An Overview of Vision Transformers for Image Processing: A Survey

Ch.Sita Kameswari[1], Kavitha J[2], T. Srinivas Reddy[3], Balaswamy Chinthaguntla[4],
Senthil Kumar Jagatheesaperumal[5], Silvia Gaftandzhieva[6], Rositsa Doneva[7]
Department of Computer Science and Engineering (AI&ML), Keshav Memorial Institute of Technology, Hyderabad, India[1]
Department of Information Technology, BVRIT HYDERABAD College of Engineering for Women, Hyderabad, India[2]
Department of Electronics and Communication Engineering, Malla Reddy Engineering College, Secunderabad, India[3]
Department of Electronics and Communication Engineering, Sheshadri Rao Gudlavalleru Engineering College, Gudlavalleru, India[4]
Department of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Sivakasi 626005, India[5]
University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria[6, 7]

***Abstract*—Using image processing technology has become increasingly essential in the education sector, with universities and educational institutions exploring innovative ways to enhance their teaching techniques and provide a better learning experience for their students. Vision transformer-based models have been highly successful in various domains of artificial intelligence, including natural language processing and computer vision, which have generated significant interest from academic and industrial researchers. These models have outperformed other networks like convolutional and recurrent networks in visual benchmarks, making them a promising candidate for image processing applications. This article presents a comprehensive survey of vision transformer models for image processing and computer vision, focusing on their potential applications for student verification in university systems. The models can analyze biometric data like student ID cards and facial recognition to ensure that students are accurately verified in real-time, becoming increasingly vital as online learning continues to gain traction. By accurately verifying the identity of students, universities and educational institutions can guarantee that students have access to relevant learning materials and resources necessary for their academic success.**

Keywords*—Vision transformers; image processing; natural language processing; image

I. INTRODUCTION

In recent years, deep neural networks such as convolutional neural networks (CNNs) [1], recurrent neural networks (RNNs) [2], graph neural networks (GNNs) [3], and attention neural networks [4] have been widely applied to a variety of artificial intelligence (AI) tasks. In contrast to previous non-neural models, which relied heavily on hand-crafted features and statistical methods, neural models can automatically learn low-dimensional continuous vectors as task-specific features from data, avoiding the need for complex feature engineering. Despite the popularity of deep neural networks, many studies have discovered that one of their fundamental limitations is their data-hungry nature. Due to many parameters in deep neural networks, they are prone to overfitting and have poor generalization capacity without appropriate training data [5].

CNNs are a fundamental component of modern computer vision systems. The advantage of CNNs was that they eliminated the need for manually constructed visual elements instead of learning to execute tasks "end to end" from data. The CNNs minimize manual feature extraction, and the CNN architecture is optimized for images and can be computationally expensive. Recent arguments have claimed that need goes beyond convolutions to represent long-range relationships. These initiatives aim to enhance convolutional models with content-based interactions, such as self-attention and non-local means, to improve performance in various vision tasks [6]. Transformers [7] are models that focus entirely on the self-attention process to establish global dependencies between input and output, and they have dominated natural language modelling in recent years [8-9]. Transformers and their variations have been thoroughly explored and used in natural language processing tasks such as machine translation [10], light-weight transformers [11], dynamic mask attention networks [12], language modelling [13], routing transformers [14], positional encoding schemes [15], and named entity identification [16-17]. The contrasts in size of visual elements and the high quality of pixels in images compared to words in text provide challenges in converting transformer from language to vision. The standard transformer is intended to process sequence data and is expected to receive a 1D series of token embedding. Many applications, including video understanding [18], image recognition [19], image super-resolution [20], object detection [21], segmentation [22], text-image synthesis [23] and visual question-answering [24], have been successfully implemented using transformer models and their variants in a variety of fields.

The survey in [25] explores recent advancements in visual transformers, an architecture originally designed for natural language processing but increasingly applied in computational visual media. The survey categorizes visual transformers based on task scenarios and analyzes their key ideas, with a particular focus on low-level vision and generation. The study reviews in detail backbone design approaches, offers quantitative comparisons, showcases image results, and includes information on computational costs and source code links to facilitate future development. In another recent survey by Jamil

et al. [26], the authors presented the first application of ViTs in -computer vision, providing an overview of their usage and performance in various applications such as image classification, object detection, segmentation, compression, super-resolution, denoising, and anomaly detection, along with a comprehensive analysis of existing models, insights, and future research directions. Liu et al. [27] provided a comprehensive review of over one hundred visual transformers, attention-based encoder- decoder models inspired by the Transformer architecture in computer vision. It analyzes their effectiveness in fundamental tasks (classification, detection, segmentation) and different data stream types, presents a taxonomy to organize the methods, evaluates and compares them under various configurations, identifies unexploited aspects for further improvement, and suggests three promising research directions for future development. Subsequently, the survey [28] examines the advancements and trends in utilizing Transformers for video modeling, addressing their limitations with inductive biases and scalability. It analyzes how videos are handled at the input level, architectural modifications to enhance efficiency and capture temporal dynamics, various training regimes, self- supervised learning strategies, and provides a performance comparison against 3D ConvNets, demonstrating the superior performance of Video Transformers in action classification with reduced computational complexity.

Additional work in this approach may aid in a better understanding of Transformer models and detecting any ererroneous behaviour or biases in the decision-making process. Since Transformer designs do not incorporate inductive biases (previous knowledge) to deal with visual input, transformers generally require a substantial quantity of training data in pre-training to determine the underlying modality-specific rules [29]. Several neural network architectures are known, including CNN, RNN, and transformer. CNNs were once the standard [30] in the Computer Vision domain, but transformers are gaining popularity [29]. While CNNs may capture inductive biases such as translation equivariance and localization, Vision Transformer overcomes inductive bias through large-scale training. According to the existing research [31], CNNs excel at small datasets, whereas transformers excel at massive datasets. The following fundamental issue is whether to employ in future CNN or a transformer.

Fig. 1 shows the number of publications on different image processing techniques using vision transformers [40].

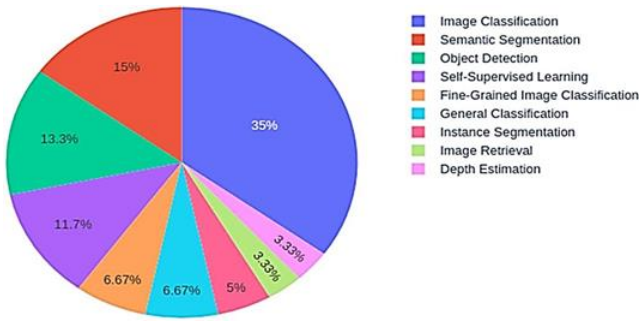

Fig. 1. The number of publications on different image processing techniques using vision transformers.

TABLE I. SUMMARY OF CONTRIBUTIONS FROM RECENT SURVEYS ON VISION TRANSFORMERS

Reference	Year	Scope	Contributions
Han et al. [32]	2022	General overview	Provides a comprehensive introduction to Vision Transformers
Chen et al. [33]	2021	Image classificati on	Focuses on the application of Vision Transformers for image classification tasks
Jamilet al. [26]	2023	General overview	Offers an in-depth analysis of Vision Transformers in various domains
Selvaet al. [28]	2023	Recent advancem ents	Highlights the latest research trends and advance ments in Vision Transformers
Gehrig et al. [34]	2023	Object detection	Discusses the utilization of Vision Transformers for object detection tasks
Zhai et al. [35]	2022	NLP to computer vision transition	Explores the adaptation of Vision Transformers from natural language processing to computer vision
Yang et al. [36]	2022	Comprehe nsive review	Provides an extensive analysis of Vision Transformers and their applications -
Guo et al. [37]	2022	Compariso n with CNNs	Compares the performance and characteristics of Vision Transformers with CNNs
He et al. [38]	2022	Medical image analysis	Examines the use of Vision Transformers in the field of medical image analysis
Aleissaee et al. [39]	2023	Remote sensing ap plications	Surveys the application of Vision Transformers in remote sensing tasks

Table I summarizes significant contributions from the existing survey articles on vision transformers.

The contributions of this article are as follows:

- An overview of the background and preliminaries of vision transformers, widely used in natural language processing, and how they can be adapted for image processing.
- Discusses how vision transformers have been used for image classification and enhancement, which involves improving the quality of images by removing noise, enhancing contrast, and increasing resolution.
- Explores how vision transformers can be used for object detection, which is the process of identifying and locating objects within an image, and how they can achieve state-of-the-art performance on this task.
- Highlights the role of vision transformers in education and university systems, specifically in student verification, where they can automate the process of verifying student identities, making it faster and more accurate.
- Discusses how vision transformers can deal with multimodal tasks, where they can process and fuse information from multiple modalities, such as text, image, and audio.

