Transformers in Medical Imaging: A Survey

Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu

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, 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 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.

Index Terms—Transformers, medical image analysis, vision transformers, deep neural networks, clinical report generation

1 INTRODUCTION

CONVOLUTIONAL Neural Networks (CNNs) [1]–[4] 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 [5], Endoscopy [6], Computed Tomography (CT) [7], [8], Mammography Images (MG) [9], Ultrasound Images [10], Magnetic Resonance Imaging (MRI) [11], [12], and Positron Emission Tomography (PET) [13], 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 [14]–[16] in CNN-inspired architectures [17]–[22]. 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 [23]. Recent works have shown that these Transformer modules can fully replace the standard convolutions in deep neural networks by operating on a sequence of image patches, giving rise to Vision Transformers (ViTs) [22]. Since their inception, ViT models have been shown to push the state-of-the-art in numerous vision tasks, including image classification [22], object detection [24], semantic segmentation [25], image colorization [26], low-level vision [27], and video understanding [28] to name a few. Furthermore, recent research indicate that the prediction errors of ViTs are more consistent with those of humans than CNNs [29]–[32]. 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 [33].

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 dominant in prestigious medical imaging conferences and journals, 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
Figure 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, Recons: reconstruction, Reg: registration, Synth: synthesis, Det: detection, Rep: report generation, US: ultrasound.

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

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 Sec. 9).

**Paper Organization.** The rest of the paper is organized as follows. In Sec. 2, we provide background of the field with a focus on salient concepts underlying Transformers. From Sec. 3 to Sec. 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. Sec. 11 presents open problems and future directions about the field as a whole. Finally, in Sec. 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 three categories: hand-crafted, CNN-based, and ViT-based. For the hand-crafted and 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 Hand-Crafted Approaches

Conventional algorithms to solve medical imaging tasks are based on hand-crafted mathematical models designed by field experts using domain knowledge. The development of these hand-crafted models with a focus on refining discriminative features and efficient optimization algorithms for a range of medical imaging problems has been the central research topic in the past [40], [41]. Successful hand-crafted models in medical imaging include total-variation [42], non-local self-similarity [43], sparsity/structured sparsity [44], Markov-tree models on wavelet coefficients [45], and untrained neural networks [46]–[48]. These models have been extensively leveraged in medical domain for image segmentation [49], reconstruction [50], disease classification [51], enhancement [52], and anomaly detection [53] due to their interpretability with solid mathematical foundations and theoretical supports on the robustness, recovery, and complexity [54], [55]. Further, unlike deep learning-based approaches, they do not require large annotated medical imaging datasets for training. This reduced reliance on labeled datasets is crucial to the medical research community.
Figure 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) [34], (b) [35], (c) [36], (d) [37], (e) [38], (f) [39].

as collecting voluminous, reliable, and labeled data in the medical domain is difficult due to the lack of expert annotators, high time consumption, ethical considerations, and financial costs. However, due to inadequacy to leverage the expressive power of large medical imaging data sets, these hand-crafted models often suffer from poor discriminative capability [56]. Consequently, these models often fail to represent nuances of high-dimensional complex medical imaging data that can hamper the performance of the medical imaging diagnosis systems [57], [58]. To circumvent the poor discriminability and generalization issue, learned hand-crafted models have been proposed to exploit data better. The representative approaches include optimal directions [59], K-SVD [60], data-driven tight frame [61], low-rank models [62], and piece-wise smooth image model [63]. Next, we explain the popular data-driven approaches explored in the literature.

2.2 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 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 [64]–[66]. We refer interested readers to comprehensive survey papers related to CNNs applications in medical imaging [56], [67]–[75]. Despite considerable perfor-
Figure 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.

In contrast, 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 [76]. These hybrid methods contain special domain-specific layers, and include unrolled optimization [77], generative models [78], and learned denoiser-based approaches [79]. 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 [33] (see Fig. 3). This demands an alternative architectural design capable of modeling long-range pixel relationships for better representation learning.

2.3 Transformers

Transformers were introduced by Vaswani et al. [14] 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 [80]. 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) [22] that are built on vanilla Transformer model [14] by cascading multiple transformer layers to capture the global context of an input image. Specifically, Dosovitskiy et al. [22] 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 from models in an effort to leverage the availability of larger datasets coupled with increased computational 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 [76]. These hybrid methods contain special domain-specific layers, and include unrolled optimization [77], generative models [78], and learned denoiser-based approaches [79]. 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 [33] (see Fig. 3). This demands an alternative architectural design capable of modeling long-range pixel relationships for better representation learning.

2.3.1 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 [80]. For 2D images, we first reshape the image $x \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N \times (P^2C)}$, 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 $X \in \mathbb{R}^{N \times D}$. 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 $X$ into Queries (via $W^Q \in \mathbb{R}^{D \times D_Q}$), Keys (via $W^K \in \mathbb{R}^{D \times D_K}$) and Values (via $W^V \in \mathbb{R}^{D \times D_V}$), where $D_q = D_k$. The input sequence $X$ is first projected onto these weight matrices to get $Q = XW^Q$, $K = XW^K$ and $V = XW^V$. The corresponding attention matrix $A \in \mathbb{R}^{N \times N}$ can be written as,

$$A = \text{softmax}\left( \frac{QK^T}{\sqrt{D_q}} \right).$$

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

$$Z = \text{SA}(X) = AV.$$

### 2.3.2 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 $\{W^Q_i, W^K_i, W^V_i\}$, where $i = 0 \cdots (h-1)$ and $h$ denotes total number of heads in MHSA block. Specifically, we can write,

$$\text{MHSA}(Q, K, V) = [Z_0, Z_1, \ldots, Z_{h-1}]W^O,$$

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

$$Z_i = \text{softmax}\left( \frac{QW^Q_i(KW^K_i)^T}{\sqrt{D_q/h}} \right)W^V_i.$$

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 [85], linearization attention [86], low-rank attention [87], memory compression based approaches [88], and improved MHSA [89]. 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 [90–93] 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 [94]), therefore, learning global context between the pixels corresponding to the background can help the model in preventing misclassification.

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, oclusions, and contrast of skin lesion areas, resulting in ambiguous boundaries [98] and consequently deterioration in segmentation performance. To address the issue of ambiguous boundaries, Wang et al. [97] 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 [99], [100] and ISIC 2018 [101] validate the efficacy of their boundary-wise prior, as shown in Fig. 6. Similarly, Wu et al. [102] 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 [99], [101], [103], and PH2 [100] 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) [104]. However, it is challenging due to blurry boundaries and overexposed and underexposed images. To address these challenges, Li et al. [105] propose Group Transformer U-Net (GT U-Net) that consists of transformer and convolutional layers to encode global and local context, respectively. A shape-sensitive Fourier Descriptor loss function [106] has been proposed to deal with the fuzzy tooth boundaries. Furthermore, grouping and bottleneck structure has been introduced in the GT U-Net to significantly reduce the computational cost. Experiments on their in-house Tooth Root segmentation dataset with six evaluation metrics demonstrate the effectiveness of GT U-Net architectural components and Fourier-based loss function. In another work, Li et al. [107] propose anatomy-guided multibranch Transformer (AGMB-Transformer) to incorporate the strengths of group convolutions [108] 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 et al. [109] 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 [110] and group convolutions to significantly reduce the number of parameters. Extensive experiments on large-scale left ventricle segmentation dataset, echo-cardiographs [111] demonstrate the benefit of TransBridge over CNNs and Transformer-based baseline approaches [112].

**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. [117] 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 [118] 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) [119] demonstrate the effectiveness COTR-Net.

**Cell Segmentation.** Inspired from the Detection Transformers (DETR) [132], Zhang et al. proposed Cell-DETR [133], a Transformer-based framework for instance segmentation of cellular conditions. 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 [134] and Sorenson dice loss [132] are used for the segmentation branch. To enhance performance, they integrate three residual decoder blocks [118] in Cell-DETR to generate accurate instance masks. Experiments on their in-house yeast cells dataset demonstrate the effectiveness of Cell-DETR relative to U-Net based baselines [114]. Similarly, existing medical imaging segmentation approaches generally struggle for Corneal endothelial cells due to blurry edges caused by the subject’s movement [135]. This demands preserving more local details and making full use of the global context. Considering these attributes, Zhang et al. [136] 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 [137] of Corneal Endothelial Cells show the effectiveness of MBT-Net architectural components.

**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. [131] systematically evaluate the performance of six latest ViTs, and CNNs-based approaches on whole slide images of the PAIP liver histopathological dataset [130]. 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.

### 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 such as Alzheimer’s disease. Recently, ViT-based models have been proposed to segment brain tumors effectively. Wang et al. [138] 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 final segmentation map. To further boost the performance, they make use of test-time augmentation. Extensive experimentation on BraTS 2019^2 and BraTS 2020^2 datasets show the effectiveness of their proposed approach compared to CNN-based methods. 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 [138], Jia et al. [139] propose Bi-Transformer U-Net (BiTr-UNet) that performs relatively better on BraTS 2021 [140] segmentation challenge. Different from TransBTS [138], 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 [141] followed by model ensemble via majority voting [142]. Similarly, Peiris et al. [143] 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 [140] show that VT-UNet is robust to data artifacts and exhibits strong generalization ability. In another similar work, Hatamizadeh et al. [145] propose Swin UNETR 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 [140] validation set. In Table 2, we provide dice score and other parameters of various Transformer based models for the 3D multimodal BraTS 2021 dataset [140].

**Histopathology.** Boxiang et. al [116] propose Spectral Transformer (SpecTr) for hyperspectral pathology image

![Figure 7: Segmentation results for hyperspectral pathology dataset using Spectral Transformer (SpecTr). From left to right: Input image, Ground truth label, HIS Hyper (CNN-based) [113], UNet (CNN-based) [114], Attn UNet (CNN-based) [115], UNet++ (CNN-based) [95], and SpecTr (ViT-based) [116]. Image adapted from [116].](image)

**Table 1: Evaluation of Transformer-based semantic segmentation methods for pathological image segmentation in terms of average jaccard index on PAIP liver histopathological dataset [130]. It can be seen that transformer-based models tend to outperform CNNs with the exception of Swin-UNet. Results are from [131], 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 [120]          | 0.58±0.35         | 0.49±0.27         |
| UNet [115]            | 0.65±0.24         | 0.60±0.31         |
| DeepLabV3 [124]       | 0.63±0.28         | 0.67±0.24         |
| FPN [122]             | 0.64±0.20         | 0.72±0.22         |
| PAN [123]             | 0.63±0.24         | 0.69±0.23         |
| LinkNet [124]         | 0.58±0.33         | 0.54±0.25         |

| Transformer-based Models | 40x Magnification | 20x Magnification |
|--------------------------|-------------------|-------------------|
| TransUNet [10]          | 0.77±0.12         | 0.77±0.13         |
| Swin-UNet [125]         | 0.53±0.23         | 0.42±0.23         |
| Swin-Transformer (Base) [126] | 0.79±0.14   | 0.71±0.26         |
| Segmenter [127]         | 0.80±0.14         | 0.85±0.11         |
| Medical Transformer [128] | 0.71±0.14      | 0.62±0.17         |
| BEIT [129]              | 0.72±0.21         | 0.66±0.28         |

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

---

Figures and tables are referenced as follows:

1. https://www.med.upenn.edu/cbica/brats2019
2. https://www.med.upenn.edu/cbica/brats2020
Table 2: Segmentation results and parameters of various Transformer-based models on 3D Multimodal Brain Tumor BraTS 2021 dataset [140].

| Method         | #params | Flops | Dice Score (Avg.) |
|----------------|---------|-------|-------------------|
| TransBTS [138] | 33 M    | 333 G | 84.99             |
| BiTr-UNET [139]|          |       |                   |
| UNETR [35]     | 102.5 M | 193.5 G | 84.51             |
| nnFormer [144] | 39.7 M  | 110.7 G | 86.56             |
| Swin UNETR [145]| 61.98 M| 394.84 G | 88.97             |
| VT-UNET-T [143]| 5.4 M   | 52 G   | 86.82             |
| VT-UNET-S [143]| 11.6 M  | 100.8 G | 87.00             |
| VT-UNET-B [143]| 20.8 M  | 165 G  | 88.07             |
| VT-UNET-B [143]| 20.8 M  | 165 G  |                   |

Figure 8: Overview of TransUNet architecture [96] proposed for multi-organ segmentation. It is one of the first transformer-based architecture proposed for medical image segmentation and merits both transformer and UNet. It employs a hybrid CNN-Transformer architecture for encoder, followed by multiple upsampling layers in decoder to output final segmentation mask. Image adapted from [96].

