Medical Image Analysis 88 (2023) 102802

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Survey paper

Transformers in medical imaging: A survey

Fahad Shamshad [a,*], Salman Khan [a,b], Syed Waqas Zamir [c], Muhammad Haris Khan [a], Munawar Hayat [d], Fahad Shahbaz Khan [a,e], Huazhu Fu [f]

[a] MBZ University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[b] CECS, Australian National University, Canberra ACT 0200, Australia
[c] Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[d] Faculty of IT, Monash University, Clayton VIC 3800, Australia
[e] Computer Vision Laboratory, Linköping University, Sweden
[f] Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore

ARTICLE INFO

Keywords:
Transformers
Medical image analysis
Vision transformers
Deep neural networks
Clinical report generation

ABSTRACT

Following unprecedented success on the natural language tasks, Transformers have been successfully applied to several computer vision problems, achieving state-of-the-art results and prompting researchers to reconsider the supremacy of convolutional neural networks (CNNs) as de facto operators. Capitalizing on these advances in computer vision, the medical imaging field has also witnessed growing interest for Transformers that can capture global context compared to CNNs with local receptive fields. Inspired from this transition, in this survey, we attempt to provide a comprehensive review of the applications of Transformers in medical imaging covering various aspects, ranging from recently proposed architectural designs to unsolved issues. Specifically, we survey the use of Transformers in medical image segmentation, detection, classification, restoration, synthesis, registration, clinical report generation, and other tasks. In particular, for each of these applications, we develop taxonomy, identify application-specific challenges as well as provide insights to solve them, and highlight recent trends. Further, we provide a critical discussion of the field's current state as a whole, including the identification of key challenges, open problems, and outlining promising future directions. We hope this survey will ignite further interest in the community and provide researchers with an up-to-date reference regarding applications of Transformer models in medical imaging. Finally, to cope with the rapid development in this field, we intend to regularly update the relevant latest papers and their open-source implementations at https://github.com/fahadshamshad/awesome-transformers-in-medical-imaging.

1. Introduction

Convolutional Neural Networks (CNNs) (Goodfellow et al., 2016; LeCun et al., 1989; Krizhevsky et al., 2012; Liu et al., 2022b) have significantly impacted the field of medical imaging due to their ability to learn highly complex representations in a data-driven manner. Since their renaissance, CNNs have demonstrated remarkable improvements on numerous medical imaging modalities, including Radiography (Lakhani and Sundaram, 2017), Endoscopy (Min et al., 2019), Computed Tomography (CT) (Würfl et al., 2016; Lell and Kachelrieß, 2020), Mammography Images (MG) (Hamidinekoo et al., 2018), Ultrasound Images (Liu et al., 2019), Magnetic Resonance Imaging (MRI) (Lundervold and Lundervold, 2019; Akkus et al., 2017), and Positron Emission Tomography (PET) (Reader et al., 2020), to name a few. The workhorse in CNNs is the *convolution* operator, which operates locally and provides translational equivariance. While these properties help in developing efficient and generalizable medical imaging solutions, the local receptive field in convolution operation limits capturing long-range pixel relationships. Furthermore, the convolutional filters have stationary weights that are not adapted for the given input image content at inference time.

Meanwhile, significant research effort has been made by the vision community to integrate the attention mechanisms (Vaswani et al., 2017; Devlin et al., 2018; Fedus et al., 2021) in CNN-inspired architectures (Wang et al., 2018b; Yin et al., 2020; Ramachandran et al., 2019; Bello et al., 2019; Vaswani et al., 2021; Dosovitskiy et al., 2020). These attention-based 'Transformer' models have become an attractive solution due to their ability to encode long-range dependencies and learn highly effective feature representations (Chaudhari et al., 2019). Recent works have shown that these Transformer modules can fully replace the standard convolutions in deep neural networks by operating

* Corresponding author.
E-mail address: fahad.shamshad@mbzuai.ac.ae (F. Shamshad).

https://doi.org/10.1016/j.media.2023.102802
Received 3 February 2022; Received in revised form 11 March 2023; Accepted 23 March 2023
Available online 5 April 2023

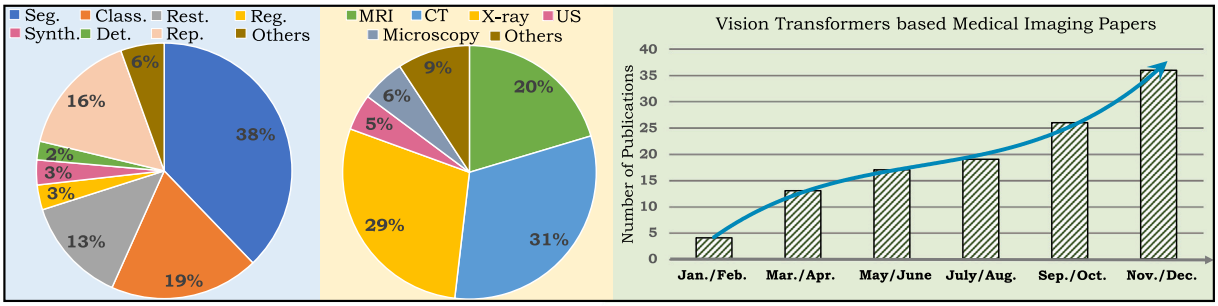

Fig. 1. (Left) The pie-charts show statistics of the papers included in this survey according to medical imaging problem settings and data modalities. The rightmost figure shows consistent growth in the recent literature (for year 2021). Seg: segmentation, Class: classification, Rest: restoration, Reg: registration, Synth: synthesis, Det: detection, Rep: report generation, US: ultrasound.

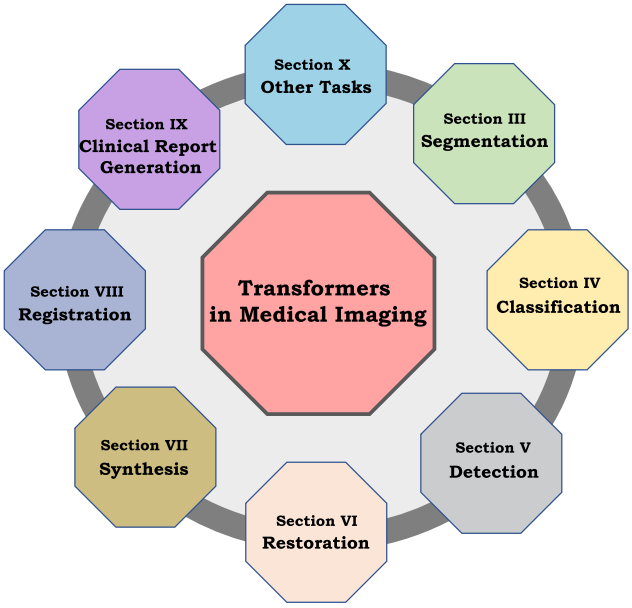

Fig. 2. A diverse set of application areas of Transformers in medical imaging covered in this survey.

on a sequence of image patches, giving rise to Vision Transformers (ViTs) (Dosovitskiy et al., 2020). Since their inception, ViT models have been shown to push the state-of-the-art in numerous vision tasks, including image classification (Dosovitskiy et al., 2020), object detection (Zhu et al., 2020), semantic segmentation (Zheng et al., 2021b), image colorization (Kumar et al., 2021), low-level vision (Chen et al., 2021j), and video understanding (Arnab et al., 2021) to name a few. Furthermore, recent research indicates that the prediction errors of ViTs are more consistent with those of humans than CNNs (Naseer et al., 2021a; Portelance et al., 2021; Geirhos et al., 2021; Tuli et al., 2021). These desirable properties of ViTs have sparked great interest in the medical community to adapt them for medical imaging applications, thereby mitigating the inherent inductive biases of CNNs (Matsoukas et al., 2021).

Motivation and Contributions: Recently, medical imaging community has witnessed an exponential growth in the number of Transformer based techniques, especially after the inception of ViTs (see Fig. 1). The topic is now gaining more attention in the medical imaging community, and it is getting increasingly difficult to keep pace with the recent progress due to the rapid influx of papers. As such, a survey of the existing relevant works is timely to provide a comprehensive account of new methods in this emerging field. To this end, we provide a holistic overview of the applications of Transformer models in medical imaging. We hope this work can provide a roadmap for the researchers to explore the field further. Our major contributions include:

- This is the first survey paper that comprehensively covers applications of Transformers in the medical imaging domain, thereby bridging the gap between the vision and medical imaging community in this rapidly evolving area. Specifically, we present a comprehensive overview of more than 125 relevant papers to cover the recent progress.
- We provide a detailed coverage of the field by categorizing the papers based on their applications in medical imaging as depicted in Fig. 2. For each of these applications, we develop a taxonomy, highlight task-specific challenges, and provide insights about solving them based on the literature reviewed.
- Finally, we provide a critical discussion of the field's current state as a whole, including identifying key challenges, highlighting open problems, and outlining promising future directions.
- Although the main focus of this survey is on Vision Transformers, we are also the first since the inception of the original Transformer, about half a decade ago, to extensively cover its language modeling capabilities in the clinical report generation task (see Section 9).

Paper Organization. The rest of the paper is organized as follows. In Section 2, we provide background of the field with a focus on salient concepts underlying Transformers. From Sections 3 to 10, we comprehensively cover applications of Transformers in several medical imaging tasks as shown in Fig. 2. In particular, for each of these tasks, we develop a taxonomy and identify task-specific challenges. Section 11 presents open problems and future directions about the field as a whole. Finally, in Section 12, we give recommendations to cope with the rapid development of the field and conclude the paper.

2. Background

Medical imaging approaches have undergone significant advances over the past few decades. In this section, we briefly provide a background of these advancements and broadly group them into CNN-based and ViT-based approaches. For the CNN-based methods, we describe the underlying working principles along with their major strengths and shortcomings in the context of medical imaging. For the ViT-based methods, we highlight core concepts behind their success and defer further details to later sections.

2.1. CNN-based methods

CNNs are effective at learning discriminative features and extracting generalizable priors from large-scale medical datasets, thus providing excellent performance on medical imaging tasks, making them an integral component of modern AI-based medical imaging systems. The

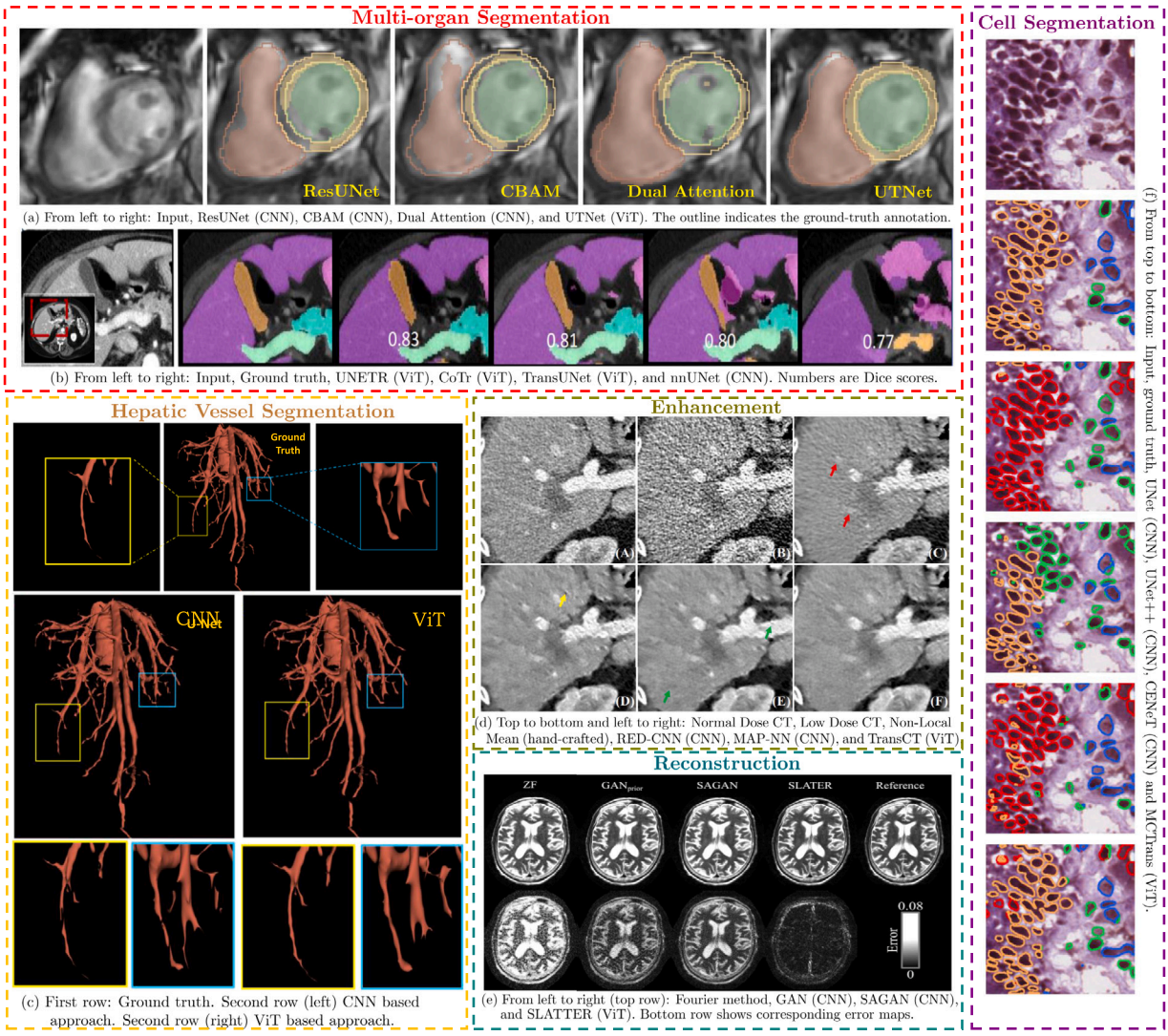

Fig. 3. Applications of ViTs in various medical imaging problems along with the baseline CNN-based approaches. ViT-based approaches give superior performance as compared to CNN-based methods due to their ability to model the global context. Figure sources: (a) Gao et al. (2021c), (b) Hatamizadeh et al. (2021), (c) Wu et al. (2021c), (d) Zhang et al. (2021h), (e) Korkmaz et al. (2021a), (f) Ji et al. (2021).

advancements in CNNs have been mainly fueled by novel architectural designs, better optimization procedures, availability of special hardware (*e.g.,* GPUs) and purpose-built open source software libraries (Gibson et al., 2018; Pérez-García et al., 2021; Beers et al., 2021). We refer interested readers to comprehensive survey papers related to CNNs applications in medical imaging (Yi et al., 2019; Litjens et al., 2017; Greenspan et al., 2016; Zhou et al., 2017; Shen et al., 2017; Cheplygina et al., 2019; Hesamian et al., 2019; Duncan et al., 2019; Haskins et al., 2020; Zhou et al., 2021a). Despite considerable performance gains, the reliance of CNNs on large labeled datasets limits their applicability over the full spectrum of medical imaging tasks. Furthermore, CNNs-based approaches are generally more challenging to interpret and often act as black box solutions. Therefore, there has been an increasing effort in the medical imaging community to amalgamate the strengths of hand-crafted and CNNs based methods resulting in the prior information-guided CNNs models (Shlezinger et al., 2020). These hybrid methods contain special domain-specific layers, and include unrolled optimization (Monga et al., 2021), generative models (Ongie et al., 2020), and learned denoiser-based approaches (Ahmad et al., 2020). Despite these architectural and algorithmic advancements, the decisive factor behind CNNs success has been primarily attributed to their image-specific inductive bias in dealing with scale invariance and modeling local visual structures. While this intrinsic locality (limited receptive field) brings efficiency to CNNs, it impairs their ability to capture long-range spatial dependencies in an input image, thereby stagnating performance (Matsoukas et al., 2021) (see Fig. 3). This demands an alternative architectural design capable of modeling long-range pixel relationships for better representation learning.

2.2. Transformers

Transformers were introduced by Vaswani et al. (2017) as a new attention-driven building block for machine translation. Specifically, these attention blocks are neural network layers that aggregate information from the entire input sequence (Bahdanau et al., 2014). Since their inception, these models have demonstrated state-of-the-art performance on several Natural Language Processing (NLP) tasks, thereby becoming the default choice over recurrent models. In this section, we will focus on Vision Transformers (ViTs) (Dosovitskiy et al., 2020) that are built on vanilla Transformer model (Vaswani et al., 2017) by cascading multiple transformer layers to capture the global context of an input image. Specifically, Dosovitskiy et al. (2020) interpret an image as a sequence of patches and process it by a standard transformer encoder as used in NLP. These ViT models continue the long-lasting trend of removing hand-crafted visual features and inductive biases

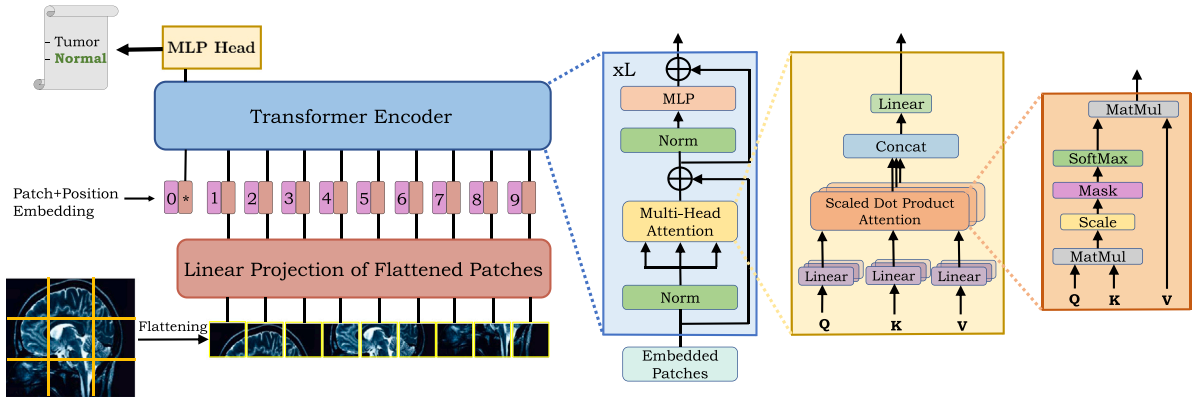

Fig. 4. Architecture of the Vision Transformer (*on the left*) and details of the Vision Transformer encoder block (*on the right*). Vision transformer first splits the input image into patches and projects them (after flattening) into a feature space where a transformer encoder processes them to produce the final classification output.

Algorithm 1: ViT Working Principle.

1: Split a medical image into patches of fixed sizes
2: Vectorize image patches via flattening operation
3: Create lower-dimensional linear embedding from vectorized patches via trainable linear layer
4: Add positional encoding to lower dimensional linear embeddings
5: Feed the sequence to ViT encoder as shown in Fig. 4
6: Pre-train the ViT model on a large-scale image dataset
7: Fine-tune on the down stream medical image classification task

from models in an effort to leverage the availability of larger datasets coupled with increased computational capacity. ViTs have garnered immense interest in the medical imaging community, and a number of recent approaches have been proposed which build upon ViTs. We highlight the working principle of ViT in a step-by-step manner in Algorithm 1 for medical image classification (also see Fig. 4).

Below, we briefly describe the core components behind the success of ViTs that are *self-attention* and *multi-head self-attention*. For a more in-depth analysis of numerous ViT architectures and applications, we refer interesting readers to the recent relevant survey papers (Chaudhari et al., 2019; Han et al., 2020; Khan et al., 2021; Tay et al., 2020; Lin et al., 2021b).

2.2.1. Self-attention

The success of the Transformer models has been widely attributed to the self-attention (SA) mechanism due to its ability to model long-range dependencies. The key idea behind the SA mechanism is to learn self-alignment, that is, to determine the relative importance of a single token (patch embedding) with respect to all other tokens in the sequence (Bahdanau et al., 2014). For 2D images, we first reshape the image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2C)}$, where H and W denotes height and width of the original image respectively, C is the number of channels, $P\times P$ is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches. These flattened patches are projected to D dimension via trainable linear projection layer and can be represented in matrix form as $\mathbf{X} \in \mathbb{R}^{N\times D}$. The goal of self-attention is to capture the interaction amongst all these N embeddings, that is done by defining three learnable weight matrices to transform input $\mathbf{X}$ into Queries (via $\mathbf{W}^Q \in \mathbb{R}^{D\times D_q}$), Keys (via $\mathbf{W}^K \in \mathbb{R}^{D\times D_k}$) and Values (via $\mathbf{W}^V \in \mathbb{R}^{D\times D_v}$), where $D_q = D_k$. The input sequence $\mathbf{X}$ is first projected onto these weight matrices to get $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}^K$ and $\mathbf{V} = \mathbf{X}\mathbf{W}^V$. The corresponding attention matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ can be written as,

$$\mathbf{A} = \mathbf{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D_q}}\right).$$

The output $\mathbf{Z} \in \mathbb{R}^{N\times D_v}$ of the SA layer is then given by,

$$\mathbf{Z} = \mathrm{SA}(\mathbf{X}) = \mathbf{A}\mathbf{V}.$$

2.2.2. Multi-head self-attention

Multi-Head Self Attention (MHSA) consists of multiple SA blocks (heads) concatenated together channel-wise to model complex dependencies between different elements in the input sequence. Each head has its own learnable weight matrices denoted by $\{\mathbf{W}^{Q_i}, \mathbf{W}^{K_i}, \mathbf{W}^{V_i}\}$, where $i = 0\cdots(h-1)$ and h denotes total number of heads in MHSA block. Specifically, we can write,

$$\mathrm{MHSA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = [\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_{h-1}]\mathbf{W}^O,$$

whereas $\mathbf{W}^O \in \mathbb{R}^{h.D_v\times N}$ computes linear transformation of heads and $\mathbf{Z}_i$ can be written as,

$$\mathbf{Z}_i = \mathbf{softmax}\left(\frac{\mathbf{Q}\mathbf{W}^{Q_i}(\mathbf{K}\mathbf{W}^{K_i})^T}{\sqrt{D_q/h}}\right)\mathbf{V}\mathbf{W}^{V_i}.$$

Note that the complexity of computing the softmax for SA block is quadratic with respect to the length of the input sequence that can limit its applicability to high-resolution medical images. Recently, numerous efforts have been made to reduce complexity, including sparse attention (Rao et al., 2021), linearization attention (Katharopoulos et al., 2020), low-rank attention (Xiong et al., 2021), memory compression based approaches (Choromanski et al., 2020), and improved MHSA (Shazeer et al., 2020). We will discuss the efficient SA in the context of medical imaging in the relevant sections.

Further, we find it important to clarify that several alternate attention approaches (Jin et al., 2020; Schlemper et al., 2019; Maji et al., 2022; Guo et al., 2021) have been explored in the literature based on convolutional architectures. In this survey, we focus on the specific attention used in transformer blocks (MHSA) which has recently gained significant research attention in medical image analysis. Next, we outline these methods categorized according to specific application domains.

3. Medical image segmentation

Accurate medical image segmentation is a crucial step in computer-aided diagnosis, image-guided surgery, and treatment planning. The global context modeling capability of Transformers is crucial for accurate medical image segmentation because the organs spread over a large receptive field can be effectively encoded by modeling the relationships between spatially distant pixels (e.g., lungs segmentation). Furthermore, the background in medical scans is generally scattered (e.g., in ultrasound scan (Avola et al., 2021)); therefore, learning global context between the pixels corresponding to the background can help the model in preventing misclassification.

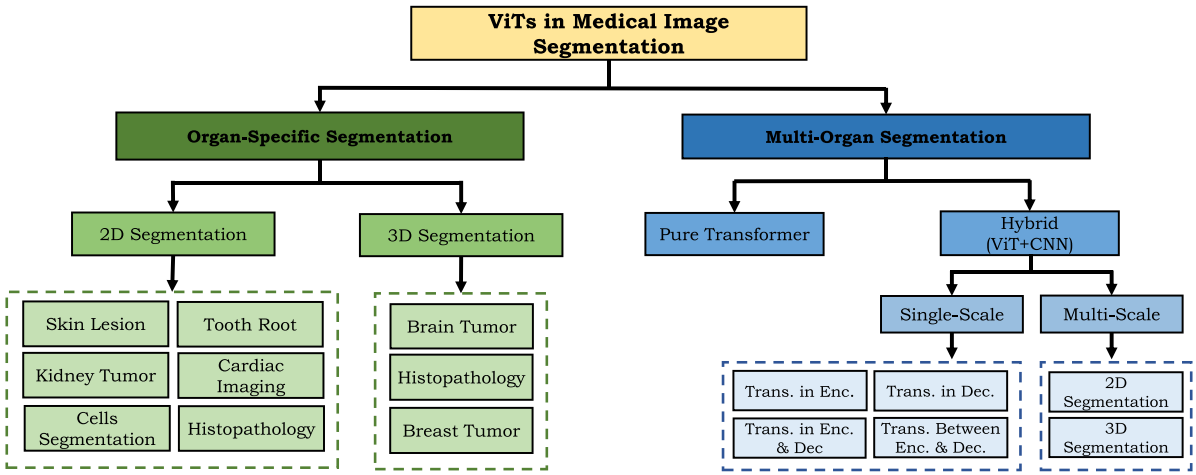

Fig. 5. Taxonomy of ViT-based medical image segmentation approaches.

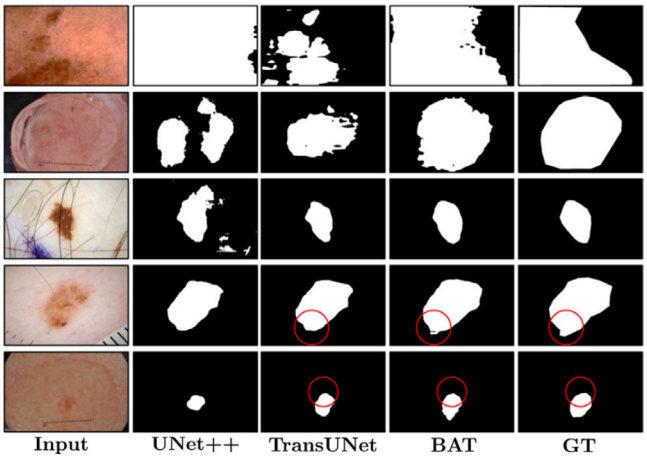

Fig. 6. Comparison of different skin lesion segmentation approaches. From left to right: Input image, CNN based UNet++ (Zhou et al., 2019b), ViT-based TransUNet (Chen et al., 2021g), Boundary aware transformer (BAT) (Wang et al., 2021e), and the ground truth (GT) image. Red circles highlight small regions with an ambiguous boundary where BAT can perform well due to the use of boundary-wise prior knowledge.
Source: Image taken from Wang et al. (2021e).

Below, we highlight various attempts to integrate ViT-based models for medical image segmentation. We broadly classify the ViT-based segmentation approaches into *organ-specific* and *multi-organ* categories, as depicted in Fig. 5, due to the varying levels of context modeling required in both sets of methods.

3.1. Organ-specific segmentation

ViT-based organ-specific approaches generally consider a specific aspect of the underlying organ to design architectural components or loss functions. We mention specific examples of such design choices in this section. We have further categorized organ-specific categories into 2D and 3D-based approaches depending on the input type.

3.1.1. 2D segmentation

Here, we describe the organ-specific ViT-based segmentation approaches for 2D medical scans.

Skin Lesion Segmentation. Accurate skin lesion segmentation for identifying melanoma (cancer cells) is crucial for cancer diagnosis and subsequent treatment planning. However, it remains a challenging task due to significant variations in color, size, occlusions, and contrast of skin lesion areas, resulting in ambiguous boundaries (Yuan, 2017) and consequently deterioration in segmentation performance. To address the issue of ambiguous boundaries, Wang et al. (2021e) propose a novel Boundary-Aware Transformer (BAT). Specifically, they design a boundary-wise attention gate in Transformer architecture to exploit the prior knowledge about boundaries. The auxiliary supervision of the boundary-wise attention gate provides feedback to train BAT effectively. Extensive experiments on ISIC 2016+PH2 (Gutman et al., 2016; Mendonça et al., 2013) and ISIC 2018 (Codella et al., 2019) validate the efficacy of their boundary-wise prior, as shown in Fig. 6. Similarly, Wu et al. (2021a) propose a dual encoder-based feature adaptive transformer network (FAT-Net) that consists of CNN and transformer branches in the encoder. To effectively fuse the features from these two branches, a memory-efficient decoder and feature adaptation module have been designed. Experiments on ISIC 2016–2018 (Gutman et al., 2016; Berseth, 2017; Codella et al., 2019), and PH2 (Mendonça et al., 2013) datasets demonstrate the effectiveness of FAT-Net fusion modules.

Tooth Root Segmentation. Tooth root segmentation is one of the critical steps in root canal therapy to treat periodontitis (gum infection) (Gao and Chae, 2010). However, it is challenging due to blurry boundaries and overexposed and underexposed images. To address these challenges, Li et al. (2021e) propose Group Transformer UNet (GT UNet) that consists of transformer and convolutional layers to encode global and local context, respectively. A shape-sensitive Fourier Descriptor loss function (Zahn and Roskies, 1972) has been proposed to deal with the fuzzy tooth boundaries. Furthermore, grouping and bottleneck structure has been introduced in the GT UNet to significantly reduce the computational cost. Experiments on their in-house Tooth Root segmentation dataset with six evaluation metrics demonstrate the effectiveness of GT UNet architectural components and Fourier-based loss function. In another work, Li et al. (2021g) propose anatomy-guided multibranch Transformer (AGMB-Transformer) to incorporate the strengths of group convolutions (Chollet, 2017) and progressive Transformer network. Experiments on their self-collected dataset of 245 tooth root X-ray images show the effectiveness of AGMB-Transformer.

Cardiac Image Segmentation. Despite their impressive performance in medical image segmentation, Transformers are computationally demanding to train and come with a high parameter budget. To handle these challenges for cardiac image segmentation task, Deng (Deng et al., 2021) propose TransBridge, a lightweight parameter-efficient hybrid model. TransBridge consists of Transformers and CNNs based encoder–decoder structure for left ventricle segmentation in echocardiography. Specifically, the patch embedding layer of the Transformer has been re-designed using the shuffling layer (Zhang and Yang, 2021) and group convolutions to significantly reduce the number of

Table 1
Evaluation of Transformer-based semantic segmentation methods for pathological image segmentation in terms of average jaccard index on PAIP liver histopathological dataset (Kim et al., 2021). It can be seen that transformer-based models tend to outperform CNNs with the exception of Swin-UNet. Results are from Nguyen et al. (2021a), which is one of the first study to systematically evaluate the performance of transformers on pathological image segmentation task.

CNN-based Models	40x magnification	20x magnification
PSPNet (Zhao et al., 2017)	0.58 ± 0.33	0.49 ± 0.27
UNet (Ronneberger et al., 2015)	0.65 ± 0.24	0.60 ± 0.31
DeepLabV3 (Chen et al., 2017)	0.63 ± 0.28	0.67 ± 0.24
FPN (Lin et al., 2017a)	0.64 ± 0.20	0.72 ± 0.22
PAN (Li et al., 2018)	0.63 ± 0.24	0.69 ± 0.23
LinkNet (Chaurasia and Culurciello, 2017)	0.35 ± 0.33	0.54 ± 0.25
Transformer-based models	**40x magnification**	**20x magnification**
TransUNet (Chen et al., 2021g)	0.77 ± 0.12	0.77 ± 0.13
Swin-UNet (Cao et al., 2021)	0.53 ± 0.23	0.42+0.23
Swin Transformer (Base) (Liu et al., 2021a)	0.79 ± 0.14	0.71 ± 0.26
Segmenter (Strudel et al., 2021)	0.80 ± 0.14	0.82 ± 0.11
Medical Transformer (Valanarasu et al., 2021)	0.71 ± 0.14	0.62 ± 0.17
BEiT (Bao et al., 2021)	0.72 ± 0.21	0.66 ± 0.28

parameters. Extensive experiments on large-scale left ventricle segmentation dataset, echo-cardiographs (Ouyang et al., 2020) demonstrate the benefit of TransBridge over CNNs and Transformer-based baseline approaches (Xie et al., 2021a).