The rest of this article is organized as follows. The second part introduces the background details of the transformer, and the third part explores the usage of the visual transformer variants. The fourth part throws light on multimodal variants of

using vision transformers. It also discusses the future research directions of visual transformers and the fifth part concludes the paper.

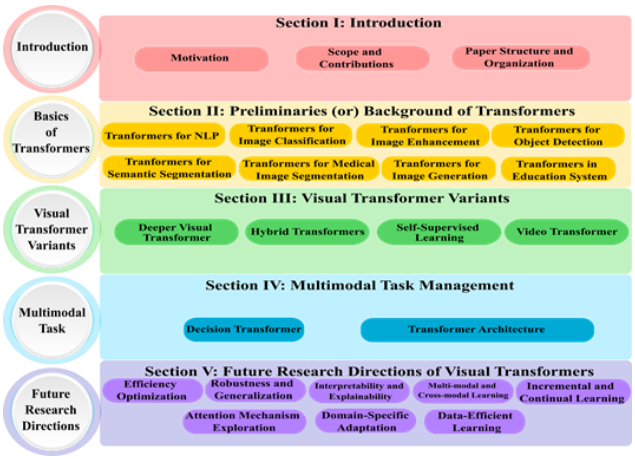

Fig. 2. Overall organization of the article.

Fig. 2 shows the overall organization of the sections presented in this article.

II. Preliminaries (or) Background of Transformers

The transformer is made of L layers, each of which has two major blocks: a Multi-Headed Self Attention (MSA) layer that performs a self-attention operation on various projections of the input tokens and a Multi-Layer Perceptron (MLA). Both the MSA and MLA layers are preceded by layer normalization and followed by a skip connection [41]. The attention mechanism was initially applied for 1-D data processing in natural language processing [42-43]. It has recently expanded to handle two-dimensional images and three-dimensional video data [44].

The fundamental components of a transformer include Multihead Self-Attention (MSA), Multi-Layer Perceptron (MLP), and Layer Normalization (LN) [7]. The authors [45] proposed the Gaussian Error Linear Unit (GELU) used to a great extent as one of the high-performing activation functions for neural networks. The work in [46] discusses the challenges of applying batch normalization to RNNs and introduces a new technique (called layer normalization) which addresses these challenges. Layer normalization computes mean and variance for normalization from all summed inputs to neurons in a layer on a single training case. It is effective in stabilizing hidden state dynamics in recurrent networks. It significantly reduces training time compared to previous techniques.

A. Transformers for NLP

In recent years, the transformer has evolved into a fundamental component of numerous cutting-edge natural language processing (NLP) models. Like RNN, the transformer is a robust performance model helpful for standard NLP applications such as intent identification in a search engine, text creation in a chatbot engine, and classification. The authors proposed a feed-forward network design that relies entirely on attention processes and avoids convolutions and recurrence. It achieved state-of-the-art performance on several tasks significantly and generalized exceptionally well to other NLP tasks, even with limited data. This design served as the foundation for numerous NLP models. GPT [47-49] and BERT [8] are two pioneering Transformer- based pre-trained models (PTMs) that employ autoregressive and autoencoding language modeling as pre-training objectives, respectively. Different Pre-trained models XLNet [50], RoBERTa [51], ALBERT [52], and T-NLG [53] are used in NLP tasks. Fig. 3 shows the structural difference between Transformer, GPT, and BERT [54].

Devlin et al. utilized the Transformer encoder (and only the encoder) to pre-train deep bidirectional representations from the unlabeled text. This pre-trained BERT model is fine-tuned with just one extra output layer to reach state-of-the-art performance for various NLP tasks without significant task-specific architectural changes. GPT [47] is a framework and training technique for natural language processing problems based on the Transformer architecture. The process for training is twofold. First, unlabeled data was used to learn the initial parameters of a neural network model using a language modeling aim. Then, using the associated supervised goal, these parameters are modified to a target task.

B. Vision Transformers for Image Classification

There have been many efforts to apply Transformers to vision tasks. These works are divided into two categories. The first category comprises models of pure attention. These models frequently use self-attention and strive to create convolution-free vision models. The second category encompasses networks developed using self-attention and convolutions [55]. Self-attention networks have revolutionized NLP and rapidly advanced image analysis tasks such as image classification and object recognition [56-57].

In computer vision, attention is employed in conjunction with or instead of CNN. This reliance on CNN is not required, as a pure transformer applied straight to sequences of image patches can do quite well on image classification tasks. The original text Transformer accepts a series of words as input and then uses them for classification, translation, or other natural language processing tasks. Dosovitskiy et al. [58] made the fewest feasible changes to the Transformer architecture to work directly on images rather than words for the vision transformer. Fig. 4 shows the architecture of the vision transformer.

Vision transformer generates a grid of square patches from an image. Each patch is converted to a single vector by concatenating the channels of all its pixels and then linearly projecting it to the chosen input dimension. Because transformers are structure-independent, they can add learnable position embedding to each patch, allowing the model to learn about the structure of the images. Vision transformer does not know the relative location of patches in the image or even if the image has a two-dimensional structure a priori. It must learn this information from training data and encode it in the position embeddings. Feed the sequence as an input to a state-of-the-art transformer encoder. Pre-train the vision transformer model with image labels, fully supervised then on an extensive dataset. Fine-tune the downstream dataset for image classification.

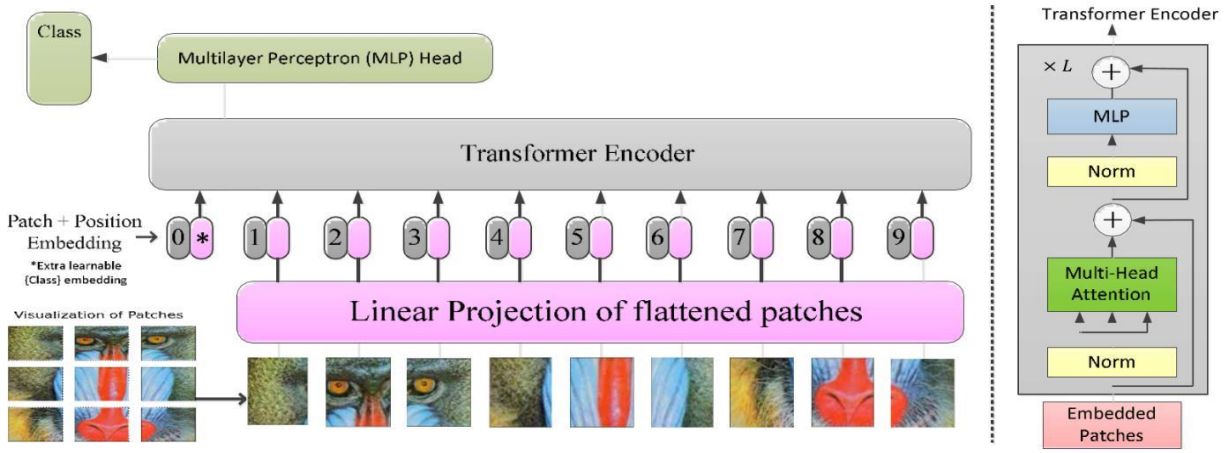

Fig. 3. Structure of transformer, GPT, and BERT.

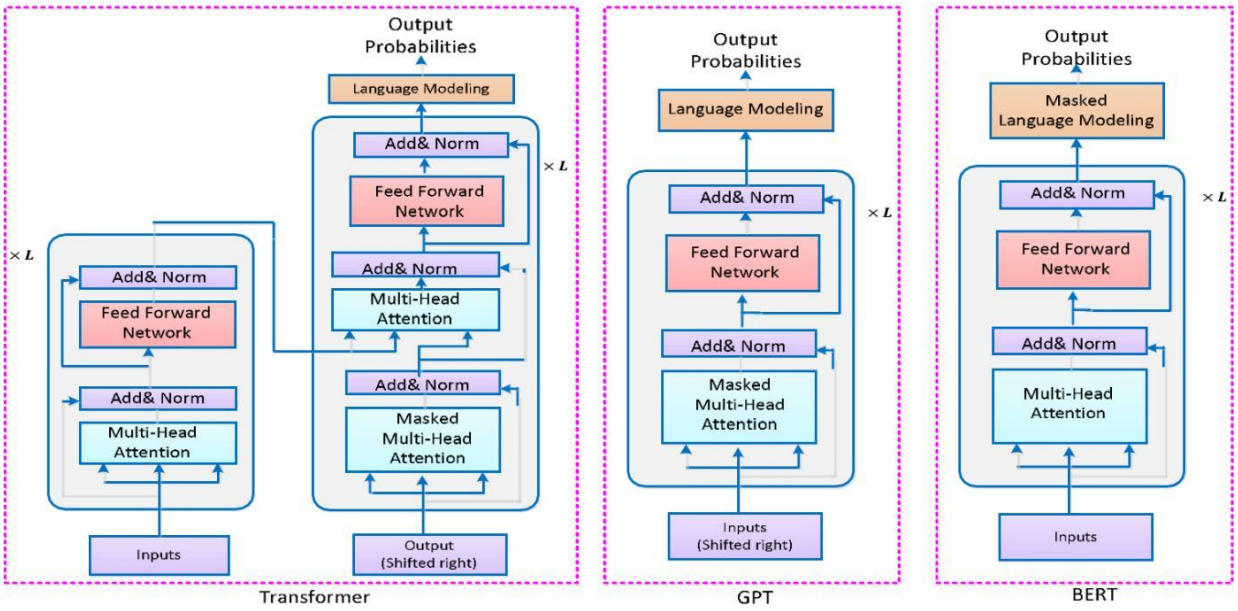

Fig. 4. Vision transformer architecture.

