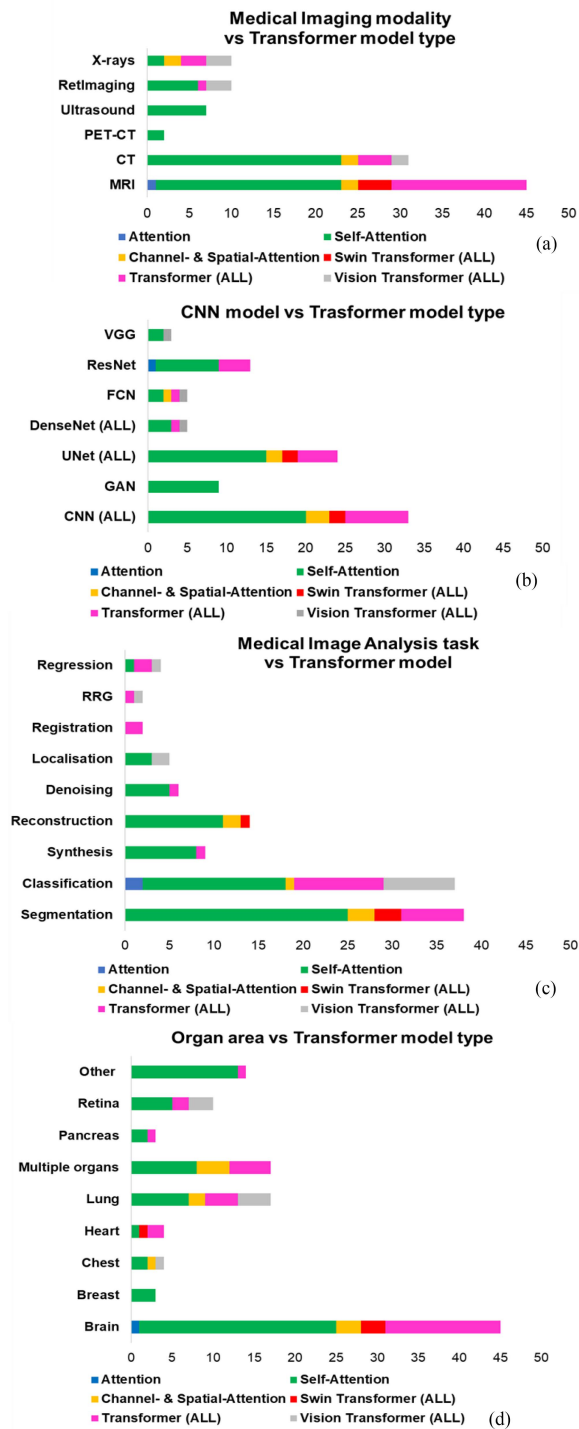


Fig. 4. Publication record showing combinations between the Transformer model/component type and (a) the Medical Imaging modality, (b) the CNN model type, (c) the Medical Image Analysis task and (d) the Organ area. RRG: Radiology report generation.

across all CNN model types, apart from VGG. ST was the third most common type and was only used in conjunction with standard CNN and UNet structures. Based on our findings, mainly “light” (simplified) Transformer blocks, encoders or layers were used across studies (Table I). Novel transfer learning strategies, multi-centre data and/ or increasingly available supercomputer

TABLE IV

DATASET DIVERSITY IN TERMS OF PATIENT SIZE AND DATA SIZE FOR ALL THE ARTICLES THAT SATISFIED OUR GENERALIZATION CRITERIA

Patient size	Data size	Studies
≤ 100 patients	≤ 2,000 images	[20]
	> 2,000 images and ≤ 10,000 images	[57, 118]
	> 10,000 images	[124]
>100 patients and ≤ 1,000 patients	≤ 2,000 images	[35, 53]
	> 2,000 images and ≤ 10,000 images	[111]
	> 10,000 images	[56, 66, 125]
> 1,000 patients and ≤ 10,000 patients	≤ 2,000 images	-
	> 2,000 images and ≤ 10,000 images	[108]
	> 10,000 images	[47, 55]
> 10,000 patients	> 10,000 images	[39, 41, 52, 81, 86, 92]
	NR	-
NR	≤ 2,000 images	[18, 75]
	> 2,000 images and ≤ 10,000 images	[45, 105, 112]
	> 10,000 images	[89, 90, 93]

NR: not reported.

facilities may encourage the use of full Transformer architectures in future work [140], [141]. However, the current hybrid models have showed performance breakthroughs across studies, highlighting them as powerful and relatively simplified (against large pre-trained models) techniques on the MIA tasks reviewed.

For standard CNN and UNet structures, all Transf/Attention mechanisms were used, except for standard attention mechanisms and ViT (Fig. 4(b)). For GAN models, only self-attention mechanisms were implemented. These results demonstrate that there was a large degree of variability in terms of CNN-Transf/Attention combinations across studies. Moreover, large variability was observed in the data augmentations, loss functions and metrics used to evaluate findings (Table II).

8) *Downstream Tasks and Clinical Applications:* Fig. 4(c) illustrates all Transf/Attention components used across each downstream task. Self-attention was implemented across all organ areas (Fig. 4(d)). Transformer architectures were used in the brain, lung, multiple organs, heart, retina, neck, and pancreas. ViT and ST were mainly applied to a relatively limited clinical application space: lung and retina, and brain and heart, respectively. Similarly, channel- and spatial attention have been used in multiple organs, brain, lung and chest.

C. Performance and Generalization Opportunities

Most proposed hybrid models have outperformed baseline and previous SOTA comparison methods, across downstream tasks. Although the evaluation metrics used differed considerably across image analysis tasks and studies making direct comparisons challenging (Table II), there was a clear performance improvement when Transf/Attention mechanisms were used across studies. Some of the studies demonstrated either large ($\geq 5\%$) differences against the best baseline models [21], [35], [46], [79], [101], [108], [117], [121], [122], [126], [127], [135], or moderate ($<5\%$) but consistent improvements across different metrics evaluated [13], [18], [39], [53], [54], [56], [57], [62], [70], [78], [91], [94], [105] and/ or data used [98], [100], [103], [105], [108].

In the following paragraphs, we detail studies that followed our 2 objective generalisation criteria (see Methods): whether a model was a) trained on large data (>2000 images, Table I) and/ or b) analysed data from heterogeneous modalities, and/ or multiple modalities and/ or multiple organ areas and/ or multiple datasets of the same modality and organ. Table IV shows the dataset diversity in terms of patient size and data size across all articles that satisfied our generalisation criteria. Although there were studies that involved large patient data (>1000 patients),

some articles presented analyses from relatively small cohorts or did not report patient size.

1) *Segmentation*: Image segmentation is an important aspect in the field MIA, as it is a necessary intermediate step towards extracting a region of interest within the organ under investigation [142], [143], [144], [145], [146]. Although UNet models revolutionised medical image segmentation [147], [148], image segmentation remains an open challenge as it relies on strong supervision, hence, a large fraction of labelled data are required. However, there is a considerable “data challenge” barrier, as labels are commonly limited for MI data [2], [146]. To address this, several approaches have been proposed, such as disentangled representations for semi-supervised learning which can generate accurate segmentations by only using a small fraction of labelled data [146], or GAN techniques to obtain accurate paired synthesis of images and segmentation masks [149].

Cheng et al. proposed a multi-task methodology for simultaneous glioma segmentation from MRI images and parallel classification of genetic profiles for neuro-oncology patients [53]. They developed a CNN model with serial ResNet blocks in the encoder and decoder. Between the encoder and decoder, 2 Transformer layers were engineered. Unlike most of the MRI and CNN studies, the authors used multi-parametric MRI data for image segmentation (4 different MRI sequences). The authors compared their method against 10 baseline CNN models and demonstrated superior performance for both tasks. In the context of small but heterogeneous data analyses, Wang et al. designed a CNN encoder-decoder model with residual connections and self-attention modules connected with CNN layers in the encoder [57]. The authors demonstrate that their method outperformed all baseline models in identifying COVID-19 lung abnormalities from CT images. They also developed a zero-shot learning strategy based on the same hybrid model, in which a UNet model was applied to predict pseudo-labels in a non-labelled CT dataset, which in turn guided semi-supervised learning. Rajamani et al [18] engineered a deformable attention module into a UNet model. Their model (called “DDANet”) was trained and tested on a large publicly available CT COVID-19 dataset, achieving superior performance for lung infection segmentation compared to baseline models. As future work, the authors discuss that their model can be further validated to detect small and irregular lesions for other MI segmentation problems.