Kidney Tumor Segmentation. Accurate segmentation of kidney tumors via computer diagnosis systems can reduce the effort of radiologists and is a critical step in related surgical procedures. However, it is challenging due to varying kidney tumor sizes and the contrast between tumors and their anatomical surroundings. To address these challenges, Shen et al. (2021b) propose a hybrid encoder–decoder architecture, COTR-Net, that consists of convolution and transformer layers for end-to-end kidney, kidney cyst, and kidney tumor segmentation. Specifically, the encoder of COTR-Net consists of several convolution-transformer blocks, and the decoder comprises several up-sampling layers with skip connections from the encoder. The encoder weights have been initialized using a pre-trained ResNet (He et al., 2016) architecture to accelerate convergence, and deep supervision has been exploited in the decoder layers to boost segmentation performance. Furthermore, the segmentation masks are refined using morphological operations as a post-processing step. Extensive experiments on the Kidney Tumor Segmentation dataset (KiTS21) (KiTS, 2021) demonstrate the effectiveness COTR-Net. Although COTR-Net is able to surpass the vanilla U-Net in terms of dice score, its rank on the KiTS21 challenge leader board is 22nd out of 25 teams, with the top 3 teams having their model based on the advance variants of CNN.

Cell Segmentation. Inspired from the Detection Transformers (DETR) (Carion et al., 2020; Prangemeier et al., 2020) propose Cell-DETR, a Transformer-based framework for instance segmentation of biological cells. Specifically, they integrate a dedicated attention branch to the DETR framework to obtain instance-wise segmentation masks in addition to box predictions. During training, focal loss (Lin et al., 2017b) and Sorenson dice loss (Carion et al., 2020) are used for the segmentation branch. To enhance performance, they integrate three residual decoder blocks (He et al., 2016) in Cell-DETR to generate accurate instance masks. Experiments on their in-house yeast cells dataset demonstrate the effectiveness of Cell-DETR relative to UNet based baselines (Ronneberger et al., 2015). Similarly, existing medical imaging segmentation approaches generally struggle for Corneal endothelial cells due to blurry edges caused by the subject's movement (Van den Bogerd et al., 2019). This demands preserving more local details and making full use of the global context. Considering these attributes, Zhang et al. (2021c) propose a Multi-Branch hybrid Transformer Network (MBT-Net) consisting of convolutional and transformer layers. Specifically, they propose a body-edge branch that provides precise edge location information and promotes local consistency. Extensive ablation studies on their self-collected TM-EM3000 and public Alisarine dataset (Ruggeri et al., 2010) of Corneal Endothelial Cells show the effectiveness of MBT-Net architectural components. On publicly available Alisarine dataset, although MBT-Net is able to outperform UNet, UNet++, and a baseline transformer-based approach in terms of dice score, its performance is significantly lower compared to advanced CNN-based approaches.

Histopathology. Histopathology refers to the diagnosis and study of the diseases of tissues under a microscope and is the gold standard for cancer recognition. Therefore accurate automatic segmentation of histopathology images can substantially alleviate the workload of pathologists. Recently, Nguyen et al. (2021a) systematically evaluate the performance of six latest ViTs, and CNNs-based approaches on whole slide images of the PAIP liver histopathological dataset (Kim et al., 2021). Their results (shown in Table 1) demonstrate that almost all Transformer-based models indeed exhibit superior performance as compared to CNN-based approaches due to their ability to encode the global context.

Retinal Vessel Segmentation. Accurate segmentation of the retina vessel is essential for the early diagnosis of eye-related diseases. Recently, Yu et al. (2022) proposes a new retina segmentation network called CAViT-DAGC, which incorporates a channel attention vision transformer (CAViT) and a deep adaptive gamma correction (DAGC) module. The CAViT block consists of an efficient channel attention (ECA) module and a ViT. The ECA module analyzes the interdependencies among feature channels, while the ViT extracts significant edge structures from the feature map weighted by the ECA module by focusing on global context. They demonstrate the effectiveness of their proposed modules on CHASE DB1 and DRIVE dataset. Similarly, Philippi et al. (2023) proposed ViT-based method that efficiently combines the long-range feature extraction and aggregation capabilities of transformers with the data-efficient training of CNNs. In another work, Wang et al. (2022c) proposed a dual branch transformer module that takes advantage of both the global context of the image and the local information of the patch level for effective segmentation of the retinal vessels. The decoder of our DA-Net uses an adaptive strip sampling block to capture context information that flexibly and effectively adjusted to the distribution of retinal vessels. Similarly, Huang et al. (2022) proposes a relation transformer block (RTB) employing attention mechanisms at two levels: a self-attention transformer to analyze global dependencies among lesion features, and a cross-attention transformer to integrate vascular information and alleviate ambiguity in lesion detection caused by complex fundus structures. In addition, a global transformer block has also been introduced to capture the small lesion pattern. In other work, Li et al. (2022) propose a global transformer network based on dual local network that not only captures the long-range dependency, but also mitigates discontinuity in blood vessel segmentation. In particular, their proposed architecture fuses features at different scales to mitigate the loss of information during feature fusion.

3.1.2. 3D medical segmentation

Here, we describe ViT-based segmentation approaches for volumetric medical data.

Brain Tumor Segmentation. An automatic and accurate brain tumor segmentation approach can lead to the timely diagnosis of neurological disorders. Recently, ViT-based models have been proposed to accurately segment brain tumors. Wang et al. (2021a) have made the first attempt to leverage Transformers for 3D multimodal brain tumor segmentation by effectively modeling local and global features in both spatial and depth dimensions. Specifically, their encoder–decoder architecture, TransBTS, employs a 3D CNN to extract local 3D volumetric spatial features and Transformers to encode global features. Progressive upsampling in the 3D CNN-based decoder has been used to predict the

Table 2
Segmentation results and parameters of various Transformer-based models on 3D Multimodal Brain Tumor BraTS 2021 dataset (Baid et al., 2021). As validation and test data for BraTS 2021 challenge are not available, so few works further split the train set (1251 MRI scans) of BraTS into validation (208 scans) and test set (209 scans). We report results on the these 209 scans in the top table. The dice score of the top-rank methods on the challenge validation dataset are provided in the below table. Note that the BraTS challenge uses aggregate ranking based on the mean dice and Hausdorff score. In the table, we have provided only the dice scores. The transformer-based approach Swin UNETR is ranked 7th on the validation set leader board.

Dice Score on 209 test set images, splitted from 1251 training set			
Method	#params	Flops	Dice score (avg.)
[a]3D U-Net (Çiçek et al., 2016)	11.9 M	557.9 G	86.42
[a]Residual U-Net (Kerfoot et al., 2018)	–	–	87.48
[a]nnU-Net (Isensee et al., 2018)	–	–	**89.68**
TransBTS (Wang et al., 2021a)	33 M	333 G	84.99
BiTr-UNet (Jia and Shu, 2021)	–	–	86.20
UNETR (Hatamizadeh et al., 2021)	102.5 M	193.5 G	88.03
nnFormer (Zhou et al., 2021b)	39.7 M	110.7 G	86.56
VT-UNET-T (Peiris et al., 2021)	**5.4 M**	**52 G**	86.82
VT-UNET-S (Peiris et al., 2021)	11.8 M	100.8 G	87.00
VT-UNET-B (Peiris et al., 2021)	20.8 M	165 G	88.07
Dice Score on the provided validation set of challenge			
Method	#params	Flops	Dice score (avg.)
[a]HNF-Netv2 (Jia et al., 2022)	–	–	88.43
[a]Ensemble 3D nnU-Net (Kotowski et al., 2022)	–	–	87.51
[a]Kaist MRI (Luu and Park, 2021)	–	–	88.36
[a]3D SegNet (Jabareen and Lukassen, 2022)	–	–	88.43
[a]NVAUTO (Siddiquee and Myronenko, 2021)	–	–	**89.11**
Swin UNETR (Hatamizadeh et al., 2022a)	61.98 M	394.84 G	88.97

[a]Indicates the approach is CNN-based.

final segmentation map. To further boost the performance, they make use of test-time augmentation. Extensive experimentation on BraTS 2019[1] and BraTS 2020[2] validation set show the effectiveness of the TransBTS with the baseline approaches. However, the performance in terms of dice score is inferior compared to the CNN-based top performing BraTS 2019 and 2020 leaderboard approaches. Unlike most of the ViT-based image segmentation approaches, TransBTS does not require pre-training on large datasets and has been trained from scratch. In another work, inspired from the architectural design of TransBTS (Wang et al., 2021a; Jia and Shu, 2021) propose Bi-Transformer UNet (BiTr-UNet) that performs relatively better on BraTS 2021 (Baid et al., 2021) segmentation challenge. Different from TransBTS (Wang et al., 2021a), BiTr-UNet consists of an attention module to refine encoder and decoder features and has two ViT layers (instead of one as in TransBTS). Furthermore, BiTr-UNet adopts a post-processing strategy to eliminate a volume of predicted segmentation if the volume is smaller than a threshold (Isensee et al., 2018) followed by model ensemble via majority voting (Lam and Suen, 1997). Similarly, Peiris et al. (2021) propose a light-weight UNet shaped volumetric transformer, VT-UNet, to segment 3D medical image modalities in a hierarchical manner. Specifically, two self-attention layers have been introduced in the encoder of VT-UNet to capture both global and local contexts. Furthermore, the introduction of window-based self-attention and cross-attention modules and Fourier positional encoding in the decoder significantly improve the accuracy and efficiency of VT-UNet. Experiments on BraTs 2021 (Baid et al., 2021) show that VT-UNet is robust to data artifacts and exhibits strong generalization ability. In another similar work, Hatamizadeh et al. (2022a) propose Swin UNet based architecture, Swin UNETR, that consists of Swin transformer as the encoder and a CNN-based decoder. Specifically, Swin UNETR computes self-attention in an efficient shifted window partitioning scheme and is a top-performing model on BraTs 2021 (Baid et al., 2021) validation set. In Table 2, we provide dice score and other parameters of various Transformer based models for the 3D multimodal BraTs 2021 dataset (Baid et al., 2021).

Histopathology. Yun et al. (2021) propose Spectral Transformer (SpecTr) for hyperspectral pathology image segmentation, which employs transformers to learn the contextual feature across the spectral dimension. To discard the irrelevant spectral bands, they introduce a sparsity-based scheme (Correia et al., 2019). Furthermore, they employ separate group normalization for each band to eliminate the interference caused by distribution mismatch among spectral images. Extensive experimentation on the hyperspectral pathology dataset, Cholangiocarcinoma (Zhang et al., 2019b), shows the effectiveness of SpecTr as also shown in Fig. 7.

Breast Tumor Segmentation. Detection of breast cancer in the early stages can reduce the fatality rate by more than 40% (Huang et al., 2017b). Therefore, automatic breast tumor detection is of immense importance to doctors. Recently, Zhu et al. (2021) propose a region aware transformer network (RAT-Net) to fuse the Breast tumor region information into multiple scales to obtain precise segmentation. Extensive experiments on a large ultrasound breast tumor segmentation dataset show that RAT-Net outperforms CNN and transformer-based baselines. Similarly, Liu et al. (2021c) also propose a hybrid architecture consisting of transformer layers in the decoder part of 3D UNet (Çiçek et al., 2016) to accurately segment tumors from volumetric breast data.

3.2. Multi-organ segmentation

Multi-organ segmentation aims to segment several organs simultaneously and is challenging due to inter-class imbalance and varying sizes, shapes, and contrast of different organs. ViT models are particularly suitable for the multi-organ segmentation due to their ability to model global relations and differentiate multiple organs. We have categorized multi-organ segmentation approaches based on the architectural design, as these approaches do not consider any organ-specific aspect and generally focus on boosting performance by designing effective and efficient architectural modules (Lei et al., 2020). We categorize multi-organ segmentation approaches into *Pure Transformer* (only ViT layers) and *Hybrid Architectures* (both CNNs and ViTs layers).

3.2.1. Pure Transformers

Pure Transformer based architectures consist of only ViT layers and have seen fewer applications in medical image segmentation compared to hybrid architectures as both global and local information is crucial for dense prediction tasks like segmentation (Chen et al., 2021g). Recently, Karimi et al. (2021) propose a pure Transformer-based model for 3D medical image segmentation by leveraging self-attention (Wang et al., 2018b) between neighboring linear embedding of 3D medical image patches. They also propose a method to pre-train their model when only a few labeled images are available. Extensive experiments show the effectiveness of their convolution-free network on three benchmark 3D medical imaging datasets related to brain cortical plate (Dou et al., 2020), pancreas, and hippocampus. One of the drawbacks of using Pure Transformer-based models in segmentation is the quadratic complexity of self-attention with respect to the input image dimensions. This can hinder the ViTs applicability in the segmentation of high-resolution medical images. To mitigate this issue, Cao et al. (2021) propose Swin-UNet that, like Swin Transformer (Liu et al., 2021a), computes self-attention within a local window and has linear computational complexity with respect to the input image. Swin-UNet also contains a patch

[1] https://www.med.upenn.edu/cbica/brats2019
[2] https://www.med.upenn.edu/cbica/brats2020

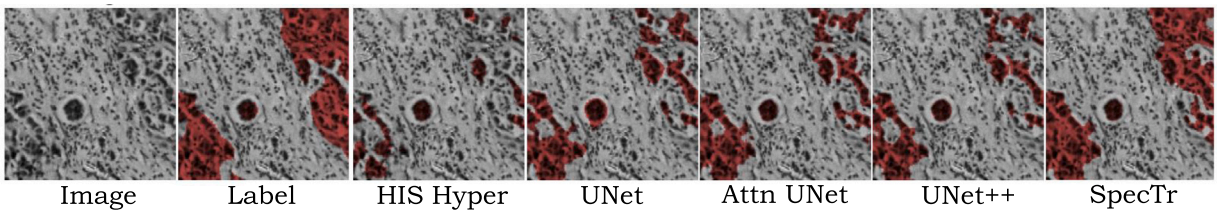

Fig. 7. Segmentation results for hyperspectral pathology dataset using Spectral Transformer (SpecTr). From left to right: Input image, Ground truth label, HIS Hyper (CNN-based) (Wang et al., 2020c), UNet (CNN-based) (Ronneberger et al., 2015), Attn UNet (CNN-based) (Oktay et al., 2018), UNet++ (CNN-based) (Zhou et al., 2019b), and SpecTr (ViT-based) (Yun et al., 2021).
Source: Image adapted from Yun et al. (2021).

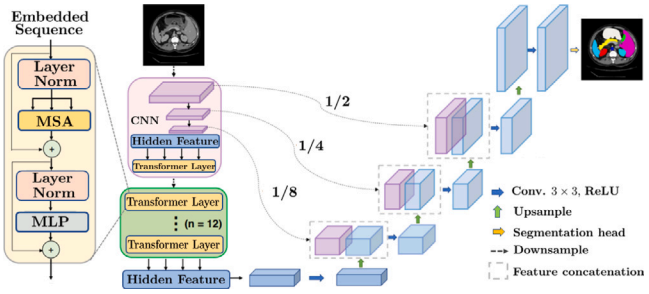

Fig. 8. Overview of TransUNet architecture (Chen et al., 2021g) proposed for multi-organ segmentation. It is one of the first transformer-based architectures proposed for medical image segmentation and merits both transformer and UNet. It employs a hybrid CNN-Transformer architecture for the encoder, followed by multiple upsampling layers in the decoder to output final segmentation mask.
Source: Image adapted from Chen et al. (2021g).

expanding layer for upsampling decoder's feature maps and shows superior performance in recovering fine details compared to bilinear upsampling. Experiments on Synapse multi-organ segmentation and ACDC (Bernard et al., 2018) dataset demonstrate the effectiveness of the Swin-UNet architectural design.

3.2.2. Hybrid architectures

Hybrid architecture-based approaches combine the complementary strengths of Transformers and CNNs to effectively model global context and capture local features for accurate segmentation. We have further categorized these hybrid models into single and multi-scale approaches.

Single-scale architectures. These methods process the input image information at one scale only and have seen widespread applications in medical image segmentation due to their low computational complexity compared to multi-scale architectures. We can sub-categorized single-scale architectures based on the position of the Transformer layers in the model. These sub-categories include *Transformer in Encoder*, *Transformer between Encoder and Decoder*, *Transformer in Encoder and Decoder*, and *Transformer in Decoder*.

Transformer in Encoder. Most initially developed Transformer-based medical image segmentation approaches have Transformer layers in the model's encoder. The first work in this category is TransUNet (Chen et al., 2021g) that consists of 12 Transformer layers in the encoder as shown in Fig. 8. These Transformer layers encode the tokenized image patches from the CNN layers. The resulting encoded features are upsampled via up-sampling layers in the decoder to output the final segmentation map. With skip-connection incorporated, TransUNet sets new records (at the time of publication) on and automated cardiac diagnosis challenge (ACDC) (Bernard et al., 2018). In other work, Zhang et al. (2021d) propose TransFuse to fuse features from the Transformer and CNN layers via BiFusion module. The BiFusion module leverages the self-attention and multi-modal fusion mechanism to selectively fuse the features. Extensive evaluation of TransFuse on multiple modalities (2D and 3D), including polyp segmentation, skin lesion segmentation, hip segmentation, and prostate segmentation, demonstrate its efficacy. Both TransUNet (Chen et al., 2021g) and TransFuse (Zhang et al., 2021d) require pre-training on ImageNet dataset (Deng et al., 2009) to learn the positional encoding of the images. To learn this positional bias without any pre-training, Valanarasu et al. (2021) propose a modified gated axial attention layer (Wang et al., 2020e) that works well on small medical image segmentation datasets. Furthermore, to boost segmentation performance, they propose a Local-Global training scheme to focus on the fine details of input images. Extensive experimentation on brain anatomy segmentation (Wang et al., 2018a), gland segmentation (Sirinukunwattana et al., 2017), and MoNuSeg (microscopy) (Kumar et al., 2019) demonstrate the effectiveness of their proposed gated axial attention module.

In another work, Tang et al. (2021) introduce Swin UNETR, a novel self-supervised learning framework with proxy tasks to pre-train Transformer encoder on 5,050 images of CT dataset. They validate the effectiveness of pre-training by fine-tuning the Transformer encoder with a CNN-based decoder on the downstream task of MSD and Synapse multi-organ segmentation datasets. Similarly, Sobirov et al. (2022) show that transformer-based models can achieve comparable results to state-of-the-art CNN-based approaches on the task of head and neck tumor segmentation. Few works have also investigated the effectiveness of Transformer layers by integrating them into the encoder of UNet-based architectures in a plug-and-play manner. For instance, Chang et al. (2021) propose TransClaw UNet by integrating Transformer layers in the encoding part of the Claw UNet (Yao et al., 2020) to exploit multi-scale information. TransClaw-UNet achieves gain of 0.6% in dice score compared to Claw-UNet on Synapse multi-organ segmentation dataset and shows excellent generalization. Similarly, inspired from the LeViT (Graham et al., 2021; Xu et al., 2021) propose LeViT-UNet which aims to optimize the trade-off between accuracy and efficiency. LeViT-UNet is a multi-stage architecture that demonstrates good performance and generalization ability on Synapse multi-organ segmentation and ACDC benchmarks.

Transformer between Encoder and Decoder. In this category, Transformer layers are between the encoder and decoder of a U-Shape architecture. These architectures are more suitable to avoid the loss of details during down-sampling in the encoder layers. The first work in this category is TransAttUNet (Chen et al., 2021e) that leverages guided attention and multi-scale skip connection to enhance the flexibility of traditional UNet. Specifically, a robust self-aware attention module has been embedded between the encoder and decoder of UNet to concurrently exploit the expressive abilities of global spatial attention and transformer self-attention. Extensive experiments on five benchmark medical imaging segmentation datasets demonstrate the effectiveness of TransAttUNet architecture. Similarly, Yan et al. (2021c) propose Axial Fusion Transformer UNet (AFTer-UNet) that contains a computationally efficient axial fusion layer between encoder and decoder to fuse inter and intra-slice information for 3D medical image segmentation. Experimentation on BCV (Simpson et al., 2019), Thorax-85 (Chen et al., 2021i), and SegTHOR (Lambert et al., 2020) datasets demonstrate the effectiveness of their proposed fusion layer.

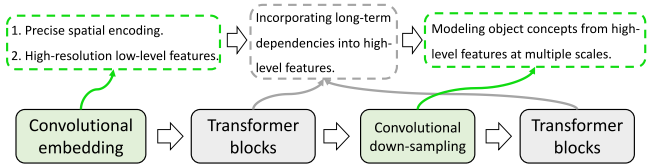

Fig. 9. Overview of the interleaved encoder not-another transFormer (nnFormer) (Zhou et al., 2021b) for volumetric medical image segmentation. Note that convolution and transformer layers are interleaved to give full play to their strengths.
Source: Image taken from Zhou et al. (2021b).

Transformer in Encoder and Decoder. Few works integrate Transformer layers in both encoder and decoder of a U-shape architecture to better exploit the global context for medical image segmentation. The first work in this category is UTNet (Gao et al., 2021c) that efficiently reduces the complexity of the self-attention mechanism from quadratic to linear (Wang et al., 2020a). Furthermore, to accurately model the image content, UTNet exploits the two-dimensional relative position encoding (Bello et al., 2019). Experiments show strong generalization ability of UTNet on multi-label and multi-vendor cardiac MRI challenge dataset cohort (Campello et al., 2021). Similarly, to optimally combine convolution and transformer layers for medical image segmentation, Zhou et al. (2021b) propose nnFormer, an interleave encoder–decoder based architecture, where convolution layer encodes precise spatial information and Transformer layer encodes global context as shown in Fig. 9. Like Swin Transformers (Liu et al., 2021a), the self-attention in nnFormer has been computed within a local window to reduce the computational complexity. Moreover, deep supervision in the decoder layers has been employed to enhance performance. Experiments on ACDC and Synapse multi-organ segmentation datasets show that nnFormer surpass Swin-UNet (Cao et al., 2021) (transformer-based medical segmentation approach) by over 7% (dice score) on Synapse multi-organ segmentation dataset. In other work, Lin et al. (2021a) propose Dual Swin Transformer UNet (DS-TransUNet) to incorporate the advantages of Swin Transformer in U-shaped architecture for medical image segmentation. They split the input image into non-overlapping patches at two scales and feed them into the two Swin Transformer-based branches of the encoder. A novel Transformer Interactive Fusion module has been proposed to build long-range dependencies between different scale features in encoder. DS-TransUNet outperforms CNN-based methods on four standard datasets related to Polyp segmentation, ISIC 2018, GLAS, and Datascience bowl 2018.

Transformer in Decoder. Li et al. (2021a) investigate the use of Transformer as an upsampling block in the decoder of the UNet for medical image segmentation. Specifically, they adopt a window-based self-attention mechanism to better complement the upsampled feature maps while maintaining efficiency. Experiments on MSD Brain and Synapse multi-organ segmentation datasets demonstrate the superiority of their architecture compared to bilinear upsampling. In another work, Li et al. (2021d) propose SegTran, a Squeeze-and-Expansion Transformer for 2D and 3D medical image segmentation. Specifically, the squeeze block regularizes the attention matrix, and the expansion block learns diversified representations. Furthermore, a learnable sinusoidal positional encoding has been proposed that helps the model to encode spatial relationships. Extensive experiments on Polyp, BraTS19, and REFUGE20 (fundus images) segmentation challenges demonstrate the strong generalization ability of SegTran (see Table 3).

Multi-scale architectures. These architectures process input at multiple scales to segment organs having irregular shapes and different sizes. Here, we highlight various attempts to integrate the multi-scale architectures for medical image segmentation. We further group these approaches into 2D and 3D segmentation categories based on the input image type.

2D Segmentation. Most ViT-based multi-organ segmentation approaches struggle to capture information at multiple scales as they partition the input image into fixed-size patches, thereby losing useful information. To address this issue, Zhang et al. (2021g) propose a pyramid medical transformer, PMTrans, that leverage multi-resolution attention to capture correlation at different image scales using a pyramidal architecture (Ghiasi and Fowlkes, 2016). PMTrans works on multi-resolution images via an adaptive partitioning scheme of patches to access different receptive fields without changing the overall complexity of self-attention computation. Extensive experiments on three medical imaging datasets of GLAS (Sirinukunwattana et al., 2017), MoNuSeg (Kumar et al., 2017), and HECKTOR (Andrearczyk et al., 2020) show the effectiveness of exploiting multi-scale information. In other work, Ji et al. (2021) propose a Multi-Compound transformer (MCTrans) that learns not only feature consistency of the same semantic categories but also capture correlation among different semantic categories for accurate segmentation (Yu et al., 2020). Specifically, MCTrans captures cross-scale contextual dependencies via the Transformer self-attention module and learned semantic correspondence among different categories via Transformer Cross-Attention module. An auxiliary loss has also been introduced to improve feature correlation of the same semantic category. Extensive experiments have been performed on six benchmark segmentation datasets. In particular, experiments on Pannuke dataset (Gamper et al., 2020) shows that MCTrans achieves average dice score of 68.90, surpassing the performance of AttentionUNet (Oktay et al., 2018) (64.97) and CENet (Gu et al., 2019) (66.50).

3D Segmentation. The majority of multi-scale architectures have been proposed for 2D medical image segmentation. To directly handle volumetric data, Hatamizadeh et al. (2021) propose a ViT-based architecture (UNETR) for 3D medical image segmentation. UNETR consists of a pure transformer as the encoder to learn sequence representations of the input volume. The encoder is connected to a CNN-based decoder via skip connections to compute the final segmentation output. UNETR achieves impressive performance on Synapse multi-organ segmentation dataset (Landman et al., 2015) and MSD (Simpson et al., 2019) segmentation datasets as shown in Fig. 10. One of the drawbacks of UNETR is its large computational complexity in processing large 3D input volumes. To mitigate this issue, Xie et al. (2021a) propose a computationally efficient deformable self-attention module (Dai et al., 2017) that casts attention only to a small set using multi-scale features, as shown in Fig. 11, to reduce the computational and spatial complexities. Experiments on Synapse multi-organ segmentation dataset demonstrate that their approach is able to beat the TransUNet (Chen et al., 2021g) approach in terms of average dice score.

3.3. Discussion

From the extensive literature reviewed in this section, we note that the medical image segmentation area is heavily impacted by transformer-based models, with more than 50 publications within one year since the inception of the first ViT model (Dosovitskiy et al., 2020). We believe such interest is due to the availability of large medical segmentation datasets and challenge competitions associated with them in top conferences compared to other medical imaging applications. As shown in Fig. 12, a recent transformer-based hybrid architecture is able to achieve 13% performance gain in terms of dice score compared to a simple baseline transformer model, indicating rapid progression of the field. In short, ViT-based architectures have achieved good results over benchmark medical datasets, as also shown in Table 4.

As mentioned before, the high computational cost associated with extracting features at multiple levels hinders the applicability of multi-scale architectures in medical segmentation tasks. These multi-scale architectures exploit processing input image information at multiple levels and achieve superior performance than single-scale architectures. Therefore, designing

Table 3
An overview of ViT-based approaches for medical image segmentation.

Method	Organ	Modality	Type	Datasets	Metrics	Arch.	P.T.	Highlights
TransUNet (Chen et al., 2021g)	Multi-organ	CT, MRI	2D	Synapse (Synapse, 2015) , ACDC (Bernard et al., 2018)	Dice, Hausdorff distance	Hybrid	Yes	Encodes strong global context by treating the image features as sequences but also well utilizes the low-level CNN features via a u-shaped hybrid architectural design.
TransFuse (Zhang et al., 2021d)	Multi-organ	Colonoscopy	2D, 3D	KVASIR (Jha et al., 2020), Clinic DB (Bernal et al., 2015), Colon DB (Tajbakhsh et al., 2015), EndoScene (Vázquez et al., 2017), ETIS (Silva et al., 2014), ISIC 2017 (Berseth, 2017), MSD (Simpson et al., 2019)	Dice, Jaccard index	Hybrid	Yes	Leverages the inductive bias of CNNs on modeling spatial correlation and the powerful capability of Transformers on modeling global relationship. Novel Bi-Fusion module to fuse CNN and Transformers features for segmentation.
MedT (Zhang et al., 2021g)	Multi-organ	Ultrasound, Microscopy	2D	Brain US (Private), GLAS (Sirinukunwattana et al., 2017), MoNuSeg (Kumar et al., 2017)	F1	Hybrid	No	Gated axial attention layer for positional encoding. Local global training for training on both full resolution images as well in patches.
Conv. Free (Karimi et al., 2021)	Multi-organ	MRI, CT	3D	Brain crotial, (private)	Dice, Hausdorff Distance, Average Symmetric Surface Distance	Pure	Yes	Convolutional free medical segmentation model. Based on self-attention between neighboring 3D patches.
CoTr (Xie et al., 2021a)	Multi-organ	CT	3D	BCV (Simpson et al., 2019)	Dice	Hybrid	Yes	Deformable self-attention mechanism to reduce the computational and spatial complexities of modeling the long range dependency.
SpecTr (Yun et al., 2021)	Bile-duct	Hyperspectral	3D	Choledock (Zhang et al., 2019b)	Dice, IoU, Hausdorff distance	Hybrid	No	First application to hyperspectral and learn contextual feature across spectral dimension.
TransBTS (Wang et al., 2021a)	Brain	MRI	3D	BraTS 19 (BraTS, 2019a), BraTS 20 (bra, 2022b)	Dice, Hausdorff distance	Hybrid	Yes	3D CNN for capturing local volumetric features and transformers for encoding global features.
U-Transformer (Petit et al., 2021)	Multi-organ	CT	2D	TCIA (Clark et al., 2013), Private multi-organ	Dice	Hybrid	No	Propose self and cross-attention modules to model long-range interactions and spatial dependencies.
UNETR (Hatamizadeh et al., 2021)	Brain, Spleen	MRI, CT	3D	Synapse, MSD (Simpson et al., 2019)	Dice, Hausdorff distance	Hybrid	No	Transformer as the encoder to learn sequence representations of the input volume and effectively capture the global multi-scale information
PMTrans (Zhang et al., 2021g)	Multi-organ	Microscopy, CT, PET	2D	GLAS (Sirinukunwattana et al., 2017), MoNuSeg (Kumar et al., 2017), HECKTOR (Andrearczyk et al., 2020)	Dice	Hybrid	No	Integrate multi-scale attention and CNN feature extraction using a pyramidal network architecture. An adaptive partitioning scheme was implemented to retain informative relations and to access different receptive fields efficiently.
Swin-UNet (Cao et al., 2021)	Multi-organ	CT	2D	Synapse (Synapse, 2015), ACDC (Bernard et al., 2018)	Dice	Pure	Yes	Swin transformer based pure transformer architecture to segmentation with patch expanding layer design in decoder.
SegTran (Li et al., 2021d)	Multi-organ	Fundus, Colonoscopy, MRI	2D, 3D	REFUGE 20 (Orlando et al., 2020), BraTS 19 (BraTS, 2019a), CVC (Fan et al., 2020), KVASIR (Jha et al., 2020)	Dice	Hybrid	Yes	Squeeze-and-Expansion transformer where squeeze block regularize the self-attention module and expansion block learns diversified represetations.
MBT-Net (Zhang et al., 2021c)	Eye	Pathology	2D	TM-EM3000 (private), Alisarine (Ruggeri et al., 2010)	Dice, F1, Sensitivity, Specificity	Hybrid	No	Body edge branch for precise edge location information for corneal endothelial cells segmentation.
DS-TransUNet (Lin et al., 2021a)	Multi-organ	Colonoscopy, Histology	2D	Kvasir (Jha et al., 2020), Colon DB (Tajbakhsh et al., 2015) and Clinic DB (Bernal et al., 2015), EndoScene (Vázquez et al., 2017), ETIS (Silva et al., 2014), ISIC 18 (Codella et al., 2019), GLAS (Sirinukunwattana et al., 2017), Data Science Bowl 18 (Caicedo et al., 2019)	Mean Dice, Mean IoU, Precision, Recall	Hybrid	Yes	Dual-branch Swin Transformer in encoder and decoder to extract multiscale representation. Transformer Interactive Fusion module to build long-range dependencies between features of different scales
MCTrans (Ji et al., 2021)	Multi-organ	Colonoscopy, Pathology	2D	Pannuke dataset, Colon DB (Tajbakhsh et al., 2015), Clinic DB (Bernal et al., 2015), ETIS (Silva et al., 2014), KVASIR (Jha et al., 2020), ISIC 2018 (Codella et al., 2019)	Dice	Hybrid	No	Transformer self-attention for cross-scale contextual dependencies Transformer cross attention layer for semantic correspondence.

(*continued on next page*)

Table 3 (*continued*).