The Image GPT (iGPT) method [59] is an unsupervised generative pre-training technique for developing robust visual representations. By directly applying a GPT-2 [48] model to the image pixels, compelling image completions and samples were obtained, demonstrating that a completely Transformer-based architecture is viable for some visual tasks, regardless of the input image quality.

It does not need to prepare a large dataset to properly train the network in Data-efficient Image Transformers (DeiT) [6]. Instead, student-teacher setup and more intensive data augmentation and regularization are employed, such as stochastic depth [60] or repeated augmentation [61]. The teacher is a neural network designed to guide its student induction bias for convolutions [19].

LeViT is based on the architecture of the vision transformer [58] and the training technique of DeiT [19]. Regarding the speed/accuracy trade-off, LeViT considerably outperforms previous convents and vision transformers [62]. LeViT is five times faster than EfficientNet on the CPU at 80.

To improve image classification accuracy, Chen et al. [51] describe Cross Vision Transformer (CrossViT) [63], a dual-branch vision transformer learning multi-scale features. The proposed technique analyses separately small-patch and large-patch tokens using two distinct branches with varying computational costs. These tokens are subsequently merged numerous times repeatedly by attention to complement one another. It also created an efficient token fusion module based on cross-attention using a single token for each branch as a query to exchange information with other branches. Cross-attention needs linear time for computational and memory complexity when it usually requires quadratic time.

Transformer-iN-Transformer (TNT) [64] combines both patch-level and pixel-level representation by utilizing an outer Transformer block that processes patch embedding and an inner Transformer block that models the relation between pixel embedding.

C. Transformer for Image Enhancement

Chen et al. developed a pre-trained image processing model based on the transformer design, namely *Image Processing Transformer (IPT)* [65]. The IPT model features multiple heads, multiple tails, and a standard transformer body for performing various image processing tasks such as super-resolution and denoising. The IPT model was trained using supervised and unsupervised methods, demonstrating a significant capacity to capture intrinsic characteristics for low-level image processing. Experiments indicate that IPT can outperform state-of-the-art techniques using a single pre-trained model following a brief fine-tuning phase.

Yang et al. [20] proposed a novel *Texture Transformer Network for Image Super-Resolution (TTSR)* in which the low-resolution (LR) and high-resolution (Ref) images are expressed as queries and keys, respectively, in a transformer. TTSR is a collection of closely linked modules designed for image generation tasks, comprising a learnable texture extractor based on deep neural networks, a relevance embedding module, a hard-attention module for texture transfer, and a soft-attention module texture synthesis.

D. Transformer for Object Detection

Carion et al. [21] introduced *DEtection Transformer (DETR)* to eliminate the requirement for such hand-crafted components and developed the first fully end-to-end object detector with highly competitive performance. DETR is a basic architecture shown in Fig. 5 that combines CNNs with Transformer encoder-decoders [66]. They use Transformer's versatile and robust relation modelling capabilities to substitute hand-crafted rules when appropriately prepared training signals are used. DETR is a novel approach to object recognition based on transformers and bipartite matching loss for direct set prediction. Applied to the problematic COCO dataset, the method obtains results equivalent to an improved Faster R-CNN baseline. DETR is simple to construct and offers a modular design easily extendable to panoptic segmentation, resulting in competitive performance. Additionally, it outperforms Faster R-CNN on big objects, most likely because of the global information processing produced by self-attention.

However, it has its own range of difficulties. These difficulties are primarily due to the Transformer's attention deficiencies in handling image feature maps as essential elements: (1) DETR's ability to identify small objects is relatively poor. Modern object detectors use high-resolution feature maps to identify small objects more accurately. However, high-resolution feature maps would impose an excessive complexity level on the self-attention module of DETR's Transformer encoder, which scales quadratically with the spatial dimension of the input feature maps. (2) Compared to current object detectors, DETR takes more training epochs to converge. DETR is primarily due to the difficulty of training the attention modules that analyze visual characteristics.

Deformable (DETR) [21] remove the requirement for several handmade components in object detection while exhibiting acceptable performance. However, because of the limitations of Transformer attention modules in processing visual feature maps, it has a sluggish convergence rate and a restricted feature spatial resolution. To address these concerns, authors [67] suggested Deformable DETR, in which the attention modules focus exclusively on a limited number of critical sampling points surrounding a reference. Deformable DETR is a technique for object detection that seeks to address DETR's delayed convergence and high complexity problems. It combines the advantages of deformable convolution sparse spatial sampling with the relation modelling capability of transformers. Deformable DETR introduced a deformable attention module that uses a few sample sites as a pre-filter for conspicuous key components among all feature map pixels. Without relying on FPN, the module may be organically expanded to aggregate multi-scale characteristics. With ten fewer training epochs, deformable DETR can outperform DETR (particularly on tiny objects). Extensive trials on the COCO [68] benchmark validate this method.

Zheng et al. propose a novel transformer variation called the Adaptive Clustering Transformer (ACT) to reduce the computation cost associated with high-resolution input [69]. ACT uses Locality Sensitive Hashing (LSH) to cluster query characteristics adaptively and approximates the query-key interaction using the prototype-key interaction. ACT is capable of reducing the quadratic complexity inherent in self-attention.

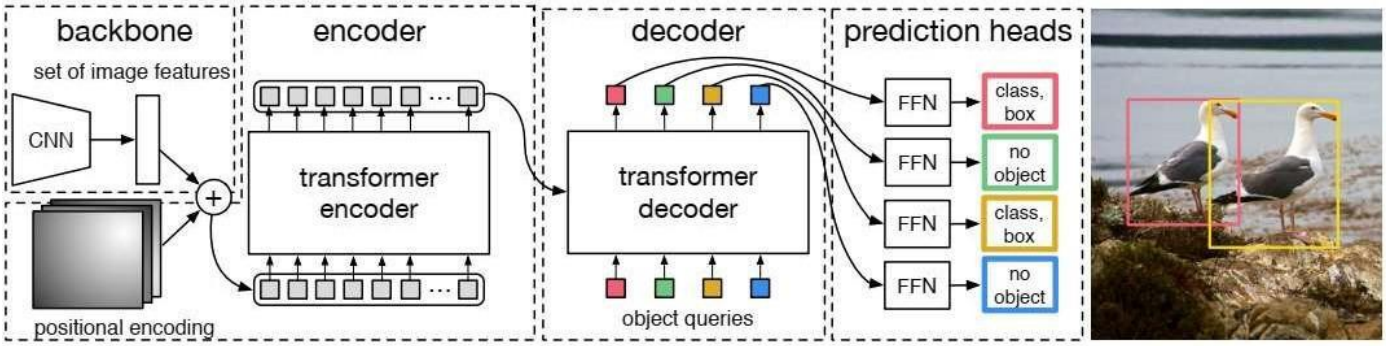

Fig. 5. DETR's general architecture. The image is from [21].

Inspired by the tremendous success of pre-training transformers in natural language processing, Dai et al. [70] proposed Unsupervised Pre-train DETR (UP-DETR) for object detection. The proposed UP-DETR model includes pre-training and fine-tuning procedures: (a) the transformers are unsupervised trained on a large-scale dataset without human annotations, and (b) the complete model is fine-tuned using labelled data, similar to the original DETR. Precisely clip random regions from the provided image and send them to the decoder as queries. Pre-trained on these query patches from the original image, the model detects them. The authors solve two critical difficulties during pre-training: multitask learning and multi-query localization. To balance classification and localization preferences in the pretext task, freeze the CNN backbone and propose a patch feature reconstruction branch optimized for conjunction with patch detection. (2) To accomplish multi-query localization, expand UP-DETR from single-query patches to multi-query patches by including object query shuffling and an attention mask. To expedite DETR's training convergence and prediction capability in object detection, Sun et al. [71] conduct extensive experiments and suggest two innovative methods, namely TSP-FCOS (transformer-based Set Prediction with FCOS) and TSP-RCNN (transformer-based Set Prediction with RCNN). These techniques converge considerably quicker than the original DETR and significantly outperform DETR and other baselines regarding detection accuracy

Beal et al. [72] developed ViT-FRCNN, a competitive object detection solution that uses a transformer backbone, implying that sufficiently distinct architectures from the well-studied CNN backbone are viable for advancement on complex vision problems. Transformer-based models have proven capacity to pre-train with large datasets without reaching saturation and fast fine-tune to different tasks, both observed with ViT-FRCNN.