Next to limited labelled data, another challenge in medical image segmentation is the analysis of “less anatomical” and more “biophysical” imaging data, in which imaging physics are modified so that anatomical information at the pixel level is “sacrificed”, to “emphasize” perfusion, functional, temporal or other biophysical information [2], [150], [151], [152], [153]. Most segmentation algorithms are focusing on imaging sequences that contain enough anatomic (to efficiently guide semantic) representations during training [2], [153]. Shi et al. developed a powerful method that is capable to analyse 4 different parametric perfusion maps: a) cerebral blood volume, b) cerebral blood flow, c) time to maximum peak and d) mean transit time (of contrast enhancement) [78]. They developed two parallel subnetworks to analyse blood flow (a, b) and time (c, d) parameters, simultaneously. Each subnetwork included a CNN model with skip connections between the encoder and decoder. A cross-attention module was incorporated between the encoder and decoder for feature fusion. The model was

compared against baseline methods (achieving higher and comparable performance, depending on the metric) and evaluated on both public and in-house data.

On a retinal MIA study, Mou et al. developed a versatile curvilinear structure segmentation network, based on dual self-attention modules which can address both 2D and 3D retinal imaging data [105]. In their model named as “CS2Net”, they devised two channel- and spatial self-attention mechanisms to generate attention-aware features and capture long-range contextual information. By performing extensive experiments on 9 (2D and 3D) datasets, they demonstrated SOTA performances in detecting curvilinear structures from different imaging modalities. They showed that their technique can work as a generalized approach for retinal morphology analyses. Of note, such hybrid models can be impactful, since retinal imaging is not only used to assess ophthalmic pathologies, but also changes in retinal morphologies that may occur early in a broad spectrum of diseases, such as Alzheimer’s [154], cardiac pathologies [155], cerebral small vessel disease [156] and others. Wang et al. have proposed the MstGANet, a UNet model enhanced with a Transformer block that consists of a series of multi-head self-attention mechanisms incorporated in the encoder, to capture both local and global pixel interactions early in the learning process [45]. A series of channel- and spatial attention modules were also inserted between different positions of the encoder and decoder, to efficiently fuse feature semantics during training. At inference, the model predicted labels in non-labelled data, which were then used as pseudo-labels to augment the dataset, as a semi-supervised learning strategy (in which pseudo-labels were then used to guide semi-supervision). The model outperformed previous SOTA methods in supervised and semi-supervised segmentation tasks.

Xu et al. replaced 2 layers in the encoder and 1 layer in the decoder of a UNet model with self-attention mechanisms [108]. Their hybrid model achieved SOTA performance in segmenting several fetal anatomies, when compared to 6 other models. Segmenting fetal structures from ultrasound is particularly challenging due to moving and fuzzy anatomical organ boundaries [157]. Sinha et al. developed a generalizable hybrid model for segmentation of numerous abdominal, cardiovascular and brain structures by analysing different MRI sequences [111]. The authors used a ResNet model for initial feature extractions which were then fed into a stack of spatial and channel self-attention mechanisms. They demonstrated superior performance against 6 previous SOTA baselines. The model was capable to perceive a broad spectrum of anatomical (different organs) and semantic (different MRI sequences) information and can therefore be potentially useful to be further validated in future single- and multi-centre MRI studies [2], [29].

Xie et al developed a 3D UNet architecture which consisted of 2 cascading UNets both enhanced with self-attention [121]. The overall model was trained on a chronic obstructive pulmonary disease (COPD) CT dataset. Following training, the hybrid model was evaluated on COPD data and on an unseen COVID-19 dataset. The model outperformed previous techniques in detecting several lung nodules in COPD and COVID-19 data. Following further validation using CT data from other organs, this hybrid approach can potentially have applicability in terms of detecting small and irregular lesions across different diseases and organ areas. Other studies focused on segmentation of large-scale MRI data [58], [59], [60], [61], [62].

2) *Classification*: In their noteworthy study, Zhou et al. proposed a cross-supervised method called REFERS, which generates X-ray image labels from radiology reports, to perform lung pathology detection through image classification [52]. The authors employed ViT blocks composed of multi-head self-attention mechanisms, to learn joint representations from multiple radiograph views and corresponding radiology reports. Subsequently, the model performs feature fusion and employs two additional subnetworks for bidirectional visual to textual feature mapping. REFERS was first pre-trained on a source domain X-ray dataset and then fine-tuned on 4 well-established datasets (target domain with text labels). During fine-tuning, the authors performed fully supervised learning on the target domain (using structured radiograph labels). Differing from other models, their technique did not require labels during pre-training. The authors also showed that their model outperformed powerful baseline models on all datasets under extremely limited supervision (1% labelled images during fine-tuning). Their model was consistently accurate in detecting several lung pathologies thus, having tremendous potential for real-world applications where labelling is substantially limited.

To address large-scale analysis from different domains, Wood et al. developed a DenseNet-based supervised learning framework for detecting clinically relevant abnormalities from clinical T2-weighted and diffusion-weighted head MRI scans [39]. The DenseNet model was trained using a Transformer-based neuro-radiology report classifier to generate a labelled dataset of 70206 examinations from 2 U.K. hospitals. The Transformer model was trained using a small dataset ($N = 5000$) of neuroradiology reports. The authors showed accurate, fast and generalisable classification of abnormal against normal brain MRI between hospitals. This work demonstrated the merit of CNN and Transformer synergy when combined under the same MIA pipeline.

Zhang et al. devised a 3D ResNet block that operated as initial feature extractor before feeding feature information into a self-attention block [21]. The authors performed several classification experiments for identifying Alzheimer's disease and mild cognitive impairment from MRI data, showing superior performance against baseline methods. Despite they focused on using T1-weighted data (mainly used for anatomical imaging and does not contain "functional" [158] or "perfusion" [159], [160] tissue information), they analysed data from both 1.5 T and 3T MRI scanners, which is known that they have differences in the signal-to-noise ratio, imaging content and artefacts [1], [2], [158], [159], [160]. Since their technique was assessed on public data (ADNI), for 2 different brain pathologies and analysed data from different field MR scanners, it can potentially be useful to be validated across further MRI data.

Another study by Let et al. [35], proposed a CNN encoder-decoder network connected with a self-attention mechanism (called PreSANet) to detect cancer recurrence, distant metastasis and overall patient survival for head and neck cancer patients. The model was trained on public data and was validated on various unseen datasets demonstrating good ($\sim 70\%$) generalisability. Mondal et al. pre-trained a ViT encoder connected with a FCN layer, to discriminate COVID-19 positive cases from other pneumonia types and normal controls [55]. The model was trained on the ImageNet dataset, fine-tuned on a large collection of chest X-ray and tested on both CT and X-ray lung data.

Zhao et al. proposed a UNet model with residual blocks enhanced with self-attention, to classify malignant from benign thyroid nodules from ultrasound images. The model was

evaluated on a large-scale dataset via extensive experiments and achieved high performance (89%) [86]. Wu et al. developed a ViT encoder and performed accurate diabetic retinopathy grading from retinal images using a large Kaggle dataset [89]. Duong et al. developed an Efficient backbone model connected with a full ViT and demonstrated accurate and generalisable detection of tuberculosis from heterogeneous X-ray public sources [90]. Lin et al. developed a deformable ResNet model with self-attention incorporated to detect irregular and diffused lung nodules due to COVID-19 infection and showed SOTA performance in large and diverse public datasets [92]. Shome et al. developed a Transformer encoder connected with an MLP block to perform multi-classification of COVID-19 infection against other pneumonia types and normal lung, from large X-ray datasets [93]. Other studies focused on large-scale analysis of MRI brain (schizophrenia, Alzheimer's Disease) [36], [63], chest X-ray (tuberculosis) [42] and retinal diseases [133]. There were also studies demonstrating innovative architectures and high diagnostic accuracies in the setting of image classification, however, using smaller datasets [49], [84].