Li et al. (2021a)	Multi-organ	MRI, CT	2D	Synapse (Synapse, 2015), MSD Brain (Simpson et al., 2019)	Dice, Hausdorff distance	Hybrid	No	Investigate the use of transformer decoder for medical image segmentation and its usage in upsampling.
UTNet (Gao et al., 2021c)	Heart	MRI	2D	MRI Challenge Cohort (Campello et al., 2021)	Dice, Hausdorff distance	Hybrid	No	Self-attention modules in encoder and decoder. Design relative position encoding to reduce the complexity of self-attention from quadratic to linear.
TransClaw UNet (Chang et al., 2021)	Multi-organ	CT	2D	Synapse (Synapse, 2015)	Dice, Hausdorff distance	Hybrid	No	Integrated transformer layer in the encoder path of Claw-UNet to extract shallow spatial features.
TransAttUNet (Chen et al., 2021e)	Multi-organ	Xray, CT	2D	ISIC 2018 (Codella et al., 2019), JSRT (Shiraishi et al., 2000), Montgomery (Jaeger et al., 2014), NIH (Tang et al., 2019), Clean-CC-CCII (He et al., 2020), Data Science Bowl 18 (Caicedo et al., 2019), GLAS (Sirinukunwattana et al., 2017)	Dice, F1	Hybrid	No	Multi-level guided attention and multi-scale skip connections to mitigate information recession problem.
LeViT-UNet (Xu et al., 2021)	Multi-organ	CT, MRI	2D	Synapse (Synapse, 2015), ACDC (Bernard et al., 2018)	Dice, Hausdorff distance	Hybrid	Yes	Integrate multiscale LeViT architecture as the encoder in UNet.
Polyp-PVT (Dong et al., 2021)	Polyp on organs	Colonoscopy	2D	KVASIR (Jha et al., 2020), Clinic DB (Bernal et al., 2015), Colon DB (Tajbakhsh et al., 2015), Endoscene (Vázquez et al., 2017), ETIS (Silva et al., 2014)	Dice, IoU, MAE, Weighted F-measure, S-measure, E-measure	Hybrid	No	Pyramid vision transformer backbone as encoder to extract robust features. Proposed architectural components to handle noise, occlusions, and capturing global semantic cues.
COTRNet (Shen et al., 2021b)	Kidney	CT	2D	KITS21 Challenge (KiTS, 2021)	Dice, Surface Dice	Hybrid	Yes	CNN and transformer based interleaved encoder–decoder. Supervision of decoder's hidden layers.
nnFormer (Zhou et al., 2021b)	Multi-organ	CT, MRI	3D	Synapse (Synapse, 2015), ACDC (Bernard et al., 2018)	Dice	Hybrid	Yes	Interleaved convolution and self-attention based encoder–decoder architecture.
MISSFormer (Huang et al., 2021b)	Multi-organ	CT, MRI	2D	Synapse (Synapse, 2015), ACDC (Bernard et al., 2018)	Dice, Hausdorff distance	Hybrid	No	Hierarchical encoder–decoder network with enhanced transformer block to mitigate the problem of feature inconsistency
TransBridge (Deng et al., 2021)	Heart	Echocardiography	2D	EchoNet-Dynamic (Ouyang et al., 2020)	Dice, Hausdorff distance	Hybrid	No	Shuffling layer and group convolution for patch embedding to significantly reduce the number of parameters.
BiTr-UNet (Jia and Shu, 2021)	Brain	MRI	3D	BraTS 21 (Baid et al., 2021)	Dice, Hausdorff distance	Hybrid	No	Refined version of TransBTS with two sets of ViT layers instead of one.
GT UNet (Li et al., 2021e)	Tooth	X-ray	2D	Tooth root dataset (private)	Dice, Accuracy, Sensitivity, Specificity, Jaccard similarity	Hybrid	No	Group transformer layers to reduce computational cost. Fourier descriptor based loss function to integrate shape prior.
BAT (Wang et al., 2021e)	–	Dermoscopy	2D	ISIC 2016+PH2 (Gutman et al., 2016), ISIC 2018 (Codella et al., 2019)	Dice, IoU	Hybrid	Yes	Boundary-wise attention gate is added at the end of each transformer encoder layer to tackle challenging cases with ambiguous boundaries.
AFTer-UNet (Yan et al., 2021c)	Multi-organ	CT	3D	BCV (Simpson et al., 2019), Thorax-85 (Chen et al., 2021i), SegTHOR (Lambert et al., 2020)	Dice	Hybrid	No	Axial fusion mechanism to fuse intra-slice and inter-slice contextual information to guide segmentation.
VT-UNet (Peiris et al., 2021)	Multi-organ	CT, MRI	3D	BraTS 21 (Baid et al., 2021), MSD (Simpson et al., 2019)	Dice, Hausdorff distance	Hybrid	Yes	U-shaped encoder–decoder design. Encoder has two consecutive self-attention layers to encode local and global cues, and our decoder has novel parallel shifted window based self and cross attention blocks to capture fine details.
Swin UNETR (Hatamizadeh et al., 2022a)	Brain	MRI	3D	BraTS 21 (Baid et al., 2021)	Dice, Hausdorff distance	Hybrid	Yes	Swin UNet based architecture that consists of Swin transformer as the encoder and a CNN-based decoder. Computes self-attention in an efficient shifted window partitioning scheme.

P.T: pretraining.

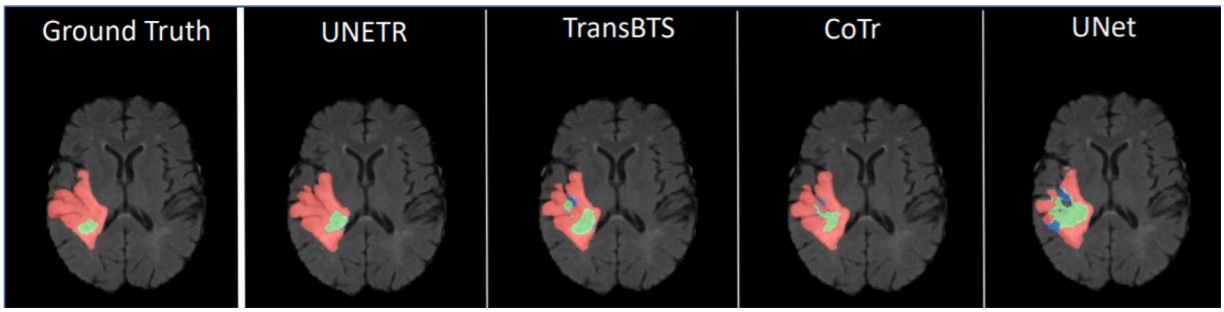

Fig. 10. Qualitative results of brain tumor segmentation task using transformer. From left to right: Ground truth image, UNETR (Hatamizadeh et al., 2021) (ViT-based), TransBTS (Wang et al., 2021a) (ViT-based), CoTr (Shen et al., 2021a) (ViT-based), and UNet (Ronneberger et al., 2015) (CNN based). Note that transformer-based approaches demonstrate better performance in capturing the fine-grained details of brain tumors as compared to CNN-based method.
Source: Image courtesy Hatamizadeh et al. (2021).

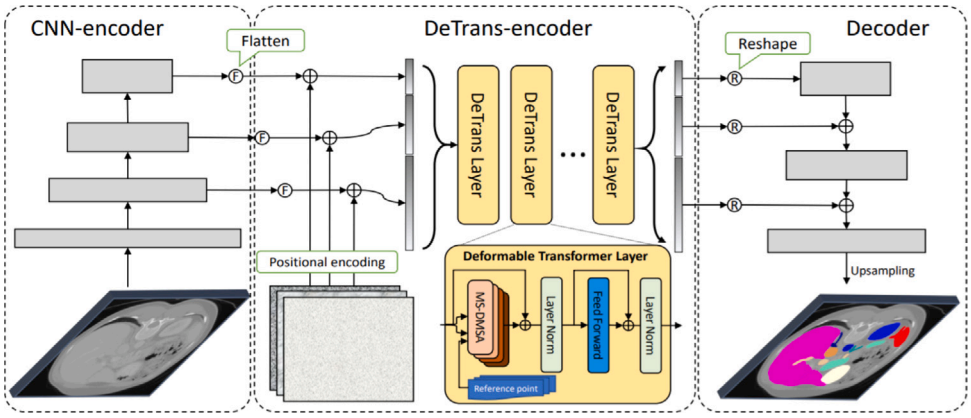

Fig. 11. Overview of CNN and a Transformer (CoTr) architecture (Xie et al., 2021a) proposed for 3D medical image segmentation. It consists of CNN encoder (left) to extract multi-scale features from the input, followed by DeTrans-encoder (yellow blocks) to process the flattened multi-scale feature maps. Output features from encoder are fed to the CNN decoder (right) for segmentation mask prediction.
Source: Image courtesy Xie et al. (2021a).

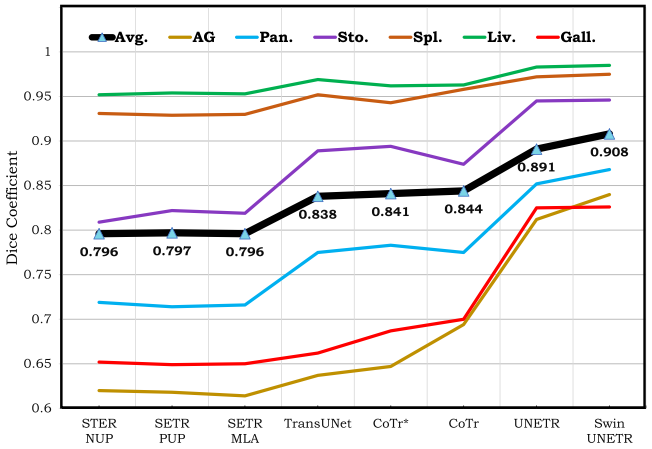

Fig. 12. Dice results of Synapse multi-organ segmentation challenge for various transformer based methods. It can be seen that Swin-UNETR is able to achieve on average 13% improvement in dice coefficient score compare to SETR method, indicating rapid pace of research in the field. Transformer based approaches used for the comparison include SETR NUP (Zheng et al., 2021b), SETR PUP (Zheng et al., 2021b), SETR MLA (Zheng et al., 2021b), TransUNet (Chen et al., 2021g), CoTr* (Xie et al., 2021a) (small CNN encoder compared to CoTr), CoTr (Xie et al., 2021a), UNETR (Hatamizadeh et al., 2021), and Swin UNETR (Tang et al., 2021). Note: Avg: Average results (over 12 organs), AG: left and right adrenal glands, Pan: pancreas, Sto: stomach, Spl: spleen, Liv: liver, Gall: gallbladder.

Table 4
Top-performing methods for Medical Image Segmentation on Synapse multi-organ CT.

Method	Dice score	Extra training data
Swin UNETR (Tang et al., 2022)	90.80	✓
UNETR (Hatamizadeh et al., 2022b)	89.10	✓
nnUNet (Isensee et al., 2018)	88.80	×
PaNN (Zhou et al., 2019a)	85.40	×
MISSFormer (Huang et al., 2021b)	81.96	×
SETR (Zheng et al., 2021b)	79.60	×
SwinUNet (Cao et al., 2021)	79.13	×
UCTransNet (Wang et al., 2022a)	78.99	×
TransUNet (Chen et al., 2021g)	77.48	×

efficient transformer architectures for multi-scale processing requires more attention.

Most of the proposed ViT-based models are pre-trained on the ImageNet dataset for the downstream task of medical image segmentation. This approach is sub-optimal due to the large domain gap between natural and medical image modalities. Recently, few attempts have been made to investigate the impact of self-supervised pre-training on medical imaging datasets on the ViTs segmentation performance. However, these works have shown that ViT pre-trained on one modality (CT) gives unsatisfactory performance when applied directly to other medical imaging modalities (MRI) due to the large domain gap making it an exciting avenue to explore. We defer detailed

discussion related to pre-training ViTs for downstream medical imaging tasks to Section 11.1.

Moreover, recent ViT-based approaches mainly focus on 2D medical image segmentation. Designing customized architectural components by incorporating temporal information for efficient high-resolution and high-dimensional segmentation of volumetric images has not been extensively explored. Recently, few efforts have been made, e.g., UNETR (Hatamizadeh et al., 2021) uses Swin Transformer (Liu et al., 2021a) based architectures to avoid quadratic computing complexity. Extending these efforts to design lightweight transformer-based segmentation approaches to mitigate the issue of high computation cost for temporal sequence processing requires further work.

In addition to focusing on the scale of datasets, with the advent of ViTs, we note there is a need to collect more diverse and challenging medical imaging datasets. Although diverse and challenging datasets are also crucial to gauge the performance of ViTs in other medical imaging applications, they are particularly relevant for medical image segmentation due to a major influx of ViT-based models in this area. We believe these datasets will play a decisive role in exploring the limits of ViTs for medical image segmentation.

4. Medical image classification

Accurate classification of medical images plays an essential role in aiding clinical care and treatment. In this section, we comprehensively cover applications of ViTs in medical image classification. We have broadly categorized these approaches into COVID-19, tumor, and retinal disease classification based methods due to a different set of challenges associated with these categories as shown in Fig. 13.

4.1. COVID-19 diagnosis

Studies suggest that COVID-19 can potentially be better diagnosed with radiological imaging as compared to tedious real-time polymerase chain reaction (RT-PCR) test (Ai et al., 2020; Fang et al., 2020; Chen et al., 2021c). Recently, ViTs have been successfully employed for diagnosis and severity prediction of COVID-19, showing good performance. In this section, we briefly describe the impact of ViTs in advancing recent efforts on automated image analysis for the COVID-19 diagnosis process. Most of these works use three modalities, including Computerized tomography (CT), Ultrasound scans (US), and X-ray. We have further categorized ViT-based COVID-19 classification approaches into *2D* and *3D* classification categories, depending on the input image type. Below, we briefly describe these approaches:

2D COVID-19 Classification: The High computational cost of ViTs hinders their deployment on portable devices, thereby limiting their applicability in real-time COVID-19 diagnosis. Perera et al. (2021) propose a lightweight **Point-of-Care Transformer (POCFormer)** to diagnose COVID-19 from lungs images captured via portable devices. Specifically, POCFormer leverages Linformer (Wang et al., 2020a) to reduce the space and time complexity of self-attention from quadratic to linear. POCFormer has two million parameters that are about half of MobileNetv2 (Sandler et al., 2018), and an average accuracy of 91% at 70 frames per second. Experiments on COVID-19 lungs POCUS dataset (Born et al., 2020; Cohen et al., 2020) demonstrate the effectiveness of their proposed architecture with above 90% classification accuracy. In other work, Liu and Yin (2021) proposed ViT-based model for COVID-19 diagnosis by exploiting a new attention mechanism named Vision Outlooker (VOLO) (Yuan et al., 2021b). VOLO is effective for **encoding fine-level features** into ViT token representation, thereby improving classification performance. Further, they leverage the transfer learning approach to handle the issue of insufficient and generally unbalanced COVID-19 datasets. Experiments on two publicly available COVID-19 CXR datasets (Chowdhury et al., 2020; Cohen et al., 2020) demonstrate the effectiveness of their architecture. Similarly, Jiang and Lin (2021) leverage **Swin Transformer** (Liu et al., 2021a) and **Transformer-in-Transformer** (Han et al., 2021) to classify

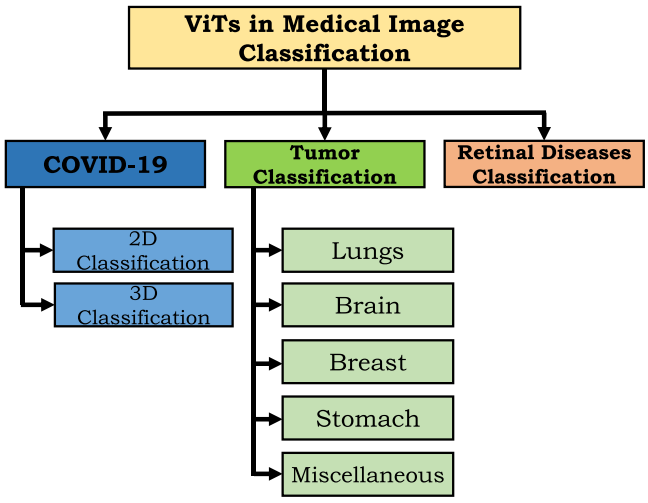

Fig. 13. Taxonomy of ViT-based medical image classification approaches. The influx of ViT-based COVID-19 classification approaches makes it a dominating category in the taxonomy.

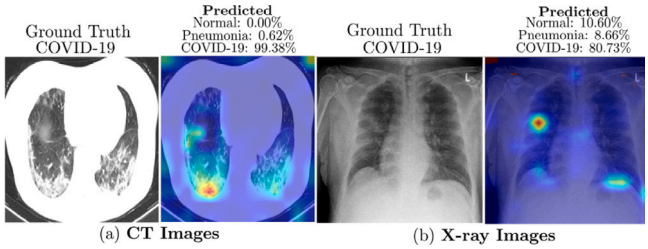

Fig. 14. CT scans (a) and X-ray (b) images along with their ground truth labels (left) and saliency maps (right). For figure (a), xViTCOS-CT localized suspicious lesion regions exhibiting ground glass opacities, consolidation, reticulations in bilateral postero basal lung. xViTCOS-CT (Mondal et al., 2021) is able predict these regions correctly. For figure (b), radiologist's interpretation is: *thick walled cavity in right middle zone with surrounding consolidation*. As shown in last column, xViTCOS-CXR (Mondal et al., 2021) is able predict it correctly.
Source: Figure courtesy Mondal et al. (2021).

COVID-19 images from pneumonia and normal images. To further boost the accuracy, they employ model ensembling using a weighted average. Research progress in ViT-based COVID-19 diagnosis approaches is heavily impeded due to the requirement of a large amount of labeled COVID-19 data, thereby demanding collaborations among hospitals. This collaboration is difficult due to limited consent by patients, privacy concerns, and ethical data usage (Dou et al., 2021). To mitigate this issue, Park et al. (2021a) proposed a **Federated Split Task-Agnostic (FESTA) framework that leveraged the merits of Federated and Split Learning** (Yang et al., 2019; Vepakomma et al., 2018) in utilizing ViT to simultaneously process multiple chest X-ray tasks, including the diagnosis in COVID-19 Chest X-ray images on a massive decentralized dataset. Specifically, they split ViT into the shared transformer body and task-specific heads. The transformer body is shared across multiple tasks by leveraging multitask-learning (MTL) strategy (Caruana, 1997), as shown in Fig. 16. They affirm the suitability of ViTs for collaborative learning in medical imaging applications via extensive experiments on the CXR dataset.

Few authors also show the features that influence the decision of ViT-based COVID-19 classification model, generally via visualization techniques like saliency-based methods (Cong et al., 2018), Grad-CAM (Selvaraju et al., 2017), etc. **Saliency Based Visualization:** Park et al. (2021b), propose a ViT-based method for COVID-19 diagnosis by exploiting the low-level CXR features extracted from the pre-trained

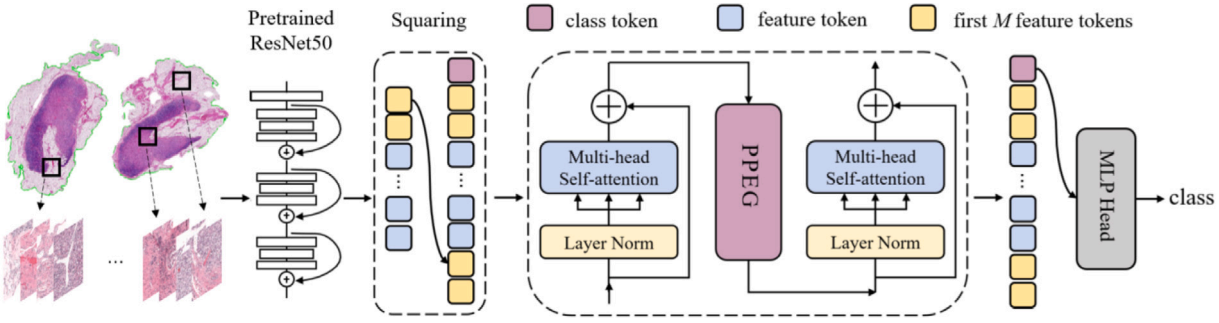

Fig. 15. Overview of Transformer based Multiple Instance Learning (TransMIL) architecture (Shao et al., 2021a) for whole slide brain tumor classification. Patches of WSI are embedded in the feature space of ResNet-50. The sequence of embedded features are then processed by their proposed pipeline that include: squaring of sequence, Correlation modeling of the sequence, conditional position encoding (via Pyramid Position Encoding Generator (PPEG) module) and local information fusion, feature aggregation, and mapping from transformer space to label space.
Source: Image taken from Shao et al. (2021a).

backbone network. The backbone network has been trained in a self-supervised manner (using contrastive-learning based SimCLR (Chen et al., 2020a) method) to extract abnormal CXR features embeddings from large and well-curated CXR dataset of CheXpert (Irvin et al., 2019). These feature embeddings have been leveraged by ViT model for high-level diagnosis of COVID-19 images. Extensive experiments on three CXR test datasets acquired from different hospitals demonstrate the superiority of their approach compared to CNN-based models. They also validated the generalization ability of their proposed approach and adopted saliency map visualizations (Chefer et al., 2021) to provide interpretable results. In another work, Mondal et al. (2021) introduce xViTCOS for COVID-19 screening from lungs CT and X-ray images. Specifically, they pre-train xViTCOS on ImageNet to learn generic image representations and fine-tune the pre-trained model on a large chest radiographic dataset. Further, xViTCOS leverage the explainability-driven saliency-based approach (Chefer et al., 2021) with clinically interpretable visualizations to highlight the role of critical factors in the resulting predictions, as shown in Fig. 14. Experiments on COVID CT-2 A (Gunraj et al., 2021) and their privately collected Chest X-ray dataset demonstrate the effectiveness of xViTCOS. **Grad-CAM Based Visualization:** Shome et al. (2021) propose a ViT-based model to diagnose COVID-19 infection at scale. They combine several open-source COVID-19 CXR datasets to form a large-scale multi-class and binary classification dataset. For better visual representation and model interpretability, they further create Grad-CAM based visualization (Selvaraju et al., 2017).

3D COVID-19 Classification: Most of the ViT-based approaches for COVID-19 classification operate on 2D information only. However, as suggested by Kwee and Kwee (2020), the symptoms of COVID-19 might be present at different depths (slices) for different patients. To exploit both 2D and 3D information, Hsu et al. (2021) propose a hybrid network consisting of transformers and CNNs. Specifically, they determine the importance of slices based on significant symptoms in the CT scan via Wilcoxon signed-rank test (Woolson, 2007) with Swin Transformer (Liu et al., 2021a) as backbone network. To further exploit the intrinsic features in the spatial and temporal dimensions, they propose a Convolutional CT Scan Aware Transformer module to fully capture the context of the 3D scans. Extensive experiments on the COVID-19-CT dataset show the effectiveness of their proposed architectural components. Similarly, Zhang and Wen (2021b,a) also proposed Swin Transformer based two-stage framework for the diagnosis of COVID-19 in the 3D CT scan dataset (Kollias et al., 2021). Specifically, their framework consists of UNet based lung segmentation model followed by the image classification with Swin Transformer (Liu et al., 2021a) backbone. Similarly, Gao et al. (2021b) propose COVID-ViT to classify COVID from non-COVID images as part of the MIA-COVID19 challenge (Kollias et al., 2021). Their experiments on 3D CT lungs images demonstrated the superiority of ViT-based approach over DenseNet (Huang et al., 2017a) baseline in terms of F1 score. However, they are not able to beat the CNN-based approaches on the challenge leader board (BraTS, 2019b), where there rank was 12th out of 12 teams that outperform the baseline.

4.2. Tumor classification

A tumor is an abnormal growth of body tissues and can be cancerous (malignant) or noncancerous (benign). Early-stage malignant tumor diagnosis is crucial for subsequent treatment planning and can greatly improve the patient's survival rate. In this section, we review ViT-based models for tumor classification. We categorize these models based on the underlying organs.

Lungs. Similarly, other works employ hybrid Transformer-CNN architectures to solve medical classification problem for different organs. For instance, Khan and Lee (2021) propose Gene-Transformer to predict the **lung** cancer subtypes. Experiments on TCGA-NSCLC (Napel and Plevritis, 2014) dataset demonstrates the superiority of Gene Transformer over CNN baselines. To diagnose **lung** tumors, Zheng et al. (2021a) propose graph transformer network (GTN) to leverage the graph-based representation of WSI. GTN consists of a graph convolutional layer (Kipf and Welling, 2016), a transformer layer, and a pooling layer. GTN further employs GraphCAM (Chefer et al., 2021) to identify regions that are highly associated with the class label. Extensive evaluations on TCGA dataset (Napel and Plevritis, 2014) show the effectiveness of GTN.

Brain: Later, Lu et al. (2021) propose a two-stage framework that first performs contrastive pre-training on glioma sub-type classification in the **brain** followed by the feature aggregation via proposed transformer-based sparse attention module. Ablation studies on TCGA-NSCLC (Napel and Plevritis, 2014) dataset show the effectiveness of their two-stage framework.

Breast: For the task of **breast** cancer classification, Gheflati and Rivaz (2021) systematically evaluate the performance of pure and hybrid pre-trained ViT models. Experiments on two breast ultrasound datasets provided by Al-Dhabyani et al. (2020), Yap et al. (2017) shows that ViT-based models provide better results than those of the CNNs for classifying images into benign, malignant, and normal categories.

Stomach: Chen et al. (2021d) present a multi-scale GasHis-Transformer to diagnose gastric cancer in the **stomach**. GasHis-Transformer combines the advantages of the CNNs and ViTs to extract local and global information respectively. GasHis-Transformer shows improve robustness against adversarial noise, and demonstrate good generalizability ability.

Miscellaneous. Since the annotation procedure is expensive and laborious, one label is assigned to a set of instances (bag) in whole

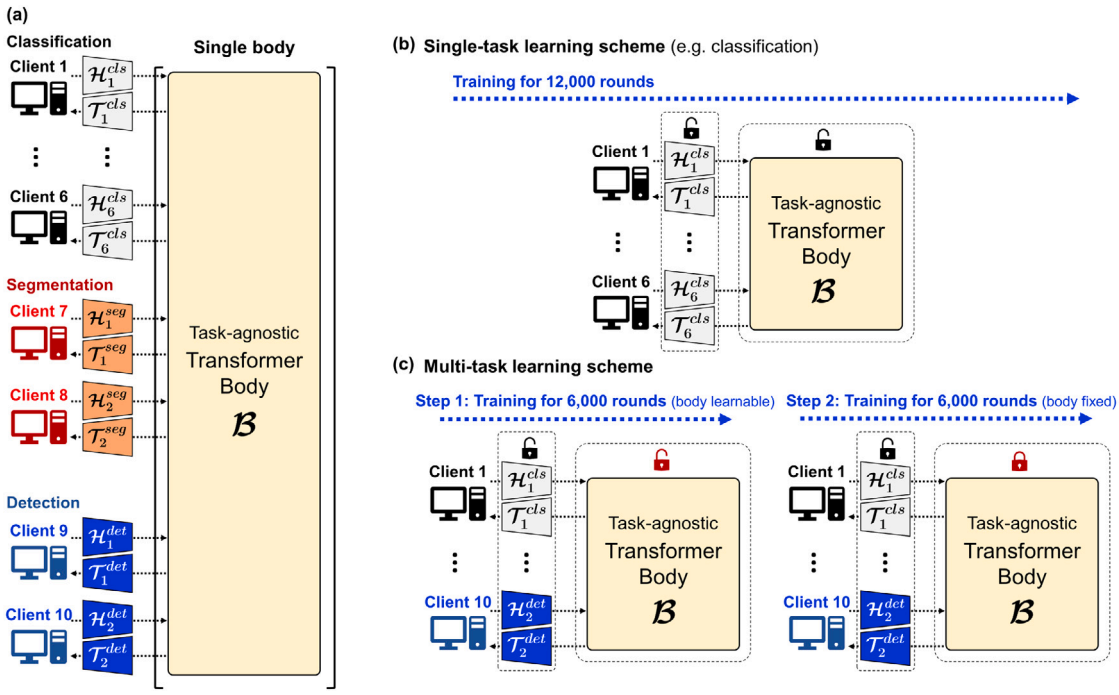

Fig. 16. Implementation details of Federated Split Task-Agnostic (FESTA) framework (Park et al., 2021a) equipped with a Transformer to simultaneously process multiple chest X-ray tasks including diagnosis of COVID-19. (a) Experimental setting for multi-task learning of classification, segmentation, and detection of chest X-ray images. The clients only train the head ($\mathcal{H}$) and tail ($\mathcal{T}$) parts of the network, whereas the transformer body ($\mathcal{B}$) is shared across multiple clients. In the second step, the embedded features in the head are utilized by transformers for processing of individual tasks. (b) shows training scheme for single task. (c) shows training scheme for multi-task learning.
Source: Image courtesy Park et al. (2021a).

slide imaging (WSI) based pathology diagnosis. This type of weakly supervised learning is known as Multiple Instance Learning (Fung et al., 2007), where a bag is labeled positive if at least one instance is positive or labeled negative when all instances in a bag are negative. Most of the current MIL methods assume that the instances in each bag are independent and identically distributed, thereby neglecting the correlation among different instances. Shao et al. (2021a) present TransMIL to explore both morphological and spatial information in weakly supervised WSI classification. Specifically, TransMIL aggregates morphological information with two transformer-based modules and a position encoding layer as shown in Fig. 15. To encode spatial information, a pyramid position encoding generator is proposed. Further, the attention scores from the TransMIL have been visualized to demonstrate interpretability, as shown in Fig. 17. TransMIL performs well on three different computational pathology datasets CAMELYON16 **(breast)** (Bejnordi et al., 2017), TCGA-NSCLC **(lung)** (Napel and Plevritis, 2014), and TCGA-R **(kidney)** (TCGA, 2013). Similarly, Li et al. (2021f) present a novel embedded-space MIL model based on deformable transformer architecture and convolutional layers for histopathological image analysis. Experiments on histopathological image analysis tasks demonstrate that DT-MIL performs well compared with other transformer based MIL architectures. In another work, TransMed (Dai et al., 2021) leverages ViTs for medical image classification. TransMed is a hybrid CNN and transformer-based architecture that is capable of classifying **parotid** tumors in the multi-modal MRI medical images. TransMed also employs a novel image fusion strategy to capture mutual information from images of different modalities, thereby achieving competitive results on their privately collected parotid tumor classification dataset. Similarly, Jiang et al. (2021) propose a hybrid model consisting of convolutional and transformer layers to diagnose acute lymphocytic leukemia by using symmetric cross-entropy loss function. In another work, Xia et al. (2021) explore detecting pancreatic cancer from non-contrast CT scans, as a relatively cheap and safe imaging modality. Specifically,

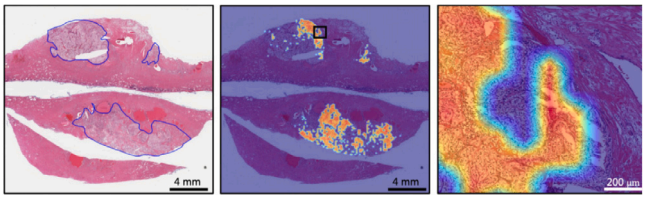

Fig. 17. Left: The area within the blue region is the cancer region. Middle: Attention scores from TransMIL are visualized as a heatmap (red for tumor and blue for normal) to interpret the important morphology used for diagnosis. Right: Zoomed-in view of the black square in the middle figure.
Source: Figure courtesy Shao et al. (2021a).

they propose hybrid transformer model that is able to achieve high specificity and sensitivity as compared to mean radiologists.

4.3. Retinal disease classification

Yu et al. (2021a) propose MIL-ViT model which is first pre-trained on a large fundus image dataset and later fine-tuned on the downstream task of the retinal disease classification. MIL-ViT architecture uses MIL-based head that can be used with ViT in a plug-and-play manner. Evaluation performed on APTOS2019 (APTOS, 2019) and RFMiD2020 (Quellec et al., 2020) datasets shows that MIL-ViT is achieving more favorable performance than CNN-based baselines. Most data-driven approaches treat diabetic retinopathy (DR) grading and lesion discovery as two separate tasks, which may be sub-optimal as the error may propagate from one stage to the other. To jointly handle both these tasks, Sun et al. (2021) propose lesion aware transformer (LAT) that consists of a pixel relation based encoder and a lesion-aware transformer decoder. In particular, they leverage transformer decoder to formulate lesion discovery as a weakly supervised lesion localization problem. LAT model sets state-of-the-art on Messidor-1 (Decencière et al.,

2014), Messidor-2 (Decencière et al., 2014), and EyePACS (Cuadros and Bresnick, 2009) datasets. Yang et al. (2021a) propose a hybrid architecture consisting of convolutional and Transformer layers for fundus disease classification on OIA dataset (OIA, 2019). Similarly, Wu et al. (2021b), AlDahoul et al. (2021) also verify that ViT models are more accurate in DR grading than their CNNs counterparts.