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 effectively 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 [152]. 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 [96]. Recently, Karimi et. al [153] propose a pure Transformer-based model for 3D medical image segmentation by leveraging self-attention [17] between neighboring linear embedding of 3D medical image patches. They also propose a method to effectively 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 [154], 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. [125] propose Swin-UNET that, like Swin Transformer [126], computes self-attention within a local window and has linear computational complexity with respect to the input image. Swin-UNET also contains a patch expanding layer for upsampling decoder’s feature maps and shows superior performance in recovering fine details compared to bilinear upsampling. Experiments on Synapse and ACDC [155] 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.

3.2.2.1 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-categorize 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 [96] that consists of 12 Transformer layers in the encoder as shown in Figure 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 synapse multi-organ segmentation dataset [156] and automated cardiac diagnosis challenge (ACDC) [155].

In other work, Zhang et al. propose TransFuse [157] to effectively 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 [96] and TransFuse [157] require pre-training on ImageNet dataset [158] to effectively learn the positional encoding of the images. To learn this positional bias without any pre-training, Valanarasu et al. [128] propose a modified gated axial attention layer [159] 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 [160], gland segmentation [161], and MoNuSeg (microscopy) [162] demonstrate the effectiveness of their proposed gated axial attention module.

In another work, Tang et al. [163] introduce SwinUNETR, 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 BTCV segmentation datasets. Similarly, Sobirov et al. [164] 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, Cheng et al. [165] propose TransClaw UNet by integrating Transformer layers in the encoding part of the Claw UNet [166] to exploit multi-scale information. TransClaw-UNet achieves an absolute 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 [167], Xu et al. [168] 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 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 downsampling in the encoder layers. The first work in this category is TransAttUNet [169] 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. [170] propose Axial Fusion Transformer UNet (AFTer-UNet) that contains a computationally efficient axial fusion layer between encoder and decoder to effectively fuse inter and intra-slice information for 3D medical image segmentation. Experiments on BCV [171], Thorax-85 [172], and SegTHOR [173] datasets demonstrate the effectiveness of their proposed fusion layer.

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 that efficiently reduces the complexity of the self-attention mechanism from quadratic to linear [174]. Furthermore, to model the image content effectively, UTNet exploits the two-dimensional relative position encoding [20]. Experiments show strong generalization ability of UTNet on multi-label and multi-vendor cardiac MRI challenge dataset cohort [175]. Similarly, to optimally combine convolution and transformer layers for medical image segmentation, Zhou et al. [144] 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 [126], 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 datasets show that nnFormer surpass Swin-UNet [125] (transformer-based medical segmentation approach) by over 7% (dice score) on Synapse dataset. In other work, Lin et al. propose Dual Swin Transformer UNet (DS-TransUNet) [176] 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. [177] 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

Figure 9: Overview of the interleaved encoder not-another Transformer (nnFormer) [144] for volumetric medical image segmentation. Note that convolution and transformer layers are interleaved to give full play to their strengths. Image taken from [144].
| Method          | Organ      | Modality       | Type        | Datasets                                           | Metrics                      | Arch. | PT | Highlights |
|-----------------|------------|----------------|-------------|---------------------------------------------------|------------------------------|-------|----|------------|
| TransUNet [96]  | Multi-organ| C1, MRI        | 2D          | Synapse [59], ACDC [159]                          | Dice, Hausdorff distance     | Hybrid|    |            |
| TransFuse [107] | Multi-organ| Colonoscopy    | 213, 3D     | Kvasir [175], Colon DB [179], EndoScope [181], ETS [182], RSIC 18 [190], MSD [191] | Dice, Jaccard index         | Hybrid|    |            |
| MedT [178]      | Multi-organ| Ultrasound, Microscopy | 2D | HIP [165], PROMIS-FAN [166], MNOseg [168] | H | Hybrid|    |            |
| Conv-Free [151] | Multi-organ| MRI, CT        | 3D          | Brain, Crotial (private)                          | Dice, Hausdorff Distance, Average Surface Distance | Pure  |    |            |
| COFI [12]       | Multi-organ| CT             | 3D          | Bv, V [12]                                      | Dice                          | Hybrid|    |            |
| SpecFl [145]    | Bile-duct   | Hyperspectral  | 3D          | Cholecrof [159]                                 | Dice, IoU, Hausdorff distance | Hybrid|    |            |
| Tissue-IS [158] | Brain       | MRI            | 3D          | Brats 19 [158], Brats 20 [169]                   | Dice, Hausdorff distance     | Hybrid|    |            |
| U-Transformer [64] | C1             | MRI            | Dice        | No Proposal self and cross-attention modules to model long-range interactions and spatial dependencies |
| UNI-KE [93]    | Brain, Spine | MRI, CT        | 2D          | BVL, MRI [171]                                  | Dice                          | Hybrid|    |            |
| PM-Trans [191] | Multi-organ | Colonoscopy, PET | 2D | GLASS [181], MNOseg [193], HECKTOR [189] | Dice                          | Hybrid|    |            |
| SpecTr [116] | Multi-organ | MRI, CT        | 2D          | Synapse [59], AJDC [159]                         | Dice                          | Pure  |    |            |
| MBT-Net [136]  | Multi-organ | Synapse, Colonoscopy, MRI | 213, 3D | REF9CE [190], CVC [192], Kvasir [175] | Dice                          | Hybrid|    |            |
| X-Net [158]    | Eye         | Retinopathy    | 2D          | LMEMO1 [private], Antares [155]                  | Dice, F1, Sensitivity, Specificity | Hybrid|    |            |
| DS-transUNet [76] | Multi-organ | Colonoscopy, Histology | 2D | KRAS [85], Colon DB [68], and Clinic DB [139], EndoScope [181], ETS [182], RSIC 18 [190], GLASS [161], Data Science Bowl 18 [191] | Mean Dice, Mean IoU, Precision, Recall | Hybrid|    |            |
| M-Trans [171]  | Multi-organ | Colonoscopy, Pathology | 2D | PanCanine database, Colon DB [38], Colon DB [179], ETS [182], Kvasir [175], RSIC 2018 [190] | Dice                          | Hybrid|    |            |
| X-Net [158]    | Eye         | Retinopathy    | 2D          | LMEMO1 [private], Antares [155]                  | Dice, F1, Sensitivity, Specificity | Hybrid|    |            |
| U-Net [154]    | Heart       | MRI            | 2D          | MRI Challenge Cohort [179]                       | Dice, Hausdorff distance     | Hybrid|    |            |
| Free-lane U-Net [86] | C1             | MRI            | 2D          | Synapse [59]                                   | Dice, Hausdorff distance     | Hybrid|    |            |
| TransARDUINet [159] | Multi-organ | Xray, C1 | 2D | ISIC 2018 [180], JSRT [194], Montgomery [195], NIH [196], Core-CC-CCII [199], Data Science Bowl 18 [191], GLASS [181] | Dice, F1                      | Hybrid|    |            |
| LeVi-UNet [199] | Multi-organ | C1, MRI        | 2D          | Synapse [59], AJDC [159]                         | Dice, Hausdorff distance     | Hybrid|    |            |
| Polyst-PVT [178] | Multi-organ | Colonoscopy   | 2D          | Kvasir [175], Colon DB [179], EndoScope [181], ETS [182], Kvasir [175], RSIC 2018 [190] | Dice, JIO, MAC, Weighted F-measure, Sensum, E-measure | Hybrid|    |            |
| COBYNet [111]  | Kidney      | CT             | 2D          | KHSU1 Challenge [199]                           | Dice, Surface Dice            | Hybrid|    |            |
| reFormer [151] | C1, MRI     | Synapse, AJDC  | 2D          | Synapse [59], AJDC [159]                         | Dice                          | Hybrid|    |            |
| MS-Transformer [199] | Multi-organ | C1, MRI | 2D | Synapse [59], AJDC [159] | Dice, Hausdorff distance | Hybrid|    |            |
| HAO-Trans [101] | Heart       | Echocardiography | 2D | EchoNATraHydrant [111] | Dice, Hausdorff distance | Hybrid|    |            |
| B3-U-Net [171] | Brain       | MRI            | 3D          | Brats 21 [184]                                  | Dice                          | Hybrid|    |            |
| CT U-Net [199] | Tooth        | X-ray          | 2D          | Tooth root dataset (private)                    | Dice, Accuracy, Sensitivity, Specificity, Jaccard similarity | Hybrid|    |            |
| BAT [179]      | –           | Dermoscopy     | 2D          | ISIC 2016-JPIT [95], JSRT [194], ETS [182]      | Dice, JIO                     | Hybrid|    |            |
| APNet-UNet [174] | C1           | MRI            | 2D          | BVC [81], Thomas-SS [175], SegTHOR [170]       | Dice                          | Hybrid|    |            |
| P-ATNet [143]  | Multi-organ | CT, MRI        | 3D          | Brain 21 [184], MSIT [174]                      | Dice                          | Hybrid|    |            |
| Swin UNETR [145] | Brain       | MRI            | 3D          | Brats 21 [184]                                  | Dice                          | Hybrid|    |            |
efficiency. Experiments on MSD Brain and Synapse datasets demonstrate the superiority of their architecture compared to bilinear upsampling. In another work, Li et al. [190] 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.

3.2.2.2 Multi-Scale Architectures: These architectures process input at multiple scales to effectively 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. [183] propose a pyramid medical transformer, PMTrans, that leverage multi-resolution attention to capture correlation at different image scales using a pyramidal architecture [201]. 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 [161], MoNuSeg [184], and HECKTOR [189] show the effectiveness of exploiting multi-scale information. In other work, Ji et al. [39] 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 [202]. 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 on six benchmark segmentation datasets demonstrate the effectiveness of the architectural components of MCTrans.

3D Segmentation. The majority of multi-scale architectures have been proposed for 2D medical image segmentation. To directly handle volumetric data, Hatamizadeh et al. [35] 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. The encoder is connected to a CNN-based decoder via skip connections to compute the final segmentation output. UNETR achieves impressive performance on BTCV [203] and MSD [171] 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. [112] propose a computationally efficient deformable self-attention module [204] that casts attention only to a small set using...
multi-scale features, as shown in Figure 11, to reduce the computational and spatial complexities. Experiments on BTCV [203] demonstrate the effectiveness of their deformable self-attention module for 3D multi-organ segmentation.

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 [22]. 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 the baseline transformer model, indicating rapid progression of the field. In short, ViT-based architectures have achieved impressive results over benchmark medical datasets, competing and most of the time improving over CNN-based segmentation approaches (see Table 3 for details). Below, we briefly describe some of the challenges associated with ViTs based medical segmentation methods and give possible solutions based on insights from the relevant papers discussed.

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 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 Sec. 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 [35] uses Swin Transformer [126] based architectures to avoid quadratic computing complexity; however, it requires further attention from the community.

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 [205]–[207]. Recently, ViTs have been successfully employed for diagnosis and severity prediction of COVID-19, showing SOTA 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 Black-box models and Interpretable models according to the level of explainability offered.