3) *Reconstruction*: Medical image reconstruction aims to form an image representation from raw signals acquired by the scanner [2]. Reconstruction of fast acquisitions (of periodically moving organs such as the heart) and/ or low doses (e.g., CT), has important clinical applications. Using relatively small but highly diverse data, Zhou et al. developed a CNN-based method enhanced with self-attention for ultrasound image reconstruction of various organs and tissues [117]. Another study demonstrated accurate brain reconstruction by using a CNN with Transformer layers on large MRI data (>30000 MR images) [56]. Tan et al. devised a CNN model with residual connections in which channel- and spatial-attention modules were engineered to reconstruct X-ray images of the lung, from a large dataset (>55000 images) [80]. Other studies focused on MRI reconstruction and demonstrated accurate and generalisable hybrid models by analysing large and diverse imaging data [15], [114], [129], [139].

4) *Synthesis*: Image synthesis is an important field as it can address the need of data augmentation across different modalities [2], [161]. Yang et al. developed a CycleGAN with self-attentions for unsupervised MR-to-CT synthesis, outperforming 2 plain CycleGAN baselines [16]. In the field of MR-to-CT synthesis, Dalmaz et al. developed a series of residual Transformer blocks between the encoder and decoder of a CNN [66] and Tomar et al. developed a GAN model with ResNet blocks and self-attention modules for cardiac and brain image synthesis [107].

Wei et al. developed a first-of-its kind GAN model with self-attention in the generator and discriminator, that was able to synthesise PET-derived myelin content through the analysis of multi-sequence MRI data [20].

5) *Denosing*: Denosing is an important step prior to image quantification as it can enhance signal-to-noise-ratio and remove artefacts [2], [26], [27], [28], [29]. Li et al. combined a 3D CNN model with self-attention blocks and an autoencoder perceptual loss (used as a self-supervised learning module) with CNN-based and GAN-based models. They achieved improved denoising performance against baseline models for chest and abdominal CT images [124]. Huang et al. proposed an end-to-end CycleGAN model with criss-cross self-attention and channel-attention mechanisms to reduce noise, remove artefacts and preserve anatomical structures in low-dose dental and

abdominal CT images [19]. To denoise low-count PET images, Xue et al. developed a 3D GAN model with self-attention, achieving improved performance against baseline methods [118]. Their method was evaluated on large-scale PET data and showed that it can improve PET image quality, reduce motion artefacts and provide accurate diagnostic information.

6) Localisation: Image localisation focuses on detecting the location of an area of interest within MI data [26], [27]. Tao et al. proposed a ResNet model for initial feature extraction followed by a series of self-attention and cross-attention mechanisms for vertebrae CT localisation and segmentation [13]. They demonstrated accurate and generalisable performance across 2 CT datasets. Li et al. developed a DenseNet model parallelised with a ViT block to extract local and global pixel dependencies which were fused before fed into a CNN model. Their technique outperformed baseline models on classification and localisation of several lung abnormalities when trained and tested in a large X-ray dataset (of >112000 images) [81]. Xie et al. used a pre-trained VGG model enhanced with self-attention to enhance feature extraction before feeding this information into 2 subsequent CNN models [112]. They demonstrated accurate fovea localisation in 2 different retinal imaging datasets.

7) Regression: In the task of brain age estimation from MRI, He et al. developed a hybrid model consisting of a CNN and a full Transformer, to capture the relationships between pairs of images with different chronological ages from patients [47]. Their method outperformed 8 SOTA baselines as tested on 8 public datasets (N = 6049 patients in total).

8) Registration: Image registration is the process of aligning the spatial coordinates of different images into a common geometrical coordinate system. Image registration has wide applications in multi-modal and longitudinal MIA [162], [163]. Yang et al. developed a plain Transformer encoder with an attention-based decoder model for brain MRI registration, demonstrating accurate results against baselines across 3 different datasets [75]. Song et al. proposed a CNN model with Transformer blocks consisted of modified multi-head self-attention for brain MRI registration, producing SOTA registration performance [77]. Although analysing brain images from different MRI sequences is challenging, the brain is a static organ that is less prone to misregistration across modalities. Further work is required to expand towards organ areas that are subject to periodic (e.g., heart) and non-periodic (e.g., abdomen) motion, and to register images from different modalities.

IV. DISCUSSION

A. Current Opportunities and Challenges

We studied all the articles from the perspective of 4 professionals (co-authors GP, ND, CW, GY) with extensive experience in deep learning and MI. We identify general challenges and opportunities, from the multi-disciplinary perspective of developers and end-users of these hybrid models in MI. To the best of our knowledge, there is no previous review focusing on these topics and given the heterogeneous architectures of these models, more extensive studies are required in the future to develop data-driven generalization best practices for both developers and end-users. The following points can therefore guide future work and systematic reviews towards solidifying these hybrid models in further, larger and multi-centre studies.

Challenges (with mitigations, where applicable): 1) We highlighted studies that have the potential to work as generalisation frameworks. However, additional validation is required to transfer a method to real-world data for the same organ/ imaging modality or from one organ/ imaging modality to another, due to data content differences. Hybrid models in studies that involved small ($N \leq 100$, Table IV) or even medium size patient ($N \leq 1000$) cohorts may have been susceptible to within-subject image correlations and must be carefully validated in further diversified cohorts. 2) Model architectures varied considerably when similar hybrid models were compared. For example, in studies for which a UNet with self-attention were developed, there were large disparities in terms of how these individual components were combined. 3) The previous point indicates that a trial-and-error logic is currently followed for model development, based on which architecture performs optimally for a given dataset. Nevertheless, this is in the opposite direction from developing generalised models and best practices. It is important for the community to initiate discussions about the development of generalisation frameworks, based on certain data-driven boundary conditions: e.g., UNet-full Transformer for cardiac segmentation would be a preferred design if a particular data size, data content (e.g., T1-weighted MRI) and in-house computational capabilities are satisfied. Thus, solid domain generalization strategies to methodically address “why” and “how” to develop model X for data Y are required. 4) Developing harmonised implementation protocols is particularly challenging. Implementation aspects such as data augmentation, optimisers, loss functions and pre-processing differed substantially even between studies working on the same problem (e.g., CT for lung segmentation). 5) It will be challenging to develop robust interpretation mechanisms for complex local-global pattern recognition models that are not solely based on visualization maps. 6) There is an increasing trend in terms of developing causal logic in novel deep learning models, a field known as “Causal Representation Learning, CRL” [164]. The aim of CRL is to address open problems in the field such as model generalisation and transfer learning [164], [165]. Central to CRL is the discovery of high-level causal variables (objects in an image) from low-level observations (embeddings) [164]. One of the main challenges that must be addressed is how to factorise causal structures from deep learning embeddings [164], [165]. CNN-Transf/Attention models have an additional level of complexity due to learning embeddings from both local and global interactions. Thus, there must be a careful consideration regarding how to combine CNN-Transf/Attention models with causality and benefit from the advances of each other [164].

Opportunities: 1) Based on performance gains achieved, hybrid model studies can give emphasis on studying generalisation perspectives and standardisation protocols for multi-centre large-scale analyses. 2) Given diagnostic performance improvements across diverse studies, there is a potential to enhance early diagnosis and preventative medicine. 3) As of 2022, cardiovascular diseases, cancer, stroke, COVID-19, chronic respiratory diseases, diabetes, neurological diseases are the leading causes of death in the USA [166]. Most studies (>90%) in our review focused on at least 1 of the organ/ pathology areas corresponding to these leading causes, showing the potential to improve diagnosis and patient outcomes. 4) Technical versatility on multi-modal analyses can be achieved through CNN-Transf/Attention (images, natural language, molecular profiles,

clinical history), which can yield useful complementary information. 5) By combining CNN and attention models to learn local-global information e.g., from images and text [39], [52] or images and genetics [53], it is possible to enhance precise diagnosis via potentially extracting accurate complementary patient-level information from different modalities. 6) We strive to inspire and guide benchmark studies to extensively evaluate promising hybrid methods described in our review. For instance, benchmark studies can focus on a specific pathology assessment/imaging modality (e.g., stroke from brain MRI) and methodically compare promising hybrid models (as demonstrated in our review) and pure CNN counterparts. Particular attention needs to be paid to hybrid methods from studies that involved small or medium sized cohorts (see Table IV). 7) Focus on integrating CNN-Transf/Attention with CRL to enhance model generalisability and trustworthiness in the clinical domain. For example, it is known that causal generative models (such as conditional GANs and diffusion models) can be powerful causal inference engines for generating counterfactuals [164], [165]. Integrating Transf/Attention modules into generative models can be useful to better understand how attention mechanisms contribute to model (factual) outcomes and counterfactual generation. 8) Develop robust transfer learning methods to fully explore the benefits of CNN-Transf/Attention models on out-of-distribution datasets.