4.4. Discussion

In this section, we provide a comprehensive overview of about 25 papers related to applications of ViTs in medical image classification. In particular, we see a surge of Transformer-based architectures for diagnosing COVID-19, compelling us to develop taxonomy accordingly.

The lack of large COVID-19 datasets hindered the applicability of ViT models to diagnose COVID-19. A recent work by Shome et al. (2021) attempts to mitigate this issue by combining three open-source COVID-19 datasets to create a large dataset comprising 30,000 images. Still, creating diverse and large COVID-19 datasets is challenging and requires significant effort from the medical community.

*More attention must be given to design **interpretabile** (to gain end-users trust) and **efficient** (for point-of-care testing) ViT models for COVID-19 diagnosis to make them a viable alternative of RT-PCR testing in the future.*

We notice that most works have used the original ViT model (Dosovitskiy et al., 2020) as a plug-and-play manner to boost the medical image classification performance. In this regard, we believe that integrating domain-specific context and accordingly designing architectural components and loss functions can enhance performance and provide more insights in designing effective ViT-based classification models in the future.

*Finally, let us highlight the exciting work of Matsoukas et al. (2021) that, for the first time, demonstrates that ViTs pre-trained on ImageNet perform comparably to CNNs for the medical image classification task as shown in Table 6. This also raises an interesting question "**Can ViT models pre-trained on medical imaging datasets perform better than ViT models pre-trained on ImageNet for medical image classification?**". A recent work by Xie et al. (2021b) attempts to answer this by pre-training the ViT on large-scale 2D and 3D medical images. On the medical image classification problem, their model obtains substantial performance gain over the ViT model pre-trained on ImageNet, indicating that this area is worth exploring further. A brief overview of ViT-based medical image classification approaches has been provided in Table 5.*

5. Medical object detection

In medical image analysis, object detection refers to localization and identification of a region of interest (ROIs) such as lung nodules from X-ray images and is typically an essential aspect of diagnosis. However, it is one of the most time-consuming tasks for clinicians, thereby demanding the accurate computer-aided diagnosis (CAD) system to act as a second observer that may accelerate the process. Following the success of CNNs in medical image detection (Liao et al., 2019; Ganatra, 2021), recently few attempts have been made to improve performance further using Transformer models. These approaches are mainly based on the detection transformer (DETR) framework (Zhu et al., 2020).

Shen et al. (2021a) propose the first hybrid framework COTR, consisting of convolutional and transformer layers for end-to-end polyp detection. Specifically, the encoder of COTR contains six hybrid convolution-in-transformer layers to encode features. Whereas, the decoder consists of six transformer layers for object querying followed by a feed-forward network for object detection. COTR performs better than DETR on two different datasets ETIS-LARIB and CVC-ColonDB. The DETR model (Zhu et al., 2020) is also adapted in other works (Liu et al., 2021e; Mathai et al., 2021) for the end-to-end polyp detection (Liu et al., 2021e), and detecting lymph nodes in T2 MRI scans for the assessment of lymphoproliferative diseases (Mathai et al., 2021). In other work, Tao and Zheng (2021) propose a transformers-based 3D object detector called Spine-Transformers and applied to the task of automatic detection and localization of vertebrae in arbitrary field-of-view spine CT. Experiments on one in-house and two public datasets demonstrate good performance.

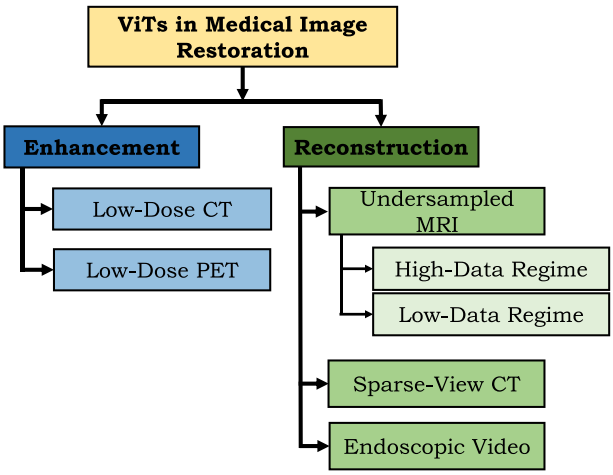

Fig. 18. Taxonomy of ViT-based medical image restoration approaches.

5.1. Discussion

Overall, the frequency of new Transformer-based approaches for the problem of medical image detection is lesser than those for the segmentation and classification. This is in contrast to the early years of CNN-based designs that were rapidly developed for the medical image detection, as indicated in Fig. 32. A recent work by Maaz et al. (2021) shows that generic class-agnostic detection mechanism of multi-modal ViTs (like MDETR (Kamath et al., 2021)) pre-trained on natural images-text pairs performs poorly on medical datasets. Therefore, investigating the performance of multi-modal ViTs by pre-training them on modality-specific medical imaging datasets is a promising future direction to explore. Furthermore, since the recent ViT-based methods yield competitive results on medical image detection problems, we expect to see more contributions in the near future.

6. Medical image restoration

The goal of medical image restoration is to obtain a clean image from a degraded input. For example, recovering a high-resolution MRI image from its under-sampled version. It is a challenging task due to its ill-posed nature. Moreover, exact analytic inverse transforms in many practical medical imaging scenarios are unknown. Recently, ViTs have been shown to address these challenges effectively. We categorize the relevant works into *medical image enhancement* and *medical image construction* areas, as depicted in Fig. 18.

6.1. Medical image enhancement

ViTs have achieved impressive success in the enhancement of medical images, mostly in the application of Low-Dose Computed Tomography (LDCT) (Gopal et al., 2010; Sadate et al., 2020). In LDCT, the X-ray dose is reduced to prevent patients from being exposed to high radiation. However, this reduction comes at the expense of CT image quality degradation and requires effective enhancement algorithms to improve the image quality and, subsequently, diagnostic accuracy.

6.1.1. LDCT enhancement

Zhang et al. (2021h) propose an hybrid architecture TransCT that leverages the internal similarity of the LDCT images to enhance them. TransCT first decomposes the LDCT image into high-frequency (HF) (containing noise) and low-frequency (LF) parts. Next it removes the noise from the HF part with the assistance of latent textures. To reconstruct the final high-quality LDCT images, TransCT further integrates features from the LF part to the output of the transformer decoder. Experiments on Mayo LDCT dataset (McCollough et al., 2017)

Table 5
An overview of ViT-based approaches for medical image classification.

Method	Organ	Modality	Type	Datasets	Metrics	Arch.	Highlights
TransMed (Dai et al., 2021)	Ear	MRI (T1,T2)	3D	MRI (private)	Accuracy Precision	Pure	First ViT-based multi-modal medical image classification approach with novel multi-modal fusion strategy
TransMIL (Shao et al., 2021a)	Multi-organ	Pathology	2D	Camelyon16 (Bejnordi et al., 2017) TCGA-NSCLC (Bakr et al., 2018) TCGA-RCC (TCGA, 2013)	Accuracy AuC	Hybrid	Transformer based architecture to explore morphological and spatial information for Whole Slide Image classification.
Matsoukas et al. (2021)	Multi-organ	Mammograms Dermoscopy	2D	APTOS-2019 (APTOS, 2019) ISIC-2019 (ISIC, 2019) CBIS-DDSM (Lee et al., 2017)	Recall AuC	Pure	Systematic study of whether one should replace CNNs with ViTs for medical image classification.
Gheflati and Rivaz (2021)	Breast	Ultrasound	2D	BUSY (Al-Dhabyani et al., 2020) Yap et al. (Yap et al., 2017)	Accuracy AuC	Pure	First application of ViTs to ultrasound images classification.
GTN (Zheng et al., 2021a)	Lung	Microscopy	2D	TCGA dataset (Napel and Plevritis, 2014)	Accuracy Precision Sensitivity Specificity Recall	Hybrid	Consists of a graph convolutional layer, a transformer module, and a pooling layer for accurate classification of WSI images.
MIL-ViT (Yu et al., 2021a)	Eye	Fundus	2D	APTOS-2019 (APTOS, 2019) RFMiD2020 (Quellec et al., 2020)	Accuracy AuC, F1 Precision Recall	Pure	First pretrained on a large fundus image dataset and later fine-tuned on the downstream task of the retinal disease classification.
LAT (Sun et al., 2021)	Eye	Fundus	2D	Messidor-1 (Decencière et al., 2014) Messidor-2 (Decencière et al., 2014) EyePACS (Cuadros and Bresnick, 2009)	AuC Kappa	Hybrid	Formulate lesion discovery as a weakly supervised lesion localization problem via a transformer decoder. Jointly solve diabetic retinopathy grading and lesion discovery.
				COVID-19			
Park et al. (Park et al., 2021b)	Chest	X-ray	2D	CheXpert (Irvin et al., 2019)	AuC	Hybrid	Leveraging a backbone network trained to find low-level abnormal CXR findings in pre-built large-scale dataset to embed feature corpus suitable for high-level disease classification.
POCFormer (Perera et al., 2021)	Chest	Ultrasound	2D	POCUS (Born et al., 2020)	Recall, F1 Specificity Sensitivity Accuracy	Pure	Proposed light weight transformer architecture that demonstrates the efficiency and performance improvements.
FESTA (Park et al., 2021a)	Chest	X-ray	2D	CheXpert (Irvin et al., 2019) SIIM-ACR (sii, 2022c) RSNA (RSNA, 2018)	Recall, F1 Specificity Sensitivity AuC	Pure	Utilize ViT to simultaneously process multiple chest X-ray tasks, including the diagnosis in COVID-19 Chest X-ray images on a massive decentralized dataset.
Liu and Yin (2021)	Chest	X-ray	2D	COVID-19-1 (Chowdhury et al., 2020) COVID-19-2 (Cohen et al., 2020)	Accuracy	Pure	Explore VOLO tailored with transfer learning technique to effectively encodes fine-level features into the token representations.
COVID-VIT (Gao et al., 2021b)	Chest	CT	3D	MIA-COV19 (Kollias et al., 2021)	Accuracy, F1	Pure	Propose ViT based architecture to classify COVID-19 CT images in MIA-COV19 competition.
xViTCOS (Mondal et al., 2021)	Chest	X-ray, CT	2D	xViTCOS-CT (Gunraj et al., 2021) xViTCOS-CXR (Wang et al., 2020b)	Precision Recall, F1 Specificity, NPV	Pure	Propose ViT based multi-stage transfer learning technique to address the issue of COVID-19 data scarcity. The approach is clinically interpretable.
Hsu et al. (2021)	Chest	CT	3D	COV19 CT DB (Kollias et al., 2021)	Accuracy Recall,F1 Precision	Hybrid	Importance of slices are determined in CT scan via Wilcoxon signed-rank test. Then, spatial and temporal features are exploited via proposed convolutional CT scan Aware Transformer module.
Zhang and Wen (2021b)	Chest	CT	3D	MIA-COV19 (Kollias et al., 2021)	F1 score	Hybrid	Swin Transformer based two stage framework for diagnosis of COVID-19 in 3D CT scans.

(*continued on next page*)

demonstrate the effectiveness of TransCT over CNN-based approaches. To perform LDCT image enhancement, Wang et al. (2021f) propose a convolution-free ViT-based encoder–decoder architecture TED-Net. It employs Token-to-token block (Yuan et al., 2021a) to enrich the image tokenization via a cascaded process. To refine contextual information, TED-Net introduces dilation and cyclic-shift blocks (Cao et al., 2021)

Table 5 (*continued*).

Method	Organ	Modality	Type	Datasets	Metrics	Arch.	Highlights
COViT-GAN (Ambita et al., 2021)	Chest	CT	2D	COVID-CT (Kollias et al., 2021) Sars-CoV-2 (Angelov and Almeida Soares, 2020)	Accuracy Sensitivity Precision, F1	Hybrid	Generate synthetic images using a self-attention generative adversarial network and use it as a data augmentation method to alleviate the problem of limited data and improve performance.
COVID-Trans. (Shome et al., 2021)	Chest	X-ray	2D	El-Shafai et al. (El-Shafai and Abd El-Samie, 2020) Sait et al. (Sait et al., 2020) Qi et al. (Qi et al., 2021)	Accuracy Precision AuC, Recall, F1	Pure	Propose a ViT model to diagnose COVID-19 at scale. Combines several open source COVID-19 chest X-ray datasets to form large dataset for binary and multi-class classification.

Table 6
Comparison of vanilla CNNs vs. ViTs with different initialization strategies on *medical imaging* classification tasks. For APTOS 2019 (APTOS, 2019) and ISIC 2019 (Tschandl et al., 2018) datasets quadratic Cohen Kappa and recall score have been reported.
Source: Table taken from Matsoukas et al. (2021).

Initialization	Model	APTOS2019, $\kappa \uparrow$	ISIC2019, Recall↑
Random	ResNet50	0.849 ± 0.022	0.662 ± 0.018
	DeiT-S	0.687 ± 0.017	0.579 ± 0.028
ImageNet (supervised)	ResNet50	0.893 ± 0.004	0.810 ± 0.008
	DeiT-S	0.896 ± 0.005	0.844 ± 0.021
ImageNet (supervised) + Self-supervised with DINO (Caron et al., 2021)	ResNet50	0.894 ± 0.008	0.833 ± 0.007
	DeiT-S	0.896 ± 0.010	0.853 ± 0.009

First row: For randomly initialized networks, CNNs outperform ViTs. *Second row*: ViTs appear to benefit significantly from pre-training on ImageNet dataset. *Third row*: Both ViTs and CNNs perform better with self-supervised pretraining.

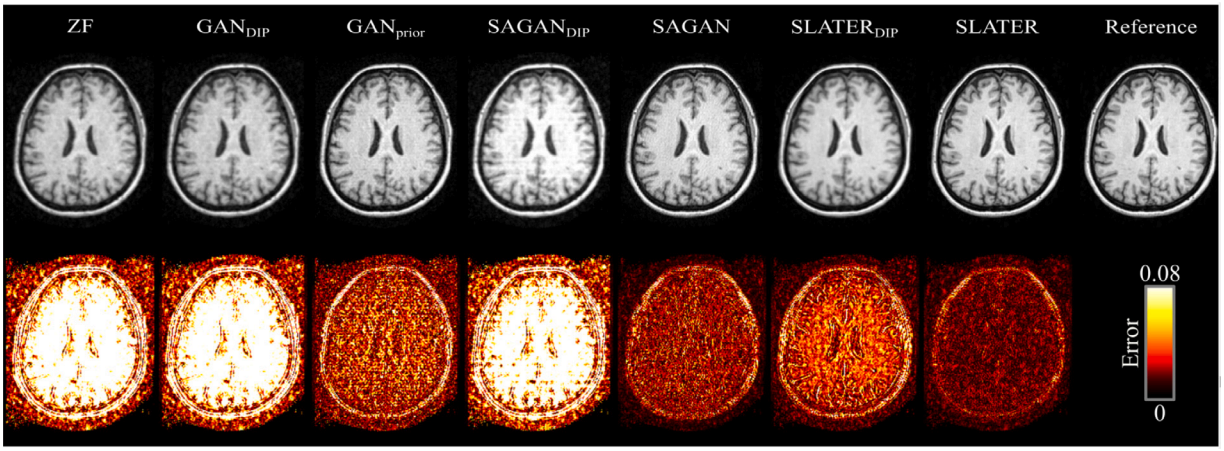

Fig. 19. Unsupervised under-sampled MRI reconstruction results of CNN based and transformer based approaches. From left to right (top row): Fourier method (ZF) (Fessler, 2010), GAN$_{DIP}$ (unsupervised CNN), GAN$_{Prior}$ (Narnhofer et al., 2019) (unsupervised CNN with pre-training), Self-Attention GAN$_{DIP}$ (Ulyanov et al., 2018) (unsupervised CNN), Self-Attention GAN (Narnhofer et al., 2019) (unsupervised CNN with pre-training), SLATER$_{DIP}$ (Korkmaz et al., 2021a) (unsupervised transformer), SLATER (Korkmaz et al., 2021a) (unsupervised transformer with pre-training), and reference image. Bottom row shows corresponding error maps. It can be seen that SLATTER outperforms all other approaches in term of quality of reconstruction. Results taken from Korkmaz et al. (2021a).

in tokenization. TED-Net shows favorable performance on the Mayo Clinic LDCT dataset (McCollough et al., 2017). In another work, Luthra et al. (2021) propose Eformer which is Transformer-based residual learning architecture for LDCT images denoising. To focus on edges, Eformer uses the power of Sobel–Feldman operator (Liang et al., 2020; Irwin et al., 1968) in the proposed edge enhancement block to boost denoising performance. Moreover, to handle the over-smoothness issue, the multi-scale perceptual loss (Liang et al., 2020) is used. Eformer achieves impressive image quality gains in terms PSNR, SSIM, and RMSE on the AAPM-Mayo Clinic dataset (McCollough et al., 2017).

6.1.2. LDPET enhancement

Like LDCT, Low-dose positron emission tomography (LDPET) images reduce the harmful radiation exposure of standard-dose PET (SD-PET) at the expense of sacrificing diagnosis accuracy. To address this challenge, Luo et al. (2021) propose an end-to-end generative adversarial network (GAN) based method integrated with Transformers, namely Transformer-GAN, to effectively reconstruct SDPET images from the corresponding LDPET images. Specifically, the generator of Transformer-GAN consists of a CNN-based encoder to learn compact feature representation, a transformer network to encode global context, and a CNN-based decoder to restore feature representation. They also

introduce adversarial loss with aims to obtain reliable images. Extensive experiments on their in-house collected clinical human brain PET dataset show the effectiveness of Transformer-GAN quantitatively and qualitatively.

6.2. Medical image reconstruction

Medical image reconstruction entails transforming signals collected by acquisition hardware (like MRI scanners) into interpretable images that can be used for diagnosis and treatment planning. Recently, ViT-based models have been proposed for multiple medical image restoration tasks, including undersampled MRI restoration, Sparse-View CT image reconstruction, and endoscopic video reconstruction. These models have pushed the boundaries of existing learning-based systems in terms of reconstruction accuracy. Next, we briefly highlight these approaches.

6.2.1. Undersampled MRI reconstruction

Reducing the number of MRI measurements can result in faster scan times and a reduction in artifacts due to patients movement at the expense of aliasing artifacts in the image (Hyun et al., 2018).

High-Data Regime Approaches. Approaches in this category assume the availability of large MRI training datasets to train the ViT model. Feng et al. (2021a) propose Transformer-based architecture, MTrans, for accelerated multi-modal MR imaging. The main component of MTrans is the cross attention module that extracts and fuses complementary features from the auxiliary modality to the target modality. Experiments on fastMRI and uiMRI datasets for reconstruction and super-resolution tasks show that MTrans achieve good performance gains over previous methods. However, MTrans requires separate training for MR reconstruction and super-resolution tasks. To jointly reconstruct and super-resolve MRI images, Feng et al. (2021b) propose Task-Transformer that leverages the power of multi-task learning to fuse complementary information between the reconstruction branch and the super-resolution branch. Experiments are performed on the public IXI and private MRI brain datasets. Similarly, Mahapatra and Ge (2021) propose a hybrid architecture to super-resolve MRI images by exploiting the complementary advantages of both CNNs and ViTs. They also propose novel loss functions (Park et al., 2020) to preserve semantic and structural information in the super-resolved images.

Low-Data Regime Approaches. Most of the aforementioned approaches required massive paired dataset of undersampled and corresponding fully sampled MRI acquisitions to train ViT models. To alleviate the data requirement issue, Korkmaz et al. (2021a,b) propose a zero-shot framework, SLATER, that leverages prior induced by randomly initialized neural networks (Ulyanov et al., 2018; Qayyum et al., 2021) for unsupervised MR image reconstruction. Specifically, during inference, SLATER inverts its transformer-based generative model via iterative optimization over-network weights to minimize the error between the network output and the under-sampled multi-coil MRI acquisitions while satisfying the MRI forward model constraints. SLATER yields quality improvements on single and multi-coil MRI brain datasets over other unsupervised learning-based approaches as shown in Fig. 19. Similarly, Lin and Heckel (2021) show that a ViT model pre-trained on ImageNet, when fine-tuned on only 100 fastMRI images, not only yields sharp reconstructions but is also more robust towards anatomy shifts compared to CNNs as shown in Fig. 20. Furthermore, their experiments indicate that ViT benefits from higher throughput and less memory consumption than the UNet baseline.

6.2.2. Sparse-view CT reconstruction

Sparse-view CT (Han and Ye, 2018) can reduce the effective radiation dose by acquiring fewer projections. However, a decrease in the number of projections demands sophisticated image processing algorithms to achieve high-quality image reconstruction (Kudo et al., 2013). Wang et al. (2021c) present a hybrid CNN-Transformer, named Dual-Domain Transformer (DuDoTrans), by considering the global nature of sinogram's sampling process to better restore high-quality images. In the first step, DuDoTrans reconstructs low-quality reconstructions of sinogram via filtered back projection step and learnable DuDo consistency layer. In the second step, a residual image reconstruction module performs enhancement to yield high-quality images. Experiments are performed on the NIH-AAPM dataset (McCollough et al., 2017) to show generalizability, and robustness (against noise and artifacts) of DuDoTrans. Experimental results on the NIH-AAPM dataset and COVID-19 dataset show that DuDoTrans achieves favorable reconstruction results compared to FBPConvNet (Jin et al., 2017) and DudoNet (Lin et al., 2019), however has higher inference time.

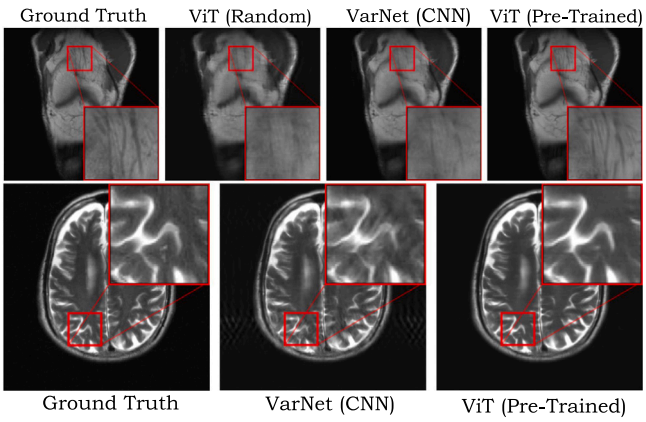

Fig. 20. Top row: Example reconstructions of models trained on 100 images (low-data regime) of fastMRI dataset. It can be seen that ViT model pre-trained on ImageNet produce sharp results as compared to recent CNN-based model (Sriram et al., 2020) and randomly initialized ViT model. **Bottom row**: Example reconstructions of Brain images by models pre-trained on ImageNet and fine-tuned on Knee MRI dataset. Results show that pre-trained ViT models are more robust to anatomical shifts.
Source: Figures adapted from Lin and Heckel (2021).

6.2.3. Endoscopic video reconstruction

Reconstructing surgical scenes from a stereoscopic video is challenging due to surgical tool occlusion and camera viewpoint changes. Long et al. (2021) propose E-DSSR to reconstruct surgical scenes from stereo endoscopic videos. Specifically, E-DSSR contains a lightweight stereo Transformer module to estimate depth images with high confidence (Poggi et al., 2021) and a segmentor network to accurately predict the surgical tool's mask. Extensive experiments on Hamlyn Centre Endoscopic Video Dataset (Ye et al., 2017) and privately collected DaVinci robotic surgery dataset demonstrate the robustness of E-DSSR against abrupt camera movements and tissue deformations in real-time.

6.3. Discussion

In this section, we have reviewed about a dozen papers related to the applications of ViT models in medical image restoration, as shown in Table 7.

*Recently, an interesting work (Lin and Heckel, 2021) has investigated the impact of pre-training ViT on the task of MRI image reconstruction. Their results indicate that pre-trained ViT yields sharp reconstructions and is robust towards anatomy shifts (see Fig. 20). The robustness of ViTs can be of particular relevance to the pathology image reconstruction as the range of pathology can vary significantly in the anatomy being imaged. Further, it raises an interesting question "**Are ViTs pre-trained on medical image dataset able to provide any advantages in terms of reconstruction performance and robustness against anatomy shifts compared to their counterparts pre-trained on ImageNet**"? Extensive and systematic experiments are required to answer this question. Another promising future direction is to investigate the impact on the performance of ViT pre-trained*

Table 7
An overview of ViT-based approaches for medical image reconstruction.

Method	Highlights	Modality	Input type	Datasets	Metric
TransCT (Zhang et al., 2021h)	Transformer for LDCT enhancement with high and low frequency decomposition.	CT	2D	NIH-AAPM (McCollough et al., 2017)	RMSE, SSIM, VIF (Sheikh and Bovik, 2006)
SLATER (Korkmaz et al., 2021a)	Transformer based approach for zero shot MRI image reconstruction.	MRI	3D	IXI (ixi, 2022a) fastMRI (Zbontar et al., 2018)	PSNR, SSIM
TED-Net (Wang et al., 2021f)	Pure transformer based encoder decoder dilation architecture for LDCT denoising.	CT	2D	NIH-AAPM (McCollough et al., 2017)	RMSE, SSIM
Eformer (Luthra et al., 2021)	Transformers based residual image denoising. Incorporate learnable Sobel filters for edge enhancement.	CT	2D	NIH-AAPM (McCollough et al., 2017)	PSNR, SSIM, RMSE
Transformer-GAN (Luo et al., 2021)	End-to-end GAN-based method integrated with Transformers to enhance LDPET images.	PET	3D	Private	PSNR, SSIM, MSE
MTrans (Feng et al., 2021a)	Leverage cross-attention module to fuse complementary features from the auxiliary modality to the target modality for fast multi-modal MRI image reconstruction.	MRI	2D	fastMRI (Zbontar et al., 2018) uiMRI (private)	PSNR, SSIM, NMSE
Task-Transformer (Feng et al., 2021b)	Simultaneously, reconstruct and super-resolve MRI images via multi-task learning.	MRI	2D	IXI (ixi, 2022a)	PSNR, SSIM, NMSE
Mahapatra and Ge (2021)	Hybrid architecture to super-resolve MRI images by exploiting the complementary advantages of CNNs and ViTs.	MRI	2D	fastMRI (Zbontar et al., 2018) IXI (ixi, 2022a)	PSNR, SSIM, NMSE
Lin and Heckel (2021)	ViT pretrained on ImageNet, when fine-tuned on only 100 fastMRI images, yields sharp reconstructions and is robust towards anatomy shifts.	MRI	2D	fastMRI (Zbontar et al., 2018)	SSIM
DuDoTrans (Wang et al., 2021c)	A hybrid CNN-Transformer architecture that consider the global nature of sinogram's sampling process to restore high-quality CT images from sparse views.	CT	2D	NIH-AAPM (McCollough et al., 2017)	PSNR, SSIM
MIST-Net (Pan et al., 2021)	Swin-transformer based projection and image domain two-stage framework to reconstruct high-quality CT images from sparse views.	CT	2D	NIH-AAPM (McCollough et al., 2017)	PSNR, SSIM, RMSE
E-DSSR (Long et al., 2021)	Leverage lightweight stereo Transformer module to estimate depth images with high confidence and a segmentor network to accurately predict the surgical tool's mask.	Endo.	2D	Hamlyn (Ye et al., 2017) DaVinci (private)	PSNR, SSIM
TranSMS (Güngör et al., 2021)	ViT-based data consistency module to super-resolve magnetic particle imaging (MPI) system matrices for accelerated calibration.	MPI	2D, 3D	Open MPI (Knopp et al., 2020) Private datasets	RMSE

on one image modality (like CT) and fine-tuned on another modality (like MRI) for image reconstruction tasks.

We notice that most of the Transformer-based approach focus on MRI and CT image reconstruction tasks, and their applicability to other modalities are yet to be explored. In addition, proposed architectures are mostly generic and have not fully exploited the application-specific aspects. We believe that designing architectural components and formulating loss functions according to the task at hand can significantly boost performance.

We want to highlight one particular work that uses the Transformer-layer architecture to regularize the challenging problem of MRI image reconstruction from under-sampled measurements (Korkmaz et al., 2021a). This work is inspired by the strong prior induced by the structure of untrained neural networks (Ulyanov et al., 2018; Qayyum et al., 2021). These untrained network priors have recently garnered much attention from the medical image community as they do not need labeled training data. Considering advances in the untrained neural network area, we believe this direction requires further attention from medical imaging researchers in the context of Transformers (Wu et al., 2022).

We also observe that compared to the early years of CNNs (one paper from 2012 to 2015), Transformers have rapidly gained widespread attention in the medical image reconstruction community (more than a dozen papers in 2021), potentially due to the recent advancement in image-to-image translation frameworks.

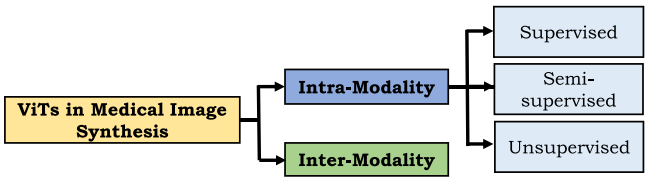

Fig. 21. Taxonomy of ViT-based medical image synthesis approaches.

7. Medical image synthesis

In this section, we provide an overview of the applications of ViTs in medical image synthesis. Most of these approaches incorporate adversarial loss to synthesize realistic and high-quality medical images, albeit at the expense of training instability (Liu et al., 2020). We have further classified these approaches into *intra-modality synthesis* and *inter-modality synthesis* due to a different set of challenges in both categories, as shown in Fig. 21.

7.1. Intra-modality approaches

The goal of intra-modality synthesis is to generate higher-quality images from the relatively lower quality input images of the same modality. Next, we describe the details of ViT-based intra-modality medical image synthesis approaches.

7.1.1. Supervised methods

Supervised image synthesis methods require paired source and target images to train ViT-based models. Paired data is difficult to obtain due to annotation cost and time constraints, thereby generally hindering the applicability of these models in medical imaging applications. Zhang et al. (2021b) focus on synthesizing infant brain structural MRIs (T1w and T2w scans) using both transformer and performer (simplified self-attention) layers (Choromanski et al., 2020). Specifically, they design a novel multi-resolution pyramid-like UNet framework, PTNet, utilizing performer encoder, performer decoder, and transformer bottleneck to synthesize high-quality infant MRI. They demonstrate the superiority of PTNet both qualitatively and quantitatively compared to pix2pix (Isola et al., 2017), and pix2pixHD (Wang et al., 2018c) on large-scale infant MRI dataset (Makropoulos et al., 2018). Furthermore, in addition to better synthesis quality, PTNet has a reasonable execution time of around 30 slices per second.

7.1.2. Semi-supervised methods

Semi-supervised approaches typically require small amounts of labeled data along with large unlabeled data to train models. Kamran et al. (2021) propose a multi-scale conditional generative adversarial network (GAN) (Isola et al., 2017) using ViT as a discriminator. They train their proposed model in a semi-supervised way to simultaneously synthesize Fluorescein Angiography (FA) images from fundus photographs and predict retinal degeneration. They use softmax activation after MLP head output and a categorical CE loss for classification. Besides adversarial loss, they also use MSE and perceptual losses to train their network. For ViT discriminator, they use embedding feature loss calculated using positional and patch features from the transformer encoder layers by successfully inserting the real and synthesized FA images. Their quantitative results in terms of Frechet inception Distance (Heusel et al., 2017) and Kernel Inception Distance (Bińkowski et al., 2018) demonstrate the superiority of their approach over baseline methods on diabetic retinopathy dataset provided by Hajeb et al. (Hajeb Mohammad Alipour et al., 2012).

7.1.3. Unsupervised methods

These approaches are particularly suitable for medical image synthesis tasks as they do not require paired training datasets. Recently, Ristea et al. (2021) propose a cycle-consistent generative adversarial transformer (CyTran) to translate unpaired contrast CT scans to non-contrast CT scans and volumetric image registration of contrast CT scans to non-contrast CT scans. To handle high-resolution CT images, they propose hybrid convolution and multi-head attention-based architecture shown in Fig. 22. CyTran is unsupervised due to the integration of cyclic loss. Moreover, they introduce the Coltea-Lung-CT100 W dataset formed of 100 3D anonymized triphasic lung CT scans of female patients.

7.2. Inter-modality approaches

The inter-modality approaches aim to synthesize targets to capture the useful structural information in the source images of different modalities. Examples include CT to MRI translation or vice-versa. Due to challenges associated with inter-modal translation, only supervised approaches have been explored.