4.1.1 Black-Box Models

ViT-based Black-box models for COVID-19 imaging classification generally focus on improving accuracy by designing novel and efficient ViT architectures. However, these models are not easily interpretable, making it challenging to gain user-trust. We have further sub-categorized black-box models into 2D and 3D categories, depending on the input image type. Below, we briefly describe these approaches:
ViTs in Medical Image Classification

COVID-19
Tumor Classification
Retinal Diseases Classification
Black Box
Interpretable
Saliency Based
Grad-CAM Based

Figure 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.

2D: 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. [208] propose a lightweight Point-of-Care Transformer (POCFormer) to diagnose COVID-19 from lungs images captured via portable devices. Specifically, POCFormer leverages Linformer [174] 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 [209], thus making it suitable for real-time diagnosis. Experiments on COVID-19 lungs POCUS dataset [210], [211] demonstrate the effectiveness of their proposed architecture with above 90% classification accuracy. In other work, Liu et al. [212] proposed ViT-based model for COVID-19 diagnosis by exploiting a new attention mechanism named Vision Outlooker (VOLO) [213]. 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 [211], [214] demonstrate the effectiveness of their architecture. Similarly, Jiang et al. [215] leverage Swin Transformer [126] and Transformer-in-Transformer [216] to classify 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 [217]. To mitigate this issue, Park et al. [218] proposed a Federated Split Task-Agnostic (FESTA) framework that leveraged the merits of Federated and Split Learning [219], [220] 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 and demonstrate the suitability of ViT body to be shared across relevant tasks by leveraging multitask-learning (MTL) [221] strategy as shown in Fig. 14. They affirm the suitability of ViTs for collaborative learning in medical imaging applications via extensive experiments on the CXR dataset.

3D: Most of the ViT-based approaches for COVID-19 classification operate on 2D information only. However, as suggested by Kwee et al. [222], 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. [223] 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 [224] with Swin Transformer [126] 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 et al. [225], [226] also proposed Swin Transformer based two-stage framework for the diagnosis of COVID-19 in the 3D CT scan dataset [227]. Specifically, their framework consists of UNet based lung segmentation model followed by the image classification with Swin Transformer [126] backbone.

4.1.2 Interpretable Models
Interpretable models aim to show the features that influence the decision of a model the most, generally via visualization techniques like saliency-based methods, Grad-CAM, etc. Due to their interpretable nature, these models are well suited to gain the trust of physicians and patients and therefore have paved their way for clinical deployment. We have further divided interpretable models into saliency-based [228] and Grad-CAM based visualization approaches.

Saliency Based Visualization. Park et al. [231], propose a ViT-based method for COVID-19 diagnosis by exploiting the low-level CXR features extracted from the pre-trained backbone network. The backbone network has been trained in a self-supervised manner (using contrastive-learning based SimCLR [232] method) to extract abnormal CXR features embeddings from large and well-curated CXR dataset of CheXpert [233]. These feature embeddings have been leveraged by ViT model for high-level diagnosis of COVID-19 images. Extensive experiments on three CRX 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 [234] to provide interpretable results. Similarly, Gao et al. [235] propose COVID-ViT to classify COVID from non-COVID images as part of the MIA-COVID19 challenge [227]. Their experiments on 3D CT lungs images demonstrated the superiority of ViT-based approach over DenseNet [236] baseline in terms of F1 score. In another work, Mondal et al. [230] 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
Figure 14: Implementation details of Federated Split Task-Agnostic (FESTA) framework [218] 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 (H) and tail (T) parts of the network, whereas the transformer body (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. Image courtesy [218].

Figure 15: 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 [230] is able to 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 [230] is able to predict it correctly. Figure courtesy [230].

saliency-based approach [234] with clinically interpretable visualizations to highlight the role of critical factors in the resulting predictions, as shown in Figure 15. Experiments on COVID CT-2A [237] and their privately collected Chest X-ray dataset demonstrate the effectiveness of xViTCOS.

Grad-CAM Based Visualization. Shome et al. [238] 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 [229].

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. These models can be mainly categorized into Black-box models and Interpretable models. We highlight the relevant anatomies in bold.

Black-Box Models. TransMed [240] is the first work that leverages ViTs for medical image classification. It 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 effectively capture mutual information from images of different modalities, thereby achieving competitive results on their privately collected parotid tumor classification dataset. Later, Lu et al. [241] 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 [242] dataset show the effectiveness of their two-stage framework. For the task of breast cancer classification, Ghelfati et al. [243] systematically evaluate the performance of pure and hybrid pre-trained ViT models. Experiments on two breast ultrasound datasets provided by Al-Dhabyani et al. [244]
15

Figure 16: Overview of Transformer based Multiple Instance Learning (TransMIL) architecture [239] 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 modelling 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. Image taken from [239].

Figure 17: Left: The area within the blue region is the cancer region. Middle: Attention scores from TransMIL are visualised 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. Figure courtesy [239].

and Yap et al. [245] shows that Vit-based models provide better results than those of the CNNs for classifying images into benign, malignant, and normal categories. Similarly, other works employ hybrid Transformer-CNN architectures to solve medical classification problem for different organs. For instance, Khan et al. [246] propose Gene-Transformer to predict the lung cancer subtypes. Experiments on TCGA-NSCLC [242] dataset demonstrates the superiority of Gene Transformer over CNN baselines. Chen et al. [247] present a multi-scale GasHis-Transformer to diagnose gastric cancer in the stomach. Jiang et al. [248] propose a hybrid model to diagnose acute lymphocytic leukemia by using symmetric cross-entropy loss function.

Interpretable Models. Since the annotation procedure is expensive and laborious, one label is assigned to a set of instances (bag) in whole slide imaging (WSI) based pathology diagnosis. This type of weakly supervised learning is known as Multiple Instance Learning [249], 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. [239] 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. 16. 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 shows state-of-the-art performance on three different computational pathology datasets CAMELYON16 (breast) [250], TCGA-NSCLC (lung) [242], and TCGA-R (kidney) [251]. To diagnose lung tumors, Zheng et al. [252] propose graph transformer network (GTN) to leverage the graph-based representation of WSI. GTN consists of a graph convolutional layer [253], a transformer layer, and a pooling layer. GTN further employs GraphCAM [234] to identify regions that are highly associated with the class label. Extensive evaluations on TCGA dataset [242] show the effectiveness of GTN.

4.3 Retinal Disease Classification

Yu et al [258] 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 [255] and RFMiD2020 [259] 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. [260] 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 [261], Messidor-2 [261], and EyePACS [262] datasets. Yang et al. [273] propose a hybrid architecture consisting of convolutional and Transformer layers for fundus disease classification on OIA dataset [274]. Similarly, Wu et al. [275] and Aldahou et al. [276] also verify that ViT models are more accurate in DR grading than their CNNs counterparts.
| Method                        | Organ       | Modality | Type               | Datasets                          | Metrics                  | Arch. | Highlights                                                                 |
|------------------------------|-------------|----------|-------------------|-----------------------------------|--------------------------|-------|-----------------------------------------------------------------------------|
| ixMim [238]                  | Ear         | MRI      | 2D                | MRI [142]                         | Accuracy                 | Pure  | First multi-modal medical image classification approach with novel multi-modal fusion strategy |
| ixMim [239]                  | Multi-organ | Pathology| 2D                | Camelyon16 [250] TCGA-NSCLC [254] TCGA-RCC [254]  | Accuracy, AuC           | Hybrid| Transformer-based architecture to explore morphological and spatial information for Whole Slide Image classification. |
| Mahakas et al. [23]          | Multi-organ | Dermatology| 2D               | APTOS-2019 [235] ISIC-2019 [256]  | Recall, AuC              | Pure  | Systematic study of whether one should replace CNNs with ViTs for medical image classification. |
| Castellani et al. [24]       | Breast      | US       | 2D                | BUSY [247] Yap et al. [247]       | Accuracy                 | Pure  | First application of ViTs to ultrasound images classification.              |
| GTN [224]                    | Lung        | Microscopy| 2D                | TCGA database [242]               | Accuracy, 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 [238]                | Eye         | Fundus   | 2D                | APTOS-2019 [235] RFMD2020 [259]   | Accuracy, AuC, F1        | Pure  | Formulates lesion discovery as a weakly supervised lesion localization problem via a transformer decoder. Jointly solve diabetic retinopathy grading and lesion discovery. |
| LAT [260]                    | Eye         | Fundus   | 2D                | Messidor-1 [231] EyePACS [263]    | AuC, Kappa               | Hybrid| First pre-trained on a large fundus images dataset and later fine-tuned on the downstream task of the retinal disease classification. |
| COVID-19                     |             |          |                   |                                   |                          |       |                                                                             |
| Park et al. [231]            | Chest       | X-ray    | 2D                | CheXpert [247]                   | 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. |
| MOCFormer [238]             | Chest       | Ultrasound| 2D               | POLUS [250]                      | Recall, F1 Sensitivity, Accuracy | Pure  | Proposed light-weight transformer architecture that demonstrates the efficiency and performance improvements. |
| FESTA [238]                  | Chest       | X-ray    | 2D                | CheXpert [252] SIM-ACR [26] RSNA [36] | Recall, F1 Sensitivity, Accuracy | Pure  | Utilize ViTs to simultaneously process multiple chest X-ray tasks, including the diagnosis in COVID-19 Chest X-ray images on a massive decentralized dataset. |
| Liu et al. [212]             | Chest       | X-ray    | 2D                | COVID-19-1 [214] COVID-19-2 [211] | Accuracy                 | Pure  | Explore VOLO tailored with transfer learning technique to effectively encode fine-level features into the token representations. |
| COVID-ViT [235]              | Chest       | C        | 3D                | MIA-COV19 [234]                  | Accuracy, F1             | Pure  | Propose ViT-based architecture to classify COVID-19 CT images in MIL-COV19 competition. |
| xViT [238]                   | Chest       | X-ray, C | 2D                | xViT-COS-CXR-CT [237] xViT-COS-CXR [265] | Precision, Recall, F1 Sensitivity, 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. [235]             | Chest       | C        | 3D                | COVID-19 C1 DB [22]              | Accuracy, Recall, F1 Precision | Hybrid| Importance of slices are determined in C1 scan via Wilcoxon signed-rank test. Then, spatial and temporal features are exploited via proposed convolutional CT scan Aware Transformer module. |
| Zhang et al. [225]           | Chest       | C        | 3D                | MIA-COV19 [226]                  | F1 score                 | Hybrid| SVM: Transformer based two stage framework for diagnosis of COVID-19 in 3D CT scans. |
| COVIDGAN [236]               | Chest       | C        | 2D                | COVID-19-1 [227] Sars-CoV2 [266] | Accuracy, Sensitivity, F1 | Hybrid| Leverages 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. [238]           | Chest       | X-ray    | 2D                | El-Shalal et al. [260] Satt et al. [261] Qb et al. [270] | Accuracy, 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 4: An overview of ViT-based approaches for medical image classification. Teal color indicates that the model is interpretable.

| Initialization          | Model          | APTOS2019, ×↑ | ISIC2019, Recall ↑ |
|-------------------------|----------------|--------------|--------------------|
| Random                  | ResNet50       | 0.849 ± 0.022 | 0.662 ± 0.018      |
|                         | DeiT-S         | 0.687 ± 0.017 | 0.579 ± 0.022      |
| ImageNet (supervised)   | ResNet50       | 0.893 ± 0.004 | 0.810 ± 0.008      |
|                         | DeiT-S         | 0.866 ± 0.005 | 0.844 ± 0.021      |
| ImageNet (supervised) + | ResNet50       | 0.894 ± 0.008 | 0.833 ± 0.007      |
|                         | DeiT-S         | 0.864 ± 0.010 | 0.853 ± 0.009      |

Table 5: Comparison of vanilla CNNs vs. ViTs with different initialization strategies on medical imaging classification tasks. For APTOS 2019 [255] and ISIC 2019 [272] datasets quadratic Cohen Kappa and recall score have been reported. 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. Table taken from [33].