1) *Importance and Drawbacks*: The combination of local and global receptive fields together with reasonable computational power requirements highlights the development of CNN-Transf/Attention models as an important research direction in MIA. The large diversity of architectures even under the same downstream tasks or applications, means that for some of these methods, limited scalability may be one of the main drawbacks [52]. Furthermore, full Transformer architectures were limited in our reviewed work, mainly due to relatively small data analysed in some studies, limited computational power and/ or lack of solid transfer learning approaches for pixel-level predictions [52], [55], [141], [167]. Further work is required in the field of transfer learning techniques for model generalisation on out-of-distribution data, to utilise the benefits of full Transformer-based hybrids.

B. Future Perspectives of the Post-Hybrid Model Era

1) *Full Transformers, ChatGPT and Beyond*: The recent developments of ChatGPT large language models (LLM) induced a phenomenal disruption in the field of data analysis and AI. To date, the latest ChatGPT version is based on the GPT-4 (launched on March 2023), reported as the largest LLM trained (>170 trillion parameters) [168], [169], [170]. The main strength of GPT-4 model is that it has been trained on a diverse and broad (in terms of topics) set of internet text including books, articles and websites, using reinforcement learning from human feedback that either rewards or “punishes” the model [170]. One of its main capabilities, is that it can perform data predictions through conversational tasks (“responses” to user “queries”). ChatGPT models perform Transformer-based and self-supervised learning-derived predictions [170], [171].

There have been some first promising approaches involving GPT models for MIA, although mainly limited to image-to-text mapping [104], [172], [173]. Wang et al. used pre-trained CNN models to extract outputs from X-rays of the lung and applied report generator GPT models to summarise the results and derive

a diagnosis in text [172]. Another study by Chen et al. used a pre-trained GPT-2 model with a visual encoding part that involved attention, to perform accurate image captioning as evaluated on natural images and X-ray data [173].

Although it can be anticipated that GPT models may expand towards MIA, as e.g., to enhance image-level predictions based on large-scale radiology reports or clinical notes [52], there are several limitations that need to be considered. First, to the best of our knowledge, there is no GPT-based MIA yet on dense image-level predictions for the MIA tasks we have presented. Local receptive fields that are based on CNN feature extractors may be necessary to perform detailed image analyses, pointing towards the direction of heavier “hybrid models” in the future (CNN-GPT). On that note, it is unknown whether existing self-supervision modules within GPT models may be enough to predict complex organ and tissue pathologies from “high-content” data such as medical images, without the incorporation of “computer vision” CNN components. Furthermore, one important limitation of GPT models is the so called “hallucination effect”, which describes the tendency of GPT models to “invent” a term eventually giving “incorrect” responses [174]. This can be the case for domains in which GPT models have been less or not yet specialized. Due to regulatory, ethical and organisational considerations from clinical and private MI data owners, we are still at infant stages regarding multi-centre large-scale data analyses that need to be available as data sources for such open code or multi-centre fine-tuning strategies. In addition, the co-existence of available MI and text data is commonly low.

2) *Transfer Learning Coming From the Future*: An important yet unsolved aspect in MIA is the democratisation of modelling techniques and data. Transfer learning strategies focusing on increasing performance while reducing computational power [141], can serve as democratisation vehicles. Transfer learning could also potentially aim towards improving model generalisability in large-scale multi-centre trials (opportunity 1), where data across different centres could be used for pre-training/ fine-tuning and external validation. However, it is known that in the absence of large publicly available (e.g., tomographic data: MRI, CT, PET) databases, transfer learning has been limited in MIA, compared to “smaller new” models trained de novo. Moreover, most transfer learning studies are based on models pre-trained on the large ImageNet [141], which may degrade model robustness to distribution shifts between natural and medical images [2], [167].

Among a large amount of studies demonstrating new models, we highlighted articles that showed robust pre-training with wide fine-tuning on large domain datasets with SOTA performance on testing data [52], [55]. In their influential study, Liu et al. recently proposed “ConvNext” as a new pure CNN technique which involved several ST-inspired adaptations in the model design and transfer learning method [141]. Some of these ST-inspired adaptations were: same augmentation protocol, network width increase, bottleneck model inversion, kernel size enlargement, use of fewer activation functions and normalisation layers. Using ConvNext, they outperformed ST on ImageNet classification tasks while using comparable computing resources. Radford et al. adapted Transformer and ResNet/ ViT models for jointly pre-training paired text and images, respectively [167]. By training on web data of 400 million image-text pairs, they demonstrate that can learn image captions from text which can be used as labels for image classification, showcasing a scalable and efficient process to learn image representations.

Following pre-training, the text model can describe new visual concepts allowing zero-shot transfer to new tasks and data. Democratizing data access (especially for tomographic data such as MRI, CT, PET that are less widely available) could stimulate further work on large hybrid models for single (i.e., MIA) [141] and multi-modal (e.g., MIA, radiology reports) [167] analyses respectively, and could support future work to improve model design, pre-training and fine-tuning (both on domain data) techniques.

V. CONCLUSION

In conclusion, hybrid models led to performance gains while demonstrating a big range of generalisation opportunities based on either their large-scale, multi-modal, heterogeneous and/ or broad span of clinical applications. The main challenge of these techniques is to align their large architectural diversity with the current technical and clinical needs in precision and preventative medicine. Based on the opportunities that we have emphasised, we aim to encourage further work on data-driven generalisation frameworks, to develop criteria for the future design of these powerful hybrid techniques. We also seek to inspire further work in the field of transfer learning for generalisation on out-of-distribution data so that models and data can be further democratised.

Our review demonstrates the benefits from the co-pollination of CNN and Transformer-inspired models which can open new horizons to further exploit CNN and full Transformers and LLM. Next to these opportunities, our review demonstrated that the benefits of CNN-Transf/Attention outweigh the challenges and may therefore be “all you need” for future validation and standardisation processes in clinical imaging.