Dalmaz et al. (2021) introduce a novel synthesis approach, ResViT, for the multi-modal imaging based on a conditional deep adversarial network with ViT-based generator. Specifically, ResViT combines the sensitivity of vision transformers to global context and the localization power of CNNs. Furthermore, adversarial loss has been leveraged to preserve the realism of the generated images. The bottleneck comprises novel aggregated residual transformer blocks to synergistically preserve local and global context, with a weight-sharing strategy to minimize model complexity. The effectiveness of ResViT model is demonstrated on two multi-contrast brain MRI datasets, BraTS (Menze et al., 2014), and a multi-modal pelvic MRI-CT dataset (Nyholm et al., 2018).

7.3. Discussion

In this section, we have reviewed the applications of ViT models in medical image synthesis. Realistic synthesis of medical images is particularly important as, in many practical applications, a certain modality is desired but infeasible to acquire due to cost and privacy issues. For instance bone segmentation is easy on CT, and therefore translating MR to CT allows indirect bone segmentation using existing architectures on CT. Recent transformer-based approaches can help towards this goal due to their ability to generate more realistic images than GAN-based methods.

Furthermore, most Transformer-based medical image synthesis approaches use the adversarial loss to generate realistic images. The adversarial loss can cause mode-collapse, and effective strategies must be employed to mitigate this issue (Wang et al., 2021d).

Lastly, to the best of our knowledge, no work has been done using transformer-based models for inter-modality image synthesis approaches in an unsupervised setting. This can be due to the highly challenging nature of the problem (e.g., CT and MRI images of the same subject have significantly different appearances, as shown in Fig. 23). Whether transformer-based models are suitable for this challenging task is an interesting direction to explore.

8. Medical image registration

Medical image registration aims to find dense per-voxel displacement and establish alignment between a pair of fixed and moving images. In medical imaging, registration may be necessary when analyzing a pair of images acquired at different times, from different viewpoints, or using different modalities (like MRI and CT) (Haskins et al., 2020). Accurate medical image registration is a challenging task due to difficulties in extracting discriminative features from multimodal medical images, complex motion, and lack of robust outlier rejection approaches (Alam et al., 2018). In this section, we briefly highlight the applications of ViTs in medical image registration.

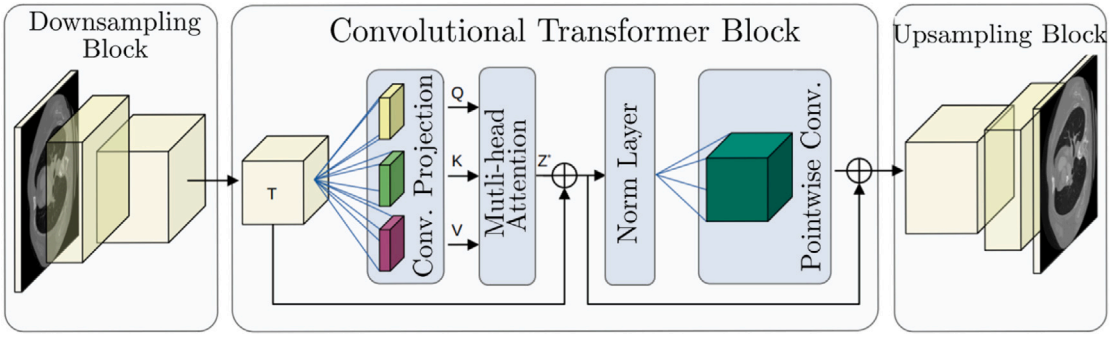

Fig. 22. Hybrid convolutional-transformer network for CT image generation as proposed in Ristea et al. (2021). It consists of down-sampling convolutional layers to extract features from input images, a convolutional-transformer block comprising a multi-head self-attention mechanism, and an upsampling block to generate output images.

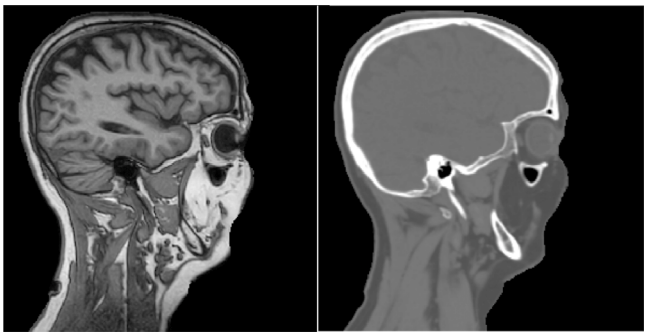

Fig. 23. A pair of MRI (left) and CT (right) images of the same subject showing the significant appearance gap between the two modalities making medical image synthesis from MRI to CT a challenging task.
Source: Image is from Wolterink et al. (2017).

The first study to investigate the usage of transformers for self-supervised medical volumetric image registration has been proposed by Chen et al. (2021b). Their model, ViT-V-Net, consists of a hybrid architecture composed of convolutional and transformer layers. Specifically, ViTs are applied to the high-level features of fixed and moving images extracted via a series of convolutional and max-pooling layers. The output from ViT is then reshaped and decoded using a V-Net style decoder (Milletari et al., 2018). To efficiently propagate the information, ViT-V-Net uses long skip connections between the encoder and decoder. The output of the ViT-V-Net decoder is a dense displacement field, which is fed to the spatial transformer network for warping. Experiments on in-house MRI dataset show superiority of ViT-V-Net over other competing approaches in terms of Dice score. Chen et al. (2021a) further extends ViT-V-Net and propose TransMorph model for volumetric medical image registration. Particularly, TransMorph makes use of Swin Transformer in the encoder to capture the semantic correspondence between input fixed and moving images, followed by long skip connections-based convolutional decoder to predict dense displacement field. For uncertainty estimation, they also introduce Bayesian deep learning by applying variational inference on the parameters of the encoder in TransMorph. Extensive evaluation is performed to compare TransMorph with other approaches for the medical image registration task. Specifically, experiments on inter-patient brain MRI registration provided by John-Hopkin university and XCAT-to-CT registration demonstrate the superiority of TransMorph against twelve different hand-crafted, CNN-based, and transformer-based approaches. Similarly, Zhang et al. (2021f) present a novel dual transformer architecture (DTN) for volumetric diffeomorphic registration by establishing correspondences between anatomical structures in an unsupervised manner as shown in Fig. 24. The DTN consists of two CNN-based 3D UNet encoders to extract embeddings of separate and concatenated volumetric MRI images. To further refine and enhance the embeddings, they propose encoder–decoder-based dual transformers to encode the cross-volume dependencies. Given the enhanced embeddings, the CNN decoder infers the deformation fields. Qualitative and quantitative results in terms of Dice similarity coefficient and negative Jacobian determinant on OASIS dataset (Marcus et al., 2007) of MRI scans demonstrate the effectiveness of their proposed architecture. In other work, Tulder et al. (2021) propose a novel cross-view transformer method to transfer information between unregistered views at the level of spatial feature maps, and shows its effectiveness for multi-view mammography and chest X-ray datasets.

8.1. Discussion

Application of transformers to medical image registration problem is still at early stages, and it is difficult to draw any conclusion at this stage. However, seeing the rapid development of Transformer-based registration approaches in generic computer vision, we expect to see the same trend in this field in the near future.

9. Clinical report generation

Recently, immense progress has been made to automatically generate clinical reports from medical images using deep learning (Pavlopoulos et al., 2021; Monshi et al., 2020; Kougia et al., 2019; Messina et al., 2020). This automatic report generation process can help clinicians in accurate decision-making. However, generating reports (or captions) from the medical imaging data is challenging due to diversity in the reports of different radiologists, long sequence length (unlike natural image captions), and dataset bias (more normal data compared to abnormal). Moreover, an effective medical report generation model is expected to process two key attributes: (1) *language fluency* for human readability and (2) *clinical accuracy* to correctly identify the disease along with related symptoms. In this section, we briefly describe how transformer models help achieve these desired goals and help mitigate the aforementioned challenges associated with medical report generation. Specifically, these transformer-based approaches have achieved state-of-the-art performance both in terms of Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics. Also note that, unlike previous sections that mainly discuss ViTs, in this section, the focus is on the transformers as powerful language models to exploit the long-range dependencies for sentence generation. We have broadly categorized transformer-based clinical report generation approaches into reinforcement learning (RL) based and supervised/unsupervised learning methods, as shown in Fig. 25, due to differences in their underlying training mechanism.

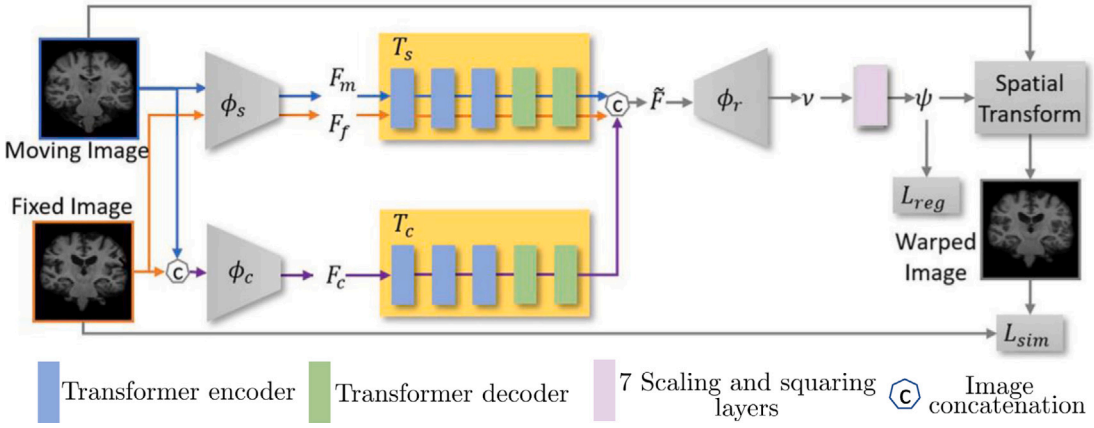

Fig. 24. Architecture of Dual Transform Network (DTN) proposed for Diffeomorphic registration (Zhang et al., 2021f). Features of moving image are extracted via 3D UNet encoder ϕ_s while features of concatenated moving and fixed image are extracted via another 3D UNet encoder ϕ_c. These features are collapsed into sequences and passed to two transformers networks, T_s and T_c, to handle the cross-image global relevance learning on separate single channel images and the image concatenation. The resulting features are concatenated together and passed to the CNN decoder (ϕ_r) to infer the velocity field (v), and subsequently registration field (ψ).

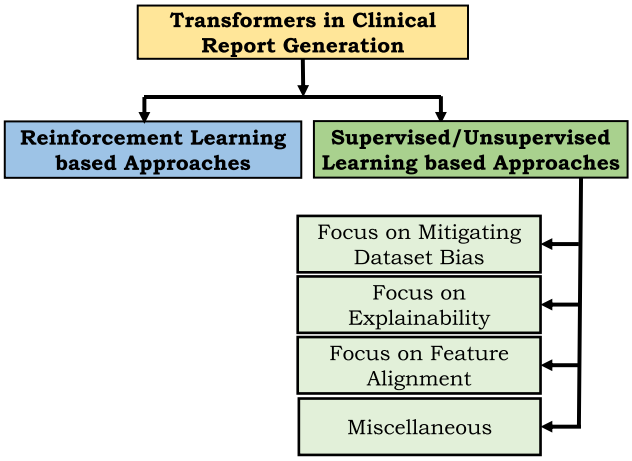

Fig. 25. Taxonomy of applications of Transformers in clinical report generation.

9.1. Reinforcement learning based approaches

RL-based medical report generation approaches can directly use the evaluation metrics of interest (like human evaluation, relevant medical terminologies, etc.) as rewards and update the model parameters via policy gradient. All approaches covered in this section use the self-critical RL (Rennie et al., 2017) approach to train models, which is more suitable for the report generation task compared to the conventional RL.

One of the first attempts to integrate transformer in clinical report generation has been made by Xiong et al. (2019). They propose Reinforced-Transformer for Medical Image Captioning (RTMIC) that consists of a pre-trained DenseNet (Huang et al., 2017a) to identify the region of interest from the input medical image, followed by a transformer-based encoder to extract visual features. These features are given as input to the captioning decoder to generate sentences. All these modules are updated via self-critical reinforcement learning method during training on IU Chest X-ray dataset (Demner-Fushman et al., 2016). Similarly, Miura et al. (2020) show that the high accuracy of automatic radiology reports as measured by natural language generation metrics such as BLEU (Papineni et al., 2002) and CIDer (Vedantam et al., 2015) are often incomplete and inconsistent. To address these challenges, Miura et al. (2020) propose a transformer-based model that directly optimizes the two newly proposed reward functions using self-critical RL. The first reward promotes the coverage of radiology domain entities with corresponding reference reports, and the second reward promotes the consistency of the generated reports with their descriptions in the reference reports. Further, they combine these reward functions with the semantic equivalence metric of BERTScore (Zhang et al., 2019a) that results in generated reports with better performance in terms of clinical metrics.

Surgical Instructions Generation. Inspired by the success of transformers in medical report generation, Zhang et al. (2021e) propose a transformer model to generate instructions from the surgical scenes. Lack of a predefined template, as in the case of medical report generation, makes generation of surgical instructions a challenging task. To handle this challenge, Zhang et al. (2021e) have proposed an encoder–decoder based architecture with a transformer backbone to effectively model the dependencies for visual features, textual features, and visual-textural relational features. In particular, their architecture is optimized via self-critical reinforcement learning (Rennie et al., 2017) to accurately generate surgical reports on the DAISI dataset (Rojas-Muñoz et al., 2020).

9.2. Supervised and unsupervised approaches

Supervised/unsupervised approaches use differentiable loss functions to train models for medical report generation and do not interact with the environment via an agent. We have categorized supervised/unsupervised approaches into methods that focus on dataset bias, explainability, feature alignment, and miscellaneous categories based on the challenges these approaches address.

9.2.1. Dataset bias

Dataset bias is a common problem in medical report generation as there are far more sentences describing normalities than abnormalities. To mitigate this bias, Srinivasan et al. (2020) propose a hierarchical classification approach using a transformer as a decoder. Specifically, the transformer decoder leverage attention between and across features obtained from reports, images, and tags for effective report generation. The architecture consists of *Abnormality Detection Network* to classify normal and abnormal images, *Tag Classification Net* to generate tags against images, and *Report Generation Net* that takes image features and tags as inputs to generate final reports. NLG metrics of the approach on IU Chest X-ray dataset (Demner-Fushman et al., 2016) are indicated in Table 9. Similarly, Liu et al. (2021b) try to imitate the work of radiologists by distilling posterior and prior knowledge to generate accurate radiology reports. Their proposed architecture consists of

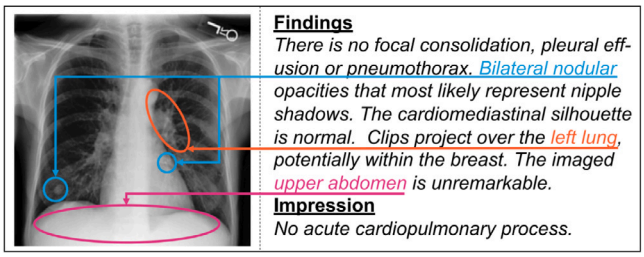

Fig. 26. A chest X-ray image and its accompanying report, which includes findings and impressions, with aligned visual and textual components highlighted in different colors.
Source: Figure taken from Chen et al. (2021h).

three modules of Posterior Knowledge Explorer (PoKE), Prior Knowledge Explorer (PrKE), and Multidomain Knowledge Distiller (MKD). Specifically, PoKE identifies the abnormal area in the input images, PrKE explores relevant prior information from the radiological reports and medical knowledge graph (in an effort to mitigate textual data bias), and MKD (based on transformer decoder) distills the posterior and prior knowledge to generate radiology report. In another work, You et al. (2021) propose AlignTransformer to generate a medical report from X-ray images. Specifically, AlignTransformer consists of two modules: align hierarchical attention module and multi-grained transformer module. Align hierarchial attention module helps to better locate the abnormal region in the input medical images. On the other hand, multi-grained transformer leverages multi-grained visual features using adaptive exploiting attention (Cornia et al., 2020) to accurately generate long medical reports. The performance of aligntransformer in terms of NLG and CE metrics on IU-Xray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) datasets are shown in Table 9.

9.2.2. Feature alignment

Feature alignment based approaches mainly focus on the accurate alignment of encoded representation of the medical images and corresponding text, which is crucial for the interaction and generation across modalities (images and text here) and subsequently for accurate report generation, as indicated in Fig. 26. To align better, Chen et al. (2021h) propose a cross-modal memory network to augment the transformer-based encoder–decoder model for radiology report generation. They design a shared memory to facilitate the alignment between the features of medical images and texts. Experiments on IU-Xray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) datasets demonstrate that the proposed model can better align image and text features as compared to baseline methods. Similarly, building on the shared-memory work of Chen et al. (2021h), Yan et al. (2021a) introduce a weakly supervised contrastive objective to favor reports that are semantically close to the target, thereby producing more clinically accurate outputs. In another work, Amjoud and Amrouch (2021) investigate the impact on the report generation performance by modifying different architectural components of the model proposed by Chen et al. (2021h) including replacing visual extractor and changing the number of layers in transformer-based decoder. Specifically, they use Densenet-121 instead of ResNet and claim that it helps in gradient flow. This is because each layer of DenseNet-121 directly connects to the gradients from the loss function. Further, they increase the number of layers of encoder and decoder to 12 instead of 3. Experiments show that they could achieve small gain (0.479 compared to 0.470 in BLEU-1, and 0.380 compared to 0.371 in ROUGH) over (Chen et al., 2021h).

9.2.3. Explainable models

Explainability in medical report generation is crucial to improve trustworthiness for deploying models in clinical settings and a mean for extracting bounding boxes for lesion localization. For model explainability, Hou et al. (2021) employ attention to identify regions of interest in the input image and demonstrate where the model is focusing for the resulting text. This attention mechanism increases the explainability of black-box models used in clinical settings and provides a method for extracting bounding boxes for disease localization. Specifically, they propose RATCHET transformer model to generate reports by using DenseNet-101 (Huang et al., 2017a) as an image feature extractor. RATCHET consists of a transformer-based RNN-Decoder for generating chest radiograph reports. They assess the model's natural language skills and the medical correctness of generated reports. Their results indicate that a transformer-based architecture can outperform traditional LSTM-based architectures in both natural language skills as well as medical pathology diagnosis. Further, the attention maps highlights the region in the image that is responsible for each of the generated text tokens. Similarly, despite the immense interest of AI and clinical medicine researchers in the automatic report generation area, benchmark datasets are scarce, and the field lacks reliable evaluation metrics. To address these challenges, Li et al. (2021b) introduce a large-scale Fundus fluorescence in Angiography images and reports dataset containing 10,790 reports describing 1,048,584 images with explainable annotations as shown in Fig. 27. The dataset comes with annotated Chinese reports and corresponding translated English reports. Further, they introduce nine reliable metrics based on human evaluation criteria.

9.2.4. Miscellaneous

In this section, we highlight several approaches that try to improve different aspects of clinical report generation from medical images. Examples include a memory-driven transformer to capture similar patterns in reports, uncertainty quantification for reliable report generation, a curriculum learning-based method, and an unsupervised approach to avoid paired training datasets.

Chen et al. (2020c) propose a **memory-driven transformer to exploit similar patterns** in the radiology image reports. Specifically, they add a module to each layer of transformer-based decoder by optimizing the original layer normalization with a novel memory-driven conditional layer normalization. Experiments on IU Chest X-ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) datasets demonstrate the superiority of their approach in terms of Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics over their self-implemented baseline that consists of vanilla Transformer, with three layers, 8 heads and 512 hidden units. Comparison with recent transformer based approach is shown in Table 9.

Extensive experiments on IU Chest X-ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) datasets demonstrate the superiority of their approach both in terms of Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics. Similarly, Lovelace and Mortazavi (2020) also leverage the transformer-based encoder and decoder for accurate medical report generation on MIMIC-CXR dataset (Johnson et al., 2019). To **emphasis on clinically relevant report generation**, they design a method to differentiate clinical information from generated reports, which they use to refine the model for clinical coherence. In another work, Alfarghaly et al. (2021) present a pre-trained transformer-based model to generate a medical report from images. Specifically, the encoder consists of a pre-retrained CheXNet model that can generate semantic features from the input medical images. These semantic features are used to **condition GPT2 decoder** (Ziegler et al., 2019; Radford et al., 2019) to generate accurate medical reports. Similarly, to judge the reliability of the automatic medical report generating model, uncertainty quantification is the key indicator. To incorporate this measure, Wang et al. (2021b) propose **transformer-based confidence guided framework** to quantify both visual and textual uncertainty. These uncertainties are subsequently used to construct an uncertainty-weighted loss to reduce misjudgment risk and improve the overall performance of the generated report.

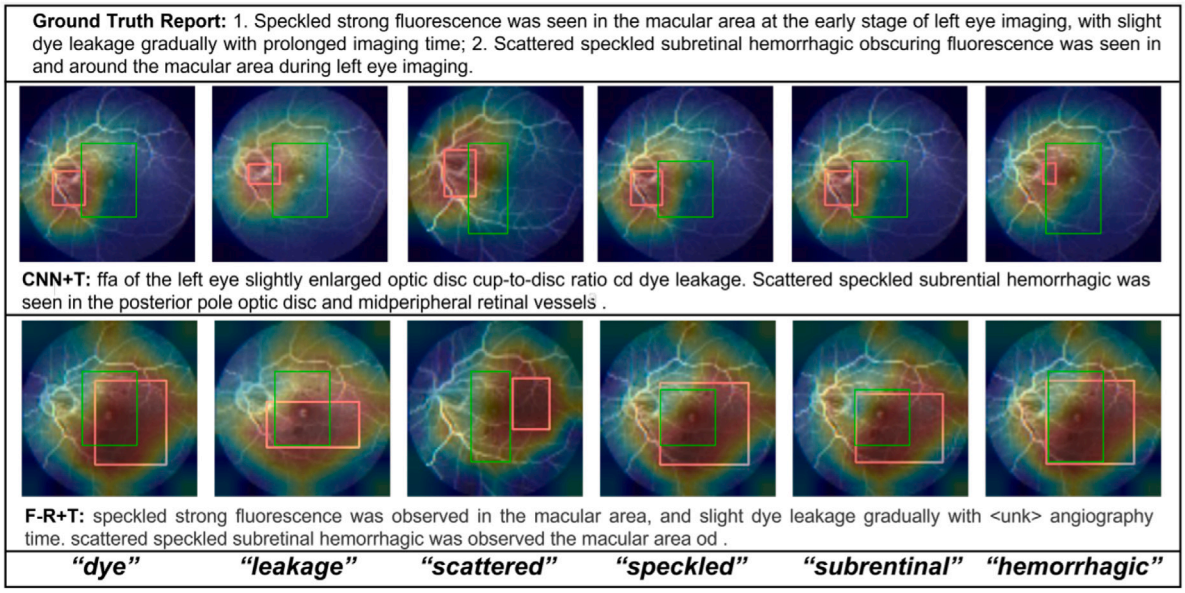

Fig. 27. The depiction of lesion-image attention mapping areas and ground truth between CNN+Transformer and Faster-RCNN+Transformer samples, where green boxes represent the annotated region for each lesion word and red boxes represent the lesion-image attention mapping regions.
Source: Image taken from Li et al. (2021b).

In other work, Nguyen et al. (2021b) propose differentiable end-to-end framework that consists of **transformer as generator for report generation**. Specifically, their proposed framework has three complementary modules: a *classifier* to learn the representation of disease features, a transformer-based *generator* model to generate the medical report, and *interpreter* to make the generated report consistent with the classifier output. They demonstrate the effectiveness of proposed components on IU-Xray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) datasets. Inspired by **curriculum learning** (Bengio et al., 2009; Nooralahzadeh et al., 2021) present a two-stage transformer architecture to progressively generate medical reports. Their progressive approach shows better performance over single-stage baselines in generating full-radiology reports. In another study, Pahwa et al. (2021) investigate the **impact of visual feature extractor** model on the performance of medical report generation. Based on insights, they propose a modified HRNet (Sun et al., 2019), MedSkip, to extract visual features for the subsequent processing by the transformer-based decoder to generate an accurate medical report. Similarly, Park et al. (2021c) **investigate the expressiveness of features** to discriminate between normal and abnormal images. They demonstrate the superiority of transformer-based decoder without global average pooling over hierarchical LSTM baseline. Existing transformer-based report generation models are mostly supervised and use paired image-report data during training. The paired data is difficult to obtain due to privacy and cost in the medical domain. To mitigate this issue, Liu et al. (2021d) propose a knowledge graph auto-encoder that works in the share latent domain of images and reports to extract useful information in an **unsupervised way**. Specifically, they use attention in the encoder to extract the knowledge representation from the knowledge graph and use a three-layer transformer in the decoder to generate reports. Their proposed framework can also be used in a semi-supervised or supervised manner in addition to the unsupervised mode. Quantitative and qualitative results, as well as evaluation by radiologists, corroborate the effectiveness of their approach.

9.3. Discussion

In this section, we have provided a comprehensive overview of the transformer's applications for clinical report generation from X-ray images. In contrast to previous sections that discuss applications of ViTs, this section focuses on transformers as powerful language models. It is also pertinent to note that even though multiple surveys exist covering the applications of deep learning in clinical report generation (Pavlopoulos et al., 2021; Monshi et al., 2020; Kougia et al., 2019; Messina et al., 2020), to the best of our knowledge, ***none of these have covered the applications of transformer models*** *in the area despite having transformers' phenomenal impact since their inception back in 2017. In this regard, we hope this section will serve as a valuable resource to the research community.*

As we have seen, transformer-based report generation models mostly rely on natural language generation (NLG) evaluation metrics such as CIDEr and BLEU to assess performance. These NLG metrics often fail to represent clinical efficacy. One recent work by Miura et al. (2020) addresses this issue by proposing two new reward functions for the transformer model in reinforcement learning framework to better capture disease and anatomical knowledge in the generated reports. Another work by Li et al. (2021b) introduces nine reliable human evaluation criteria to validate the generated reports. Despite these works, we believe that more attention from the research community is required to design reliable clinical evaluation metrics to facilitate the adoption of transformer-based medical report generation models in clinical settings.

All transformer-based approaches covered in this section use the X-ray modality for automatic report generation. Generating reports from other modalities like MRI or PET have their own challenges associated with them due to more complex nature of these sensing methodologies in 3D as compared to X-ray scans which are 2D projections of the 3D (Monshi et al., 2020). Further, few medical datasets like ROCO (Pelka et al., 2018), PEIR Gross (Jing et al., 2017), and ImageCLEF (Garcia Seco De Herrera et al., 2018) are available that consist of multiple modalities, different body parts, and corresponding captions. These datasets have the potential to become worthy benchmarks to gauge the performance of future multimodal (or unimodal like MRI, PET) transformer-based models for medical report generation. We believe that transformer-based models tailored to specific modalities to generate reports must be explored in the future with a focus on creating diverse and challenging datasets of other modalities. Details of a few existing medical report generation datasets are given in Table 8.

Further, we would like to point out the interested researchers towards the recently explored surgical instructions generation work using transformers (Zhang et al., 2021e) that could have a huge impact on surgical robotics, a market that is expected to reach USD 22.27 Billion by 2028.

Table 8
Statistics of existing medical report generation datasets where * means average number and — for datasets that do not provide relevant information. Most of the transformers based models for medical report generation use Open-IU and MIMIC-CXR for evaluating results. *Source:* Table adapted from Li et al. (2021b).

Dataset	Image			Report		
	Number	Modality	View*	Length*	Language	Cases
IU X-ray (Demner-Fushman et al., 2016)	7,470	X-ray	2	32.5	Eng.	2,955
MIMIC-CXR (Johnson et al., 2019)	377,110	X-ray	1	53.2	Eng.	276,778
PadChest (Bustos et al., 2020)	160,868	X-ray	2	–	Es	22,710
CX-CHR (Li et al., 2019)	45,598	X-ray	2	66.9	Zh	40,410
DIARETDB1 (Kälviäinen and Uusitalo, 2007)	89	CFP	1	–	Eng	89
MESSIDOR (Decencière et al., 2014)	1,200	CFP	2	–	Fr	587
FFA-IR (Li et al., 2021b)	1,048,584	FFA	87	91.2	Eng/Zh	10,790
COV-CTR (Li et al., 2020b)	728	CT-Scans	1	77.3	Eng/Zh	728
DEN (Huang et al., 2021d)	15,709	CFP+FFA	1	7	Eng	–
STARE (Hoover et al., 2000)	397	CFP+FFA	5	–	Eng	397

However, only one dataset, DAISI (Rojas-Muñoz et al., 2020), is available to evaluate models in this emerging area, demanding attention from the medical community to create diverse and more challenging datasets.

Moreover, datasets for medical report generation like IU X-ray (Demner-Fushman et al., 2016) does not contain any standard train-test split and most of the transformer-based approaches evaluate the performance on different tests data. In this regard, the results in Table 9 are not directly comparable, but they can provide an overall indication about the performance of the models. We think what seems to be missing is a set of standardized procedures for creating challenging and diverse clinical report generation datasets.

10. Other applications

In this section, we briefly highlight applications of Transformers in other medical imaging areas, including survival outcome prediction, visual question answering, and medical point cloud analysis. **Survival outcome prediction** is a challenging regression task that seeks to predict the relative risk of cancer death. Recently, transformer models have shown impressive success in predicting survival rates. Chen et al. (2021f) propose a Multimodal Co-Attention Transformer (MCAT) for the survival outcome prediction from whole-slide imaging (WSI) in pathology. MCAT learns a co-attention mapping between genomics and WSIs features to discover how histology features attend to genes while predicting patient survival outcomes. Extensive experiments on five cancer datasets demonstrate the superiority of MCAT compared to CNN-based approaches. Similarly, Kipkogei et al. (2021) propose a Transformer-based architecture, Clinical Transformer, to model the relation between clinical and molecular features to predict survival outcomes from cancerous lung dataset (Samstein et al., 2019). In other work, Huang et al. (2021a) propose transformer based model, SeTranSurv, that extracts patch features from whole slides images via self-supervised learning and adaptively aggregates these features according to their spatial information. Extensive experiments on three datasets demonstrate the effectiveness of their model. In another work, Eslami et al. (2021) propose **PubMedCLIP, a fine-tuned version of Contrastive Language-Image Pre-Training (CLIP)** (Radford et al., 2021) for the medical domain by training it on the image–caption pairs from PubMed articles. Extensive experiments show that PubMedCLIP outperforms the previous state-of-the-art by nearly 3%. Recently, Yu et al. (2021c) propose **3D Medical Point Transformer** (3DMPT) to analyze 3D medical data. 3DMPT is tested on 3D medical classification and part segmentation based tasks. Similarly, Malkiel et al. (2021) propose a **Transformer-based architecture to analyze fMRI data**. They pre-train the model on 4D fMRI data in a self-supervised manner and fine-tune it on various downstream tasks, including age and gender prediction, as well as diagnosing Schizophrenia. In another work, Yu et al. (2021b) present a method based on hybrid CNN-Transformer architecture and group contrastive learning to model **ugly duckling context for melanoma detection**. Our model is capable of concurrently performing patient-level prediction and lesions level prediction from a group of lesions. Experiments on the ISIC 2020 dataset demonstrate that our model can obtain better diagnostic result than baseline CNN trained without using lesion context. Other applications include transformers to analyze US videos (Reynaud et al., 2021), phase recognition in surgical video (Gao et al., 2021a), few-shot medical domain adaptation (Li et al., 2021c), and for automatic diagnosis of coronary artery disease (Ma et al., 2021a).

11. Open challenges and future directions

We have reviewed the exciting applications of vision transformers in medical image analysis. Despite their impressive performance, there remain several open research questions. In this section, we outline some of their limitations and highlight promising future research directions. Specifically, we will discuss the challenges of pre-training on large datasets (Section 11.1), interpretability of ViT-based medical imaging approaches (Section 11.2), robustness against adversarial attacks (Section 11.3), designing efficient ViT architectures for real-time medical applications (Section 11.4), challenges in deploying ViT-based models in distributed settings (Section 11.5), and domain adaptation (Section 11.6). Further, wherever possible, we refer interested researchers to relevant CNNs-based medical imaging resources (recent studies, datasets, software libraries, etc.) to explore previously untapped applications by ViT-based models in medical imaging like adversarial robustness.