4.4 Discussion

In particular, we see a surge of Transformer-based architectures for diagnosing COVID-19, compelling us to develop taxonomy accordingly. Below, we briefly highlight some of the challenges associated with this area, identify recent trends, and provide future directions worthy of further exploration.

The lack of large COVID-19 datasets hindered the applicability of ViT models to diagnose COVID-19. A recent work by Shome et al. [238] 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 interpretable (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 [22] 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. [33] 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 5. 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. [277] 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 4.

5 MEDICAL IMAGE 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 [278], [279], recently few attempts have been made to improve performance further using Transformer models. These approaches are mainly based on the detection transformer (DETR) framework [24].

Shen et al. [200] 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 [24] is also adapted in other works [280], [281] for the end-to-end polyp detection [280], and detecting lymph nodes in T2 MRI scans for the assessment of lymphoproliferative diseases [281].

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. 31. A recent work [282] shows that generic class-agnostic detection mechanism of multi-modal ViTs (like MDETR [283]) 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 RECONSTRUCTION

The goal of medical image reconstruction 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 restoration 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) [284], [285]. 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. [37] propose an hybrid architecture TransCT that leverages the internal similarity of the LDCT images to effectively 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 [286] demonstrate the effectiveness of TransCT over CNN-based approaches. To perform LDCT image enhancement, Wang et al. [289] propose a convolution-free ViT-based encoder-decoder architecture TED-Net. It employs Token-to-token block [290] to enrich the image tokenization via a cascaded process. To refine contextual information, TED-Net introduces dilation and cyclic-shift blocks [125] in tokenization. TED-Net shows favorable performance on the Mayo Clinic LDCT dataset [286]. In another work, Luthra et al. [291] propose Eformer which is Transformer-based residual learning architecture for LDCT images denoising. To focus on edges, Eformer uses the power of Sobel-Feldman
6.1.2 LDPET Enhancement

Like LDCT, Low-dose positron emission tomography (LD-PET) images reduce the harmful radiation exposure of standard-dose PET (SDPET) at the expense of sacrificing diagnosis accuracy. To address this challenge, Luo et al. [294] 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 to obtain reliable and clinically acceptable 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 Restoration

Medical image restoration 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 [295].

High-Data Regime Approaches. Approaches in this category assume the availability of large MRI training datasets to train ViT models. Feng et al. [296] 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. [297] 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 et al. [298] 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 [299] to preserve semantic and structural information in the super-resolved images.

Low-Data Regime Approaches. One drawback of the aforementioned approaches is the requirement of a massive paired dataset of undersampled and corresponding fully sampled MRI acquisitions to train ViT models. To alleviate the data requirement issue, Korkmaz et al. [38], [300] propose a zero-shot framework, SLATER, that leverages prior induced by randomly initialized neural networks [46], [47] 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 et al. [301] 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 U-Net baseline.

6.2.2 Sparse-View CT Reconstruction

Sparse-view CT [312] can effectively 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 [313]. Wang et al. [305] 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 [286] to show generalizability, and robustness (against noise and artifacts) of DuDoTrans.

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. [307] 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 and a segmentor network to accurately predict the surgical tool’s mask. Extensive experiments on Hamlyn Centre Endoscopic Video Dataset [308] 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 reconstruction, as shown in Table 6. Below, we highlight a few challenging problems and provide recent trends in the field.

Recently, an interesting work [301] 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 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 [38]. This work is inspired by the strong prior induced by the structure of untrained neural networks [46, 47]. 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.

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.

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 [314]. We have further classified these approaches into intra-modality synthesis and inter-modality synthesis due
Table 6: An overview of ViT-based approaches for medical image reconstruction.

| Method       | Highlights                                                                 | Modality | Input Type | Datasets                          | Metric          |
|--------------|------------------------------------------------------------------------------|----------|------------|-----------------------------------|----------------|
| TranST [305] | Transformer for LDCT enhancement with high and low frequency decomposition.  | CT       | 2D         | NIH-AAPM [286]                    | RMSE, SSIM, VIF |
| SLATER [38]  | Transformer based approach for zero shot MRI image reconstruction.           | MRI      | 3D         | IXI [303], fastMRI [304]          | PSNR, SSIM, RMSE|
| TED-Net [297]| Pure transformer based encoder decoder dilation architecture for 2D         | CT       | 2D         | NIH-AAPM [286]                    | RMSE, SSIM, VIF |
| Efformer [297]| Transformers based residual image denoising, incorporate learnable Sobel filters for edge enhancement. | CT       | 2D         | NIH-AAPM [286]                    | PSNR, SSIM, RMSE|
| Transformer-GAN [294]| End-to-end GAN-based method integrated with transformers to enhance LDPEI images. | PEI      | 3D         | Private                           | PSNR, SSIM, MSE |
| MTrans [296] | 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 [304]                      | PSNR, SSIM, NMSE|
| Task-Transformer [297]| Simultaneously, reconstruct and super-resolve MRI images via multi-task learning. | MRI      | 2D         | IXI [303]                         | PSNR, SSIM, NMSE|
| Mahapatra et al. [308]| Hybrid architecture to super-resolve MRI images by exploiting the complementary advantages of CNNs and ViTs. | MRI      | 2D         | fastMRI [304], IXI [303]           | PSNR, SSIM, NMSE|
| Lin et al. [301]| ViT pretrained on ImageNet, when fine-tuned on only 100 fastMRI images, yields sharp reconstructions and is robust towards anatomy shifts. | MRI      | 2D         | fastMRI [304]                      | SSIM           |
| DuDoTrans [305]| 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 [286]                    | PSNR, SSIM      |
| MBSF-Net [306]| Swiff-Transformer based projection and image domain two-stage framework to reconstruct high-quality CT images from sparse views. | CT       | 2D         | NIH-AAPM [286]                    | PSNR, SSIM, RMSE|
| E-LSSR [307]| 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 [308], DaVinci (private)    | PSNR, SSIM      |
| TransSMS [308]| ViT-based data consistancy module to super-resolve magnetic particle imaging (MPI) system matrices for accelerated calibration. | MPI      | 2D, 3D     | Open MPI [310], Private datasets  | RMSE            |

7.1 Intra-Modality Approaches

The goal of intra-modality synthesis is to generate high-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. [315] focus on synthesizing infant brain structural MRIs (T1w and T2w scans) using both transformer and performer (simplified self-attention) layers [88]. Specifically, they design a novel multi-resolution pyramid-like U-Net 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 [316], and pix2pixHD [317] on large-scale infant MRI dataset [318]. 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 unlabelled data to train models effectively. Kamran et al. [319] propose a multi-scale conditional generative adversarial network (GAN) [316], 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 [320] and Kernel Inception Distance [321] demonstrate the superiority of their approach over baseline methods on diabetic retinopathy dataset provided by Hajeb et al. [322].

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 [323] proposed 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...
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. [325] introduce a novel synthesis approach, ResViT, for the multi-modal imaging based on a conditional deep adversarial network with ViT-based generator. Specifically, ResViT employs convolutional and transformer branches within a residual bottleneck to preserve both local precision and contextual sensitivity along with the realism of adversarial learning. 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 [326]), and a multi-modal pelvic MRI-CT dataset [327].

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 general, more than one imaging modality is involved in accurate clinical decision-making due to their complementary strengths. Therefore, in many practical applications, a certain modality is desired but infeasible to acquire due to cost and privacy issues. Recent transformer-based approaches can effectively circumvent these issues 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 [328].

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 (like CT and MRI images of the same subject have significantly different appearances, as shown in Fig. 23), thereby making it a promising 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) [75]. 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 [329]. In this section, we briefly highlight the applications of ViTs in medical image registration.

The first study to investigate the usage of transformers for self-supervised medical volumetric image registration has been proposed by Chen et al. [330]. 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 [331]. 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. [332] 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. [333] present a novel dual transformer architecture (DTN) for volumetric diffeomorphic registration by effectively establishing correspondences between anatomical structures in an unsupervised manner. The DTN consists of two CNN-based 3D U-Net 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 [334] of MRI scans demonstrate the effectiveness of their proposed architecture.

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 [335]–[338]. 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 effectively 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. 24, due to differences in their underlying training mechanism.

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 [339] 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. [340]. They propose Reinforced-Transformer for Medical Image Captioning (RTMIC) that consists of a pre-trained DenseNet [236] 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 [341]. Similarly, Miura et al. [342] show that the high accuracy of automatic radiology reports as measured by natural language generation metrics such as BLEU [343] and CInder [344] are often incomplete and inconsistent. To address these challenges, Miura et al. [342] 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 [345] 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. [346] 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. [346] have proposed an encoder-decoder based architecture back-boned by a transformer model. Specifically, their proposed architecture, optimized via self-critical reinforcement learning [339], effectively models the dependencies for visual features, textual features, and visual-textual relational features to accurately generate surgical reports on the DAISI dataset [347].

### 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 [348] propose a hierarchical classification approach using a transformer as a decoder. Specifically, the transformer decoder leverages 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. Experiments on IU Chest X-ray dataset [341] demonstrate the effectiveness of the proposed architectural components.

Similarly, Liu et al. [349] try to imitate the work of radiologists by distilling posterior and prior knowledge to generate accurate radiology reports. Their proposed architecture consists of 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 (mitigate image data bias), PrKE explores relevant prior information from the radiological reports and medical knowledge graph (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. [350] propose AlignTransformer to generate a medical report from X-ray images. Specifically, AlignTransformer consists of two modules: align hierarchical attention and multi-grained transformer. Align hierarchical 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 [351] to accurately generate long medical reports. AlignTransformer achieves favorable performance on IU-Xray [341] and MIMIC-CXR [352] datasets.

#### 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. 25. To align better, Chen et al. [353] 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 [341] and MIMIC-CXR [352] 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. [353], Yan et al. [361] introduce a weakly supervised contrastive objective to favor reports that are semantically close to the target, thereby producing more clinically accurate outputs. Similarly, Amjoud et al. [362] investigate the impact on the report generation performance by modifying different architectural components of the model proposed by Chen et al. [353] including replacing visual extractor and changing the number of layers in transformer-based decoder.

#### 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. [363] 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 [236] 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. 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. [354] 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. 26. 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. [364] 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. Extensive experiments on IU Chest X-ray [341] and MIMIC-CXR [352] datasets demonstrate the superiority of their approach both in terms of Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics. Similarly, Lovelace et al. [369] also leverage the transformer-based encoder and decoder for accurate medical report generation on MIMIC-CXR dataset [352]. To emphasize 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. [372] 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 [373], [374] 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. [366] 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. In other work, Nguyen et al. [367] 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 [341] and MIMIC-CXR [352] datasets. Inspired by curriculum learning [375], Nooralahzadeh et al. [365] 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. [376] investigate the impact of visual feature extractor model...
Table 8: 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) [343], METEOR (MTR) [370] and ROUGE-L (RG-L) [371] and CE metrics include precision, recall and F1 score.