The authors [73] suggested a novel variation of Vision Transformer models based on focal attention, called Focal Transformer that outperforms state-of-the-art (SoTA) vision Transformers on various publicly available image classification and object detection benchmarks.

Extracting strong feature representations is a significant issue in the re-identification of objects (ReID). Although techniques based on CNNs have gained considerable success, they analyze just one local area at a time and suffer from information loss due to convolution and down-sampling operators. To address these constraints, He et al. [74] introduced Transformer for Object re-identification (TransReID), a pure transformer-based object Reid framework. First, encode each image as a sequence of patches and then construct a transformer-based strong baseline with a few essential enhancements that obtain competitive performance on many Reid benchmarks using CNN-based techniques.

E. Transformer for Semantic Segmentation

Lie et al. [75] proposed a novel vision Transformer called Swin Transformer, a hierarchical Transformer with shifted windows used for general-purpose computer vision. To improve performance, offset windowing restricts self-attention computation to non-overlapping local windows while permitting cross-window connections. This hierarchical architecture can simulate various sizes and has a linear computational cost with image scalability. Swin Transformer used for image classification 86.4 accuracy on ImageNet-1K [76], dense prediction tasks including object identification 58.7 box AP on COCO, and semantic segmentation 53.5 mIoU on ADE20K [77].

Zheng et al. [78] introduced a sequence-to-sequence prediction framework for semantic segmentation. For the first time, authors have eliminated the need for FCN and solved a restricted receptive field problem, unlike current FCN-based techniques that use dilated convolutions and attention modules at the component level. This encoder may be coupled with a primary decoder to build a robust segmentation model called SEgmentation TRansformer (SETR). SETR uses a pure transformer (no convolution or resolution reduction) to encode an image as a patch sequence. MaX-DeepLab [79] is the first end-to-end model for panoptic segmentation that automatically infers masks and classes without the need for hand-coded priors such as object centres or boxes.

The Dense Prediction Transformer (DPT) [80] is a neural network design that successfully uses visual transformers for dense prediction problems. The monocular depth estimation and semantic segmentation tests demonstrate that the given architecture generates more fine-grained and globally coherent predictions than fully convolutional networks. As with previous work on transformers, when trained on large-scale datasets, the DPT reaches its full potential.

F. Transformer for Medical Image Segmentation

Segmentation of medical images is necessary for developing healthcare systems, particularly for disease diagnosis and treatment planning. The U-shaped architecture, commonly known as U-Net [81], achieved remarkable success in various medical image segmentation tasks. However, because convolution processes are intrinsically local, U-Net typically exhibits problems when representing long-range dependence clearly. To fully use the capabilities of Transformers, Chen et al. [82] presented TransUNet, which incorporates a fully global context by considering image features as sequences and effectively uses low-level CNN features via a U- shaped hybrid architectural design. Several experiments were conducted to evaluate the proposed TransUNet system and validate its performance in various scenarios, including 1) model scaling, 2) the number of skip-connections, 3) patch size and sequence length 4) input resolution. TransUNet outperforms many competing methods, including CNN-based self-attention methods, as an alternate framework to the current FCN-based systems for medical image segmentation. Fig. 6 shows the architecture of TransUNet.

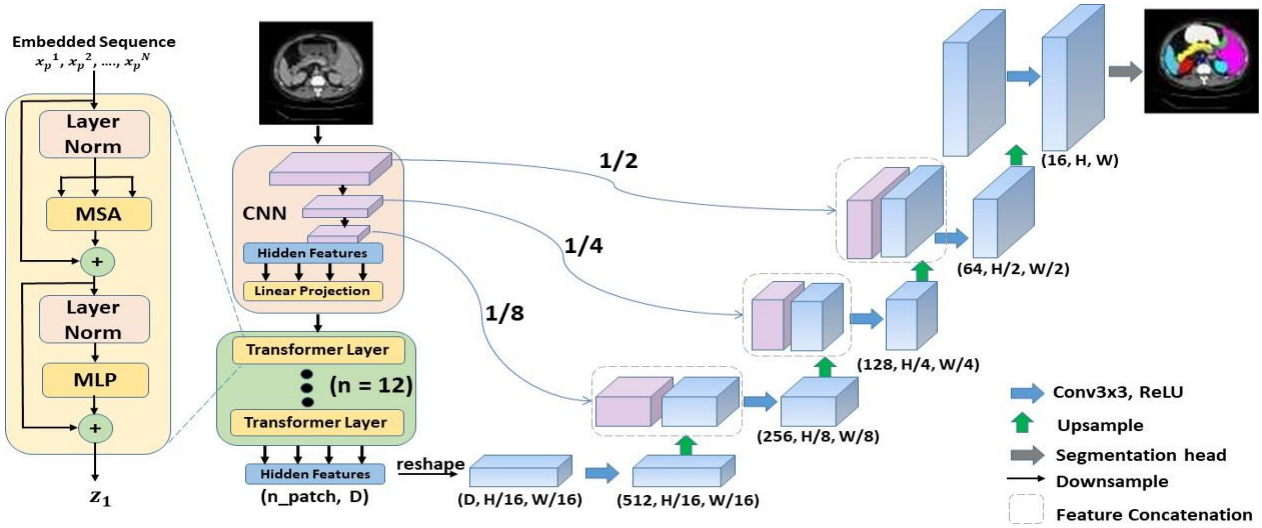

Fig. 6. An overview of the TransUNet architecture. Image credit [82].

Yun et al. [83] introduced Spectral Transformer (SpecTr), a method for segmenting hyperspectral pathology images that use transformers to learn contextual features across spectral bands. They applied two critical schemes to support context learning: (1) A sparsity strategy is used to learn context-dependent sparsity patterns and improve the model performance and interpretability. (2) A technique for spectral normalization is given that allows for independent normalizing of the feature map at each spectral location, therefore eliminating interference caused by distribution mismatches between spectral images. On a cholangiocarcinoma segmentation dataset, SpecTr was assessed. Experiments demonstrate the superiority of the suggested method for hyperspectral pathology image segmentation.

Valanarasu et al. [84] introduced MedT (Medical Transformer), an encoder that employs gated axial attention as its primary building block and is trained using the Local-Global training strategy (LoGo) approach.

The above equation is formulated by the attention model presented in [85], and q, K, k w×w is formulated by the width-wise axial attention model. The gated axial-attention model extends previous designs by incorporating a self-attention module with an extra control mechanism.

Guo et al. introduced a unique framework for point cloud learning called Point Cloud Trans- former (PCT) [86]. PCT is based on the transformer, which has achieved enormous success in natural language processing and has tremendous promise in image processing. Because it is intrinsically permutation invariant when processing a sequence of points, it is ideally suited for point cloud learning. The authors increase input embedding with the help of the furthest point sampling and closest neighbour search to better capture local context inside the point cloud. Extensive experimental evidence demonstrates that the PCT outperforms state-of-the-art shape classification, part segmentation, semantic segmentation, and normal estimation tasks.

G. Transformer for Image Generation

Inspired by CNN, Image Transformer [87] confines the self-attention receptive field to local regions. Image Transformer implements an encoder-decoder architecture in which the encoder creates a contextualized representation for each pixel-channel in the inputs, and the decoder generates one channel per pixel at each time step autoregressively.

Jiang et al. [88] suggest the first GAN completely using transformers without convolution. TransGAN has a novel grid self-attention mechanism, a memory-friendly generator and a multi-scale discriminator. These architectural components have been carefully developed to strike a compromise between memory efficiency, global feature statistics, and local fine details in the presence of spatial variations.

Lee et al. [89] incorporate the Vision Transformer architecture into adversarial generating networks (GANs). The authors discovered that conventional regularisation techniques for GANs had a poor interaction with self-attention, resulting in severe training instability. The ViTGAN model developed innovative regularisation strategies for training GANs with Vision Transformers to address the problem.

H. Role of Transformer in Education and University Systems

To confirm the identity of students, universities and other educational institutions typically require them to provide identification documents, but this process can be time-consuming and susceptible to errors. Vision Transformer can automate this process, resulting in faster and more accurate verification. To implement this approach, the educational institution can build a database containing student images and identification documents [90]. Then, the Transformer model can be trained on this database to learn how to identify students' faces in the images and match them to their identification documents. Students can capture the image with their mobile device or webcam in the verification process. The Transformer model can analyze the image to authenticate the student's identity. The model can compare the student's face in the image to the database of student images to verify that it corresponds to the identification document.