REFERENCES

- [1] J. Beutel et al., *Handbook of Medical Imaging*. Bellingham, WA, USA: SPEI Press, 2000.
- [2] S. K. Zhou et al., “A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises,” *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021.
- [3] Radiology Society of North America (RSNA), “Quantitative imaging biomarkers alliance (QIBA),” 2019. [Online]. Available: <https://www.rsna.org/research/quantitative-imaging-biomarkers-alliance>
- [4] N. M. deSouza et al., “Validated imaging biomarkers as decision-making tools in clinical trials and routine practice: Current status and recommendations from the EIBALL* subcommittee of the European society of radiology (ESR),” *Insights Imag.*, vol. 10, 2019, Art. no. 87.
- [5] European Society of Radiology (ESR), “ESR statement on the validation of imaging biomarkers,” *Insights Imag.*, vol. 11, 2020, Art. no. 76.
- [6] J. O’Connor et al., “Imaging biomarker roadmap for cancer studies,” *Nature Rev. Clin. Oncol.*, vol. 14, pp. 169–186, 2017.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [8] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, *arXiv:1409.0473*.
- [9] J. Cheng, L. Dong, and M. Lapata, “Long short-term memory-networks for machine reading,” 2016, *arXiv:1601.06733*.
- [10] A. Dosovitskiy et al., “An image is worth 16 x16 words: Transformers for image recognition at scale,” 2020, *arXiv:2010.11929*.
- [11] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
- [12] Z. Liu et al., “Swin transformer v2: Scaling up capacity and resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11999–12009.
- [13] R. Tao, W. Liu, and G. Zheng, “Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers,” *Med. Image Anal.*, vol. 75, 2022, Art. no. 102258.
- [14] S. Bera and P. K. Biswas, “Noise conscious training of non local neural network powered by self attentive spectral normalized Markovian patch GAN for low dose CT denoising,” *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3663–3673, Dec. 2021.
- [15] T. Du et al., “Adaptive convolutional neural networks for accelerating magnetic resonance imaging via k-space data interpolation,” *Med. Image Anal.*, vol. 72, 2021, Art. no. 102098.
- [16] H. Yang et al., “Unsupervised MR-to-CT synthesis using structure-constrained CycleGAN,” *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4249–4261, Dec. 2020.
- [17] F. Shamshad et al., “Transformers in medical imaging: A survey,” *Med. Image Anal.*, vol. 88, 2022, Art. no. 102802.
- [18] K. T. Rajamani, H. Siebert, and M. P. Heinrich, “Dynamic deformable attention network (DDANet) for COVID-19 lesions semantic segmentation,” *J. Biomed. Inform.*, vol. 119, 2021, Art. no. 103816.
- [19] Z. Huang et al., “CaGAN: A cycle-consistent generative adversarial network with attention for low-dose CT imaging,” *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1203–1218, 2020.
- [20] W. Wei et al., “Predicting PET-derived myelin content from multisequence MRI for individual longitudinal analysis in multiple sclerosis,” *NeuroImage*, vol. 223, 2020, Art. no. 117308.
- [21] X. Zhang, L. Han, W. Zhu, L. Sun, and D. Zhang, “An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer’s disease diagnosis using structural MRI,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5289–5297, Nov. 2022, doi: [10.1109/JBHI.2021.3066832](https://doi.org/10.1109/JBHI.2021.3066832), 2021.
- [22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
- [23] O. Dalmaz, M. Yurt, and T. Cukur, “ResViT: Residual vision transformers for multi-modal medical image synthesis,” *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2598–2614, Oct. 2022.
- [24] H. Chefer, S. Gur, and L. Wolf, “Transformer interpretability beyond attention visualization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 782–791.
- [25] L. Yu, L. Maozhen, and J. Changjun, “Generating self-attention activation maps for visual interpretations of convolutional neural networks,” *Neurocomputing*, vol. 490, pp. 206–216, 2022.
- [26] X. Yi, W. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” *Med. Image Anal.*, vol. 58, 2019, Art. no. 101552.
- [27] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, “Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” *Med. Image Anal.*, vol. 54, pp. 280–296, 2019.
- [28] J. S. Duncan, M. F. Insana, and N. Ayache, “Biomedical imaging and analysis in the age of Big Data and deep learning,” *Proc. IEEE*, vol. 108, no. 1, pp. 3–10, Jan. 2020.
- [29] G. Haskins, U. Kruger, and P. Yan, “Deep learning in medical image registration: A survey,” *Mach. Vis. Appl.*, vol. 31, no. 1, 2020, Art. no. 25400.
- [30] K. He et al., “Transformers in medical image analysis,” *Intell. Med.*, vol. 3, pp. 59–78, 2023.
- [31] N. Linna and C. E. Kahn, “Applications of natural language processing in radiology: A systematic review,” *Int. J. Med. Inform.*, vol. 163, 2022, Art. no. 104779.
- [32] X. Chen et al., “Recent advances and clinical applications of deep learning in medical image analysis,” *Med. Image Anal.*, vol. 79, 2022, Art. no. 102444.
- [33] A. Parvaiz et al., “Vision transformers in medical computer vision—A contemplative retrospection,” *Eng. Appl. Artif. Intell.*, vol. 122, 2023, Art. no. 106126.
- [34] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, “The PRISMA group. preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement,” *PLoS Med.*, vol. 6, no. 7, 2009, Art. no. e1000097.
- [35] W. T. Le et al., “Cross-institutional outcome prediction for head and neck cancer patients using self-attention neural networks,” *Sci. Rep.*, vol. 12, no. 1, 2022, Art. no. 3183.
- [36] K. H. Oh et al., “Diagnosis of schizophrenia with functional connectome data: A graph-based convolutional neural network approach,” *BMC Neurosci.*, vol. 23, no. 1, 2022, Art. no. 5.
- [37] Z. Zhou et al., “Local-global multiple perception based deep multi-modality learning for sub-type of esophageal cancer classification,” *Biomed. Signal Process. Control*, vol. 77, 2022, Art. no. 103757.
- [38] L. Wang et al., “Dementia analysis from functional connectivity network with graph neural networks,” *Inf. Process. Manage.*, vol. 59, no. 3, 2022, Art. no. 102901.