11.1. Pre-training

Due to a lack of intrinsic inductive biases in modeling local visual features, ViTs need to figure out the image-specific concepts on their own via pre-training from large-scale training datasets (Dosovitskiy

Table 9
Quantitative comparison of transformer models for the task of clinical report generation in terms of Natural Language Generation (NLG) and Clinical Efficacy (CE) on two benchmark datasets. The NLG metrics include BLEU (BL) (Papineni et al., 2002), METEOR (MTR) (Denkowski and Lavie, 2011) and ROUGE-L (RG-L) (Rouge, 2004) and CE metrics include precision, recall and F1 score. * indicates approaches that do not use Transformer block. Note that datasets for medical report generation does not contain any standard train-test split and most of the transformer-based approaches evaluate the performance on different tests data. In this regard, the results in this are not directly comparable, but they can provide an overall indication about the performance of the models.

Dataset	Model	Year	NLG Metrics						CE Metrics		
			BL-1	BL-2	BL-3	BL-4	MTR	RG-L	P	R	F1
IU X-ray (Demner-Fushman et al., 2016)	RTMIC (Xiong et al., 2019)	2019	0.350	0.234	0.143	0.096	–	–	–	–	–
	*Yuan et al. (Yuan et al., 2019)	2019	0.445	0.289	0.200	0.143	–	0.359	–	–	–
	HRG Transformer	2019	0.464	0.301	0.212	0.158	–	–	–	–	–
	KERP (Li et al., 2019)	2019	0.482	0.325	0.226	0.162	–	0.339	–	–	–
	*Zhang et al. (Zhang et al., 2020)	2020	0.441	0.291	0.203	0.147	–	0.367	–	–	–
	Hierarchical Transformer (Srinivasan et al., 2020)	2020	0.464	0.301	0.212	0.158	–	–	–	–	–
	Memory Transformer (Chen et al., 2020c)	2020	0.470	0.304	0.219	0.165	0.187	0.371	–	–	–
	*CMCL (Liu et al., 2022a)	2021	0.473	0.305	0.217	0.162	0.186	0.378	–	–	–
	$\mathcal{M}^2$ TR. (Nooralahzadeh et al., 2021)	2021	0.475	0.301	0.228	0.180	0.169	0.373	–	–	–
	$\mathcal{M}^2$ TR. Prog. (Nooralahzadeh et al., 2021)	2021	0.486	0.317	0.232	0.173	0.192	0.390	–	–	–
	Wang et al. (Wang et al., 2021b)	2021	0.481	0.309	0.223	0.169	0.193	0.365	–	–	–
	Nguyen et al (Nguyen et al., 2021b)	2021	0.515	0.378	0.293	0.235	0.219	0.436	–	–	–
	PPKED (Liu et al., 2021b)	2021	0.483	0.315	0.224	0.168	0.190	0.376	–	–	–
	Align Transformer (You et al., 2021)	2021	0.484	0.313	0.225	0.173	0.204	0.379	–	–	–
	KGAE Unsupervised (Liu et al., 2021d)	2021	0.417	0.263	0.181	0.126	0.149	0.318	–	–	–
	KGAE Semi-supervised (Liu et al., 2021d)	2021	0.497	0.320	0.232	0.171	0.189	0.379	–	–	–
	KGAE Supervised (Liu et al., 2021d)	2021	0.512	0.327	0.240	0.179	0.195	0.383	–	–	–
	*Wang et al. (Wang et al., 2022b)	2022	0.450	0.301	0.213	0.158	–	0.384	–	–	–
MIMIC-CXR (Johnson et al., 2019)	Transformers (Vaswani et al., 2017)	2017	0.409	0.268	0.191	0.144	0.157	0.318	–	–	–
	Memory Transformer (Chen et al., 2020c)	2020	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
	Clinical Transformer (Lovelace and Mortazavi, 2020)	2020	0.415	0.272	0.193	0.146	0.159	0.318	0.411	0.475	0.361
	*CMCL (Liu et al., 2022a)	2021	0.344	0.217	0.140	0.097	0.133	0.281	–	–	–
	$\mathcal{M}^2$ TR. (Nooralahzadeh et al., 2021)	2021	0.361	0.221	0.146	0.101	0.139	0.266	0.324	0.241	0.276
	$\mathcal{M}^2$ TR. Prog. (Nooralahzadeh et al., 2021)	2021	0.378	0.232	0.154	0.107	0.145	0.272	0.240	0.428	0.308
	PPKED (Liu et al., 2021b)	2021	0.360	0.224	0.149	0.106	0.149	0.284	–	–	–
	Align Transformer (You et al., 2021)	2021	0.378	0.235	0.156	0.112	0.158	0.283	–	–	–
	Ngyuen et al (Nguyen et al., 2021b)	2021	0.495	0.360	0.278	0.224	0.222	0.390	–	–	–
	MDT+WCL (Yan et al., 2021a)	2021	0.373	–	–	0.107	0.144	0.274	0.384	0.274	0.294
	$\mathcal{M}^2$ Trans (CE) (Miura et al., 2020)	2021	–	–	–	0.111	–	–	0.463	0.732	0.567
	$\mathcal{M}^2$ Trans (EN) (Miura et al., 2020)	2021	–	–	–	0.114	–	–	0.503	0.651	0.567
	KGAE Unsupervised (Liu et al., 2021d)	2021	0.221	0.144	0.096	0.062	0.097	0.208	0.214	0.158	0.156
	KGAE Semi-supervised (Liu et al., 2021d)	2021	0.352	0.219	0.149	0.108	0.147	0.290	0.360	0.302	0.307
	KGAE Supervised (Liu et al., 2021d)	2021	0.369	0.231	0.156	0.118	0.153	0.295	0.389	0.362	0.355

et al., 2020). This pose a challenge to their widespread application in medical imaging, where typically datasets are orders of magnitude smaller compared to natural image datasets due to cost, privacy concerns, and the rarity of certain diseases, thereby making ViTs difficult to train efficiently in the medical domain. Existing learning-based medical imaging approaches commonly rely on transferring learning

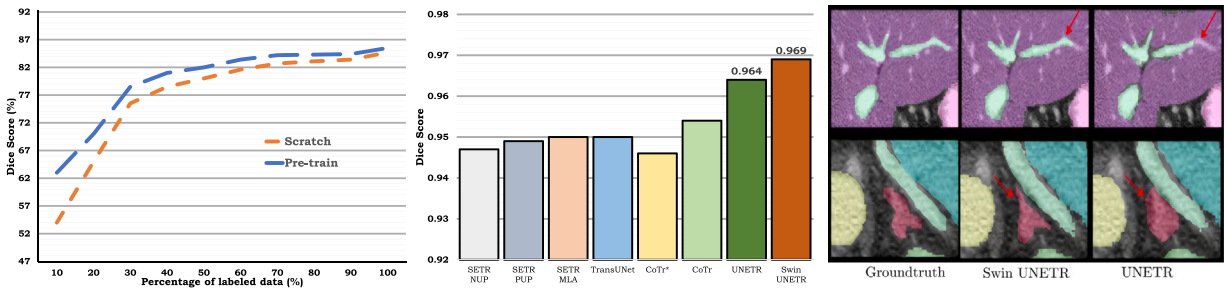

Fig. 28. Impact of pre-training ViT on domain specific medical imaging dataset. *First column:* By only using 10% labeled data, Swin UNET pre-trained on CT images (blue) is able to achieve 10% improvement in dice score over Swin UNET trained from scratch (orange). *Middle column:* Dice scores of recently proposed transformer models on the spleen segmentation task of MSD dataset. Swin UNETR achieves state-of-the-art performance due to pre-training on the domain specific (CT) medical dataset. *Last column:* Qualitative visualizations of Swin UNETR pre-trained on CT images on Synapse multi-organ segmentation challenge. It can be seen that Swin UNETR predictions are closer to the groundtruth compared to baseline UNETR approach. First and last columns are adapted from Tang et al. (2021).

via ImageNet pretraining, which may be sub-optimal due to drastically different image characteristics between medical and natural images. Recently, Matsoukas et al. (2021) has studied the impact of pre-training on ViTs performance for image classification and segmentation via a careful set of extensive experiments on several medical imaging datasets. Below, we briefly highlight major findings of their work.

- CNNs outperform ViTs for the medical image classification task when initialized with random weights.
- CNNs and ViTs benefit significantly from ImageNet initialization for medical image classification. ViTs appear to benefit more from transfer learning, as they make up for the gap observed using random initialization, performing on par with their CNN counterparts.
- CNNs and ViTs perform better with self-supervised pre-training approaches like DINO (Caron et al., 2021) and BYOL (Grill et al., 2020). ViTs appear to outperform CNNs in this setting for medical image classification by a small margin.

In short, although recent ViT-based data-efficient approaches like DeiT (Touvron et al., 2021), Token-to-Token (Yuan et al., 2021a), transformer-in-transformer (Han et al., 2021), etc., report encouraging results in the generic vision applications, the task of learning these transformer models tailored to domain-specific medical imaging applications in a data-efficient manner is challenging. Recently, Tang et al. (2021) has made an attempt to handle this issue by investigating the effectiveness of self-supervised learning as a pre-training strategy on domain-specific medical images. Specifically, they propose 3D transformer-based hierarchical encoder, Swin UNETR, and after pre-training on 5,050 CT images, demonstrates its effectiveness via fine-tuning on the downstream task of medical image segmentation. The pre-training on the medical imaging dataset also reduces the annotation effort compared to training Swin UNETR from scratch with random initialization. This is shown in Fig. 28, where it can be seen that pre-trained Swin UNETR can achieve the same performance by using only 60% of data as achieved by random initialized Swin UNETR using 100% of labeled data. This results in 40%reduction of manual annotation effort. Furthermore, as shown in Fig. 28, fine-tuning pre-trained Swin UNETR on the downstream medical image segmentation achieves better quantitative and qualitative results as compared to randomly initialized UNETR. Despite these efforts, there still remain several open challenges like Swin UNETR pre-trained on CT dataset gives unsatisfactory performance when applied directly to other medical imaging modalities like MRI due to large domain gap between CT and MRI images. Furthermore, the effectiveness of Swin UNETR on other downstream medical imaging tasks like classification and detection requires further investigation. Moreover, recent works for CNNs have shown that self-supervised pre-training on both ImageNet and medical datasets can improve the generalization performance (for classification) of the model on distribution shifted medical dataset (Azizi et al., 2021) as compared to pre-training on ImageNet only. We believe such studies for ViT-based models, along with multi-instance contrastive learning to leverage patient meta data (Vu et al., 2021), will provide further insights to the community. Similarly, combining the self-supervised and semi-supervised pre-training in the context of ViTs for medical imaging applications is also an interesting avenue to explore (Chen et al., 2020b).

11.2. Interpretability

Although the success of transformers has been empirically established in an impressive number of medical imaging applications, it has so far eluded a satisfactory interpretation. In most medical imaging applications, ViT models have been deployed as block-boxes, thereby failing to provide insights and explain their learning behavior for making predictions. This black-box nature of ViTs has hindered their deployment in clinical practice since, in areas such as medical applications, it is imperative to identify the limitations and potential failure cases of designed systems, where interpretability plays a fundamental role (Reyes et al., 2020). Although several explainable AI-based medical imaging systems have been developed to gain deeper insights into the working of CNNs models for clinical applications (Singh et al., 2020; Yan et al., 2021b; Ghoshal and Tucker, 2020), however, the work is still in its infancy for ViT-based medical imaging applications. It is despite the inherent suitability of the self-attention mechanism to interpretability due to its ability to explicitly model interactions between every region in the image a shown in Fig. 29 (Chaudhari et al., 2021). Recent efforts for interpretable ViT-based medical imaging models leverage saliency-based approaches (Chefer et al., 2021) and Grad-CAM based visualizations (Selvaraju et al., 2017). Despite these efforts, the development of interpretable and explainable ViT-based approaches, specifically tailored for life-critical medical imaging applications, is a challenging and open research problem. Furthermore, formalisms, challenges, definitions, and evaluation protocols regarding interpretable ViTs based medical imaging systems must also be addressed. We believe that progress in this direction would not only help physicians to decide whether they should follow and trust automatic ViT-based model decisions but could also facilitate the deployment of such systems from a legal perspective.

11.3. Adversarial robustness

Advances in adversarial attacks have revealed the vulnerability of existing learning-based medical imaging systems against imperceptible perturbation in the input images (Ma et al., 2021b; Papangelou et al., 2018; Finlayson et al., 2019), as shown in Fig. 30. Considering the vast amount of money that underpins the medical imaging sector, this inevitably poses a risk whereby potential attackers may seek to profit from manipulation against the healthcare system. For example,

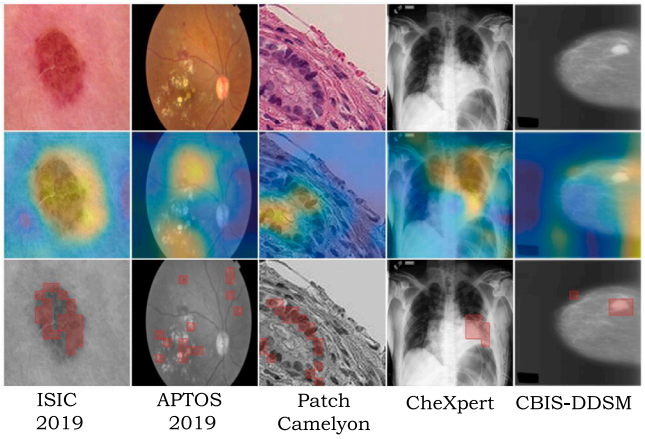

Fig. 29. Saliency maps comparison for a CNN-based ResNet-50 (He et al., 2016) (second row) and ViT-based DEIT-S (Touvron et al., 2021) (third row) on five medical imaging datasets for classification. Each column contains the original image (first row), a visualization of the ResNet-50 Grad-CAM saliency (second row), and a visualization of the DEIT-S's attention map (third row), respectively. Note that the ViTs provide a clear, localized picture of attention compared to ResNet-50, thus giving insight into how the model makes decisions.
Source: Image taken from Matsoukas et al. (2021).

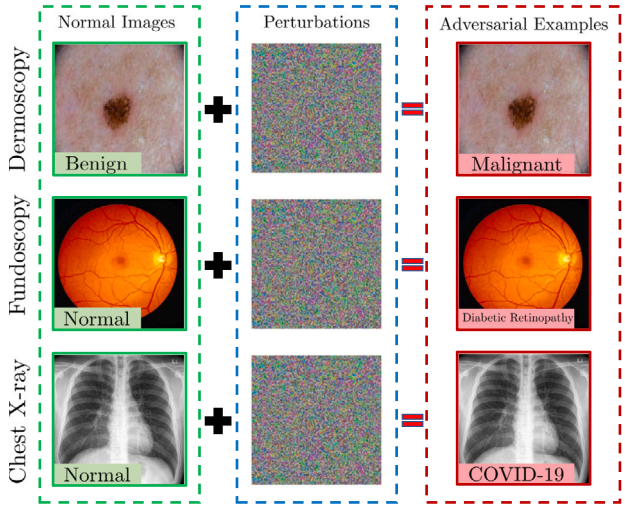

Fig. 30. Examples of adversarial attacks to fool learning-based models trained on medical image datasets. Left: normal images, Middle: adversarial perturbations, Right: adversarial images. The bottom tag is the predicted class, and green/red indicates correct/wrong predictions.

an attacker might try to manipulate the examination reports of patients for insurance fraud or a false medical reimbursement claim, thereby raising safety concerns. Therefore, ensuring the robustness of ViTs against adversarial attacks in life-critical medical imaging applications is of paramount importance. Although rich literature exists related to the robustness of CNNs in the medical imaging domain, to the best of our knowledge, no such study exists for ViTs, making it an exciting as well challenging direction to explore. Recently, few attempts have been made to evaluate the robustness of ViTs to adversarial attacks for natural images (Benz et al., 2021; Bhojanapalli et al., 2021; Wei et al., 2021; Shao et al., 2021b; Paul and Chen, 2021; Naseer et al., 2021b; Mao et al., 2021; Mahmood et al., 2021; Joshi et al., 2021; Aldahdooh et al., 2021). The main conclusions of these attempts, ignoring their nuance difference, can be summarized as *ViTs are more robust to adversarial attacks than CNNs*. However, these robust ViT models cannot be directly deployed for medical imaging applications as the variety and type of patterns and textures in medical images differ significantly from the natural domain. Therefore, a principled approach to evaluate the robustness of ViTs against adversarial attacks in the medical imaging context, which builds the groundwork for resilience, could serve as a critical model to deploy these models in clinical settings. Furthermore, theoretical understanding to provide guarantees about the performance and robustness of ViTs, like CNNs (Katz et al., 2017), can be of significant interest to medical imaging researchers. In Table 10, we provide a description of datasets used in adversarial medical learning to evaluate the robustness of CNNs for interested researchers to benchmark the robustness of ViT-based models.

Table 10
Description of datasets generally used in medical adversarial deep learning.

Dataset name	Dataset size	Modality
RSNA (RSNA, 2018)	29,700	X-ray
JSRT (Shiraishi et al., 2000)	247	X-ray
BraTS 2018 (Bakas et al., 2018)	1689	MRI
BraTS 2019 (BraTS, 2019a)	1675	MRI
OASIS (OASIS, 2019)	373–2168	MRI
HAM10000 (Tschandl et al., 2018)	10,000	Dermatoscopic
ISIC 18 (Codella et al., 2019)	3594	Dermatoscopic
LUNA 16 (LUNA, 2016)	888	CT-Scans
NIH Chest X-ray (Tang et al., 2019)	112,000	X-ray
APTOS (APTOS, 2019)	5590	Fundoscopy
Chest X-ray (Kermany et al., 2018)	5856	X-ray
NSLT (NLST, 2017)	75,000	CT-Scans
Diabetic Retinopathy (Retinopathy, 2015)	35,000	Fundoscopy

11.4. ViTs for medical imaging on edge devices

Despite the tremendous success of ViTs in numerous medical imaging applications, the intensive requirements for memory and computation hamper their deployment on resource constrained edge devices (Zhai et al., 2021; Tabani et al., 2021). Due to recent advancements in edge computing, healthcare providers can process, store and analyze complex medical imaging data on-premises, speeding up diagnosis, improving clinician workflows, enhancing patient privacy, and saving time—and potentially lives. These edge devices provide extremely fast and highly accurate processing of large amounts of medical imaging data, therefore demanding efficient hardware design to make ViT-based models suitable for edge computing-based medical imaging hardware. Recently few efforts have been made to compress transformer-based models by leveraging enhanced block-circulant matrix-based representation (Li et al., 2020a) and neural architecture search strategies (Wang et al., 2020d). Due to the exceptional performance of ViTs, we believe that there is a dire need for their domain-optimized architectural designs tailored for edge devices. It can have a tremendous impact on medical imaging-based health care applications where on-demand insights help teams make crucial and urgent decisions about patients.

11.5. Decentralized medical imaging solutions using ViTs

Building robust deep learning-based medical imaging models highly depends on the amount and diversity of the training data. The training data required to train a reliable and robust model may not be available in a single institution due to strict privacy regulations, the low incidence rate of some pathologies, data-ownership concerns, and limited numbers of patients. Federated Learning (FL) has been proposed to facilitate multi-hospital collaboration while obviating data transfer. Specifically, in FL, a shared model is built using distributed data from

Table 11
Description of some of the open-source federated learning and privacy-preserving frameworks, including some specifically developed for medical imaging applications.

Name	Framework	Description
TensorFlow Fed (Bonawitz et al., 2020)	TensorFlow	Open-source framework for machine learning and other computations on decentralized data.
CrypTen (Knott et al., 2021)	PyTorch	Framework to facilitate research in secure and privacy-preserving machine learning.
OpenMined (OpenMined, 2017)	TensorFlow, PyTorch, Keras	Open-source decentralized privacy preserving framework. Includes specialized tools like PySyft, PyGrid, and SyferText.
Opacus (Yousefpour et al., 2021)	PyTorch	Enables training PyTorch models with differential privacy. Allows the client to online track the privacy budget.
Deepee (Kaissis et al., 2021)	PyTorch	Library for differentially private deep learning for medical imaging in PyTorch.
PriMIA (Ziller et al., 2021)	PyTorch	Framework for end-to-end privacy-preserving decentralized deep learning for medical images.

multiple devices where each device trains the model using its local data and then shares the model parameters with the central model without sharing its actual data. Although a plethora of approaches exists that address FL for CNNs based medical imaging applications, the work is still in its infancy for ViTs and requires further attention. Recently few research efforts have been made to exploit the inherent structure of ViT in distributed medical imaging applications. Park et al. (2021a) propose a Federated Split Task-Agnostic (FESTA) framework that integrates the power of Federated and Split Learning (Yang et al., 2019; Vepakomma et al., 2018) in utilizing ViT to simultaneously process multiple chest X-ray tasks, including diagnosing COVID-19 CXR images on a large corpus of decentralized data. However, FESTA is just a proof-of-concept study, and its applicability in clinical trials requires further experimentation. Furthermore, challenges like privacy attacks and robustness against communication bottlenecks for ViT-based FL medical imaging systems require in-depth investigation. An interesting future direction is to explore recent privacy enhancement approaches like differential privacy (Abadi et al., 2016) to prevent gradient inversion attacks (Huang et al., 2021c) on FL-based medical imaging systems in the context of ViTs. In short, we believe that the successful implementation of distributed machine learning frameworks coupled with the strengths of ViTs could hold significant potential for enabling precision medicine at a large scale. This can lead to ViT models that yield unbiased decisions and are sensitive to rare diseases while respecting governance and privacy concerns. In Table 11, we highlight various tools and libraries that have been developed to implement distributed and secure deep learning. This can provide useful information for researchers who wish to rapidly prototype their ViT-based models for medical imaging in distributed settings.

11.6. Domain adaptation and out-of-distribution detection

Recent efforts for ViT-based medical imaging systems have primarily focused on improving the accuracy and generally lacking a principled mechanism to evaluate their generalization ability under different distribution/domain shifts. Recent studies have shown that test error generally increases in proportion to the distribution difference between training and test datasets, thereby making it a crucial issue to investigate in the context of ViTs. In medical imaging applications, these distribution shifts in data arise due to several factors that include: images acquired with a different device model at a different hospital, images of some unseen disease not in the training dataset, images that are incorrectly prepared, e.g., poor contrast, blurry images, etc. Extensive research exists on CNN-based out-of-distribution detection approaches in medical imaging (Yang et al., 2021b; Zhang et al., 2021a,i; Linmans et al., 2020; Hendrycks and Gimpel, 2016). Recently, few attempts have been made to show that large-scale pre-trained ViTs, due to their high-quality representations, can significantly improve the state-of-the-art on a range of out-of-distribution tasks across different data modalities (Fort et al., 2021; Koner et al., 2021; Radford et al., 2021). However, investigation in these works has been mostly carried out on toy datasets such as CIFAR-10 and CIFAR-100, therefore not necessarily reflecting out-of-distribution detection performance on medical images with complex textures and patterns, high variance in feature scale (like in X-ray images), and local specific features. This demands further research to design ViT-based medical imaging systems that should be accurate for classes seen during training while providing calibrated estimates of uncertainty for abnormalities and unseen classes. We believe that research in this direction using techniques from transfer learning and domain adaptation will be of interest to the practitioners working in medical imaging based life-critical applications to envision potential practical deployment. In Fig. 31, we highlight the performance gain of ViTs as compared to CNNs for out of distribution detection to inspire medical imaging researchers who wish to explore this area. Another possible direction is to explore the recent advancements in continual learning (Qu et al., 2021) to mitigate the issue of domains shift using ViTs. Few preliminary efforts have been made to explore this direction (Lenga et al., 2020); however, the work is still in its infancy and requires further attention from the community. Further, standardized and rigorous evaluation protocols also need to be established for domain adaptation in the medical imaging applications, similar to DomainBed (Gulrajani and Lopez-Paz, 2020) framework in the natural image domain. Such a framework will also help in advocating models reproducibility.

12. Discussion and conclusion

From the papers reviewed in this survey, it is evident that ViTs have pervaded every area of medical imaging (see Fig. 32). Here, we first briefly discuss the major advantages of transformers over CNNs that could be crucial towards a truly integrated AI clinical system. **Multimodality:** Digital health data is not limited to imaging modality only and can be obtained from other sources as well, including electronic health records, genetic repositories, patient family history, etc. In particular, recently, biobank studies have started to aggregate unprecedented scales of multimodal data on human health (UKBiobank, 2021). Healthcare professionals generally aim to rely on all these sources (modalities) of data before making any decision about the patient. Clearly, integrating these distinct yet complementary data modalities can help create a more holistic picture of disease traits. The inherent ability of transformers to process multimodal inputs helps them to work in modality-agnostic pipelines and garner the power of

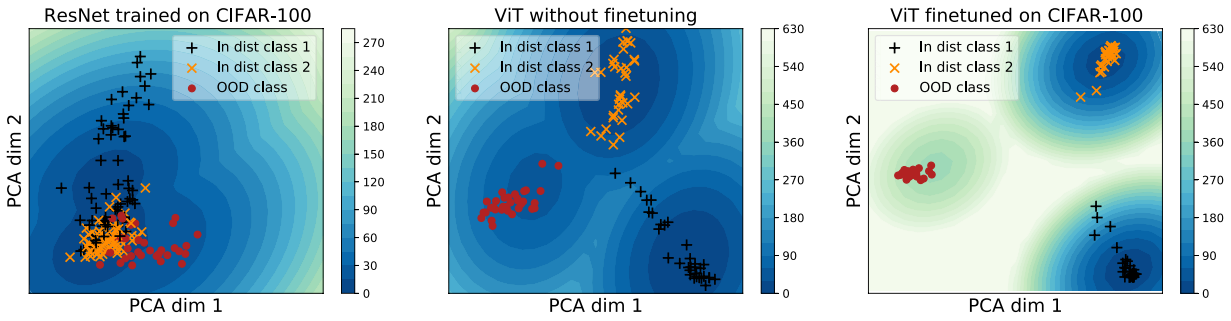

Fig. 31. A 2D PCA projection of the space of embedding vectors for three models, having two in-distribution and one out-of-distribution class. The points are projections of embeddings of the categories of in-distribution classes (yellow and black) and out-of-distribution classes (red points). The color-coding shows the Mahalanobis outlier score (Lee et al., 2018). ResNet-20 plot (left panel) leads to overlapping clusters indicating that classes are not well separated. ViT pre-trained on ImageNet-21k (middle panel) can distinguish classes from each other well but does not lead to well-separated outlier scores. ViT fine-tuned on the in-distribution dataset (right panel) is excellent at clustering embeddings based on classes and assigning high Mahalanobis distance to out-of-distribution inputs (red).
Source: Image courtesy (Fort et al., 2021).

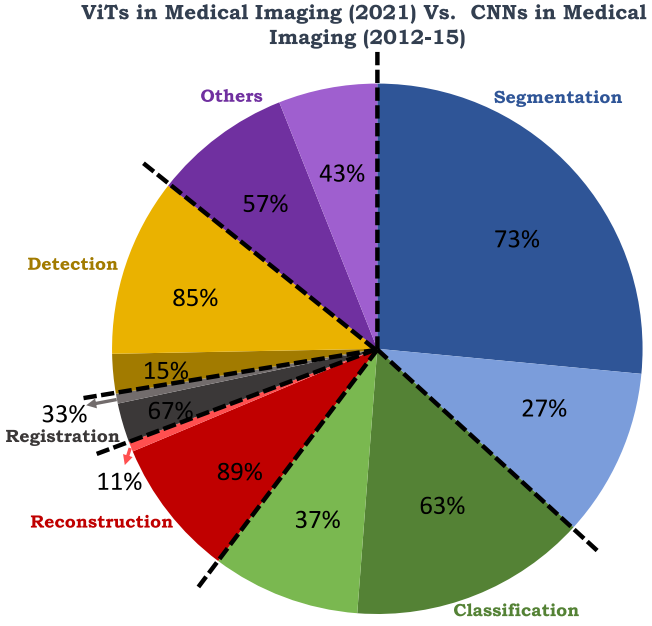

Fig. 32. ViT-based papers in medical imaging in 2021 (dark shade) vs pure CNN-based papers in medical imaging from 2012 to 2015 (light shade). It can be seen that ViTs have rapidly pervaded into almost all areas of medical imaging in a single year, with segmentation and classification being the most impacted areas. Statistics of CNNs-based papers are taken from Litjens et al. (2017).

vast amount of multimodal medical data with minimal architectural modification. **Scalability:** Despite the potential of the vast amount of multimodal data, scaling the machine learning models to incorporate several modalities can be challenging, and the training efficiency of a model is of vital importance. The training load of conventional multimodal backbones grows as the number of modalities increases since the backbone usually consists of modality-specific sub-models that need to be trained independently for each modality. Instead, Transformer models process all modalities simultaneously using a single model, which greatly reduces the training load. In addition to the *multimodality* and *scalability*, we believe a truly integrated clinical system with AI would be a system of systems that can efficiently encode multiple modalities and systematically arrange the information in short-term (quickly adaptable) and long-term (slowly adapted) knowledge bases with dedicated controller functions to modulate exchange of information while preserving the expert guidelines developed by human experts on narrow intelligence tasks.

In conclusion, we present the first comprehensive review of the applications of Transformers in medical imaging. We briefly cover the core concepts behind the success of Transformer models and then provide a comprehensive literature review of Transformers in a broad range of medical imaging tasks. Specifically, we survey the applications of Transformers in medical image segmentation, detection, classification, reconstruction, synthesis, registration, clinical report generation, and other tasks. In particular, for each of these applications, we develop taxonomy, identify application-specific challenges as well as give insights to solve them and specify recent trends. Despite their impressive performance, we anticipate there is still much exploration left to be done with Transformers in medical imaging, and we hope this survey provides a roadmap to researchers to progress this field further. We also recommend organizing the relevant workshops in top computer vision and medical imaging conferences and arranging special issues in prestigious journals to quickly disseminate the relevant research to the medical imaging community.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article

Acknowledgment

The authors would like to thank Maryam Sultana (MBZ University of Artificial Intelligence) for her help with a few figures.

Appendix A. Paper inclusion and exclusion criteria

We searched on PubMed for papers with the keyword "Transformer(s)" in their titles. We get 411 results for the year 2017 onward. We excluded papers that were not related to medical imaging. We also searched the proceedings top conferences like MICCAI, CVPR, NeurIPS, and ISBI along with the relevant workshops with the keyword "transformer(s)" in the titles of the papers and excluding the papers not related to medical imaging. ArXiv was also searched for the keyword "transformer(s)" and "medical" in the abstracts of the articles under the subject category of **Computer Vision and Pattern Recognition** and **Image and Video Processing**. For the clinical report generation section of our manuscript, we also searched the conference proceedings of ACL and EMNLP (2017 onward) with the keywords "transformer", "report", and "caption" in the title of the articles. We also search on arXiv with these keywords in the **Computer Vision and**

Pattern Recognition and **Computation and Language** categories. We excluded duplicate articles from our study.

In case of confusion, we check the abstract, conclusion, and model diagram of the paper to determine whether the paper uses Transformers for medical imaging application covered in this survey. Due to the rapid influx of papers in this area, during manuscript writing, we regularly make use of advanced search tools on Google Scholar and Google search engine by frequently checking the search results with the above keywords by setting "sort by date" on Google Scholar, and by setting "results in past 24 hours" and "results in the past week" on Google search engine. Furthermore, we checked references in all selected articles with the keywords mentioned above and also consulted with relevant colleagues to identify any articles that were missed by our initial search. Additionally, we also searched the citations of the included papers on Google Scholar to determine any missing papers.

Appendix B. Acronyms

We have described commonly used terminologies and their description in Table B.12.

Table B.12
List of frequently used acronyms in the paper.

Acronym	Expanded form
ACDC	Automated Cardiac Diagnosis Challenge
BAT	Boundary-Aware Transformers
BraTS	Brain Tumor Image Segmentation Benchmark
CE	Clinical Efficacy
CNN	Convolutional Neural Network
CT	Computed Tomography
DeTr	Detection Transformers
DL	Deep Learning
DR	Diabetic Retinopathy
DS-TransUNe	Dual Swin Transformer UNet
DNN	Deep Neural Network
DuDoTrans	Dual-Domain Transformer
FA	Fluorescein Angiography
FAT-Net	Feature Adaptive Transformer Network
FESTA	Federated Split Task-Agnostic
FL	Federated Learning
GTN	Graph Transformer Network
GT U-Net	Group Transformer U-Net
GPU	Graphical Processing Unit
KiTS	Kidney Tumor Segmentation dataset
LDCT	Low-Dose Computed Tomography
LD-PET	Low-dose Positron Emission Tomography
MCTrans	Multi-Compound Transformer
MedT	Medical Transformer
MG	Mammography
MHSA	Multi-Head Self Attention
MIA	Medical Image Analysis
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
NLG	Natural Language Generation
NLP	Natural Language Processing
PCA	Principal Component Analysis
POCFormer	Point-of-Care Transformer
PET	Positron Emission Tomography
PMTrans	Pyramid Medical Transformer
RAT-Net	Region Aware Transformer Network
SA	Self-Attention
SD-PET	Standard-dose Positron Emission Tomography
SOTA	State-of-the-Art
SpecTr	Spectral Transformer
SVD	Singular Value Decomposition
Trans-UNet	Transformer UNet
US	Ultrasound
ViT	Vision Transformer
VT-UNet	Volumetric Transformer UNet
WSI	Whole Slide Imaging

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L., 2016. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 308–318.