Moreover, such a transformer model can also detect fraudulent activity in student identification documents. The model can scrutinize the identification document and highlight any discrepancies or irregularities. For example, the model can identify if the photo on the identification document has been manipulated digitally or if the document has been tampered with in any way. This fact not only streamlines the process of verifying student identities, making it quicker and more precise but also ensures the reliability of the verification process by detecting any fraudulent activity in student identification documents.

III. Visual Transformer Variants

The development of the visual transformer has paved the way for significant advancements in computer vision. Since its inception, researchers have explored several variants of transformer architecture to further improve its performance on visual tasks. In this section, we discuss some of the notable variants of the visual transformer and their applications.

A. Deeper Visual Transformer

Zhou et al. found that in contrast to CNNs, the performance of vision transformers rapidly saturates as the number of convolutional layers increases. As the transformer progresses deeper, the attention maps become increasingly similar after a certain number of layers. The feature maps in the top layers of deep vision transformer models are often similar. It indicates that in the deeper layers of vision transformers, the self-attention mechanism cannot learn appropriate ideas for representation learning, preventing the model from achieving the predicted performance improvement. Zhou et al. [91] identified the problem of the vision transformers' attention collapsing as they progress deeper. They suggest a unique re-attention technique DeepViT to resolve it with the least amount of calculations and memory cost possible. Using Re-attention can sustain an improving performance when the depth of vision transformers increases.

The CaiT [92] network's operation involves two distinct processing phases. The first one, the self-attention stage, is similar to the vision transformer, except there is no class embedding. Second, a series of layers called the class-attention stage (CLS) compiles the patch embeddings into a class embedding CLS, given to a linear classifier.

Wang et al. [93] propose a Pyramid vision transformer (PVT), a pure Transformer backbone suitable for dense prediction applications like semantic segmentation and object detection without convolutions. The authors create a progressive pyramid shrinking algorithm and a spatial-reduction attention layer for obtaining multi-scale feature maps with minimal memory/computation resources. Extensive experimentation on semantic segmentation and object detection benchmarks demonstrates that PVT outperforms well-designed CNN when the parameters are equivalent. Fig. 7 compares the CNN architectures, the vision transformer, and the pyramid transform.

L. Yuan et al. [96] proposed a novel Token–to–Token Vision Transformer (T2T-ViT) model that can be trained entirely on ImageNet and attain performance equivalent to or better than CNN's. Using T2T-ViT, the image structure information is better modelled, and more features are provided. Thus T2T-ViT significantly exceeds the Vision Transformer features. It has a unique tokens-to-tokens (T2T) approach for tokenizing images incrementally and structurally aggregating tokens.

As a result of the improvements in computer vision and the enormous quantity of training data, many people feel Transformers are not appropriate for tiny datasets. The authors of this article [97] debunked the notion that transformers are data-hungry. The authors in [97] demonstrated that proposed Compact Convolution Transformers (CCT) can compete with state-of-the-art CNNs with appropriate data size and tokenization for the first time. Through a unique sequence pooling technique and convolutions, the suggested model eliminates the need for class tokens and positional embeddings.

Heo et al. [98] proposed a novel architecture called Pooling-based Vision Transformer (PiT) to use the pooling layers' advantages. The authors demonstrate that a commonly utilized design concept in CNN spatial dimensional transformation accomplished by pooling or convolution is ignored in transformer-based architectures, negatively affecting the model performance and the transformer architecture benefits from decreasing the spatial dimension. The authors initially examined ResNet and discovered that transforming it in terms of spatial dimension improves computing efficiency and generalization ability. To capitalize on the benefits of Vision Transformer, the authors proposed a PiT that integrates a pooling layer into Vision Transformer, and the PiT demonstrates that pooling layer benefits become effectively matched to Vision Transformer. As a result of considerably increasing the performance of the Vision Transformer architecture, the authors demonstrated that the pooling layer is critical for a self-attention-based design by considering the spatial interaction ratio. Moreover, extensive experiments showed that PiT outperforms the baseline on object detection, image classification, and robustness evaluation. Fig. 8 highlights the difference in dimensions of network architectures ResNet- 50, Vision Transformer, and PiT.

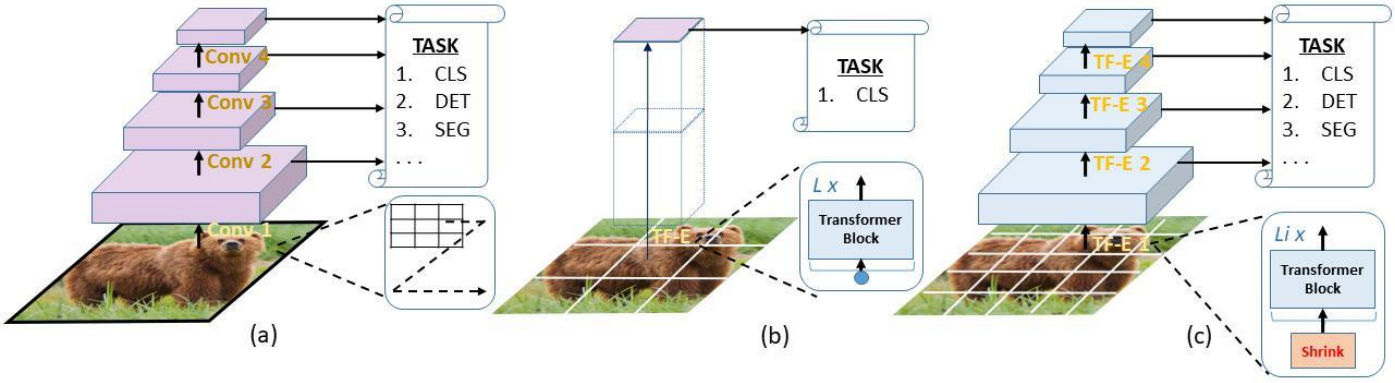

Fig. 7. (a) CNNs: ResNet [94], VGG [95], etc. (b) Vision transformer [58] (c) Pyramid vision transformer [93].

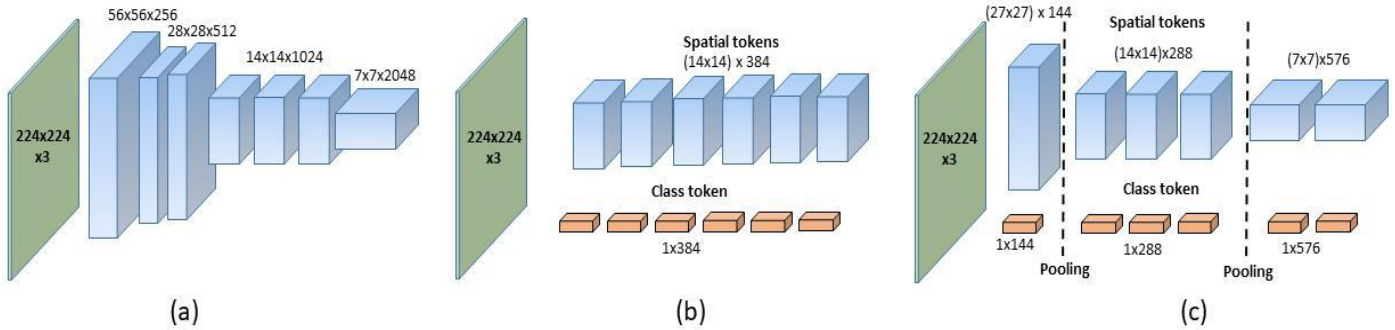

Fig. 8. A schematic diagram illustrates dimension variation in network architectures. (a) ResNet-50, (b)ViT-S/16, (c)PiT-S. image credit [98].

B. Hybrid Transformers

To improve image super-resolution, Z. Lu et al. [99] propose an *Efficient Super- Resolution Transformer (ESRT)*. ESRT is a hybrid Transformer that uses a CNN-based SR network to extract deep features. Backbones for the ESRT include the lightweight CNN (LCB) and lightweight Transformer (LTB). LCB is a low-cost SR network extracting deep SR features by dynamically changing the feature map's size. LTB consists of an efficient Transformer (ET) that consumes less GPU Memory space and benefits from the uniquely efficient multi-head attention (EMHA). The Reformer [100] is a Transformer model capable of processing context windows of up to 1 million words on a single accelerator with only 16GB of memory. Reformer combines two critical approaches for resolving the attention and memory allocation issues that limit the applicability of the transformer to lengthy context windows. The reformer uses locality-sensitive hashing (LSH) to decrease the complexity of attending to long sequences and reversible residual layers to maximize the usage of available memory.

H. Wu et al. [101] introduced a novel architecture, the *Convolutional Vision Transformer (CvT)* that enhances the performance and efficiency of the Vision Transformer by incorporating convolutions into the Vision Transformer. The author's findings indicate that positional encoding, a critical component of existing Vision Transformers, can be safely omitted from the CvT model, simplifying the design for higher-resolution vision applications.