- [39] D. A. Wood et al., "Deep learning models for triaging hospital head MRI examinations," *Med. Image Anal.*, vol. 78, 2022, Art. no. 102391.
- [40] Z. Jiang et al., "Computer-aided diagnosis of retinopathy based on vision transformer," *J. Innov. Opt. Health Sci.*, vol. 15, no. 2, 2022, Art. no. 2250009.
- [41] L. Chen, Z. You, N. Zhang, J. Xi, and X. Le, "UTRAD: Anomaly detection and localization with U-transformer," *Neural Netw.*, vol. 147, pp. 53–62, 2022.
- [42] S. Rajaraman, G. Zamzmi, L. R. Folio, and S. Antani, "Detecting tuberculosis-consistent findings in lateral chest X-rays using an ensemble of CNNs and vision transformers," *Front. Genet.*, vol. 13, 2022, Art. no. 864724.
- [43] M. M. Al Rahhal et al., "COVID-19 detection in CT/X-ray imagery using vision transformers," *J. Personalized Med.*, vol. 12, no. 2, 2022, Art. no. 310.
- [44] J. Zhang et al., "A CNN-transformer hybrid approach for decoding visual neural activity into text," *Comput. Methods Programs Biomed.*, vol. 214, 2022, Art. no. 106586.
- [45] M. Wang et al., "MsTGANet: Automatic drusen segmentation from retinal OCT images," *IEEE Trans. Med. Imag.*, vol. 41, no. 2, pp. 394–406, Feb. 2022.
- [46] G. Zheng et al., "A transformer-based multi-features fusion model for prediction of conversion in mild cognitive impairment," *Methods*, vol. 204, pp. 241–248, 2022.
- [47] S. He, Y. Feng, G. E. Grant, and Y. Ou, "Deep relation learning for regression and its application to brain age estimation," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2304–2317, Sep. 2022.
- [48] T. S. Chandraraju and A. Jeyaprkash, "Categorization of breast masses based on deep belief network parameters optimized using chaotic krill herd optimization algorithm for frequent diagnosis of Breast abnormalities," *Int. J. Imag. Syst. Technol.*, vol. 32, no. 5, pp. 1561–1576, 2022.
- [49] J. Cheng et al., "COVID-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest X-rays and clinical data," *Eur. Radiol.*, vol. 32, no. 7, pp. 4446–4456, 2022.
- [50] R. Otsuki et al., "Integrating preprocessing operations into deep learning model: Case study of posttreatment visual acuity prediction," *Adv. Biomed. Eng.*, vol. 11, pp. 16–24, 2022.
- [51] D. Chen et al., "PCAT-UNet: UNet-like network fused convolution and transformer for retinal vessel segmentation," *PLoS One*, vol. 17, no. 11, 2022, Art. no. e0262689.
- [52] H. Y. Zhou, X. Chen, Y. Zhang, R. Luo, L. Wang, and Y. Yu, "Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports," *Nature Mach. Intell.*, vol. 4, no. 1, pp. 32–40, 2022.
- [53] J. Cheng, J. Liu, H. Kuang, and J. Wang, "A fully automated multi-modal MRI-based multi-task learning for glioma segmentation and IDH genotyping," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1520–1532, Jun. 2022.
- [54] S. He, P. E. Grant, and Y. Ou, "Global-local transformer for brain age estimation," *IEEE Trans. Med. Imag.*, vol. 41, no. 1, pp. 213–224, Jan. 2022.
- [55] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, "XViT-COS: Explainable vision transformer based COVID-19 screening using radiography," *IEEE J. Transl. Eng. Health Med.*, vol. 10, 2022, Art. no. 1100110.
- [56] J. Huang et al., "Swin transformer for fast MRI," *Neurocomputing*, vol. 493, pp. 281–304, 2022.
- [57] X. Wang et al., "SSA-Net: Spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning," *Med. Image Anal.*, vol. 79, 2022, Art. no. 102459.
- [58] J. Liu, J. Zheng, and G. Jiao, "Transition Net: 2D backbone to segment 3D brain tumor," *Biomed. Signal Process. Control*, vol. 75, 2022, Art. no. 103622.
- [59] L. Wu, S. Hu, and C. Liu, "MR brain segmentation based on DE-ResUnet combining texture features and background knowledge," *Biomed. Signal Process. Control*, vol. 75, 2022, Art. no. 103541.
- [60] C. Laiton-Bonadiez, G. Sanchez-Torres, and J. Branch-Bedoya, "Deep 3D neural network for brain structures segmentation using self-attention modules in MRI images," *Sensors*, vol. 22, no. 7, 2022, Art. no. 2559.
- [61] J. Liang et al., "TransConver: Transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images," *Quantitative Imag. Med. Surg.*, vol. 12, no. 4, pp. 2397–2415, 2022.
- [62] M. Sheng et al., "Cross-attention and deep supervision UNet for lesion segmentation of chronic stroke," *Front. Neurosci.*, vol. 16, 2022, Art. no. 836412.
- [63] M. Jiang, B. Yan, Y. Li, J. Zhang, T. Li, and W. Ke, "Image classification of Alzheimer's disease based on external-attention mechanism and fully convolutional network," *Brain Sci.*, vol. 12, no. 3, 2022, Art. no. 319.
- [64] L. Wang, H. Zhu, Z. He, Y. Jia, and J. Du, "Adjacent slices feature transformer network for single anisotropic 3D Brain MRI image super-resolution," *Biomed. Signal Process. Control*, vol. 72, 2022, Art. no. 103339.
- [65] T. Dharnija, A. Gupta, S. Gupta, R. Anjum Katarya, and G. Singh, "Semantic segmentation in medical images through transfused convolution and transformer networks," *Appl. Intell.*, vol. 53, no. 1, pp. 1132–1148, 2022.
- [66] O. Dalmaz, M. Yurt, and T. Cukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2598–2614, Oct. 2022.
- [67] Z. Fu, J. Zhang, R. Luo, Y. Sun, D. Deng, and L. Xia, "TF-Unet: An automatic cardiac MRI image segmentation method," *Math. Biosciences Eng.*, vol. 19, no. 5, pp. 5207–5222, 2022.
- [68] H. Wang, X. Zhao, W. Liu, L. C. Li, and J. Ma. L. Guo, "Texture-aware dual domain mapping model for low-dose CT reconstruction," *Med. Phys.*, vol. 49, no. 6, pp. 3860–3873, 2022.
- [69] J. Wang, S. Wang, W. Liang, N. Zhang, and Y. Zhang, "The auto segmentation for cardiac structures using a dual-input deep learning network based on vision saliency and transformer," *J. Appl. Clin. Med. Phys.*, vol. 23, no. 5, 2022, Art. no. e13597.
- [70] D. Karimi, H. Dou, and A. Gholipour, "Medical image segmentation using transformer networks," *IEEE Access*, vol. 10, pp. 29322–29332, 2022.
- [71] M. Ma, H. Xia, Y. Tan, H. Li, and S. Song, "HT-Net: Hierarchical context-attention transformer network for medical ct image segmentation," *Appl. Intell.*, vol. 52, no. 9, pp. 10692–10705, 2022.
- [72] H. Cui, L. Jiang, C. Yuwen, Y. Xia, and Y. Zhang, "Deep U-Net architecture with curriculum learning for myocardial pathology segmentation in multi-sequence cardiac magnetic resonance images," *Knowl.-Based Syst.*, vol. 249, 2022, Art. no. 108942.
- [73] N. Cahan et al., "Weakly supervised attention model for RV strain classification from volumetric CTPA scans," *Comput. Methods Programs Biomed.*, vol. 220, 2022, Art. no. 106815.
- [74] X. Wang et al., "Automatic and accurate segmentation of peripherally inserted central catheter (PICC) from chest X-rays using multi-stage attention-guided learning," *Neurocomputing*, vol. 48, pp. 82–97, 2022.
- [75] T. Yang, X. Bai, X. Cui, Y. Gong, and L. Li, "TransDIR: Deformable imaging registration network based on transformer to improve the feature extraction ability," *Med. Phys.*, vol. 49, no. 2, pp. 952–965, 2022.
- [76] J. Xie, R. Zhu, Z. Wu, and J. Ouyang, "FFUNet: A novel feature fusion makes strong decoder for medical image segmentation," *IET Signal Process.*, vol. 16, no. 5, pp. 501–514, 2022.
- [77] L. Song, G. Liu, and M. Ma, "TD-Net: Unsupervised medical image registration network based on Transformer and CNN," *Appl. Intell.*, vol. 52, no. 15, pp. 18201–18209, 2022.
- [78] T. Shi, H. Jiang, and B. Zheng, "C²MA-Net: Cross-modal cross-attention network for acute ischemic stroke lesion segmentation based on CT perfusion scans," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 1, pp. 108–118, Jan. 2022.
- [79] C. Wang, J. Uh, T. E. Merchant, C. H. Hua, and S. Acharya, "Facilitating MR-guided adaptive proton therapy in children using deep learning-based synthetic CT," *Int. J. Part. Ther.*, vol. 8, no. 3, pp. 11–20, 2022.
- [80] Z. Tan, J. Li, H. Tao, S. Li, and Y. Hu, "XctNet: Reconstruction network of volumetric images from a single X-ray image," *Computerized Med. Imag. Graph.*, vol. 98, 2022, Art. no. 102067.
- [81] F. Li et al., "Modeling long-range dependencies for weakly supervised disease classification and localization on chest X-ray," *Quantitative Imag. Med. Surg.*, vol. 12, no. 6, pp. 3364–3378, 2022.
- [82] J. X. Wang, Y. Li, X. Li, and Z. H. Lu, "Alzheimer's disease classification through imaging genetic data with IGnet," *Front. Neurosci.*, vol. 16, 2022, Art. no. 846638.
- [83] J. S. Hong et al., "Acceleration of magnetic resonance fingerprinting reconstruction using denoising and self-attention pyramidal convolutional neural network," *Sensors*, vol. 22, no. 3, 2022, Art. no. 1260.
- [84] M. Al-Shabi, K. Shak, and M. Tan, "ProCAN: Progressive growing channel attentive non-local network for lung nodule classification," *Pattern Recognit.*, vol. 122, 2022, Art. no. 108309.