Ahmad, R., Bouman, C.A., Buzzard, G.T., Chan, S., Liu, S., Reehorst, E.T., Schniter, P., 2020. Plug-and-play methods for magnetic resonance imaging: Using denoisers for image recovery. IEEE Signal Process. Mag. 37 (1), 105–116.

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., Xia, L., 2020. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology 296 (2), E32–E40.

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J., 2017. Deep learning for brain MRI segmentation: state of the art and future directions. J. Digit. Imaging 30 (4), 449–459.

Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A., 2020. Dataset of breast ultrasound images. Data Brief 28, 104863.

Alam, F., Rahman, S.U., Ullah, S., Gulati, K., 2018. Medical image registration in image guided surgery: Issues, challenges and research opportunities. Biocybern. Biomed. Eng. 38 (1), 71–89.

Aldahdooh, A., Hamidouche, W., Deforges, O., 2021. Reveal of vision transformers robustness against adversarial attacks. arXiv preprint arXiv:2106.03734.

AlDahoul, N., Abdul Karim, H., Joshua Toledo Tan, M., Momo, M.A., Ledesma Fermin, J., 2021. Encoding retina image to words using ensemble of vision transformers for diabetic retinopathy grading. F1000Research 10, 948.

Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., Fahmy, A., 2021. Automated radiology report generation using conditioned transformers. Inform. Med. Unlocked 24, 100557.

Ambita, A.A.E., Boquio, E.N.V., Naval, P.C., 2021. P-GAN: Vision transformer for COVID-19 detection in CT scan imageswith self-attention GAN for DataAugmentation. In: International Conference on Artificial Neural Networks. Springer, pp. 587–598.

Amjoud, A.B., Amrouch, M., 2021. Automatic generation of chest X-ray reports using a transformer-based deep learning model. In: 2021 Fifth International Conference on Intelligent Computing in Data Sciences. ICDS, IEEE, pp. 1–5.

Andrearczyk, V., Oreiller, V., Jreige, M., Vallières, M., Castelli, J., Elhalawani, H., Boughdad, S., Prior, J.O., Depeursinge, A., 2020. Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT. In: 3D Head and Neck Tumor Segmentation in PET/CT Challenge. Springer, pp. 1–21.

Angelov, P., Almeida Soares, E., 2020. SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. medRxiv.

Anon, 2022a. Information extraction from images (IXI). http://brain-development.org/ixi-dataset/. (Accessed 20 January 2022).

Anon, 2022b. Multimodal brain tumor segmentation challenge 2020. https://www.med.upenn.edu/cbica/brats-2020/. (Accessed 20 January 2022).

Anon, 2022c. SIIM-ACR pneumothorax segmentation. https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation. (Accessed 20 January 2022).

APTOS, 2019. APTOS 2019 blindness detection: Detect diabetic retinopathy to stop blindness before it's too late. https://www.kaggle.com/c/aptos2019-blindness-detection. (Accessed 20 January 2022).

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C., 2021. Vivit: A video vision transformer. arXiv preprint arXiv:2103.15691.

Avola, D., Cinque, L., Fagioli, A., Foresti, G., Mecca, A., 2021. Ultrasound medical imaging techniques: A survey. ACM Comput. Surv. 54 (3), 1–38.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al., 2021. Big self-supervised models advance medical image classification. arXiv preprint arXiv:2101.05224.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al., 2021. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629.

Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., Zheng, H., Benson, J.A., Zhang, W., Leung, A.N., et al., 2018. A radiogenomic dataset of non-small cell lung cancer. Sci. Data 5 (1), 1–9.

Bao, H., Dong, L., Wei, F., 2021. BEiT: BERT pre-training of image transformers. arXiv preprint arXiv:2106.08254.

Beers, A., Brown, J., Chang, K., Hoebel, K., Patel, J., Ly, K.I., Tolaney, S.M., Brastianos, P., Rosen, B., Gerstner, E.R., et al., 2021. DeepNeuro: an open-source deep learning toolbox for neuroimaging. Neuroinformatics 19 (1), 127–140.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210.

Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V., 2019. Attention augmented convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3286–3295.

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 41–48.

Benz, P., Ham, S., Zhang, C., Karjauv, A., Kweon, I.S., 2021. Adversarial robustness comparison of vision transformer and MLP-mixer to CNNs. arXiv preprint arXiv:2110.02797.

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. 43, 99–111.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Med. Imaging 37 (11), 2514–2525.

Berseth, M., 2017. ISIC 2017-skin lesion analysis towards melanoma detection. arXiv preprint arXiv:1703.00523.

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A., 2021. Understanding robustness of transformers for image classification. arXiv preprint arXiv:2103.14586.

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying mmd gans. arXiv preprint arXiv:1801.01401.

Bonawitz, K., Eichner, H., Grieskamp, W., et al., 2020, TensorFlow Federated: Machine Learning on Decentralized Data. 2020, https://www.tensorflow.org/federated.

Born, J., Brändle, G., Cossio, M., Disdier, M., Goulet, J., Roulin, J., Wiedemann, N., 2020. POCOVID-Net: automatic detection of COVID-19 from a new lung ultrasound imaging dataset (POCUS). arXiv preprint arXiv:2004.12084.

BraTS, 2019a. Multimodal brain tumor segmentation challenge 2019. https://www.med.upenn.edu/cbica/brats-2019/. (Accessed 20 January 2022).

BraTS, 2019b. Multimodal brain tumor segmentation challenge 2019. https://cpb-eu-w2.wpmucdn.com/blogs.lincoln.ac.uk/dist/c/6133/files/2021/07/iccv_cov19d_leaderboard.pdf. (Accessed 20 January 2022).

Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M., 2020. Padchest: A large chest X-ray image dataset with multi-label annotated reports. Med. Image Anal. 66, 101797.

Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., et al., 2019. Nucleus segmentation across imaging experiments: the 2018 data science bowl. Nature Methods 16 (12), 1247–1253.

Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans. Med. Imaging 40 (12), 3543–3554.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294.

Caruana, R., 1997. Multitask learning. Mach. Learn. 28 (1), 41–75.

Chang, Y., Menghan, H., Guangtao, Z., Xiao-Ping, Z., 2021. TransClaw U-net: Claw U-net with transformers for medical image segmentation. arXiv preprint arXiv:2107.05188.

Chaudhari, S., Mithal, V., Polatkan, G., Ramanath, R., 2019. An attentive survey of attention models. arXiv preprint arXiv:1904.02874.

Chaudhari, S., Mithal, V., Polatkan, G., Ramanath, R., 2021. An attentive survey of attention models. ACM Trans. Intell. Syst. Technol. 12 (5), 1–32.

Chaurasia, A., Culurciello, E., 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing. VCIP, IEEE, pp. 1–4.

Chefer, H., Gur, S., Wolf, L., 2021. Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 782–791.

Chen, J., Du, Y., He, Y., Segars, W.P., Li, Y., Frey, E.C., 2021a. TransMorph: Transformer for unsupervised medical image registration. arXiv preprint arXiv:2111.10480.

Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y., 2021b. ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468.

Chen, D., Jiang, X., Hong, Y., Wen, Z., Wei, S., Peng, G., Wei, X., 2021c. Can chest CT features distinguish patients with negative from those with positive initial RT-PCR results for coronavirus disease (COVID-19)? Am. J. Roentgenol. 216 (1), 66–70.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G., 2020b. Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029.

Chen, H., Li, C., Li, X., Wang, G., Hu, W., Li, Y., Liu, W., Sun, C., Yao, Y., Teng, Y., et al., 2021d. GasHis-transformer: A multi-scale visual transformer approach for gastric histopathology image classification. arXiv preprint arXiv:2104.14528.

Chen, B., Liu, Y., Zhang, Z., Lu, G., Zhang, D., 2021e. TransAttUnet: Multi-level attention-guided U-net with transformer for medical image segmentation. arXiv preprint arXiv:2107.05274.

Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F., 2021f. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4025.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021g. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Chen, Z., Shen, Y., Song, Y., Wan, X., 2021h. Cross-modal memory networks for radiology report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5904–5914.

Chen, Z., Song, Y., Chang, T.H., Wan, X., 2020c. Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056.

Chen, X., Sun, S., Bai, N., Han, K., Liu, Q., Yao, S., Tang, H., Zhang, C., Lu, Z., Huang, Q., et al., 2021i. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiother. Oncol. 160, 175–184.

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W., 2021j. Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12299–12310.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. 54, 280–296.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al., 2020. Rethinking attention with performers. arXiv preprint arXiv:2009.14794.

Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., et al., 2020. Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 8, 132665–132676.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging 26 (6), 1045–1057.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al., 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1902.03368.

Cohen, J.P., Morrison, P., Dao, L., Roth, K., Duong, T.Q., Ghassemi, M., 2020. Covid-19 image data collection: Prospective predictions are the future. arXiv preprint arXiv:2006.11988.

Cong, R., Lei, J., Fu, H., Cheng, M.-M., Lin, W., Huang, Q., 2018. Review of visual saliency detection with comprehensive information. IEEE Trans. Circuits Syst. Video Technol. 29 (10), 2941–2959.

Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R., 2020. Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10578–10587.

Correia, G.M., Niculae, V., Martins, A.F., 2019. Adaptively sparse transformers. arXiv preprint arXiv:1909.00015.

Cuadros, J., Bresnick, G., 2009. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. J. Diabetes Sci. Technol. 3 (3), 509–516.

Dai, Y., Gao, Y., Liu, F., 2021. Transmed: Transformers advance multi-modal medical image classification. Diagnostics 11 (8), 1384.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 764–773.

Dalmaz, O., Yurt, M., Çukur, T., 2021. ResViT: Residual vision transformers for multi-modal medical image synthesis. arXiv preprint arXiv:2106.16031.

Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., et al., 2014. Feedback on a publicly distributed image database: the Messidor database. Image Anal. Stereol. 33 (3), 231–234.

Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J., 2016. Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Deng, K., Meng, Y., Gao, D., Bridge, J., Shen, Y., Lip, G., Zhao, Y., Zheng, Y., 2021. TransBridge: A lightweight transformer for left ventricle segmentation in echocardiography. In: International Workshop on Advances in Simplifying Medical Ultrasound. Springer, pp. 63–72.

Denkowski, M., Lavie, A., 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 85–91.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L., 2021. Polyp-PVT: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dou, H., Karimi, D., Rollins, C.K., Ortinau, C.M., Vasung, L., Velasco-Annis, C., Ouaalam, A., Yang, X., Ni, D., Gholipour, A., 2020. A deep attentive convolutional neural network for automatic cortical plate segmentation in fetal MRI. IEEE Trans. Med. Imaging 40 (4), 1123–1133.

Dou, Q., So, T.Y., Jiang, M., Liu, Q., Vardhanabhuti, V., Kaissis, G., Li, Z., Si, W., Lee, H.H., Yu, K., et al., 2021. Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. NPJ Digit. Med. 4 (1), 1–11.

Duncan, J.S., Insana, M.F., Ayache, N., 2019. Biomedical imaging and analysis in the age of big data and deep learning [scanning the issue]. Proc. IEEE 108 (1), 3–10.

El-Shafai, W., Abd El-Samie, F., 2020. Extensive COVID-19 X-Ray and CT chest images dataset. Mendeley data, V3.

Eslami, S., Melo, d.G., Meinel, C., 2021. Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? arXiv preprint arXiv:2112.13906.

Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L., 2020. Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 263–273.

Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., Ji, W., 2020. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology 296 (2), E115–E117.

Fedus, W., Zoph, B., Shazeer, N., 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv preprint arXiv:2101.03961.

Feng, C.M., Yan, Y., Chen, G., Fu, H., Xu, Y., Shao, L., 2021a. Accelerated multi-modal MR imaging with transformers. arXiv preprint arXiv:2106.14248.

Feng, C.M., Yan, Y., Fu, H., Chen, L., Xu, Y., 2021b. Task transformer network for joint MRI reconstruction and super-resolution. arXiv preprint arXiv:2106.06742.

Fessler, J.A., 2010. Model-based image reconstruction for MRI. IEEE Signal Process. Mag. 27 (4), 81–89.

Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S., 2019. Adversarial attacks on medical machine learning. Science 363 (6433), 1287–1289.

Fort, S., Ren, J., Lakshminarayanan, B., 2021. Exploring the limits of out-of-distribution detection. arXiv preprint arXiv:2106.03004.

Fung, G., Dundar, M., Krishnapuram, B., Rao, R.B., 2007. Multiple instance learning for computer aided diagnosis. Adv. Neural Inf. Process. Syst. 19, 425.

Gamper, J., Koohbanani, N.A., Benes, K., Graham, S., Jahanifar, M., Khurram, S.A., Azam, A., Hewitt, K., Rajpoot, N., 2020. Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:2003.10778.

Ganatra, N., 2021. A comprehensive study of applying object detection methods for medical image analysis. In: 2021 8th International Conference on Computing for Sustainable Global Development. INDIACom, IEEE, pp. 821–826.

Gao, H., Chae, O., 2010. Individual tooth segmentation from CT images using level set method with shape and intensity prior. Pattern Recognit. 43 (7), 2406–2417.

Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.-A., 2021a. Trans-SVNet: accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 593–603.

Gao, X., Qian, Y., Gao, A., 2021b. COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models. arXiv preprint arXiv:2107.01682.

Gao, Y., Zhou, M., Metaxas, D.N., 2021c. UTNet: a hybrid transformer architecture for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 61–71.

Garcia Seco De Herrera, A., Eickhof, C., Andrearczyk, V., Müller, H., 2018. Overview of the ImageCLEF 2018 caption prediction tasks. In: CEUR Workshop Proceedings.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F.A., Brendel, W., 2021. Partial success in closing the gap between human and machine vision. arXiv preprint arXiv:2106.07411.

Gheflati, B., Rivaz, H., 2021. vision transformers for classification of breast ultrasound images. arXiv preprint arXiv:2110.14731.

Ghiasi, G., Fowlkes, C.C., 2016. Laplacian pyramid reconstruction and refinement for semantic segmentation. In: European Conference on Computer Vision. Springer, pp. 519–534.

Ghoshal, B., Tucker, A., 2020. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. arXiv preprint arXiv:2003.10769.

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D.I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., et al., 2018. NiftyNet: a deep-learning platform for medical imaging. Comput. Methods Programs Biomed. 158, 113–122.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Gopal, M., Abdullah, S.E., Grady, J.J., Goodwin, J.S., 2010. Screening for lung cancer with low-dose computed tomography: a systematic review and meta-analysis of the baseline findings of randomized controlled trials. J. Thorac. Oncol. 5 (8), 1233–1239.

Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M., 2021. LeViT: a vision transformer in ConvNet's clothing for faster inference. arXiv preprint arXiv:2104.01136.

Greenspan, H., Van Ginneken, B., Summers, R.M., 2016. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Trans. Med. Imaging 35 (5), 1153–1159.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al., 2020. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733.

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J., 2019. CE-Net: Context encoder network for 2d medical image segmentation. IEEE Trans. Med. Imaging 38 (10), 2281–2292.

Gulrajani, I., Lopez-Paz, D., 2020. In search of lost domain generalization. arXiv preprint arXiv:2007.01434.

Güngör, A., Askin, B., Soydan, D.A., Saritas, E.U., Top, C.B., Çukur, T., 2021. TranSMS: Transformers for super-resolution calibration in magnetic particle imaging. arXiv preprint arXiv:2111.02163.

Gunraj, H., Sabri, A., Koff, D., Wong, A., 2021. COVID-net CT-2: Enhanced deep neural networks for detection of COVID-19 from chest CT images through bigger, more diverse learning. arXiv preprint arXiv:2101.07433.

Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M., 2021. Attention mechanisms in computer vision: A survey. arXiv preprint arXiv:2111.07624.

Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A., 2016. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397.

Hajeb Mohammad Alipour, S., Rabbani, H., Akhlaghi, M.R., 2012. Diabetic retinopathy grading by digital curvelet transform. Comput. Math. Methods Med. 2012.

Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., Zwiggelaar, R., 2018. Deep learning in mammography and breast histology, an overview and future trends. Med. Image Anal. 47, 45–67.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al., 2020. A survey on visual transformer. arXiv preprint arXiv:2012.12556.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y., 2021. Transformer in transformer. arXiv preprint arXiv:2103.00112.

Han, Y., Ye, J.C., 2018. Framing U-net via deep convolutional framelets: Application to sparse-view CT. IEEE Trans. Med. Imaging 37 (6), 1418–1429.

Haskins, G., Kruger, U., Yan, P., 2020. Deep learning in medical image registration: a survey. Mach. Vis. Appl. 31 (1), 1–18.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D., 2022a. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. arXiv preprint arXiv:2201.01266.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022b. UNETR: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584.

Hatamizadeh, A., Yang, D., Roth, H., Xu, D., 2021. UNETR: Transformers for 3d medical image segmentation. arXiv preprint arXiv:2103.10504.

He, X., Wang, S., Shi, S., Chu, X., Tang, J., Liu, X., Yan, C., Zhang, J., Ding, G., 2020. Benchmarking deep learning models and automated model design for covid-19 detection with chest CT scans. medRxiv.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hendrycks, D., Gimpel, K., 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep learning techniques for medical image segmentation: achievements and challenges. J. Digit. Imaging 32 (4), 582–596.
Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inf. Process. Syst. 30.
Hoover, A., Kouznetsova, V., Goldbaum, M., 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. Imaging 19 (3), 203–210.
Hou, B., Kaissis, G., Summers, R.M., Kainz, B., 2021. RATCHET: Medical transformer for chest X-ray diagnosis and reporting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 293–303.
Hsu, C.C., Chen, G.L., Wu, M.H., 2021. Visual transformer with statistical test for COVID-19 classification. arXiv preprint arXiv:2107.05334.
Huang, Z., Chai, H., Wang, R., Wang, H., Yang, Y., Wu, H., 2021a. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 561–570.
Huang, X., Deng, Z., Li, D., Yuan, X., 2021b. MISSFormer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162.
Huang, Y., Gupta, S., Song, Z., Li, K., Arora, S., 2021c. Evaluating gradient inversion attacks and defenses in federated learning. Adv. Neural Inf. Process. Syst. 34.
Huang, S., Li, J., Xiao, Y., Shen, N., Xu, T., 2022. RTNet: relation transformer network for diabetic retinopathy multi-lesion segmentation. IEEE Trans. Med. Imaging 41 (6), 1596–1607.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017a. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.
Huang, Q., Luo, Y., Zhang, Q., 2017b. Breast ultrasound image segmentation: a survey. Int. J. Comput. Assist. Radiol. Surg. 12 (3), 493–507.
Huang, J.H., Yang, C.H.H., Liu, F., Tian, M., Liu, Y.C., Wu, T.W., Lin, I., Wang, K., Morikawa, H., Chang, H., et al., 2021d. DeepOpht: medical report generation for retinal images via deep models and visual explanation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2442–2452.
Hyun, C.M., Kim, H.P., Lee, S.M., Lee, S., Seo, J.K., 2018. Deep learning for undersampled MRI reconstruction. Phys. Med. Biol. 63 (13), 135007.
Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01. pp. 590–597.
Irwin, F., et al., 1968. An isotropic 3x3 image gradient operator. Present. Stanf. AI Proj. 2014 (02).
Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. nnU-Net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486.
ISIC, 2019. Skin lesion analysis towards melanoma detection, 2019. https://challenge2018.isic-archive.com/. (Accessed 20 January 2022).
Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.
Jabareen, N., Lukassen, S., 2022. Segmenting brain tumors in multi-modal MRI scans using a 3D SegNet architecture. In: International MICCAI Brainlesion Workshop. Springer, pp. 377–388.
Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G., 2014. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant. Imaging Med. Surg. 4 (6), 475.
Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D., 2020. Kvasir-SEG: A segmented polyp dataset. In: International Conference on Multimedia Modeling. Springer, pp. 451–462.
Ji, Y., Zhang, R., Wang, H., Li, Z., Wu, L., Zhang, S., Luo, P., 2021. Multi-compound transformer for accurate biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 326–336.
Jia, H., Bai, C., Cai, W., Huang, H., Xia, Y., 2022. HNF-Netv2 for brain tumor segmentation using multi-modal MR imaging. arXiv preprint arXiv:2202.05268.
Jia, Q., Shu, H., 2021. BiTr-Unet: a CNN-transformer combined network for MRI brain tumor segmentation. arXiv preprint arXiv:2109.12271.
Jiang, Z., Dong, Z., Wang, L., Jiang, W., 2021. Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model. Comput. Intell. Neurosci. 2021.
Jiang, J., Lin, S., 2021. COVID-19 detection in chest X-ray images using swin-transformer and transformer in transformer. arXiv preprint arXiv:2110.08427.
Jin, K.H., McCann, M.T., Froustey, E., Unser, M., 2017. Deep convolutional neural network for inverse problems in imaging. IEEE Trans. Image Process. 26 (9), 4509–4522.
Jin, Q., Meng, Z., Sun, C., Cui, H., Su, R., 2020. RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. Front. Bioeng. Biotechnol. 8, 1471.
Jing, B., Xie, P., Xing, E., 2017. On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195.
Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S., 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042.
Joshi, A., Jagatap, G., Hegde, C., 2021. Adversarial token attacks on vision transformers. arXiv preprint arXiv:2110.04337.
Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima, I., Mancuso, J., Jungmann, F., Steinborn, M.M., et al., 2021. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat. Mach. Intell. 3 (6), 473–484.
Kälviäinen, R., Uusitalo, H., 2007. DIARETDB1 diabetic retinopathy database and evaluation protocol. In: Medical Image Understanding and Analysis, Vol. 2007. Citeseer, p. 61.
Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N., 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790.
Kamran, S.A., Hossain, K.F., Tavakkoli, A., Zuckerbrod, S.L., Sanders, K.M., Baker, S.A., 2021. VTGAN: Semi-supervised retinal image synthesis and disease prediction using vision transformers. arXiv preprint arXiv:2104.06757.
Karimi, D., Vasylechko, S., Gholipour, A., 2021. Convolution-free medical image segmentation using transformers. arXiv preprint arXiv:2102.13645.
Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F., 2020. Transformers are RNNs: Fast autoregressive transformers with linear attention. In: International Conference on Machine Learning. PMLR, pp. 5156–5165.
Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J., 2017. Towards proving the adversarial robustness of deep neural networks. arXiv preprint arXiv:1709.02802.
Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A., 2018. Left-ventricle quantification using residual U-Net. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 371–380.
Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172 (5), 1122–1131.
Khan, A., Lee, B., 2021. Gene transformer: Transformers for the gene expression-based classification of lung cancer subtypes. arXiv preprint arXiv:2108.11833.
Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2021. Transformers in vision: A survey. arXiv preprint arXiv:2101.01169.
Kim, Y.J., Jang, H., Lee, K., Park, S., Min, S.G., Hong, C., Park, J.H., Lee, K., Kim, J., Hong, W., et al., 2021. PAIP 2019: Liver cancer segmentation challenge. Med. Image Anal. 67, 101854.
Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
Kipkogei, E., GA, A.A., Kagiampakis, I., Patra, A., Jacob, E., 2021. Explainable transformer-based neural network for the prediction of survival outcomes in non-small cell lung cancer (NSCLC).
KiTS, 2021. The 2021 kidney and kidney tumor segmentation challenge. (Accessed 20 January 2022).
Knopp, T., Szwargulski, P., Griese, F., Gräser, M., 2020. OpenMPIData: An initiative for freely accessible magnetic particle imaging data. Data Brief 28, 104971.
Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., Ibrahim, M., van der Maaten, L., 2021. Crypten: Secure multi-party computation meets machine learning. Adv. Neural Inf. Process. Syst. 34.
Kollias, D., Arsenos, A., Soukissian, L., Kollias, S., 2021. MIA-COV19D: COVID-19 detection through 3-D chest CT image analysis. arXiv preprint arXiv:2106.07524.
Koner, R., Sinhamahapatra, P., Roscher, K., Günnemann, S., Tresp, V., 2021. Oodformer: Out-of-distribution detection transformer. arXiv preprint arXiv:2107.08976.
Korkmaz, Y., Dar, S.U., Yurt, M., Özbey, M., Çukur, T., 2021a. Unsupervised MRI reconstruction via zero-shot learned adversarial transformers. arXiv preprint arXiv:2105.08059.
Korkmaz, Y., Yurt, M., Dar, S.U.H., Özbey, M., Cukur, T., 2021b. Deep MRI reconstruction with generative vision transformers. In: International Workshop on Machine Learning for Medical Image Reconstruction. Springer, pp. 54–64.
Kotowski, K., Adamski, S., Machura, B., Zarudzki, L., Nalepa, J., 2022. Coupling nnU-Nets with expert knowledge for accurate brain tumor segmentation from MRI. In: International MICCAI Brainlesion Workshop. Springer, pp. 197–209.
Kougia, V., Pavlopoulos, J., Androutsopoulos, I., 2019. A survey on biomedical image captioning. arXiv preprint arXiv:1905.13302.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.
Kudo, H., Suzuki, T., Rashed, E.A., 2013. Image reconstruction for sparse-view CT and interior CT—introduction to compressed sensing and differentiated backprojection. Quant. Imaging Med. Surg. 3 (3), 147.
Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.A., Li, J., Hu, Z., et al., 2019. A multi-organ nucleus segmentation challenge. IEEE Trans. Med. Imaging 39 (5), 1380–1391.
Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A., 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging 36 (7), 1550–1560.
Kumar, M., Weissenborn, D., Kalchbrenner, N., 2021. Colorization transformer. arXiv preprint arXiv:2102.04432.

Kwee, T.C., Kwee, R.M., 2020. Chest CT in COVID-19: what the radiologist needs to know. RadioGraphics 40 (7), 1848–1865.
Lakhani, P., Sundaram, B., 2017. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 284 (2), 574–582.
Lam, L., Suen, S., 1997. Application of majority voting to pattern recognition: an analysis of its behavior and performance. IEEE Trans. Syst. Man Cybern. A 27 (5), 553–568.
Lambert, Z., Petitjean, C., Dubray, B., Kuan, S., 2020. SegTHOR: Segmentation of thoracic organs at risk in CT images. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications. IPTA, IEEE, pp. 1–6.
Landman, B., Xu, Z., Igelsias, J.E., Styner, M., Langerak, T., Klein, A., 2015. MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI: Multi-Atlas Labeling beyond Cranial Vault-Workshop Challenge.
LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1 (4), 541–551.
Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L., 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci. Data 4 (1), 1–9.
Lee, K., Lee, K., Lee, H., Shin, J., 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Adv. Neural Inf. Process. Syst. 31.
Lei, Y., Fu, Y., Wang, T., Qiu, R.L., Curran, W.J., Liu, T., Yang, X., 2020. Deep learning in multi-organ segmentation. arXiv preprint arXiv:2001.10619.
Lell, M.M., Kachelrieß, M., 2020. Recent and upcoming technological developments in computed tomography: high speed, low dose, deep learning, multienergy. Invest. Radiol. 55 (1), 8–19.
Lenga, M., Schulz, H., Saalbach, A., 2020. Continual learning for domain adaptation in chest X-ray classification. In: Medical Imaging with Deep Learning. PMLR, pp. 413–423.
Li, Y., Cai, W., Gao, Y., Hu, X., 2021a. More than encoder: Introducing transformer decoder to upsample. arXiv preprint arXiv:2106.10637.
Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., et al., 2021b. FFA-IR: Towards an explainable and reliable medical report generation benchmark. In: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).
Li, C.Y., Liang, X., Hu, Z., Xing, E.P., 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01. pp. 6666–6673.
Li, B., Pandey, S., Fang, H., Lyv, Y., Li, J., Chen, J., Xie, M., Wan, L., Liu, H., Ding, C., 2020a. FTRANS: energy-efficient acceleration of transformers using FPGA. In: Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design. pp. 175–180.
Li, S., Sui, X., Fu, J., Fu, H., Luo, X., Feng, Y., Xu, X., Liu, Y., Ting, D.S., Goh, R.S.M., 2021c. Few-shot domain adaptation with polymorphic transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 330–340.
Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., Goh, R.S.M., 2021d. Medical image segmentation using squeeze-and-expansion transformers. arXiv preprint arXiv:2105.09511.
Li, M., Wang, F., Chang, X., Liang, X., 2020b. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. arXiv preprint arXiv:2006.03744.
Li, Y., Wang, S., Wang, J., Zeng, G., Liu, W., Zhang, Q., Jin, Q., Wang, Y., 2021e. GT U-net: A U-net like group transformer network for tooth root segmentation. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 386–395.
Li, H., Xiong, P., An, J., Wang, L., 2018. Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180.
Li, H., Yang, F., Zhao, Y., Xing, X., Zhang, J., Gao, M., Huang, J., Wang, L., Yao, J., 2021f. DT-MIL: Deformable transformer for multi-instance learning on histopathological image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 206–216.
Li, Y., Zeng, G., Zhang, Y., Wang, J., Jin, Q., Sun, L., Zhang, Q., Lian, Q., Qian, G., Xia, N., et al., 2021g. AGMB-transformer: Anatomy-guided multi-branch transformer network for automated evaluation of root canal therapy. IEEE J. Biomed. Health Inf.
Li, Y., Zhang, Y., Liu, J.Y., Wang, K., Zhang, K., Zhang, G.S., Liao, X.F., Yang, G., 2022. Global transformer and dual local attention network via deep-shallow hierarchical feature fusion for retinal vessel segmentation. IEEE Trans. Cybern.
Liang, T., Jin, Y., Li, Y., Wang, T., 2020. EDCNN: Edge enhancement-based densely connected network with compound loss for low-dose CT denoising. In: 2020 15th IEEE International Conference on Signal Processing, Vol. 1. ICSP, IEEE, pp. 193–198.
Liao, F., Liang, M., Li, Z., Hu, X., Song, S., 2019. Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network. IEEE Trans. Neural Netw. Learn. Syst. 30 (11), 3484–3495.
Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., 2021a. DS-TransUNet: Dual swin transformer U-net for medical image segmentation. arXiv preprint arXiv:2106.06716.
Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.
Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.
Lin, K., Heckel, R., 2021. Vision transformers enable fast and robust accelerated MRI.
Lin, W.A., Liao, H., Peng, C., Sun, X., Zhang, J., Luo, J., Chellappa, R., Zhou, S.K., 2019. Dudonet: Dual domain network for ct metal artifact reduction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10512–10521.
Lin, T., Wang, Y., Liu, X., Qiu, X., 2021b. A survey of transformers. arXiv preprint arXiv:2106.04554.
Linmans, J., van der Laak, J., Litjens, G., 2020. Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. In: MIDL. pp. 465–478.
Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.
Liu, F., Ge, S., Wu, X., 2022a. Competence-based multimodal curriculum learning for medical report generation. arXiv preprint arXiv:2206.14579.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021a. Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV.
Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022b. A ConvNet for the 2020s. arXiv preprint arXiv:2201.03545.
Liu, C., Salzmann, M., Lin, T., Tomioka, R., Süsstrunk, S., 2020. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. arXiv preprint arXiv:2006.08403.
Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D., Wang, T., 2019. Deep learning in medical ultrasound analysis: a review. Engineering 5 (2), 261–275.
Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y., 2021b. Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13753–13762.
Liu, Y., Yang, Y., Jiang, W., Wang, T., Lei, B., 2021c. 3D deep attentive U-net with transformer for breast tumor segmentation from automated breast volume scanner. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society. EMBC, IEEE, pp. 4011–4014.
Liu, C., Yin, Q., 2021. Automatic diagnosis of COVID-19 using a tailored transformer-like network. In: Journal of Physics: Conference Series, Vol. 2010, No. 01. IOP Publishing, 012175.
Liu, F., You, C., Wu, X., Ge, S., Sun, X., et al., 2021d. Auto-encoding knowledge graph for unsupervised medical report generation. Adv. Neural Inf. Process. Syst. 34.
Liu, S., Zhou, H., Shi, X., Pan, J., 2021e. Transformer for polyp detection. arXiv preprint arXiv:2111.07918.
Long, Y., Li, Z., Yee, C.H., Ng, C.F., Taylor, R.H., Unberath, M., Dou, Q., 2021. E-DSSR: Efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 415–425.
Lovelace, J., Mortazavi, B., 2020. Learning to generate clinically coherent chest X-ray reports. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. pp. 1235–1243.
Lu, M., Pan, Y., Nie, D., Liu, F., Shi, F., Xia, Y., Shen, D., 2021. SMILE: Sparse-attention based multiple instance contrastive learning for glioma sub-type classification using pathological images. In: MICCAI Workshop on Computational Pathology. PMLR, pp. 159–169.
LUNA, 2016. Lung nodule analysis, 2016. https://luna16.grand-challenge.org/Data/. (Accessed 20 January 2022).
Lundervold, A.S., Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Z. Med. Phys. 29 (2), 102–127.
Luo, Y., Wang, Y., Zu, C., Zhan, B., Wu, X., Zhou, J., Shen, D., Zhou, L., 2021. 3D transformer-GAN for high-quality PET reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 276–285.
Luthra, A., Sulakhe, H., Mittal, T., Iyer, A., Yadav, S., 2021. Eformer: Edge enhancement based transformer for medical image denoising. arXiv preprint arXiv:2109.08044.
Luu, H.M., Park, S.H., 2021. Extending nn-UNet for brain tumor segmentation. arXiv preprint arXiv:2112.04653.
Ma, X., Luo, G., Wang, W., Wang, K., 2021a. Transformer network for significant stenosis detection in CCTA of coronary arteries. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 516–525.
Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F., 2021b. Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognit. 110, 107332.
Maaz, M., Rasheed, H., Khan, S., Khan, F.S., Anwer, R.M., Yang, M.H., 2021. Multi-modal transformers excel at class-agnostic object detection. arXiv preprint arXiv:2111.11430.
Mahapatra, D., Ge, Z., 2021. MR image super resolution by combining feature disentanglement CNNs and vision transformers.
Mahmood, K., Mahmood, R., Van Dijk, M., 2021. On the robustness of vision transformers to adversarial examples. arXiv preprint arXiv:2104.02610.
Maji, D., Sigedar, P., Singh, M., 2022. Attention Res-UNet with guided decoder for semantic segmentation of brain tumors. Biomed. Signal Process. Control 71, 103077.