| Dataset   | Model                     | Year | BL-1 | BL-2 | NLG Metrics | Year | BL-3 | MTR | RG-L | CE Metrics |
|-----------|---------------------------|------|------|------|-------------|------|------|-----|------|------------|
| IU X-RAY  [341] | Transformers [14]         | 2017 | 0.409 | 0.268 | 0.191       | 2020 | 0.353 | 0.218 | 0.145 | 0.103     |
|           | Memory Transformer [364]  | 2020 | 0.415 | 0.272 | 0.193       | 2020 | 0.353 | 0.218 | 0.145 | 0.103     |
|           | Clinical Transformer [369] | 2020 | 0.361 | 0.221 | 0.146       | 2020 | 0.353 | 0.218 | 0.145 | 0.103     |
|           | M2 TR. [365]             | 2021 | 0.361 | 0.221 | 0.146       | 2021 | 0.353 | 0.218 | 0.145 | 0.103     |
|           | PPKE [349]                | 2021 | 0.378 | 0.232 | 0.154       | 2021 | 0.353 | 0.218 | 0.145 | 0.103     |
|           | ALIGN Transformer [350]   | 2021 | 0.457 | 0.307 | 0.278       | 2021 | 0.361 | 0.221 | 0.146 | 0.103     |
| MIMIC-CXR [352] | NGUYEN ET AL [367]       | 2017 | 0.495 | 0.360 | 0.278       | 2020 | 0.378 | 0.235 | 0.156 | 0.112     |
|           | MDT-WCL [361]            | 2021 | 0.373 | -     | -           | 2021 | 0.378 | 0.235 | 0.156 | 0.112     |
|           | M2 TR. (CE) [342]        | 2021 | 0.378 | 0.232 | 0.154       | 2021 | 0.378 | 0.232 | 0.154 | 0.107     |
|           | M2 TRANS (EN) [342]      | 2021 | -     | -     | -           | 2021 | -     | -     | -     | -          |
|           | KGAE UNSUPERVISED [368]  | 2021 | 0.221 | 0.144 | 0.096       | 2021 | 0.352 | 0.219 | 0.149 | 0.080     |
|           | KGAE SIMI-SUPERVISED [368]| 2021 | 0.352 | 0.219 | 0.149       | 2021 | 0.352 | 0.219 | 0.149 | 0.080     |
|           | KGAE SUPERVISED [368]    | 2021 | 0.369 | 0.231 | 0.156       | 2021 | 0.369 | 0.231 | 0.156 | 0.118     |

on the performance of medical report generation. Based on insights, they propose a modified HRNet [377], MedSkip, to extract visual features for the subsequent processing by the transformer-based decoder to generate an accurate medical report. Similarly, Park et al. [378] 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. [368] propose a knowledge graph auto-encoder that works in the share latent domain. To mitigate this issue, Liu et al. [368] propose a knowledge graph auto-encoder that works in the share latent domain. To mitigate this issue, Liu et al. [368] propose a knowledge graph auto-encoder that works in the share latent domain.

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 [335]–[338], 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. Below, we briefly highlight a few challenges associated with transformer-based models for report generation and outline promising future directions to explore.

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. [342] 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. [354] 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 their specific nature and distinct characteristics. Further, few medical datasets like ROCO [379], PEIR Gross [380], and ImageCLEF [381] 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...
Further, we would like to point out the interested researchers toward the recently explored surgical instructions generation work using transformers [346] that could have a huge impact on surgical robotics, a market that is expected to reach USD 22.27 Billion by 2028. However, only one dataset, DAISSI [347], 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 [341] 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 8 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. [382] 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 state-of-the-art CNN-based approaches. Similarly, Kipkogei et al. [383] propose a Transformer-based architecture, Clinical Transformer, to model the relation between clinical and molecular features to predict survival outcomes from cancerous lung dataset [384]. In another work, Eslami [385] propose PubMedCLIP, a fine-tuned version of Contrastive Language-Image Pre-Training (CLIP) [386] 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, Liu et al. [10] propose 3D Medical Point Transformer (3DMPT) to analyse 3D medical data. 3DMPT is tested on 3D medical classification and part segmentation based tasks. Similarly, Malkiel et al. [387] propose a Transformer-based architecture to analyse 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.

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:

- 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 [271] and BYOL [388]. ViTs appear to outperform CNNs in this setting for medical image classification by a small margin.

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 (Sec. 11.1), interpretability of ViT-based medical imaging approaches (Sec. 11.2), robustness against adversarial attacks (Sec. 11.3), designing efficient ViT architectures for real-time medical applications (Sec. 11.4), challenges in deploying ViT-based models in distributed settings (Sec. 11.5), and domain adaptation (Sec. 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 [22]. This may be a barrier 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 via ImageNet pre-training, which may be sub-optimal due to drastically different image characteristics between medical and natural images. Recently, Matsoukas et al. [33] 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 [271] and BYOL [388]. 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 [389], Token-to-Token [290], transformer in transformer [216], 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. [163] 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
We believe such studies for ViT-based models, along with the success of transformers has been empirically shown in clinical applications [394]–[396], 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 as shown in Fig. 28 [397]. Recent efforts for interpretable ViT-based medical imaging models leverage saliency-based approaches [234] and Grad-CAM based visualizations [229]. Despite these efforts, the development of interpretable and explainable ViT-based approaches, specifically tailored for life-critical medical imaging applications, is a challenging 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.

### 11.2 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 black-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 [393]. Although several explainable AI-based medical imaging systems have been developed to gain deeper insights into the working of CNNs models for clinical applications [394]–[396], it underpins the medical imaging sector, this inevitably poses a risk whereby potential attackers may seek to profit from manipulation against the healthcare system, as shown in Fig. 29. For example, 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
Figure 28: Saliency maps comparison for a CNN-based ResNet-50 [118] (second row) and ViT-based DEIT-S [389] (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. Image taken from [33].

Table 9: Description of datasets generally used in medical adversarial deep learning.

| Dataset Name          | Dataset Size | Modality      |
|-----------------------|--------------|---------------|
| RSNA [264]            | 29,700       | X-ray         |
| JSRT [194]            | 247          | X-ray         |
| BraTS 2018 [398]      | 1689         | MRI           |
| BraTS 2019 [398]      | 1675         | MRI           |
| OASIS [399]           | 37,216,682   | MRI           |
| HAM10000 [272]        | 10,000       | Dermatoscopic |
| ISIC 18 [101]         | 3594         | Dermatoscopic |
| LUNA 16 [400]         | 888          | CT-Scans      |
| NIH Chest X-ray [196] | 112,000      | X-ray         |
| APTOS [255]           | 5590         | Fundoscopy    |
| Chest X-ray [301]     | 3856         | X-ray         |
| NSLT [402]            | 75,000       | CT-Scans      |
| Diabetic Retinopathy  | 40,000       | Fundoscopy    |

Figure 29: 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.

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 constraint edge devices [423], [424]. Due to recent advancements in edge computing, healthcare providers can process, store and analyze complex medical imaging data on-premises, speeding 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 [425] and neural architecture search strategies [426]. 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 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. [218] propose a Federated Split Task-Agnostic (FESTA) framework that integrates the power of Federated and Split Learning [219], [220] in utilizing ViT to simultaneously process multiple chest X-ray tasks, including diagnosing COVID-19 CXR images on a large corpus of decentralized data. Specifically, they split ViT into shared body and task-specific heads and demonstrate that ViT body with sufficient capacity can be shared across relevant tasks by leveraging the multitask-learning (MTL) [221] strategy. 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 [427] to prevent gradient inversion attacks [328] 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 10, 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 [431]–[435]. 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. 30, 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 [437] to effectively mitigate the issue of domains shift using ViTs.

Few preliminary efforts have been made to explore this direction [438]; 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 [439] framework in the natural image domain. Such a framework will also help in advocating models reproducibility.

| Name            | Framework       | Description                                                                 |
|-----------------|-----------------|------------------------------------------------------------------------------|
| TensorFlow Fed  | TensorFlow      | Open-source framework for machine learning and other computations on decentralized data. |
| CrypTen [419]   | PyTorch         | Framework to facilitate research in secure and privacy-preserving machine learning. |
| OpenMined [420] | TensorFlow, PyTorch, Keras | Open-source decentralised privacy preserving framework. Includes specialized tools like PySyft, PyGrid, and SyferText. |
| Opacus [421]    | PyTorch         | Enables training PyTorch models with differential privacy. Allows the client to online track the privacy budget. |
| Deepee [422]    | PyTorch         | Library for differentially private deep learning for medical imaging in PyTorch. |
| PriMIA [422]    | PyTorch         | Framework for end-to-end privacy-preserving decentralized deep learning for medical images. |
Figure 30: 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 [429]. 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). Image courtesy [430].

Figure 31: 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 [68].

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. 31). To keep pace with this rapid development, we 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.

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.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
[2] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541–551, 1989.
[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25:1097–1105, 2012.
[4] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. arXiv preprint arXiv:2201.03545, 2022.
[5] Parash Lakshmi and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology, 284(2):574–582, 2017.
[6] Jun Ki Min, Min Seob Kwak, and Jae Myung Cha. Overview of deep learning in gastrointestinal endoscopy. Gut and liver, 13(4):388, 2019.
deep image priors. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 636–646. Springer, 2006.

[49] KKD Ramesh, G Kiran Kumar, K Swapna, Debabrata Datta, and S Suman Rajest. A review of medical image segmentation algorithms. EAI Endorsed Transactions on Pervasive Health and Technology, 2021.

[50] Saiprasad Ravishankar, Jong Chul Je, and Jeffrey A Fessler. Image reconstruction: From sparsity to data-adaptive methods and machine learning. Proceedings of the IEEE, 108(1):86–109, 2019.

[51] Eka Miranda, Mediana Aryuni, and E Irwansyah. A survey of medical image classification techniques. In 2016 International Conference on Information Management and Technology (ICIMTech), pages 56–61. IEEE, 2016.

[52] Anand Jawdekar and Manish Dixit. A review of image enhancement techniques in medical imaging. Machine Intelligence and Smart Systems, pages 25–33, 2021.

[53] Maximilian E Tschuchnig and Michael Gadermayr. Anomaly detection in medical imaging—a mini review. arXiv preprint arXiv:2108.11986, 2021.

[54] Stefan Rüping et al. Learning interpretable models. 2006.

[55] Muhammad Aurangzeb Ahmad, Carl Eckert, and Ankur Terederai. Interpretable machine learning in healthcare. In Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics, pages 359–360, 2018.

[56] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. IEEE, 2021.

[57] Chinmay Hegde. Algorithmic aspects of inverse problems using generative models. In 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 166–172. IEEE, 2018.

[58] Wenyi Lin, Kyle Hasenstab, Guilherme Moura Cunha, and Armin Schwartzman. Comparison of handcrafted features and convolutional neural networks for liver mr image adequacy assessment. Scientific Reports, 10(1):1–11, 2020.

[59] Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings. ICAASP99 (Cat. No. 99CH36258), volume 3, pages 2443–2446. IEEE, 1999.

[60] I Wolf, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing, 54(11):4311–4322, 2006.

[61] Jian-Feng Cai, Hui Ji, Zuowei Shen, and Gui-Bo Ye. Data-driven tight frame construction and image denoising. IEEE Transactions on Image Processing, 19:221–248, 2010.

[62] Jian-Feng Cai, Hui Ji, Zuowei Shen, and Weichuan Yu. Low-rank modeling and its applications in image analysis. ACM Computing Surveys (CSUR), 47(2):1–33, 2014.

[63] Jungha An, Mikael Rousson, and Chenyang Xu. γ-convergence approximation to piecewise smooth medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 497–502. Springer, 2007.

[64] Eli Grousson, Wenhui Li, Carole Suh, Lucas Fidon, H Shakhshahi, I Shafik, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, et al. Niftynet: a deep-learning platform for medical imaging. Computer methods and programs in biomedicine, 158:113–122, 2018.

[65] Jean-Paul F. G. Marti, Louis-Francois Arnaud, Aurelien Astier, Yves Clavel, Paul J. Bruckman, and Sebastien Renault. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine, page 102636, 2021.

[66] Andrew Beers, James Brown, Ken Chang, Katharina Hoefel, Jay Patel, K Ina Ly, Sara M Tolaney, Priscilla Brastianos, Bruce Rosen, Elizabeth R Gerstner, et al. Deepneuro: an open-source deep learning toolbox for neuroimaging. Neuroinformatics, 19(1):127–140, 2021.

[67] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. Medical image analysis, 58:101552, 2019.

[68] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohnesh Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. Medical image analysis, 42(6):88–98, 2017.

[69] Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE transactions on medical imaging, 35(5):1153–1159, 2016.

[70] S Kevin Zhou, Hayit Greenspan, and Dinggang Shen. Deep learning for medical image analysis. Academic Press, 2017.