- [85] D. Liu, F. Liu, Y. Tie, L. Qi, and F. Wang, "Res-trans networks for lung nodule classification," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 17, no. 6, pp. 1059–1068, 2022.
- [86] S. Zhao, Y. Chen, K. -F. Yang, Y. Luo, B. -Y. Ma, and Y. -J. Li, "A local and global feature disentangled network: Toward classification of benign-malignant thyroid nodules from ultrasound image," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1497–1509, Jun. 2022.
- [87] J. Zhao et al., "Semantic consistency generative adversarial network for cross-modality domain adaptation in ultrasound thyroid nodule classification," *Appl. Intell.*, vol. 52, no. 9, pp. 10369–10383, 2022.
- [88] L. Zhang et al., "Spatial adaptive and transformer fusion network (STFNet) for low-count PET blind denoising with MRI," *Med. Phys.*, vol. 49, no. 1, pp. 343–356, 2022.
- [89] J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, "Vision transformer-based recognition of diabetic retinopathy grade," *Med. Phys.*, vol. 48, no. 12, pp. 7850–7863, 2021.
- [90] L. T. Duong, N. H. Le, T. B. Tran, V. M. Ngo, and P. T. Nguyen, "Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning," *Expert Syst. with Appl.*, vol. 184, 2021, Art. no. 115519.
- [91] Z. AlNazi, F. Rabbi Mashrur, M. A. Islam, and S. Saha, "Fibro-CoSANet: Pulmonary fibrosis prognosis prediction using a convolutional self attention network," *Phys. Med. Biol.*, vol. 66, no. 22, 2021, Art. no. 225013.
- [92] Z. Lin et al., "AANet: Adaptive attention network for COVID-19 detection from chest X-ray images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4781–4792, Nov. 2021.
- [93] D. Shome et al., "Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare," *Int. J. Environ. Res. Public Health*, vol. 18, no. 21, 2021, Art. no. 11086.
- [94] Z. Wang, D. Peng, Y. Shang, and J. Gao, "Autistic spectrum disorder detection and structural biomarker identification using self-attention model and individual-level morphological covariance brain networks," *Front. Neurosci.*, vol. 15, 2021, Art. no. 756868.
- [95] Y. Fu, P. Xue, and E. Dong, "Densely connected attention network for diagnosing COVID-19 based on Chest CT," *Comput. Biol. Med.*, vol. 137, 2021, Art. no. 104857.
- [96] F. Rundo et al., "Three-dimensional deep noninvasive radiomics for the prediction of disease control in patients with metastatic urothelial carcinoma treated with immunotherapy," *Clin. Genitourinary Cancer*, vol. 19, no. 5, pp. 396–404, 2021.
- [97] H. Sun, A. Wang, W. Wang, and C. Liu, "An improved deep residual network prediction model for the early diagnosis of Alzheimer's disease," *Sensors*, vol. 21, no. 12, 2021, Art. no. 4182.
- [98] S. Estrada et al., "Automated olfactory bulb segmentation on high resolution T2-weighted MRI," *NeuroImage*, vol. 242, 2021, Art. no. 118464.
- [99] H. Xie et al., "Super-resolution of pneumocystis carinii pneumonia CT via self-attention GAN," *Comput. Methods Programs Biomed.*, vol. 212, 2021, Art. no. 106467.
- [100] Y. Li, J. Yang, J. Ni, A. Elazab, and J. Wu, "TA-Net: Triple attention network for medical image segmentation," *Comput. Biol. Med.*, vol. 137, 2021, Art. no. 104836.
- [101] X. Qu, G. Yan, D. Zheng, S. Fan, Q. Rao, and J. Jiang, "A deep learning-based automatic first-arrival picking method for ultrasound sound-speed tomography," *IEEE Trans. Ultrasonics, Ferroelect., Freq. Control*, vol. 68, no. 8, pp. 2675–2686, Aug. 2021.
- [102] M. Kim and B. D. Lee, "Automatic lung segmentation on Chest X-rays using self-attention deep neural network," *Sensors*, vol. 21, no. 2, pp. 1–12, 2021.
- [103] Q. Sun, N. Fang, Z. Liu, L. Zhao, Y. Wen, and H. Lin, "HybridCTrm: Bridging CNN and transformer for multimodal brain image segmentation," *J. Healthcare Eng.*, vol. 31, 2021, Art. no. 7467261.
- [104] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Inform. Med. Unlocked*, vol. 24, 2021, Art. no. 100557.
- [105] L. Mou et al., "CS2-Net: Deep learning segmentation of curvilinear structures in medical imaging," *Med. Image Anal.*, vol. 67, 2021, Art. no. 101874.
- [106] C. Z. Wu, J. Sun, J. Wang, L. F. Xu, and S. Zhan, "Encoding-decoding network with pyramid self-attention module for retinal vessel segmentation," *Int. J. Automat. Comput.*, vol. 18, no. 6, pp. 973–980, 2021.
- [107] D. Tomar, M. Lortkipanidze, G. Vray, B. Bozorgtabar, and J. P. Thiran, "Self-attentive spatial adaptive normalization for cross-modality domain adaptation," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2926–2938, Oct. 2021.
- [108] L. Xu et al., "Exploiting vector attention and context prior for ultrasound image segmentation," *Neurocomputing*, vol. 454, pp. 461–473, 2021.
- [109] Y. Dai, Y. Gao, and F. Liu, "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, 2021, Art. no. 1384.
- [110] T. Zhong et al., "DIKA-Nets: Domain-invariant knowledge-guided attention networks for brain skull stripping of early developing macaques," *NeuroImage*, vol. 227, 2021, Art. no. 117649.
- [111] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 1, pp. 121–130, Jan. 2021.
- [112] R. Xie et al., "End-to-end fovea localisation in colour fundus images with a hierarchical deep regression network," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 116–128, Jan. 2021.
- [113] Z. Zhang, T. Zhao, H. Gay, W. Zhang, and B. Sun, "Weaving attention U-net: A novel hybrid CNN and attention-based method for organs-at-risk segmentation in head and Neck CT images," *Med. Phys.*, vol. 48, no. 11, pp. 7052–7062, 2021.
- [114] W. Zhou, H. Du, W. Mei, and L. Fang, "Spatial orthogonal attention generative adversarial network for MRI reconstruction," *Med. Phys.*, vol. 48, no. 2, pp. 627–639, 2021.
- [115] Z. Hu, H. Liu, Z. Li, and Z. Yu, "Cross-model transformer method for medical image synthesis," *Complexity*, vol. 31, 2021, Art. no. 5624909.
- [116] S. Lee, E. Kim, J. S. Bae, J. H. Kim, and S. Yoon, "Robust end-to-end focal liver lesion detection using unregistered multiphase computed tomography images," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 2, pp. 319–327, Apr. 2023.
- [117] Z. Zhou, Y. Wang, Y. Guo, X. Jiang, and Y. Qi, "Ultrafast plane wave imaging with line-scan-quality using an ultrasound-transfer generative adversarial network," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 4, pp. 943–956, Apr. 2020.
- [118] H. Xue et al., "A 3D attention residual encoder-decoder least-square GAN for low-count PET denoising," *Nucl. Instruments Methods Phys. Res., Sect. A: Accelerators, Spectrometers, Detectors Assoc. Equip.*, vol. 983, 2020, Art. no. 164638.
- [119] J. M. J. Valanarasu, R. Yasarla, P. Wang, I. Hacıhaliloglu, and V. M. Patel, "Learning to segment brain anatomy from 2D ultrasound with less data," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1221–1234, Oct. 2020.
- [120] Z. Kuang, X. Deng, L. Yu, H. Wang, T. Li, and S. Wang, "Ψ -Net: Focusing on the border areas of intracerebral hemorrhage on CT images," *Comput. Methods Programs Biomed.*, vol. 194, 2020, Art. no. 105546.
- [121] W. Xie, C. Jacobs, J. P. Charbonnier, and B. Van Ginneken, "Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2664–2675, Aug. 2020.
- [122] B. Lei et al., "Self-co-attention neural network for anatomy segmentation in whole breast ultrasound," *Med. Image Anal.*, vol. 64, 2020, Art. no. 101753.
- [123] X. Jia, Y. Liu, Z. Yang, and D. Yang, "Multi-modality self-attention aware deep network for 3D biomedical segmentation," *BMC Med. Inform. Decis. Mak.*, vol. 20, 2020, Art. no. 119.
- [124] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "SACNN: Self-attention convolutional neural network for low-dose ct denoising with self-supervised perceptual loss network," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2289–2301, Jul. 2020.
- [125] S. Li, C. Ge, X. Sui, Y. Zheng, and W. Jia, "Channel and spatial attention regression network for cup-to-disc ratio estimation," *Electronics*, vol. 9, no. 6, 2020, Art. no. 909.
- [126] T. Fan, G. Wang, Y. Li, and H. Wang, "MA-net: A multi-scale attention network for Liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179656–179665, 2020.
- [127] S. Gou, N. Tong, S. Qi, S. Yang, R. Chin, and K. Sheng, "Self-channel-and-spatial-attention neural network for automated multi-organ segmentation on head and Neck CT images," *Phys. Med. Biol.*, vol. 65, no. 24, 2020, Art. no. 245034.
- [128] Y. Wu, D. Li, L. Xing, and G. Gold, "Deriving new soft tissue contrasts from conventional MR images using deep learning," *Magn. Reson. Imag.*, vol. 74, pp. 121–127, 2020.
- [129] Z. Yuan et al., "SARA-GAN: Self-attention and relative average discriminator based generative adversarial networks for fast compressed sensing MRI reconstruction," *Front. Neuroinform.*, vol. 14, 2020, Art. no. 611666.