Chu et al. [102] proposed a highly efficient and direct implementation of two architectures - *Twins-PCPVT* and *Twins-SVT* vision transformer designs. Twins-PCPT is based on PVT [93] and CPVT [103] and utilizes global attention. Twins-SVT is based on the proposed SSSA, consisting of two distinct attention operations: locally-grouped self-attention (LSA) and globally subsampled attention (GSA). Both transformer models established new benchmarks for image classification, semantic/instance segmentation, and object detection.

Zhang et al. introduced the *Nested Transformer (NesT)* [104]. The block aggregation function is essential for enabling non-local information transmission across blocks. The accuracy of a NesT trained on ImageNet for 100/300 epochs is 82.3 per cent compared to other techniques [19], [105] that achieved up to 57% parameter reduction. A NesT with 6M parameters trained from scratch on CIFAR10 [106] achieves 96% accuracy using a single GPU.

Visual Transformers (VT) [107] defines the problem in the semantic token space, intending to represent and process high-level concepts in images using visual tokens. Moreover, different parts of the image have different meanings due to their different content. Note that this is entirely different from the transformer that processes information in pixel space (such as Vision Transformer, DeiT, IPT, etc.) because the amount of calculation differs by multiple orders of magnitude. The author [107] uses the spatial attention mechanism to convert the feature map into compact semantic tokens. Then input these tokens into a Transformer, and use the unique functions of the transformer to capture the connection between the tokens. In this way, VT can 1) Focus on those relatively important areas instead of treating all pixels equally like CNN. 2) Encode

semantic concepts in visual tokens instead of modelling all concepts in all images. 3) Use the Transformer to model the relationship between tokens. The VT model is used in classification tasks (Model Base: ResNet, Dataset: ImageNet, reduced by 6.9 times, increased by 4.6- 7 Accuracy) and semantic segmentation tasks (Model Base: FPN, Dataset: LIP and COCO-stuff reduced by 6.4 times the amount of calculation, the increase point 0.35 mIoU) has achieved excellent performance.

C. Supervised Learning in Vision Transformer

Supervised learning enables the transformer to learn a bottleneck representation in which the content and context are mixed around the class token. This results in a relatively superficial data model, and its association with labels needs many training examples. On the other hand, unsupervised learning uses the information redundancy and complementarity inherent in image data by learning to rebuild local content through context integration [108].

In self-supervised learning, no concept whatsoever of labelled data for the training. Self-supervised techniques can be classified broadly as generative or discriminative [109]. Generative methods learn to predict the data distribution. However, data modelling is inherently computationally expensive and may not be required in all cases for representation learning. Discriminative methods, generally implemented in a contrastive learning framework [110] or through pretext tasks [111], have the capacity to create more generalized representations with minimal computing needs.

Auto et al. [112] proposed a *Self-supervised vision Transformer (SiT)*, a unique approach for learning visual representations without supervision. Using the autoencoder transformer's inherent capacity to perform multitask learning, it created a robust self-supervised system that optimizes reconstruction, rotation classification, and contrastive losses concurrently. The last utilizes the strength of the transformer to train SiT to perform three distinct tasks: image reconstruction, rotation prediction, and contrastive learning.

Bao et al. [113] introduced a self-supervised vision representation model *BEiT*, which stands for Bidirectional Encoder representation from image Transformers. The authors proposed a masked image modelling task to pre-train vision Transformers in a self-supervised manner. In pre-training, each image contains two perspectives - image patches and visual tokens. First, "tokenize" the original image into visual tokens. Then, using a random masking technique, feed specific image patches into the backbone Transformer. The purpose of the pre-training is to reconstruct the original visual tokens from the damaged image patches. After pre-training BEIT, fine-tune model parameters directly on downstream tasks by superimposing task layers on the pre-trained encoder. Experiments on image classification and semantic segmentation demonstrate that our model outperforms prior pre-training approaches. For example, base-size BEIT achieves 83.2% top-1 accuracy on ImageNet-1K with the same configuration, considerably surpassing DeiT training from scratch at 81.8% [19].

D. Video Transformer

Following the recent success of vision transformer models in image classification, Arnab et al. [114] presented pure-transformer-based video classification models *Video Vision Transformer (ViViT)*. To efficiently handle a high count of Spatiotemporal tokens, the authors in [114] constructed multiple model variations that factorize the transformer encoder's many components across spatial and temporal dimensions. The authors in [113] demonstrated how to use additional regularisation and pre-trained models to compensate for the fact that video datasets are often smaller than the image datasets on which Vision Transformer was trained.

The *VisTR*, a new video instance segmentation framework based on Transformers, considers the video in-stance segmentation (VIS) problem an end-to-end concurrent sequence decoding/prediction issue [115]. Fig. 9 shows the architecture of VisTR. The paradigm is qualitatively distinct from previous techniques, streamlining the whole process significantly. VisTR approaches the VIS problem from a novel similarity-based perspective. Segmentation was used to determine pixel-level similarity, whereas tracking was used to determine instance-to-instance similarity. Thus, tracking instances occurs naturally and smoothly in the instance segmentation context. VisTR's success is developing a novel technique, such as sequence matching and segmentation, optimized for the framework. This well-designed method enables monitoring and segmenting instances at the sequence level in their entirety. ViSTR is composed of four major components:1) a CNN backbone that extracts feature representations from multiple images, 2) an encoder-decoder Transformer that models the relationships between pixel-level and instance-level features and decodes them, 3) an instance sequence matching module that supervises the model, and 4) an instance sequence segmentation module that outputs the final mask sequences. VisTR outperforms other techniques that employ a single model on the YouTube-VIS dataset, reaching 40.1% in mask mAP at 57.7 frames per second.

Fan et al. [116] introduce *Multi-scale Vision Transformers (MViT)* for video and image recognition by fusing the foundational concept of multi-scale feature hierarchies with transformer models. Multi-scale Transformers feature many scale stages with varying degrees of channel resolution.

The industry's high demand for autonomous driving has led to a surge of interest in three-dimensional object detection, resulting in several practical three-dimensional object detection algorithms [117]. Yuan et al. [55] proposed a *Temporal-Channel transformer* to represent spatial-temporal and channel domain relationships for video object detection from Lidar data. The transformer's unique architecture encodes temporal-channel information for many frames, whereas the decoder decodes spatial-channel information for the current frame voxel-by-voxel.

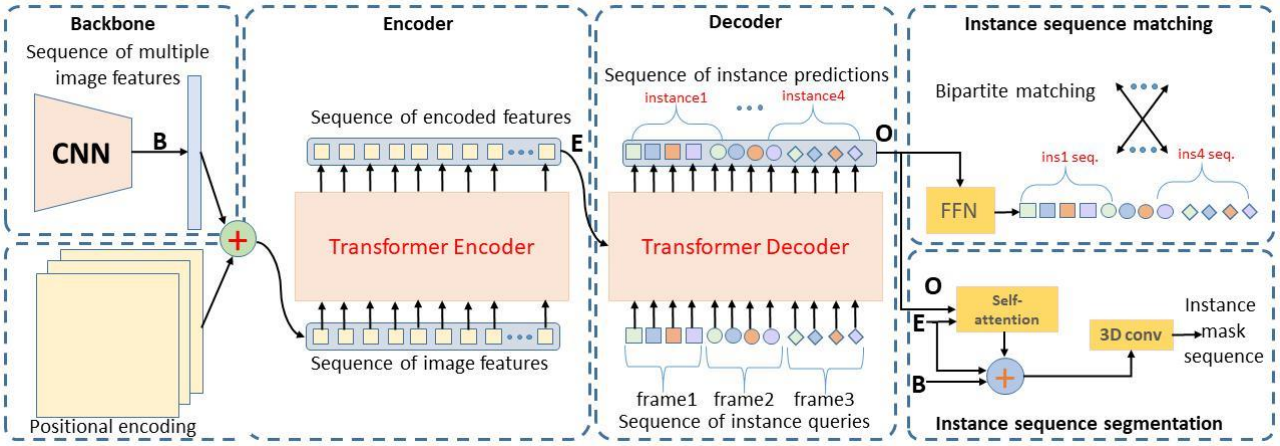

Fig. 9. The architecture of VisTR. image credit ([115]).

IV. Multimodal Task

With the increasing demand for models that can process both visual and textual inputs, the application of the visual transformer has been extended to multimodal tasks. In this section, we explore the capabilities of the visual transformer in handling multimodal inputs and review some of the recent developments in this area.