[71] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. Annual review of biomedical engineering, 19:221–248, 2017.

[72] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis, 54:280–296, 2019.

[73] Mohammad Hesam Hemayan, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging, 32(4):582–596, 2019.

[74] James S Duncan, Michael F Insana, and Nicholas Ayache. Biomedical imaging and analysis in the age of big data and deep learning [scanning the issue]. Proceedings of the IEEE, 108(1):3–10, 2019.

[75] Grant Haskins, Uwe Kruger, and Pingyun Kan. Deep learning in medical image registration: a survey. Machine Vision and Applications, 31(1):1–18, 2020.

[76] Nir Shlezinger, Jay Whang, Yonina C Eldar, and Alexandros G Dimakis. Model-based deep learning. arXiv preprint arXiv:2012.08405, 2020.

[77] Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. IEEE Signal Processing Magazine, 38(2):18–44, 2021.

[78] Gregory Ongie, Ajil Jalal, Christopher A Metzler, Richard G Baraniuk, Alexandros G Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. IEEE Journal on Selected Areas in Information Theory, 1(1):39–56, 2020.

[79] Rizwan Ahmad, Charles A Bouman, Gregory T Buzzard, Stanley Chan, Sizhao Liu, Edward T Reehorst, and Philip Schniter. Plug-and-play methods for magnetic resonance imaging: Using denoisers for image recovery. IEEE signal processing magazine, 37(1):105–116, 2020.

[80] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[81] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. arXiv preprint arXiv:2012.12556, 2020.

[82] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Afzaal, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. arXiv preprint arXiv:2101.01169, 2021.

[83] Yi Tai, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. arXiv preprint arXiv:2009.06732, 2020.

[84] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qi. A survey of transformers. arXiv preprint arXiv:2106.04554, 2021.

[85] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Chen-Hui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. arXiv preprint arXiv:2106.02034, 2021.

[86] Angelos Katharopoulos, Apoorv Vyas, Nikolas Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In International Conference on Machine Learning, pages 5156–5165. PMLR, 2020.

[87] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nystr(“) omformer: A nyst(“) om-based algorithm for approximating self-attention. arXiv preprint arXiv:2102.13902, 2021.

[88] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. arXiv preprint arXiv:2009.14794, 2020.
[89] Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, and Le Hou. Talking-heads attention. arXiv preprint arXiv:2003.02436, 2020.

[90] Qiangguo Jin, Zhaopeng Meng, Changming Sun, Hui Cui, and Ran Su. Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. Frontiers in Bioengineering and Biotechnology, 8:1471, 2020.

[91] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glover, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis, 53:197–207, 2019.

[92] Dhiraj Maji, Prarthana Sigedar, and Munendra Singh. Attention reus-net with guided decoder for semantic segmentation of brain tumors. Biomedical Signal Processing and Control, 71:103077, 2022.

[93] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. arXiv preprint arXiv:2111.07624, 2021.

[94] Danilo Avola, Luigi Cinque, Alessio Faggioli, Gianluca Foresti, and Alessio Mecca. Ultrasound medical imaging techniques: A survey. ACM Computing Surveys (CSUR), 54(3):1–38, 2021.

[95] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tadjakhs, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging, 39(6):1856–1867, 2020.

[96] Jieneng Chen, Yongli Lu, Qiangyu Xu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.

[97] Jiacheng Wang, Lan Wei, Liangsheng Wang, Qichao Zhou, Lei Zhu, and Jing Qin. Boundary-aware transformers for skin lesion segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 206–216. Springer, 2021.

[98] Yading Yuan. Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. arXiv preprint arXiv:1703.05165, 2017.

[99] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpken. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isci). arXiv preprint arXiv:1605.01397, 2016.

[100] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Ribeira. Ph 2-a dermoscopy image database for research and benchmarking. In 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pages 5437–5440. IEEE, 2013.

[101] Noel Codella, Veronica Rotemberg, Philipp Tschandl, Emre Celebi, Stephen Duszka, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isci). arXiv preprint arXiv:1902.03368, 2019.

[102] Huili Wu, Shihuai Chen, Guilian Chen, Wei Wang, Biaiyi Lei, and Zhenkun Wen. Fat-net: Feature adaptive transformers for medical image segmentation. arXiv preprint arXiv:1902.03368, 2019.

[103] Matt Berseth. Isic 2017-skin lesion analysis towards melanoma detection. arXiv preprint arXiv:1703.00523, 2017.

[104] Hui Gao and Oksam Chae. Individual tooth segmentation from ct images using level set method with shape and intensity prior. Pattern Recognition, 43(7):2406–2417, 2010.

[105] Yunxiang Li, Shuai Wang, Jun Wang, Guodong Zeng, Wenjun Liu, Qian Zhang, Qun Jin, and Yaqi Wang. Gear u-net: A u-net like group transformer network for tooth root segmentation. In International Workshop on Machine Learning in Medical Imaging, pages 386–393. Springer, 2021.

[106] Charles T Zahn and Ralph Z Roskies. Fourier descriptors for plane closed curves. IEEE Transactions on computers, 100(3):269–281, 2021.

[107] Yunxiang Li, Guodong Zeng, Yifan Zhang, Jun Wang, Qun Jin, Lingling Sun, Qian Zhang, Osi Lian, Guiping Qian, Neng Xia, et al. Agm-b transformer: Anatomy-guided multi-branch transformer network for automated evaluation of root canal therapy. IEEE Journal of Biomedical and Health Informatics, 2021.

[108] François Chollet. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1251–1258, 2017.

[109] Kaizhong Deng, Yanda Meng, Dongxu Gao, Joshua Bridge, Yaochun Shen, Gregory Lip, Yitian Zhao, and Yalin Zheng. Transunet: A lightweight transformer for left ventricle segmentation in echocardiography. In International Workshop on Advances in Simplifying Medical Ultrasound, pages 63–72. Springer, 2021.

[110] Qing-Long Zhang and Yu-Bin Yang. Sa-net: Shuffle attention for deep convolutional neural networks. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2239–2243. IEEE, 2021.

[111] David Guo, Cheng-He Hu, Amirata Ghorbani, Neal Yuan, Joseph Ebinguer, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. Nature, 580(7802):252–256, 2020.

[112] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. arXiv preprint arXiv:2103.03024, 2021.

[113] Qian Wang, Li Sun, Yan Wang, Meihou Zhou, Menghua Jin, Jianguang Chen, Ying Wen, and Qingli Li. Identification of melanoma from hyperspectral pathology image using 3d convolutional networks. IEEE Transactions on Medical Imaging, 40(1):218–227, 2020.

[114] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[115] Özcan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunori Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999, 2018.

[116] Boxiang Yun, Yan Wang, Jiengen Chen, Huiyu Wang, Wei Shen, and Qingli Li. Spectr: Spectral transformer for hyperspectral pathology image segmentation. arXiv preprint arXiv:2103.03604, 2021.

[117] Zhiqiang Shen, Hua Yang, Zhen Zhang, and Shaohua Zheng. Automated kidney tumor segmentation with convolution and transformer network. 2021.

[118] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[119] The 2021 kidney and kidney tumor segmentation challenge. https://kits21.kits-challenge.org/. Accessed: 2022-01-20.

[120] Hengshuang Zhao, Jianping Shi, Xiaoqian Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2881–2891, 2017.

[121] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 77–88, 2017.

[122] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017.

[123] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramidal attention network for semantic segmentation. arXiv preprint arXiv:1805.10180, 2018.

[124] Abbasish Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In 2017 IEEE Visual Communications and Image Processing (VCIP), pages 1–4. IEEE, 2017.

[125] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: U-net-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537, 2021.

[126] Xue Liu, Yutong Lin, Yue Cao, Han Hu, Yuxian Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In ICCV, 2021.

[127] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. arXiv preprint arXiv:2105.05633, 2021.
[128] Jeya Maria Jose Valaranasu, Poojan Oza, Iker Hachihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. arXiv preprint arXiv:2102.10662, 2021.

[129] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08524, 2021.

[130] Yao Jun Kim, Hyungjoong Jang, Kyoungbun Lee, Seongkeun Park, Sung-Gyu Min, Chyeon Heong, Jeong Hwan Park, Kanggeun Lee, Jisoo Kim, Wonjae Hong, et al. Paip 2019: Liver cancer segmentation challenge. Medical Image Analysis, 67:101854, 2021.

[131] Cam Nguyen, Zuhayr Asad, and Yuankai Huo. Evaluating transformer based semantic segmentation networks for pathological image segmentation. arXiv preprint arXiv:2108.11993, 2021.

[132] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European Conference on Computer Vision, pages 213–229. Springer, 2020.

[133] Tim Frangiemeier, Christoph Reich, and Heinz Koeppel. Attention-based transformers for instance segmentation of cells in microstructures. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 700–707. IEEE, 2020.

[134] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 29–38, 2017.

[135] Bert Van den Bergd, Nadia Zakaria, Bianca Adam, Steffi Matthyssen, Karina Koppen, and Sorcha Ni Dhunghailigh. Corneal endothelial cells over the past decade: are we missing the mark (er)? Translational vision science & technology, 8(6):13–13, 2019.

[136] Yinglin Zhang, Risa Higashita, Huazhu Fu, Yanwu Xu, Yang Zhang, Haofeng Liu, Jian Zhang, and Jiang Liu. A multi-branch hybrid transformer network for corneal endothelial cell segmentation. arXiv preprint arXiv:2106.07579, 2021.

[137] Alfredo Ruggeri, Fabio Scarpa, Massimo De Luca, Christian Meltendorf, and Jan Schroeter. A system for the automatic estimation of morphometric parameters of corneal endothelium in alizarine red-stained images. British Journal of Ophthalmology, 94(1):63–67, 2010.

[138] Wensuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiayun Li. Transbts: Multimodal brain tumor segmentation using transformer. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 109–119. Springer, 2021.

[139] Qiran Jia and Hai Shu. Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation. arXiv preprint arXiv:2109.12271, 2021.

[140] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Billello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314, 2021.

[141] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wossenral, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486, 2018.

[142] Louise Lam and SY Suen. Application of majority voting to pattern recognition: an analysis of its behavior and performance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 27(5):553–568, 1997.

[143] Himash Peiris, Munawar Hayat, Zhaolin Chen, Gary Egan, and Mehrtash Harandi. A volumetric transformer for accurate 3d tumor segmentation. arXiv preprint arXiv:2111.13300, 2021.

[144] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Lian-Sheng Wang, and Yizhou Yu. nnformer: Interleaved transformer based semantic segmentation networks for pathological image transformers. arXiv preprint arXiv:2109.12271, 2021.

[145] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers and cnns for medical image segmentation. In 39th International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 4011–4014. IEEE, 2021.

[146] Qinghua Huang, Yanzhi Wang, and Qianzi Zhang. Breast ultrasound image segmentation: a survey. International journal of computer assisted radiology and surgery, 12(3):493–507, 2017.

[147] Yiyao Liu, Huoji Han, Qiang Wang, Jiangbo Yao, Di Ou, Ding Xu, et al. Region aware transformer for automatic breast ultrasound tumor segmentation. 2021.

[148] Yiyao Liu, Yi Yang, Wei Jiang, Tianfu Wang, and Baiying Lei. 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), pages 4011–4014. IEEE, 2021.

[149] Oguzin Cicek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention, pages 424–432. Springer, 2016.

[150] Yang Lei, YaBo Fu, Tonghe Wang, Richard LJ Qiu, Walter J Curran, Tian Liu, and Xiaofeng Yang. Deep learning in multi-organ segmentation. arXiv preprint arXiv:2001.10619, 2020.

[151] Davood Karimi, Serge Vasylenko, and Ali Gholipour. Convolution-free medical image segmentation using transformers. arXiv preprint arXiv:2102.13645, 2021.