Makropoulos, A., Robinson, E.C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S.J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J., et al., 2018. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage 173, 88–112.

Malkiel, I., Rosenman, G., Wolf, L., Hendler, T., 2021. Pre-training and fine-tuning transformers for fMRI prediction tasks. arXiv preprint arXiv:2112.05761.

Mao, X., Qi, G., Chen, Y., Li, X., Ye, S., He, Y., Xue, H., 2021. Rethinking the design principles of robust vision transformer. arXiv preprint arXiv:2105.07926.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19 (9), 1498–1507.

Mathai, T.S., Lee, S., Elton, D.C., Shen, T.C., Peng, Y., Lu, Z., Summers, R.M., 2021. Lymph node detection in T2 MRI with transformers. arXiv preprint arXiv:2111.04885.

Matsoukas, C., Haslum, J.F., Söderberg, M., Smith, K., 2021. Is it time to replace CNNs with transformers for medical images? arXiv preprint arXiv:2108.09038.

McCollough, C.H., Bartley, A.C., Carter, R.E., Chen, B., Drees, T.A., Edwards, P., Holmes, III, D.R., Huang, A.E., Khan, F., Leng, S., et al., 2017. Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge. Med. Phys. 44 (10), e339–e352.

Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J., 2013. PH 2-A dermoscopic image database for research and benchmarking. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, IEEE, pp. 5437–5440.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34 (10), 1993–2024.

Messina, P., Pino, P., Parra, D., Soto, A., Besa, C., Uribe, S., Tejos, C., Prieto, C., Capurro, D., et al., 2020. A survey on deep learning and explainability for automatic image-based medical report generation. arXiv preprint arXiv:2010.10563.

Milletari, F., Navab, N., Ahmadi, S., 2018. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. June 2016.

Min, J.K., Kwak, M.S., Cha, J.M., 2019. Overview of deep learning in gastrointestinal endoscopy. Gut Liver 13 (4), 388.

Miura, Y., Zhang, Y., Tsai, E.B., Langlotz, C.P., Jurafsky, D., 2020. Improving factual completeness and consistency of image-to-text radiology report generation. arXiv preprint arXiv:2010.10042.

Mondal, A.K., Bhattacharjee, A., Singla, P., AP, P., 2021. xViTCOS: Explainable vision transformer based COVID-19 screening using radiography. TechRxiv.

Monga, V., Li, Y., Eldar, Y.C., 2021. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. IEEE Signal Process. Mag. 38 (2), 18–44.

Monshi, M.M.A., Poon, J., Chung, V., 2020. Deep learning in generating radiology reports: A survey. Artif. Intell. Med. 106, 101878.

Napel, S., Plevritis, S.K., 2014. NSCLC radiogenomics: initial stanford study of 26 cases. Cancer Imaging Arch.

Narnhofer, D., Hammernik, K., Knoll, F., Pock, T., 2019. Inverse GANs for accelerated MRI reconstruction. In: Wavelets and Sparsity XVIII, Vol. 11138. International Society for Optics and Photonics, p. 111381A.

Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., 2021a. Intriguing properties of vision transformers. arXiv preprint arXiv:2105.10497.

Naseer, M., Ranasinghe, K., Khan, S., Khan, F.S., Porikli, F., 2021b. On improving adversarial transferability of vision transformers. arXiv preprint arXiv:2106.04169.

Nguyen, C., Asad, Z., Huo, Y., 2021a. Evaluating transformer based semantic segmentation networks for pathological image segmentation. arXiv preprint arXiv:2108.11993.

Nguyen, H.T., Nie, D., Badamdorj, T., Liu, Y., Zhu, Y., Truong, J., Cheng, L., 2021b. Automated generation of accurate\& fluent medical X-ray reports. arXiv preprint arXiv:2108.12126.

NLST, 2017. National lung screening trial - the cancer data assess system. https://cdas.cancer.gov/nlst/. (Accessed 20 January 2022).

Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M., 2021. Progressive transformer-based generation of radiology reports. arXiv preprint arXiv:2102.09777.

Nyholm, T., Svensson, S., Andersson, S., Jonsson, J., Sohlin, M., Gustafsson, C., Kjellén, E., Söderström, K., Albertsson, P., Blomqvist, L., et al., 2018. MR and CT data with multiobserver delineations of organs in the pelvic area—Part of the gold atlas project. Med. Phys. 45 (3), 1295–1300.

OASIS, 2019. Open access series of imaging studies. https://www.oasis-brains.org/. (Accessed 20 January 2022).

OIA, 2019. Ophthalmic image analysis dataset. https://github.com/nkicsl/OIA. (Accessed 20 January 2022).

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Ongie, G., Jalal, A., Metzler, C.A., Baraniuk, R.G., Dimakis, A.G., Willett, R., 2020. Deep learning techniques for inverse problems in imaging. IEEE J. Sel. Areas Inf. Theory 1 (1), 39–56.

OpenMined, 2017. To lower the barrier to entry to privacy preserving technology. https://github.com/OpenMined. (Accessed 20 January 2022).

Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al., 2020. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. 59, 101570.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al., 2020. Video-based AI for beat-to-beat assessment of cardiac function. Nature 580 (7802), 252–256.

Pahwa, E., Mehta, D., Kapadia, S., Jain, D., Luthra, A., 2021. MedSkip: Medical report generation using skip connections and integrated attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3409–3415.

Pan, J., Wu, W., Gao, Z., Zhang, H., 2021, Multi-Domain Integrative Swin Transformer Network for Sparse-View Tomographic Reconstruction, Available at SSRN 3991087.

Papangelou, K., Sechidis, K., Weatherall, J., Brown, G., 2018. Toward an understanding of adversarial examples in clinical trials. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 35–51.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318.

Park, S., Kim, G., Kim, J., Kim, B., Ye, J.C., 2021a. Federated split vision transformer for COVID-19CXR diagnosis using task-agnostic training. arXiv preprint arXiv:2111.01338.

Park, S., Kim, G., Oh, Y., Seo, J.B., Lee, S.M., Kim, J.H., Moon, S., Lim, J.K., Ye, J.C., 2021b. Vision transformer for COVID-19 CXR diagnosis using chest X-ray feature corpus. arXiv preprint arXiv:2103.07055.

Park, H., Kim, K., Park, S., Choi, J., 2021c. Medical image captioning model to convey more details: Methodological comparison of feature difference generation. IEEE Access 9, 150560–150568.

Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A.A., Zhang, R., 2020. Swapping autoencoder for deep image manipulation. arXiv preprint arXiv:2007.00653.

Paul, S., Chen, P.Y., 2021. Vision transformers are robust learners. arXiv preprint arXiv:2105.07581.

Pavlopoulos, J., Kougia, V., Androutsopoulos, I., Papamichail, D., 2021. Diagnostic captioning: a survey. arXiv preprint arXiv:2101.07299.

Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M., 2021. A volumetric transformer for accurate 3D tumor segmentation. arXiv preprint arXiv:2111.13300.

Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M., 2018. Radiology objects in COntext (ROCO): a multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, pp. 180–189.

Perera, S., Adhikari, S., Yilmaz, A., 2021. POCFormer: A lightweight transformer architecture for detection of COVID-19 using point of care ultrasound. arXiv preprint arXiv:2105.09913.

Pérez-García, F., Sparks, R., Ourselin, S., 2021. TorchIO: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Programs Biomed. 106236.

Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., Soler, L., 2021. U-net transformer: self and cross attention for medical image segmentation. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 267–276.

Philippi, D., Rothaus, K., Castelli, M., 2023. A vision transformer architecture for the automated segmentation of retinal lesions in spectral domain optical coherence tomography images. Sci. Rep. 13 (1), 517.

Poggi, M., Kim, S., Tosi, F., Kim, S., Aleotti, F., Min, D., Sohn, K., Mattoccia, S., 2021. On the confidence of stereo matching in a deep-learning era: a quantitative evaluation. arXiv preprint arXiv:2101.00431.

Portelance, E., Frank, M.C., Jurafsky, D., Sordoni, A., Laroche, R., 2021. The emergence of the shape bias results from communicative efficiency. arXiv preprint arXiv:2109.06232.

Prangemeier, T., Reich, C., Koeppl, H., 2020. Attention-based transformers for instance segmentation of cells in microstructures. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 700–707.

Qayyum, A., Ilahi, I., Shamshad, F., Boussaid, F., Bennamoun, M., Qadir, J., 2021. Untrained neural network priors for inverse imaging problems: A survey. TechRxiv.

Qi, X., Brown, L.G., Foran, D.J., Nosher, J., Hacihaliloglu, I., 2021. Chest X-ray image phase features for improved diagnosis of COVID-19 using convolutional neural network. Int. J. Comput. Assis. Radiol. Surg. 16 (2), 197–206.

Qu, H., Rahmani, H., Xu, L., Williams, B., Liu, J., 2021. Recent advances of continual learning in computer vision: An overview. arXiv preprint arXiv:2109.11369.

Quellec, G., Lamard, M., Conze, P.H., Massin, P., Cochener, B., 2020. Automatic detection of rare pathologies in fundus photographs using few-shot learning. Med. Image Anal. 61, 101660.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI Blog 1 (8), 9.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J., 2019. Stand-alone self-attention in vision models. arXiv preprint arXiv:1906.05909.

Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J., 2021. DynamicViT: Efficient vision transformers with dynamic token sparsification. arXiv preprint arXiv:2106.02034.

Reader, A.J., Corda, G., Mehranian, A., da Costa-Luis, C., Ellis, S., Schnabel, J.A., 2020. Deep learning for PET image reconstruction. IEEE Trans. Radiat. Plasma Med. Sci. 5 (1), 1–25.

Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V., 2017. Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7008–7024.

Retinopathy, 2015. Diabetic retinopathy challenge. https://www.kaggle.com/c/diabetic-retinopathy-detection. (Accessed 20 January 2022).

Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.M., Tengg-Kobligk, H.v., Summers, R.M., Wiest, R., 2020. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiol. Artif. Intell. 2 (3), e190043.

Reynaud, H., Vlontzos, A., Hou, B., Beqiri, A., Leeson, P., Kainz, B., 2021. Ultrasound video transformers for cardiac ejection fraction estimation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 495–505.

Ristea, N.C., Miron, A.I., Savencu, O., Georgescu, M.I., Verga, N., Khan, F.S., Ionescu, R.T., 2021. CyTran: Cycle-consistent transformers for non-contrast to contrast CT translation. arXiv preprint arXiv:2110.06400.

Rojas-Muñoz, E., Couperus, K., Wachs, J., 2020. DAISI: Database for AI surgical instruction. arXiv preprint arXiv:2004.02809.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Rouge, L.C., 2004. A package for automatic evaluation of summaries. In: Proceedings of Workshop on Text Summarization of ACL. Spain.

RSNA, 2018. RSNA pneumonia detection challenge (2018). https://www.kaggle.com/c/rsna-pneumonia-detection-challenge. (Accessed 20 January 2022).

Ruggeri, A., Scarpa, F., De Luca, M., Meltendorf, C., Schroeter, J., 2010. A system for the automatic estimation of morphometric parameters of corneal endothelium in alizarine red-stained images. Br. J. Ophthalmol. 94 (5), 643–647.

Sadate, A., Occean, B.V., Beregi, J.P., Hamard, A., Addala, T., de Forges, H., Fabbro-Peray, P., Frandon, J., 2020. Systematic review and meta-analysis on the impact of lung cancer screening by low-dose computed tomography. Eur. J. Cancer 134, 107–114.

Sait, U., Lal, K., Prajapati, S., Bhaumik, R., Kumar, T., Sanjana, S., Bhalla, K., 2020. Curated dataset for COVID-19 posterior-anterior chest radiography images (X-Rays), 1. Mendeley Data.

Samstein, R.M., Lee, C.H., Shoushtari, A.N., Hellmann, M.D., Shen, R., Janjigian, Y.Y., Barron, D.A., Zehir, A., Jordan, E.J., Omuro, A., et al., 2019. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nature Genet. 51 (2), 202–206.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Med. Image Anal. 53, 197–207.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y., 2021a. TransMIL: Transformer based correlated multiple instance learning for whole slide image classication. arXiv preprint arXiv:2106.00908.

Shao, R., Shi, Z., Yi, J., Chen, P.-Y., Hsieh, C.J., 2021b. On the adversarial robustness of visual transformers. arXiv preprint arXiv:2103.15670.

Shazeer, N., Lan, Z., Cheng, Y., Ding, N., Hou, L., 2020. Talking-heads attention. arXiv preprint arXiv:2003.02436.

Sheikh, H.R., Bovik, A.C., 2006. Image information and visual quality. IEEE Trans. Image Process. 15 (2), 430–444.

Shen, Z., Lin, C., Zheng, S., 2021a. COTR: Convolution in transformer network for end to end polyp detection. arXiv preprint arXiv:2105.10925.

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248.

Shen, Z., Yang, H., Zhang, Z., Zheng, S., 2021b. Automated kidney tumor segmentation with convolution and transformer network.

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., Kodera, Y., Doi, K., 2000. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am. J. Roentgenol. 174 (1), 71–74.

Shlezinger, N., Whang, J., Eldar, Y.C., Dimakis, A.G., 2020. Model-based deep learning. arXiv preprint arXiv:2012.08405.

Shome, D., Kar, T., Mohanty, S.N., Tiwari, P., Muhammad, K., AlTameem, A., Zhang, Y., Saudagar, A.K.J., 2021. COVID-transformer: Interpretable COVID-19 detection using vision transformer for healthcare. Int. J. Environ. Res. Public Health 18 (21), 11086.

Siddiquee, M.M.R., Myronenko, A., 2021. Redundancy reduction in semantic segmentation of 3D brain tumor MRIs. arXiv preprint arXiv:2111.00742.

Silva, J., Histace, A., Romain, O., Dray, X., Granado, B., 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int. J. Comput. Assis. Radiol. Surg. 9 (2), 283–293.

Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063.

Singh, A., Sengupta, S., Lakshminarayanan, V., 2020. Explainable deep learning models in medical image analysis. J. Imaging 6 (6), 52.

Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al., 2017. Gland segmentation in colon histology images: The glas challenge contest. Med. Image Anal. 35, 489–502.

Sobirov, I., Nazarov, O., Alasmawi, H., Yaqub, M., 2022. Automatic segmentation of head and neck tumor: How powerful transformers are? arXiv preprint arXiv:2201.06251.

Srinivasan, P., Thapar, D., Bhavsar, A., Nigam, A., 2020. Hierarchical X-Ray report generation via pathology tags and multi head attention. In: Proceedings of the Asian Conference on Computer Vision.

Sriram, A., Zbontar, J., Murrell, T., Defazio, A., Zitnick, C.L., Yakubova, N., Knoll, F., Johnson, P., 2020. End-to-end variational networks for accelerated MRI reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 64–73.

Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation. arXiv preprint arXiv:2105.05633.

Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., Zhang, Y., 2021. Lesion-aware transformers for diabetic retinopathy grading. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10938–10947.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5703.

Synapse, 2015. Synapse multi-organ segmentation dataset. https://www.synapse.org/#!Synapse:syn3193805/wiki/217789. (Accessed 20 January 2022).

Tabani, H., Balasubramaniam, A., Marzban, S., Arani, E., Zonooz, B., 2021. Improving the efficiency of transformers for resource-constrained devices. In: 2021 24th Euromicro Conference on Digital System Design. DSD, IEEE, pp. 449–456.

Tajbakhsh, N., Gurudu, S.R., Liang, J., 2015. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging 35 (2), 630–644.

Tang, Y.B., Tang, Y.X., Xiao, J., Summers, R.M., 2019. Xlsor: A robust and accurate lung segmentor on chest X-rays using criss-cross attention and customized radiorealistic abnormalities generation. In: International Conference on Medical Imaging with Deep Learning. PMLR, pp. 457–467.

Tang, Y., Yang, D., Li, W., Roth, H., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2021. Self-supervised pre-training of swin transformers for 3D medical image analysis. arXiv preprint arXiv:2111.14791.

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. Self-supervised pre-training of swin transformers for 3D medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740.

Tao, R., Zheng, G., 2021. Spine-transformers: Vertebra detection and localization in arbitrary field-of-view spine CT with transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 93–103.

Tay, Y., Dehghani, M., Bahri, D., Metzler, D., 2020. Efficient transformers: A survey. arXiv preprint arXiv:2009.06732.

TCGA, 2013. The cancer genome atlas program. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. (Accessed 20 January 2022).

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR, pp. 10347–10357.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5 (1), 1–9.

Tulder, G.v., Tong, Y., Marchiori, E., 2021. Multi-view analysis of unregistered medical images using cross-view transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 104–113.

Tuli, S., Dasgupta, I., Grant, E., Griffiths, T.L., 2021. Are convolutional neural networks or transformers more like human vision? arXiv preprint arXiv:2105.07197.

UKBiobank, 2021. The world's largest multi-modal imaging study resumed. https://www.ukbiobank.ac.uk/explore-your-participation/contribute-further/imaging-study. (Accessed 2022-08-20).

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2018. Deep image prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9446–9454.

Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M., 2021. Medical transformer: Gated axial-attention for medical image segmentation. arXiv preprint arXiv:2102.10662.

Van den Bogerd, B., Zakaria, N., Adam, B., Matthyssen, S., Koppen, C., Dhubhghaill, S.N., 2019. Corneal endothelial cells over the past decade: are we missing the mark (er)? Transl. Vis. Sci. Technol. 8 (6), 13.

Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B., Shlens, J., 2021. Scaling local self-attention for parameter efficient visual backbones. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12894–12904.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008.

Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A., 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthc. Eng. 2017.

Vedantam, R., Lawrence Zitnick, C., Parikh, D., 2015. Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4566–4575.

Vepakomma, P., Gupta, O., Swedish, T., Raskar, R., 2018. Split learning for health: Distributed deep learning without sharing raw patient data. arXiv preprint arXiv:1812.00564.

Vu, Y.N.T., Wang, R., Balachandar, N., Liu, C., Ng, A.Y., Rajpurkar, P., 2021. MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. arXiv preprint arXiv:2102.10663.

Wang, H., Cao, P., Wang, J., Zaiane, O.R., 2022a. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 3. pp. 2441–2449.

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J., 2021a. Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 109–119.

Wang, P., Cuccolo, N.G., Tyagi, R., Hacihaliloglu, I., Patel, V.M., 2018a. Automatic real-time CNN-based neonatal brain ventricles segmentation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging. ISBI 2018, IEEE, pp. 716–719.

Wang, X., Girshick, R., Gupta, A., He, K., 2018b. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803.

Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H., 2020a. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768.

Wang, Y., Lin, Z., Tian, J., Shi, Z., Zhang, Y., Fan, J., He, Z., 2021b. Confidence-guided radiology report generation. arXiv preprint arXiv:2106.10887.

Wang, L., Lin, Z.Q., Wong, A., 2020b. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest X-ray images. Sci. Rep. 10 (1), 1–12.

Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B., 2018c. High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8798–8807.

Wang, C., Shang, K., Zhang, H., Li, Q., Hui, Y., Zhou, S.K., 2021c. DuDoTrans: Dual-domain transformer provides more attention for sinogram restoration in sparse-view CT reconstruction. arXiv preprint arXiv:2111.10790.

Wang, Z., She, Q., Ward, T.E., 2021d. Generative adversarial networks in computer vision: A survey and taxonomy. ACM Comput. Surv. 54 (2), 1–38.

Wang, Q., Sun, L., Wang, Y., Zhou, M., Hu, M., Chen, J., Wen, Y., Li, Q., 2020c. Identification of melanoma from hyperspectral pathology image using 3D convolutional networks. IEEE Trans. Med. Imaging 40 (1), 218–227.

Wang, S., Tang, L., Lin, M., Shih, G., Ding, Y., Peng, Y., 2022b. Prior knowledge enhances radiology report generation. arXiv preprint arXiv:2201.03761.

Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J., 2021e. Boundary-aware transformers for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 206–216.

Wang, H., Wu, Z., Liu, Z., Cai, H., Zhu, L., Gan, C., Han, S., 2020d. Hat: Hardware-aware transformers for efficient natural language processing. arXiv preprint arXiv:2005.14187.

Wang, D., Wu, Z., Yu, H., 2021f. TED-net: Convolution-free T2T vision transformer-based encoder-decoder dilation network for low-dose CT denoising. arXiv preprint arXiv:2106.04650.

Wang, C., Xu, R., Xu, S., Meng, W., Zhang, X., 2022c. DA-net: Dual branch transformer and adaptive strip upsampling for retinal vessels segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II. Springer, pp. 528–538.

Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C., 2020e. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision. Springer, pp. 108–126.

Wei, Z., Chen, J., Goldblum, M., Wu, Z., Goldstein, T., Jiang, Y.G., 2021. Towards transferable adversarial attacks on vision transformers. arXiv preprint arXiv:2109.04176.

Wolterink, J.M., Dinkla, A.M., Savenije, M.H., Seevinck, P.R., van den Berg, C.A., Išgum, I., 2017. Deep MR to CT synthesis using unpaired data. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 14–23.

Woolson, R.F., 2007. Wilcoxon signed-rank test. In: Wiley Encyclopedia of Clinical Trials. Wiley Online Library, pp. 1–3.

Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z., 2021a. FAT-net: Feature adaptive transformers for automated skin lesion segmentation. Med. Image Anal. 102327.

Wu, J., Hu, R., Xiao, Z., Chen, J., Liu, J., 2021b. Vision transformer-based recognition of diabetic retinopathy grade. Med. Phys.

Wu, M., Qian, Y., Liao, X., Wang, Q., Heng, P.A., 2021c. Hepatic vessel segmentation based on 3D swin-transformer with inductive biased multi-head self-attention. arXiv preprint arXiv:2111.03368.

Wu, M., Xu, Y., Xu, Y., Wu, G., Chen, Q., Lin, H., 2022. Adaptively re-weighting multi-loss untrained transformer for sparse-view cone-beam CT reconstruction. arXiv preprint arXiv:2203.12476.

Würfl, T., Ghesu, F.C., Christlein, V., Maier, A., 2016. Deep learning computed tomography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 432–440.

Xia, Y., Yao, J., Lu, L., Huang, L., Xie, G., Xiao, J., Yuille, A., Cao, K., Zhang, L., 2021. Effective pancreatic cancer screening on non-contrast CT scans via anatomy-aware transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 259–269.

Xie, Y., Zhang, J., Shen, C., Xia, Y., 2021a. CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation. arXiv preprint arXiv:2103.03024.

Xie, Y., Zhang, J., Xia, Y., Wu, Q., 2021b. Unified 2D and 3D pre-training for medical image classification and segmentation. arXiv preprint arXiv:2112.09356.

Xiong, Y., Du, B., Yan, P., 2019. Reinforced transformer for medical image captioning. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 673–680.

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V., 2021. Nystr\" omformer: A nystr\" om-based algorithm for approximating self-attention. arXiv preprint arXiv:2102.03902.

Xu, G., Wu, X., Zhang, X., He, X., 2021. LeViT-UNet: Make faster encoders with transformer for medical image segmentation. arXiv preprint arXiv:2107.08623.

Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., McAuley, J., Hsu, C.N., 2021a. Weakly supervised contrastive learning for chest X-ray report generation. arXiv preprint arXiv:2109.12242.

Yan, K., Tang, Y., Harrison, A.P., Cai, J., Lu, L., Lu, J., 2021b. Interpretable medical image classification with self-supervised anatomical embedding and prior knowledge.

Yan, X., Tang, H., Sun, S., Ma, H., Kong, D., Xie, X., 2021c. AFTer-UNet: Axial fusion transformer unet for medical image segmentation. arXiv preprint arXiv:2110.10403.

Yang, H., Chen, J., Xu, M., 2021a. Fundus disease image classification based on improved transformer. In: 2021 International Conference on Neuromorphic Computing. ICNC, IEEE, pp. 207–214.

Yang, Q., Liu, Y., Chen, T., Tong, Y., 2019. Federated machine learning: Concept and applications. ACM Trans. Intell. Syst. Technol. 10 (2), 1–19.

Yang, J., Zhou, K., Li, Y., Liu, Z., 2021b. Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334.

Yao, C., Tang, J., Hu, M., Wu, Y., Guo, W., Li, Q., Zhang, X.P., 2020. Claw U-net: A unet-based network with deep feature concatenation for scleral blood vessel segmentation. arXiv preprint arXiv:2010.10163.

Yap, M.H., Pons, G., Martí, J., Ganau, S., Sentís, M., Zwiggelaar, R., Davison, A.K., Marti, R., 2017. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J. Biomed. Health Inf. 22 (4), 1218–1226.

Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P., Yang, G.Z., 2017. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. arXiv preprint arXiv:1705.08260.

Yi, X., Walia, E., Babyn, P., 2019. Generative adversarial network in medical imaging: A review. Med. Image Anal. 58, 101552.

Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S., Hu, H., 2020. Disentangled non-local neural networks. In: European Conference on Computer Vision. Springer, pp. 191–207.

You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X., 2021. AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 72–82.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., et al., 2021. Opacus: User-friendly differential privacy library in PyTorch. arXiv preprint arXiv:2109.12298.

Yu, S., Ma, K., Bi, Q., Bian, C., Ning, M., He, N., Li, Y., Liu, H., Zheng, Y., 2021a. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 45–54.

Yu, Z., Mar, V., Eriksson, A., Chandra, S., Bonnington, P., Zhang, L., Ge, Z., 2021b. End-to-end ugly duckling sign detection for melanoma identification with transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 176–184.

Yu, H., Shim, J.h., Kwak, J., Song, J.W., Kang, S.J., 2022. Vision transformer-based retina vessel segmentation with deep adaptive Gamma correction. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1456–1460.

Yu, C., Wang, J., Gao, C., Yu, G., Shen, C., Sang, N., 2020. Context prior for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12416–12425.

Yu, J., Zhang, C., Wang, H., Zhang, D., Song, Y., Xiang, T., Liu, D., Cai, W., 2021c. 3D medical point transformer: Introducing convolution to attention networks for medical point cloud analysis. arXiv preprint arXiv:2112.04863.

Yuan, Y., 2017. Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. arXiv preprint arXiv:1703.05165.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F.E., Feng, J., Yan, S., 2021a. Tokens-to-token vit: Training vision transformers from scratch on imagenet. arXiv preprint arXiv:2101.11986.

Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S., 2021b. Volo: Vision outlooker for visual recognition. arXiv preprint arXiv:2106.13112.

Yuan, J., Liao, H., Luo, R., Luo, J., 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 721–729.

Yun, B., Wang, Y., Chen, J., Wang, H., Shen, W., Li, Q., 2021. SpecTr: Spectral transformer for hyperspectral pathology image segmentation. arXiv preprint arXiv:2103.03604.

Zahn, C.T., Roskies, R.Z., 1972. Fourier descriptors for plane closed curves. IEEE Trans. Comput. 100 (3), 269–281.

Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M.J., Defazio, A., Stern, R., Johnson, P., Bruno, M., et al., 2018. fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv preprint arXiv:1811.08839.

Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L., 2021. Scaling vision transformers. arXiv preprint arXiv:2106.04560.

Zhang, O., Delbrouck, J.B., Rubin, D.L., 2021a. Out of distribution detection for medical images. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. Springer, pp. 102–111.

Zhang, X., He, X., Guo, J., Ettehadi, N., Aw, N., Semanek, D., Posner, J., Laine, A., Wang, Y., 2021b. PTNet: A high-resolution infant MRI synthesizer based on transformer. arXiv preprint arXiv:2105.13993.

Zhang, Y., Higashita, R., Fu, H., Xu, Y., Zhang, Y., Liu, H., Zhang, J., Liu, J., 2021c. A multi-branch hybrid transformer networkfor corneal endothelial cell segmentation. arXiv preprint arXiv:2106.07557.

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2019a. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

Zhang, Q., Li, Q., Yu, G., Sun, L., Zhou, M., Chu, J., 2019b. A multidimensional choledoch database and benchmarks for cholangiocarcinoma diagnosis. IEEE Access 7, 149414–149421.

Zhang, Y., Liu, H., Hu, Q., 2021d. Transfuse: Fusing transformers and cnns for medical image segmentation. arXiv preprint arXiv:2102.08005.

Zhang, J., Nie, Y., Chang, J., Zhang, J.J., 2021e. Surgical instruction generation with transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 290–299.

Zhang, Y., Pei, Y., Zha, H., 2021f. Learning dual transformer network for diffeomorphic registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 129–138.

Zhang, Z., Sun, B., Zhang, W., 2021g. Pyramid medical transformer for medical image segmentation. arXiv preprint arXiv:2104.14702.

Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D., 2020. When radiology report generation meets knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07. pp. 12910–12917.

Zhang, L., Wen, Y., 2021a. MIA-COV19D: A transformer-based framework for COVID19 classification in chest CTs.

Zhang, L., Wen, Y., 2021b. A transformer-based framework for automatic COVID19 diagnosis in chest CTs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 513–518.

Zhang, Q.L., Yang, Y.B., 2021. Sa-net: Shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 2235–2239.

Zhang, Z., Yu, L., Liang, X., Zhao, W., Xing, L., 2021h. TransCT: Dual-path transformer for low dose computed tomography. arXiv preprint arXiv:2103.00634.

Zhang, C., Zhang, M., Zhang, S., Jin, D., Zhou, Q., Cai, Z., Zhao, H., Yi, S., Liu, X., Liu, Z., 2021i. Delving deep into the generalization of vision transformers under distribution shifts. arXiv preprint arXiv:2106.07617.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890.

Zheng, Y., Gindra, R., Betke, M., Beane, J., Kolachalama, V.B., 2021a. A deep learning based graph-transformer for whole slide image classification. medRxiv.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al., 2021b. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890.

Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M., 2021a. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE.

Zhou, S.K., Greenspan, H., Shen, D., 2017. Deep Learning for Medical Image Analysis. Academic Press.

Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y., 2021b. nnFormer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201.

Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A.L., 2019a. Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10672–10681.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2019b. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 39 (6), 1856–1867.

Zhu, X., Hu, H., Wang, H., Yao, J., Ou, D., Xu, D., et al., 2021. Region aware transformer for automatic breast ultrasound tumor segmentation.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

Ziegler, Z.M., Melas-Kyriazi, L., Gehrmann, S., Rush, A.M., 2019. Encoder-agnostic adaptation for conditional language generation. arXiv preprint arXiv:1908.06938.

Ziller, A., Usynin, D., Braren, R., Makowski, M., Rueckert, D., Kaissis, G., 2021. Medical imaging deep learning with differential privacy. Sci. Rep. 11 (1), 1–8.