Recent breakthroughs in deep learning have resulted in significant advancements in computer vision and natural language processing. These accomplishments enable the integration of vision and language and multimodal learning tasks such as image captioning [118-119], image-text matching, visual grounding [120], and visual question answering [121]. Yu et al. present a novel framework for picture captioning called the Multimodal Transformer (MT) [122]. The MT comprises an image encoder that creates visual representations using deep self-attention learning and a caption decoder that converts the encoder's visual characteristics to textual captions. Liu et al. [123] explore image captioning as a sequence-to-sequence prediction problem. Li et al. proposed CaPtion Trans- formeR (CPTR), a complete Transformer model, to replace the usual "CNN+Transformer" approach. CPTR is convolution-free and can model global context information at each encoder layer. Evaluation results on the famous MS COCO [68] dataset indicate that the CPTR technique is more successful than "CNN+Transformer" networks. Detailed visualizations illustrate that the CPTR model can use long-range dependencies from the start and that the decoder's "words-to-patches" attention can pay close attention. The Conditional Position encodings Visual Transformer (CPVT) [103] sub-statutes the predefined positional embeddings used in Vision Transformer with conditional position encodings (CPE), allowing transformers to analyze input images of any size without interpolation.

Hu and Singh [124] developed a *Unified Transformer (UniT)* encoder-decoder model that accepts pictures and(or) text as input and trains on various tasks ranging from visual perception and language comprehension to combined vision-language reasoning. UniT consists of encoding modules that encode each input modality as a sequence of hidden states, a transformer decoder over the encoded input modalities, and task-specific output heads that apply task-specific output heads to the decoder hidden states to generate the final predictions for each task. Desai and Johnson [125] proposed that visual representations from textual annotations (VirTex) are a pre-training technique for visual representations that use semantically dense captions. First, VirTex jointly trains CNN and Transformer to create natural language captions for images from scratch. Then, apply the newly acquired characteristics to subsequent visual recognition tasks. A Decision Transformer [126] architecture encodes states, actions, and returns using modality-specific linear embeddings and a positional episodic time step encoding. Fig. 10 shows the architecture of the Decision transformer. Tokens are fed into a GPT architecture, which uses a causal self-attention mask to predict behaviors auto-regressively.

The vision transformer architecture is integrated into generative adversarial networks (GANs) for image generation. Regularisation methods for GANs that are now available do not interact well with self-attention, resulting in significant training instability. GANs using Vision Transformers are trained using novel regularization techniques. ViTGAN beats the existing CNN-based StyleGAN2 method on the CIFAR-10, CelebA [127], and LSUN bedroom datasets [128].

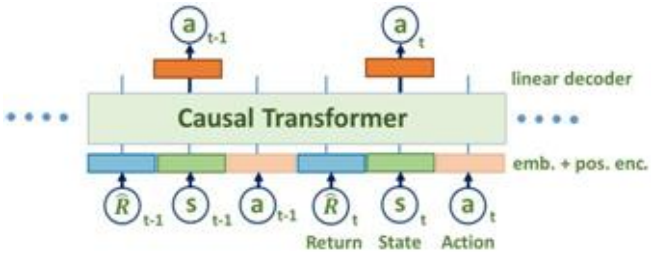

Fig. 10. Decision transformer architecture. Image credit [126].

V. Future Research Directions of Visual Transformers

As the field of Visual Transformers continues to develop, numerous potential avenues for future research exist. This section focuses on the summary of lessons learned from the discussions made throughout the article and future research directions. These directions offer valuable opportunities to enhance the performance and capabilities of Vision Transformers.

A. Lessons Learnt

1) Adaptation of vision transformers: The survey highlights the adaptability of Vision Transformers in their use for image processing, which was previously known in natural language processing. This fact demonstrates the chance to use already-developed skills and methods from one field to another, expanding the possible uses of Vision Transformers.

2) Image enhancement and classification: The survey shows how well Vision Transformers perform various image enhancement tasks, such as lowering noise, raising contrast, and boosting resolution. Additionally, their efficient use in image classification tasks demonstrates their capacity to extract significant representations from images and achieve competitive performance levels.

3) State-of-the-Art object detection: The study demonstrates the remarkable object detection performance of Vision Transformers, which outperforms conventional approaches and produces cutting-edge results. This fact shows the main contribution that Vision Transformers make to the field by enhancing robustness and accuracy in object localization and identification key processes.

4) Automation in education systems: The survey looks at how student verification processes used in educational institutions utilize Vision Transformers. Vision Transformers can streamline administrative operations, increasing speed and accuracy by automating the identity of the verification process. This fact demonstrates the crucial role that Vision Transformers have played in changing and strengthening the administrative procedures used in educational institutions.

5) Multimodal processing: The survey explores Vision Transformers' ability to handle multimodal tasks that require the processing and fusion of data from many modalities, in-including text, picture, and audio. This fact demonstrates the potential of Vision Transformers to enable thorough comprehension and analysis of complex data, leading to breakthroughs in areas like cross-modal retrieval, multimodal sentiment analysis, and visual question answering. The review demonstrates the significant contributions and developments made by Vision Transformers for image processing through an in-depth examination. The main findings from this overview highlight the flexibility of Vision Transformers, their effectiveness in image enhancement and classification, their cutting-edge performance in object detection, their potential to automate administrative tasks in educational systems, and their proficiency in handling multimodal data. These priceless insights provide direction for future study and practical use of Vision Transformers across numerous areas, encouraging further development in computer vision.

B. Future Research Directions

1) Efficiency optimization: Recent research has focused a lot of attention on finding ways to increase the effectiveness of Visual Transformers. There is a rising need to create methods to increase Visual Transformers' efficiency without sacrificing performance due to the demand for real-time and resource-constrained applications. Recent studies have looked into several solutions to this problem. For instance, researchers have studied techniques for sparse attention, which concentrate on focusing just on relevant areas of input to lighten the total computing load. Also, low-rank approximations, which approximate the attention matrices with low-rank structures and save a significant amount of computation, have been studied. Additionally, to decrease the memory footprint and inference time of Visual Transformers, researchers have looked into model reduction techniques like pruning or quantization. Recent developments in Visual Transformer efficiency show tremendous promise for enabling implementation in resource-constrained contexts while preserving their efficacy and performance.

2) Robustness and generalization: A significant area of research continues to be strengthening the robustness and generalization abilities of Visual Transformers, with current developments tackling the difficulties presented by real-world settings. Researchers have been looking into cutting-edge methods to lessen the effects of occlusions, which frequently impair accurate object detection in complex surroundings, in the pursuit of enhanced resilience. Recent research has looked at techniques to improve performance in obstructed situations, such as partial occlusion handling through attention processes or occlusion-aware training procedures. Additionally, the danger of adversarial attacks has been elevated to a top priority when using computer vision models. By using competitive training techniques or defence mechanisms against such attacks, researchers have made substantial progress toward creating robust Visual Transformers that can tolerate adversarial perturbations. In addition, recent research initiatives have focused on addressing domain transitions. It has been investigated to enhance the generalization abilities of Visual Transformers across various datasets or real-world domains using techniques like domain adaption or domain generalization. Researchers are making progress towards giving Visual Transformers the robustness and generalization skills they need to meet the challenges posed by many complicated real-world circumstances by actively taking into account these recent developments.

3) Interpretability and explainability: Various computer vision challenges have revealed impressive performance from Visual Transformers. However, it is difficult to comprehend the logic behind their forecasts because of their lack of interpretability. The recent research aims to overcome this drawback by investigating ways to make Visual Transformers easier to understand. A possible strategy is using attention visualization methods to draw attention to the areas of an image impacting judgment. Researchers and users can learn more about the particular characteristics or regions that the Visual Transformer concentrates on while making predictions by visualizing the attention maps. In addition, techniques like gradient-based attribution approaches and saliency maps have been used to pinpoint the most crucial input features influencing the result. These methods aid in identifying the

primary determinants of the Visual Transformer's choice, enhancing the predictability and interpretability of results. Future research aims to provide users with more transparent and interpretable Visual Transformers, enabling improved understanding and utilization of these potent models in practical applications. The last will be accomplished by continuing to investigate and improve these methodologies.

4) Multi-modal and cross-modal learning: A fascinating research area that will help us interpret complicated visual data better is the extension of Visual Transformers to handle multimodal and cross-modal data. Inquiries into integrating Visual Transformers with various modalities, such as text, audio, and depth information, have advanced significantly in recent years. For instance, using the strength of Visual Transformers to interpret visual data alongside textual context, researchers have created unique architectures that integrate vision and language models. Tasks such as image captioning, visual question answering, and cross-modal retrieval have all benefited from this integration. Additionally, research into using audio data in Visual Transformers has produced promising outcomes for tasks like sound event recognition or audio-visual scene analysis. Another fascinating development is integrating depth information with Visual Transformers, which enables comprehensive scene interpretation and 3D perception. Visual Transformers can deliver a thorough and holistic comprehension of complicated visual data by successfully integrating and learning from several modalities, pushing the limits of computer vision research and applications. Continued study in this field can reveal fresh perspectives and enhance Visual Transformers' capacity to handle multimodal and cross-modal input.