[152] Faoran Dou, Davood Karimi, Caitlin K Rollins, Clemence Leveque, Benjamim Bousmalis, and Eldad Gonen. Bdealhakim Oualam, Xin Yang, Dong Ni, and Ali Gholipour. A deep attentive convolutional neural network for automatic cortical plate segmentation in fetal mri. IEEE transactions on medical imaging, 40(4):1123–1133, 2020.

[153] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cerve- narsky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE transactions on medical imaging, 37(11):2514–2525, 2018.

[154] Synapse multi-organ segmentation dataset. https://www.synapse. org/#Synapse:syn3193805/wiki/217789. Accessed: 2022-01-20.

[155] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. arXiv preprint arXiv:2102.08005, 2021.

[156] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[157] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In European Conference on Computer Vision, pages 108–126. Springer, 2020.

[158] Puyang Wang, Nick G Cuccolo, Rachana Tyagi, Iker Hachihaliloglu, and Vishal M Patel. Automatic real-time cnn-based neonatal brain ventricles segmentation. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pages 716–719. IEEE, 2018.

[159] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojian Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. 3d gland segmentation in colon histology images: The glas challenge contest. Medical image analysis, 55:489–502, 2017.

[160] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Esfattatis Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. IEEE transactions on medical imaging, 39(5):1380–1391, 2019.

[161] Yucheng Tang, Dong Yang, Wengi Li, Holger Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. arXiv preprint arXiv:2111.14791, 2021.

[162] Ikboljon Sobirov, Otabek Nazarov, Hussain Alasmawi, and Mohammad Yaqub. Automatic segmentation of head and neck tumor: How powerful transformers are? arXiv preprint arXiv:2201.06251, 2022.

[163] Yao Chang, Hu Menghan, Zhai Guangtao, and Zhang Xiao-Ping. Transclaw u-net: Claw u-net with transformers for medical image segmentation. arXiv preprint arXiv:2107.05188, 2021.

[164] Chang Yao, Jingyu Tang, Menghan Hu, Yue Wu, Wenyi Guo, Qingli Li, and Xiao-Ping Zhang. Claw u-net: A unet-based
network with deep feature concatenation for scleral blood vessel segmentation. arXiv preprint arXiv:2010.10163, 2020.

[167] Ben Graham, Alaaeddin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet’s clothing for faster inference. arXiv preprint arXiv:2201.01136, 2021.

[168] Guoping Xu, Xirong Wu, Xuan Zhang, and Xinwei He. LeViT-UNet: make faster encoders with transformer for medical image segmentation. arXiv preprint arXiv:2107.08623, 2021.

[169] Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and David Zhang. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. arXiv preprint arXiv:2107.05274, 2021.

[170] Xiangyi Yan, Hao Tang, Shalin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation. arXiv preprint arXiv:2110.10403, 2021.

[171] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Kevyn Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjorn Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063, 2019.

[172] Xuming Chen, Shalin Sun, Narisai Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang, Chupeng Zhang, Zhipeng Lu, Qian Hu, Qing Chen, et al. Deep-learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiology and Oncology, 160:175–184, 2021.

[173] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In 2020 10th International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1–6. IEEE, 2020.

[174] Sinong Wang, Belinda Z Li, Madián Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768, 2020.

[175] Víctor M Campello, Fábioxeni Gkontra, Cristian Izquierdo, Carlos Martín-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiquang Hu, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE Transactions on Medical Imaging, 40(12):3543–3554, 2021.

[176] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, and Guangming Lu. Ds-transunet: Dual swim transformer u-net for medical image segmentation. arXiv preprint arXiv:2106.06716, 2021.

[177] Yijiang Li, Wenshan Cai, Ting Gao, and Xipu Hu. More than encoder: Introducing transformer decoder to upsampe. arXiv preprint arXiv:2106.10637, 2021.

[178] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Yijiang Li, Wentian Cai, Ying Gao, and Xiping Hu. More than a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules. American Journal of Roentgenology, 174(1):71–74, 2000.

[179] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Damián Gómez, Carina Rodríguez, and Fernando Villalón. Win-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics, 43:99–111, 2015.

[180] Nima Tajbakhsh, Suryakant R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging, 35(2):630–644, 2015.

[181] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michel Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering, 2017, 2017.

[182] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery, 9(2):283–293, 2014.

[183] Zhuzhuang Zhang, Baozhou Sun, and Weixiong Zhang. Pyramidal medical transformer for medical image segmentation. arXiv preprint arXiv:2104.14702, 2021.

[184] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nodule segmentation for computational pathology. IEEE transactions on medical imaging, 36(7):1550–1560, 2017.

[185] Multimodal brain tumor segmentation challenge 2019. https://www.med.upenn.edu/cbica/brats-2019/. Accessed: 2022-01-20.

[186] Multimodal brain tumor segmentation challenge 2020. https://www.med.upenn.edu/cbica/brats-2020/. Accessed: 2022-01-20.

[187] Olivier Petit, Nicolas Thomé, Clement Rambour, Loïc Temyry, Toby Collins, and Luc Soler. U-net transformer: self and cross attention for medical image segmentation. In International Workshop on Machine Learning in Medical Imaging, pages 267–276. Springer, 2021.

[188] Kenneth Clark, Bruce Vendt, Kirk Smith, Justin Kowalski, Paul Kopper, Stephen Moore, Snigdha Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging, 26(4):1045–1057, 2013.

[189] Vincent Andrearczyk, Valentijn Oreiller, Mario Freije, Martin Vallières, Joel Castelli, Hesham Elhalwani, Sarah Boughadad, John O Prior, and Adrien Depeursinge. Overview of the hecktor challenge. In Medical 2020: automated head and neck tumor segmentation in pet/cT Challenge, pages 1–21. Springer, 2020.

[190] Shaohua Li, Xiuqiao Sui, Xiongjie Lu, Xirui Xu, Yong Liu, and Rick Siow Mong Goh. Medical image segmentation using squeeze-and-expansion transformers. arXiv preprint arXiv:2105.09511, 2021.

[191] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepi R Bathula, Andrés Díaz-Pinto, Ruqiu Fan, Peng-Ann Heng, Jeyoung Kim, Jooho Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis, 59:101570, 2020.

[192] Peng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Franet: Parallel reverse attention network for polyp segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 263–273. Springer, 2020.

[193] Juan C Cai, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanne Ackerman, Marzieh Haghighi, Cher Keng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. Nature methods, 16(12):1247–1253, 2019.

[194] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsunoe Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Matite Matsui, Hiroshi Fujita, Yoshide Kodera, and Kunio Doi. 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. American Journal of Roentgenology, 174(1):71–74, 2000.

[195] Stefan Jaeger, Sema Candemir, Sameer Antani, Yi-Xiang J Wang, Stefan Mengersen, and Rainer W Diefenbach. Segthor: Segmentation of thoracic organs at risk in ct images. In Proceedings of the 2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pages 1–8. IEEE, 2014.

[196] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In European conference on computer vision, pages 519–534. Springer, 2016.

[197] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In Multi-label Image Segmentation. Springer, 2018.
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12416–12425, 2020.

[203] B. Landman, Z. Xu, J. Eugenio Igelisias, M. Syner, T. Langerak, and A. Klein. Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In Proc. MICCAI Multi-Atlas Labeling Workshop Proceedings, 2019.

[204] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 764–773, 2017.

[205] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenaoh Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyou Sun, and Liming Xia. Correlation of chest ct and rtpcr testing for coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. Radiology, 296(2):E32–E40, 2020.

[206] Yicheng Fang, Huangqi Zhang, Ticheng Xie, Minjie Lin, Lingqiu Ying, Peipei Pang, and Wenbin Ji. Sensitivity of chest ct for covid-19: comparison to rtpcr. Radiology, 296(2):E115–E117, 2020.

[207] Dandan Chen, Xinqing Jiang, Yong Hong, Zhihui Wen, Shuqian Meng, and Xinhua Wei. Can chest ct features distinguish patients with negative from those with positive initial rtpcr results for coronavirus disease (covid-19)? American Journal of Roentgenology, 216(1):66–70, 2021.

[208] Shehan Perera, Srikar Adhikari, and Alper Yilmaz. Pocformer: A lightweight transformer architecture for detection of covid-19 using point of care ultrasound. arXiv preprint arXiv:2105.09913, 2021.

[209] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4510–4520, 2018.

[210] Jannis Born, Gabriel Brändle, Manuel Cossio, Marion Disdier, Julie Goulet, Jérémie Roulin, and Nina Wiedemann. Pocovnet: automatic detection of covid-19 from a new lung ultrasound imaging dataset (pocus). arXiv preprint arXiv:2004.12084, 2020.

[211] Joseph Paul Cohen, Paul Morrison, Lan Dang, Karsten Roth, Tim Q Duong, and Marzyeh Ghamsemi. Covid-19 image data collection: Prospective predictions are the future. arXiv preprint arXiv:2006.11988, 2020.

[212] Chenguang Liu and Qingshan Yin. Automatic diagnosis of covid-19 using a tailored transformer-like network. In Journal of Physics: Conference Series, volume 10, page 012175. IOP Publishing, 2020.

[213] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlier locator. arXiv preprint arXiv:2106.13112, 2021.

[214] Muhammad EH Chowdhury, Tawsifur Rahman, Amithb Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Tao Xue, Q. Emadi, et al. Can we help in screening covid-19 pandemic? European association of emergency physicians. 8:132665–132676, 2020.

[215] Juntao Jiang and Shuyi Lin. Covid-19 detection in chest x-ray images using swin-transformer and transformer in transformer. arXiv preprint arXiv:2110.08427, 2021.

[216] Kai Han, An Xiao, Enhua Wu, Jinyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. arXiv preprint arXiv:2103.00112, 2021.

[217] Qi Dou, Tiffany Y So, Meirui Jiang, Quande Liu, Vaneet Vatsalabhati, Georgios Kassissis, Ziej Li, Weixin Si, Heather HC Lee, Kevin Yu, et al. Federated deep learning for detecting covid-19 lung abnormalities in ct: a privacy-preserving multinational validation study. NPJ digital medicine, 4(1):1–11, 2021.

[218] Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Song Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, and Jong Chul Ye. Vision transformer for covid-19 chest x-ray classification using chest x-ray feature corpus. arXiv preprint arXiv:2103.07055, 2021.

[219] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020.

[220] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silivana Ciurea-Ilicus, Chris Chute, Henrik Marklund, Behzad Haghasi, Robyn Bat, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence, volume 35, pages 3993–3997, 2021.

[221] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 782–791, 2021.

[222] Xiaohong Gao, Yu Qian, and Alice Gao. Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models. arXiv preprint arXiv:2102.07535, 2021.

[223] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708, 2017.

[224] Hayden Gunraj, Ali Sabri, David Vogel, and Alexander Wong. 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, 2021.

[225] Debaditya Shome, T Kar, Sachi Nandan Mohanty, Prayag Tewari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang, and Abdul Khader Jilani Saudagar. Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. International Journal of Environmental Research and Public Health, 18(11):586, 2021.

[226] Zhuchen Shao, Hao Bian, Yang Chen, Yifei Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based multiple instance learning for whole slide image classification. arXiv preprint arXiv:2106.00908, 2021.

[227] Zhenchao Xiao, Kai Yuan, and Hang Sun. Transmed: Transformers advance multi-modal medical image classification. Diagnostics, 1(8):1384, 2021.

[228] Mengkang Lu, Yongseng Pan, Dong Nie, Feihong Liu, Feng Shi, Yong Xia, and Dinggang Shen. SMILE: Sparse-attention based multiple instance contrastive learning for glioma subtype classification using pathological images. In MICCAI Workshop on Computational Pathology, pages 159–169. PMLR, 2021.
[322] Shirin Hajeb Mohammad Alipour, Hossein Rabbani, and Mohammad Reza Akhlaghi. Diabetic retinopathy grading by digital curvelet transform. Computational and mathematical methods in medicine, 2012, 2012.