- [130] P. Zhong, J. Wang, Y. Guo, X. Fu, and R. Wang, "Multiclass retinal disease classification and lesion segmentation in OCT B-scan images using cascaded convolutional networks," *Appl. Opt.*, vol. 59, no. 33, pp. 10312–10320, 2020.
- [131] Y. Liu et al., "CT-based multi-organ segmentation using a 3D self-attention U-net network for pancreatic radiotherapy," *Med. Phys.*, vol. 47, no. 9, pp. 4316–4324, 2020.
- [132] Y. Liu et al., "CBCT-based synthetic CT generation using deep-attention cycleGAN for pancreatic adaptive radiotherapy," *Med. Phys.*, vol. 47, no. 6, pp. 2472–2483, 2020.
- [133] V. Das, E. Prabhakararao, S. Dandapat, and P. K. Bora, "B-scan attentive CNN for the classification of retinal optical coherence tomography volumes," *IEEE Signal Process. Lett.*, vol. 27, pp. 1025–1029, 2020.
- [134] M. Klimont, M. Flieger, J. Rzeszutek, J. Stachera, A. Zakrzewska, and K. Jończyk-Potoczna, "Automated ventricular system segmentation in paediatric patients treated for hydrocephalus using deep learning methods," *BioMed Res. Int.*, vol. 31, 2019, Art. no. 3059170.
- [135] S. S. Mishra, B. Mandal, and N. B. Puhan, "Multi-level dual-attention based CNN for macular optical coherence tomography classification," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1793–1797, Dec. 2019.
- [136] X. Dong et al., "Synthetic CT generation from non-attenuation corrected PET images for whole-body PET imaging," *Phys. Med. Biol.*, vol. 64, no. 21, 2019, Art. no. 215016.
- [137] P. Song, Y. C. Eldar, G. Mazor, and M. R. D. Rodrigues, "HYDRA: Hybrid deep magnetic resonance fingerprinting," *Med. Phys.*, vol. 46, no. 11, pp. 4951–4969, 2019.
- [138] T. Wang et al., "Deep learning-based image quality improvement for low-dose computed tomography simulation in radiation therapy," *J. Med. Imag.*, vol. 6, no. 4, 2019, Art. no. 43504.
- [139] Y. Wu et al., "Self-attention convolutional neural network for improved MR image reconstruction," *Inf. Sci.*, vol. 490, pp. 317–328, 2019.
- [140] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.
- [141] Z. Liu et al., "A ConvNet for the 2020s," 2022, *arXiv:2201.03545*.
- [142] A. Chatsias et al., "Disentangle, align and fuse for multimodal and semi-supervised image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 781–792, Mar. 2021.
- [143] A. Chatsias, "Factorised spatial representation learning: Application in semi-supervised myocardial segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2018, pp. 490–498.
- [144] A. Chatsias et al., "Multimodal cardiac segmentation using disentangled representations learning," in *Proc. MICCAI STACOM: Statistical Atlases Comput. Models Heart Workshop. Lecture Notes Comput. Sci.*, 2019, vol. 12009.
- [145] H. Jiang et al., "Semi-supervised pathology segmentation with disentangled representations," in *Proc. MICCAI Workshop Domain Adapt. Representation Transfer Distrib. Collaborative Learn. Workshop. Lecture Notes Comput. Sci.*, 2020, vol. 12444.
- [146] A. Chatsias et al., "Disentangled representation learning in cardiac image analysis," *Med. Image Anal.*, vol. 58, 2019, Art. no. 101535.
- [147] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.
- [148] O. Oktay et al., "Attention u-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.
- [149] X. Xing, G. Papanastasiou, S. Walsh, and G. Yang, "Less is More: Unsupervised mask-guided annotated CT image synthesis with minimum manual segmentations," *IEEE Trans. Med. Imag.*, vol. 42, no. 9, pp. 2566–2576, Sep. 2023, doi: [10.1109/TMI.2023.3260169](https://doi.org/10.1109/TMI.2023.3260169), 2023.
- [150] G. Papanastasiou et al., "Measurement of myocardial blood flow by cardiovascular magnetic resonance perfusion: Comparison of distributed parameter and Fermi models with single and dual bolus," *J. Cardiovasc. Magn. Reson.*, vol. 17, no. 1, 2015, Art. no. 17.
- [151] G. Papanastasiou et al., "Quantitative assessment of myocardial blood flow in coronary artery disease by cardiovascular magnetic resonance: Comparison of Fermi and distributed parameter modeling against invasive methods," *J. Cardiovasc. Magn. Reson.*, vol. 18, 2016, Art. no. 57.
- [152] V. Hamy et al., "Respiratory motion correction in dynamic MRI using robust data decomposition registration-application to DCE-MRI," *Med. Image Anal.*, vol. 18, no. 2, pp. 301–313, 2014.
- [153] S. Bakas et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*.
- [154] L. Csincsik et al., "Peripheral retinal imaging biomarkers for Alzheimer's disease: A pilot study," *Ophthalmic Res.*, vol. 59, no. 4, pp. 182–119, 2018.
- [155] A. Villaplana-Velasco et al., "Fine-mapping of retinal vascular complexity loci identifies Notch regulation as a shared mechanism with myocardial infarction outcomes," *Commun. Biol.*, vol. 6, 2023, Art. no. 523.
- [156] S. J. Wiseman et al., "Retinal capillary microvessel morphology changes are associated with vascular damage and dysfunction in cerebral small vessel disease," *J. Cereb. Blood Flow Metab.*, vol. 43, no. 2, pp. 231–240, 2023.
- [157] G. Spyretta and D. D. Cokkinos, "Recent advances in vascular ultrasound imaging technology and their clinical implications," *Ultrasonics*, vol. 119, 2022, Art. no. 106599.
- [158] S. A. Tsafaris, X. Zhou, R. Tang, D. Li, and R. Dharmakumar, "Detecting myocardial ischemia at rest with cardiac phase-resolved blood oxygen level-dependent cardiovascular magnetic resonance," *Circulation: Cardiovasc. Imag.*, vol. 6, no. 2, pp. 311–319, 2013.
- [159] G. Papanastasiou et al., "Pharmacokinetic modelling for the simultaneous assessment of perfusion and 18F-flutemetamol uptake in cerebral amyloid angiopathy using a reduced PET-MR acquisition time: Proof of concept," *NeuroImage*, vol. 225, 2021, Art. no. 117482.
- [160] G. Papanastasiou et al., "Multimodality quantitative assessments of myocardial perfusion using dynamic contrast enhanced magnetic resonance and 15O-labeled water positron emission tomography imaging," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 2, no. 3, pp. 259–271, May 2018.
- [161] C. Wang et al., "DiCyc: GAN-based deformation invariant cross-domain information fusion for medical image synthesis," *Inf. Fusion*, vol. 67, pp. 147–160, 2021.
- [162] C. Wang, G. Yang, and G. Papanastasiou, "Unsupervised image registration towards enhancing performance and explainability in cardiac and brain image analysis," *Sensors*, vol. 22, no. 6, 2022, Art. no. 2125.
- [163] C. Wang, G. Yang, and G. Papanastasiou, "FIRE: Unsupervised bidirectional inter- and intra-modality registration using deep networks," in *Proc. IEEE 34th Int. Symp. Comput.-Based Med. Syst.*, 2021, pp. 510–514, doi: [10.1109/CBMS52027.2021.00101](https://doi.org/10.1109/CBMS52027.2021.00101).
- [164] B. Schölkopf et al., "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.
- [165] G. Zhou et al., "On the opportunity of causal deep generative models: A survey and future directions," 2023, *arXiv:2301.12351*.
- [166] F. B. Ahmad, J. A. Cisewski, J. Xu, and R. N. Anderson, "Provisional mortality data - United States, 2022," *Morbidity and Mortality Weekly Rep.*, vol. 72, pp. 488–492, 2023.
- [167] A. Radford et al., "Learning transferable visual models from natural language supervision," 2019, *arXiv:2103.00020*.
- [168] ChatGPT: Optimizing Language Models for Dialogue, *OpenAI*, 2022. [Online]. Available: <https://openai.com/blog/chatgpt/>
- [169] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.
- [170] OpenAI, *GPT-4 Technical Report*, 2013, *arXiv: 2303.08774*.
- [171] H. Touvron et al., "Open foundation and fine-tuned chat models," 2019, *arXiv:2307.09288*.
- [172] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "Chatcad: Interactive computer-aided diagnosis on medical image using large language models," 2023, *arXiv:2302.07257*.
- [173] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "VisualGPT: Data-efficient adaptation of pretrained language models for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18009–18019.
- [174] Y. Xiao and W. Y. Wang, "On hallucination and predictive uncertainty in conditional language generation," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics: Main Volume*, 2021, pp. 2734–2744.