[323] Nicolae-Catalin Ristea, Andreea-Iuliana Miron, Olivian Savencu, Mărgărit-Georgescu Mariana-Georgica, Nicolae Verga, Fahad Shahzad Khan, and Radu Tudor Ionescu. Cytran: Cycle-contrast transformers for non-contrast to contrast ct translation. arXiv preprint arXiv:2102.06400, 2021.

[324] Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Isgum. Deep mr to ct synthesis using unpaired data. In International workshop on simulation and synthesis in medical imaging, pages 14–25. Springer, 2017.

[325] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: Residual vision transformers for multi-modal medical image synthesis. arXiv preprint arXiv:2106.16031, 2021.

[326] Björn H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Sloothoo, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging, 34(10):1993–2024, 2014.

[327] Tufv Nyholm, Sfinna Svensson, Sebastian Andersson, Joakim Jonsson, Maja Sohlin, Christian Gustafsson, Elisabeth Kjellén, Karin Söderström, Per Albertsson, Lennart Blomqvist, et al. Mr and ct data with multiserver delineations of organs in the pelvic area—part of the gold atlas project. Medical physics, 43(5):1295–1300, 2018.

[328] Zhengwei Wang, Qi She, and Tomas E Ward. Generative adversarial networks in computer vision: A survey and taxonomy. ACM Computing Surveys (CSUR), 54(2):1–38, 2021.

[329] Fakhre Alam, Sami Ur Rajabian, Sehat Ullah, and Kamal Gulati. Medical image registration in image guided surgery: Issues, challenges and research opportunities. Bioinformatics and Biomedical Engineering, 38(1):71–89, 2018.

[330] Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468, 2021.

[331] M Milletari, N Navab, and SA Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. june 2016, 2018.

[332] Junyu Chen, Yong Du, Yufan He, William P Segars, Ye Li, and Eric C Frey. Transmorph: Transformer for unsupervised medical image registration. arXiv preprint arXiv:2111.10480, 2021.

[333] Yungeng Zhang, Turu Pei, and Hongbin Zha. Learning deep transformer network for differentiographic registration. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 129–138. Springer, 2021.

[334] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 2016.

[335] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. Improving factual completeness and consistency in image-to-text radiology report generation. arXiv preprint arXiv:2010.10042, 2020.

[336] Kirshore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318, 2002.

[337] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4566–4575, 2015.

[338] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019.

[339] Jinglu Zhang, Yinyu Nie, Jian Chang, and Jian Jun Zhang. Surgical instruction generation with transformers. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 290–299. Springer, 2021.

[340] Edgar Rojas-Muñoz, Kyle Couperus, and Juan Wachs. Dasi: Database for ai surgical instruction. arXiv preprint arXiv:2004.02809, 2020.

[341] Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Edgar Rojas-Muñoz, Kyle Couperus, and Juan Wachs. Hierarchical x-ray report generation via pathology tags and multi head attention. In Proceedings of the Asian Conference on Computer Vision, 2020.

[342] Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Zhong, Li Cai, Mengke Li, et al. Fla-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), 2021.

[343] Aurelia Bustos, Antonio Pertusa, Jose-María Salinas, and Maria de la Iglesia-Vayas. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis, 66:101797, 2020.

[344] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 6666–6673, 2019.

[345] RVPJH Kälviäinen and H Uusitalo. Diaretdb1 diabetic retinopathy database and evaluation protocol. In Medical Image Understanding and Analysis, volume 2007, page 61. Citeseer, 2007.

[346] Mingjie Li, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. arXiv preprint arXiv:2006.03744, 2020.

[347] Jia-Hong Huang, C-H Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Lu, I Lin, Kang Wang, Hiromasa Morikawa, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 2016.
Herrnghua Chang, et al. Deepoph: 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, pages 2482–2492, 2021.

[361] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating lesions in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical imaging, 19(3):203–210, 2000.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. Weakly supervised contrastive learning for chest x-ray report generation. arXiv preprint arXiv:2109.12242, 2021.

Ayoub Benali Amjoud and Mustapha Amrouch. 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), pages 1–5. IEEE, 2021.

[362] Benjamin Hou, Georgios Kaisiss, Ronald M Summers, and Bernhard Kainz. Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 293–305. Springer, 2021.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056, 2020.

Farzad Nooralehzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krumhaufler. Progressive transformer-based generation of radiology reports. arXiv preprint arXiv:2102.09777, 2021.

Yixin Wang, Zihao Lin, Jiang Tian, Zhongchao Shi, Yang Zhang, Jianping Fan, and Zhiquan He. Confidence-guided radiology report generation. arXiv preprint arXiv:2106.10887, 2021.

Hoang TN Nguyen, Dong Nie, Tavanbat Badamdorj, Youjie Liu, Yingying Zhu, Jason Truong, and Li Cheng. Automated generation of accurate & fluent medical x-ray reports. arXiv preprint arXiv:2102.12126, 2021.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. Auto-encoding knowledge graph for unsupervised medical report generation. Advances in Neural Information Processing Systems, 34, 2021.

Justin Lovelace and Bobak Mortazavi. Learning to generate clinically coherent chest x-ray reports. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 1235–1243, 2020.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In Proceedings of the sixth workshop on statistical machine translation, pages 85–91, 2011.

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

Omar Alanghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. Automated radiology report generation using conditioned transformers. Informatics in Medicine Unlocked, 24:100557, 2021.

Zachary M Ziegler, Luke Melas-Kyrizai, Sebastian Gehrmann, and Alexander M Rush. Encoder-agnostic adaptation for conditional language generation. arXiv preprint arXiv:1908.06938, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 18(9), 2019.

Joshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48, 2009.

Esha Pahwa, Dwij Mehta, Sanjida Kapadia, Devansh Jain, and Achleshwar Luthra. Medskip: Medical report generation using skip-connection and internal attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3409–3415, 2021.

Ke Sun, Bin Xiao, Dong Liu, and Jindong Wang. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5699–5708, 2019.

Hyeeryun Park, Kyungmo Kim, Seongkeun Park, and Jinwook Choi. Medical image captioning model to convey more details: Methodological comparison of feature difference generation. IEEE Access, 9:150560–150568, 2021.

Obioma Pelka, Sven Kotika, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 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, pages 183–189. Springer, 2018.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195, 2017.

Alba Garcia Seco De Herrera, Carsten Eickhof, Vincent Andreczky, and Henning Müller. Overview of the imageclef 2018 caption prediction tasks. CEUR Workshop Proceedings, 2018.

Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4025, 2021.

Elly Kipkogei, Arango Argoty GA, Ioannis Kagiampakis, Arijit Patra, and Etai Jacob. Explainable transformer-based neural network for the prediction of survival outcomes in non-small cell lung cancer (nsclc). 2021.

Robert M Samstein, Chung-Han Lee, Alexander N Shoushtari, Matthew D Hellmann, Ronglai Shen, Yelena Y Janjigian, David A Barron, Ahmet Zehir, Emmet J Jordan, Antonio Omuoro, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nature genetics, 51(2):202–206, 2019.

Sedigheh Eslam, de Gerardo Melo, and Christoph Meinerz. Does clinical benefit visual question answering in the medical domain as much as it does in the general domain? arXiv preprint arXiv:2112.13906, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural image language supervision. arXiv preprint arXiv:2103.00020, 2021.

Itzik Malkiel, Gony Rosenman, Lior Wolf, and Talma Hendler. Pre-training and fine-tuning transformers for fmri prediction tasks. arXiv preprint arXiv:2112.05761, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheslagh Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733, 2020.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR, 2021.

Shahrokhf Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikealingam, Simon Kombliith, Ting Chen, et al. Big self-supervised models advance medical image classification. arXiv preprint arXiv:2101.05224, 2021.

Yen Nhu Truong Vu, Richard Wang, Niranj Balachandar, Can Liu, Andrew Y Ng, and Pranav Rajpurkar. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. arXiv preprint arXiv:2102.10663, 2021.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029, 2020.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wisst. On the interpretability of artificial intelligence in radiology and its applications and opportunities. Radiology: Artificial Intelligence, 2(3):e190043, 2020.

Amitoideep Singh, Sowrya Sengupta, and Vasudevan Lakshminarayan. Explainable deep learning models in medical image analysis. Journal of Imaging, 6(6):52, 2020.

Ke Yan, Youbao Tang, Adam P Harrison, Jinzheng Cai, Le Lu, and Jingdong Liu. Interpretative medical image classification with self-supervised anatomical embedding and prior knowledge. arXiv preprint arXiv:2101.06524, 2021.

Biraja Ghosal and Allan Tucker. Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. arXiv preprint arXiv:2003.10769, 2020.

Sneh Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. An attentive survey of attention models. 40
[389] Open access series of imaging studies. https://www.oasis-brains.org/. Accessed: 2022-01-20.

[400] Lung nodule analysis, 2016. https://luna16.grand-challenge.org/. Accessed: 2022-01-20.

[410] Diabetic retinopathy challenge. https://www.kaggle.com/c/diabetic-retinopathy. Accessed: 2022-01-20.

[420] To lower the barrier to entry to privacy preserving technology. https://github.com/OpenMined. Accessed: 2022-01-20.

[430] Ashkan Youselpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bhargadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. arXiv preprint arXiv:2109.12298, 2021.

[440] To lower the barrier to entry to privacy preserving technology. https://github.com/OpenMined. Accessed: 2022-01-20.

[450] Georgios Kaisiss, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima, Jason Mancuso, Friederike Jungmann, Marc-Matthias Steinborn, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence, 3(6):473–484, 2021.

[460] XiaoHua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. arXiv preprint arXiv:2106.04560, 2021.

[470] Hamid Tabani, Ajay Balasubramaniam, Shabbir Marzban, Elahe Arani, and Bahram Zonooz. Improving the efficiency of transformers for resource-constrained devices. In 2021 24th Euromicro Conference on Digital System Design (DSD), pages 449–456. IEEE, 2021.

[480] Bingbing Li, Santosh Pandey, Haowen Feng, Yanjun Lyu, Ji Li, Jieyang Chen, Mimi Xie, Lipeng Wan, Hang Liu, and Caiven Ding. Energy-efficient acceleration of vision transformers using fpga. In Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design, pages 175–180, 2020.

[490] Banru Wang, Zhanghao Wu, Zhijian Liu, Han Cai, Ligeng Zhu, Chuang Gan, and Song Han. Hat: Hardware-aware transformers for efficient natural language processing. arXiv preprint arXiv:2005.14187, 2020.

[500] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pages 308–318, 2016.

[510] Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. Advances in Neural Information Processing Systems, 34, 2021.

[520] Kimin Lee, Kibok Lee, Honglak Lee, and Jinsu Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems, 31, 2018.

[530] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. arXiv preprint arXiv:2106.03004, 2021.

[540] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334, 2021.

[550] Oliver Zhang, Jean-Benoit Delbrouck, and Daniel I Rubin. Out of distribution detection for medical images. In Uncertainty in Machine Learning in Medical Imaging, and Perinatal Imaging. Placental and Preterm Image Analysis, pages 102–111. Springer, 2021.

[560] Chengzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Shuai Yi, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. arXiv preprint arXiv:2106.07617, 2021.

[570] Jasper Hendricks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136, 2016.

[580] Rajat Koner, Poulamî Sînahmâhapatra, Karsten Roscher, Stephan Gümennmann, and Volker Tresp. Oodformer: Out-of-distribution detection transformers. arXiv preprint arXiv:2107.08876, 2021.

[590] Haoxuan Qu, Hossein Rahmani, Li Xu, Bryan Williams, and Jun Liu. Recent advances of continual learning in computer vision: An overview. arXiv preprint arXiv:2109.11369, 2021.

[600] Matthias Lenga, Heinrich Schulz, and Axel Saalbach. Continuing learning for domain adaptation in chest x-ray classification. In Medical Imaging with Deep Learning, pages 413–423. PMLR, 2020.

[610] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. arXiv preprint arXiv:2007.01434, 2020.