5) Incremental and continual learning: Recent research has focused heavily on enabling Visual Transformers to continually learn from streaming or updating data since it allows models to adapt to changing contexts and evolving concepts. The flexibility and adaptability of Visual Transformer may be enhanced by recent developments in incremental learning methods. Rehearsal approaches, which save and playback a portion of previously observed data during training to reduce catastrophic forgetting, are one noteworthy strategy. Research has also looked into methods like lifelong learning, where the model gradually picks up new skills while holding on to knowledge from earlier jobs. As a result, Visual Transformers can continuously improve their skills without compromising how well they do previously mastered jobs. Strategies like adaptive learning rates, dynamic network designs, and online learning algorithms have been investigated to address the problem. Visual Transformers can effectively learn from evolving data streams, improve their performance, and keep current knowledge by concentrating on incremental learning and devising ways to adapt to new classes or concepts over time. More research is needed in this field to make Visual Transformers more flexible and adaptable in practical applications and dynamic contexts.

6) Attention mechanism exploration: Research on Visual Transformers in recent years has concentrated on understanding and improving attention mechanisms to improve their effectiveness. Different attention types that can improve the modelling skills of Visual Transformers are the subject of one area of research. For instance, non-local attention mechanisms have drawn attention to their ability to identify distant relationships in pictures or movies, facilitating a better comprehension of the whole context. Another interesting approach is sparse attention, which tries to keep good performance while reducing computing complexity by focusing only on pertinent areas or pixels inside an input. Additionally, researchers have looked at the usage of learned attention masks, in which attention weights are dynamically computed based on the input data, enabling the model to assign adaptively attention to the most informative regions. The performance and modelling skills of Visual Transformers could be significantly improved by these latest developments in attention mechanisms. Researchers can open new doors for developing computer vision and expanding the capabilities of Visual Transformers in various applications by exploring these attention variants and continuing to innovate in this area.

7) Domain-specific adaptation: Various computer vision challenges have revealed impressive performance from Visual Transformers. However, because of the particular traits and demands of such areas, its application to specific tasks or domains frequently presents difficulties. Future research efforts can concentrate on investigating domain-specific adaptation methods to modify Visual Transformers for specific application domains. Recent studies have begun to explore domain adaptation techniques that use labelled data from the target domain to align the model's representation with the domain-specific features. To adapt Visual Transformers for tasks like disease diagnosis, organ segmentation, or anomaly detection, for instance, researchers in the field of medical imaging have investigated strategies like transfer learning or fine-tuning on medical datasets. Although Visual Transformers have demonstrated potential in satellite image analysis, more study is required to create domain-specific adaptations to deal with issues like size variation, heterogeneous data sources, or a lack of labelled data. In a similar way, in robotics, Visual Transformers can be configured to perform visual perception tasks in specific robotic applications, such as robot localization, object recognition, and scene interpretation. Researchers can bridge the gap between Visual Transformers and certain application areas, enabling higher performance and overcoming the particular difficulties encountered in those domains by concentrating on domain-specific adaption strategies. The investigation of these methods holds promise for releasing Visual Transformers' full potential across many specialized fields and advancing computer vision in particular application areas.

8) Data-efficient learning: Visual Transformers have displayed outstanding performance in computer vision

applications, although their training frequently necessitates large amounts of labelled data. Recent research has concentrated on investigating data-efficient learning approaches to lessen the dependence on sizable annotated datasets and enable efficient learning with few labelled examples. In this regard, semi-supervised learning strategies have drawn interest since they use labelled and unlabeled data during training. Visual Transformers can gain from a larger training set and perform better by using the quantity of unlabeled data and incorporating it into the learning process. Another exploratory route, which seeks to learn representations solely from unlabeled data, is unsupervised learning. These techniques allow models to develop helpful presentations from unannotated data that may be applied to subsequent tasks. Unsupervised learning has recently made significant strides in several computer vision areas, including picture categorization, object recognition, and image synthesis. Researchers can harness the potential of Visual Transformers in situations with little labelled data by exploring data-efficient learning techniques, making it possible to deploy these models more frugally and widely in various applications.

V. Conclusion

This article discusses critical self-attention architectures and examines in detail transformer models for various image-processing applications. We comprehensively discuss the strengths and weaknesses of existing techniques, particularly the possible future research directions. With a particular emphasis on general image processing problems, this survey offers a unique perspective on recent advances in self-attention and Transformer-based techniques. We discuss state-of-the-art self-attention models for semantic and instance segmentation, image classification, object detection, image captioning, video analysis and classification, multi-model tasks, and three-dimensional data analysis. We hope our work will spark interest among the image-processing community in maximizing the applications of vision-transformed models. Transformer models are pretty complicated from the perspective of parameters, computing time, and resources required. Visualizing and comprehending essential parts in an image for classification purposes is still a problem in transformers, and spatially accurate activation-specific representations are necessary [129]. The use of a vision transformer model in university education systems facilitating the detection of fraudulent activities in student identification documents is also highlighted. The model can thoroughly examine the identification document, detecting inconsistencies or anomalies as highlighting fraudulent activity.

Acknowledgment

This paper is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BG-RRP-2.004-0001-C01. The paper reflects only the author's view and the Agency is not responsible for any use that may be made of the information it contains.

References

[1] Y. LeCun, "Backpropagation applied to digit recognition," Neural computation, vol. 1, no. 4, pp. 541–551,, 1989.

[2] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," Adv. Neural Inf. Process. Syst, vol. 4, no. January, pp. 3104–3112,, 2014.

[3] P. Velickovi˘c,´ A. Casanova, P. Lio,` G. Cucurull, A. Romero, and Y. Bengio, "Graph attention networks," in 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc, 2018, pp. 1–12,.

[4] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," Adv. Neural Inf. Process. Syst, vol. 2015-Janua, pp. 2017–2025,, 2015.

[5] K. Xu, M. Zhang, J. Li, S. Du, K. Kawarabayashi, and S. Jegelka, "How neural networks extrapolate: From feedforward to graph neural networks," 2020, available:. [Online]. Available: http://arxiv.org/abs/2009.11848.

[6] P. Ramachandran, I. Bello, N. Parmar, A. Levskaya, A. Vaswani, and J. Shlens, "Stand- alone self-attention in vision models," Adv. Neural Inf. Process. Syst, vol. 32, 2019.

[7] A. Vaswani, "Attention is all you need," Advances in Neural Information Processing Systems, vol. m, no. Nips, p. 5999–6009, 2017.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf, vol. 1, no. Mlm, pp. 4171–4186,, 2019.

[9] M. Peters, "Knowledge enhanced contextual word representations," emnlp- ijcnlp 2019 - 2019 conf," in Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf, 2020, pp. 43–54,

[10] C. Raffel, "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res, vol. 21, pp. 1–67,, 2020.

[11] S. Mehta, M. Ghazvininejad, S. Iyer, L. Zettlemoyer, and H. Hajishirzi, "Delight: Deep and light-weight transformer," 2020, available:. [Online]. Available: http://arxiv.org/abs/2008.00623.

[12] Z. Fan, "Mask attention networks: Rethinking and strengthen transformer," pp. 1692–1701,, 2021.

[13] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and Catanzaro, "Megatron- lm: Training multi-billion parameter language models using model parallelism," 2019, available:. [Online]. Available: http://arxiv.org/abs/1909.08053.

[14] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, "Efficient content-based sparse attention with routing transformers," Trans. Assoc. Comput. Linguist, vol. 9, pp. 53–68,, 2021.

[15] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer- xl: Attentive language models beyond a fixed-length context," in ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf, 2020, pp. 2978–2988,.

[16] H. Yan, B. Deng, X. Li, and X. Qiu, "Tener: Adapting transformer encoder for named entity recognition," 2019, available:. [Online]. Available: http://arxiv.org/abs/1911.04474.

[17] X. Li, H. Yan, X. Qiu, and X. Huang, "Flat: Chinese ner using flat-lattice transformer," pp. 6836–6842,, 2020.

[18] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit, vol. 2019-June, 2019, pp. 244–253,.

[19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and Jegou,´ "Training data- efficient image transformers distillation through attention," pp. 1–22,, 2020, available:. [Online]. Available: http://arxiv.org/abs/2012.12877.

[20] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, pp. 5790–5799,.

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and Zagoruyko, "End-to- end object detection with transformers," LNCS, vol. 12346, pp. 213–229,, 2020.

[22] Y. Wang, "Vistr: End-to-end instance segmentation with transformers," 2021, available:. [Online]. Available: http://arxiv.org/abs/2105.00637.

[23] A. Ramesh, "Zero-shot text-to-image generation," 2021, available:. [Online]. Available: http://arxiv.org/abs/2102.12092.

[24] W. Su, "Vl-bert: Pre-training of generic visual-linguistic representations," pp. 1–16,, 2019, available:. [Online]. Available: http://arxiv.org/abs/1908.08530.

[25] Y. Xu, H. Wei, M. Lin, Y. Deng, K. Sheng, M. Zhang, F. Tang, Dong, F. Huang, and C. Xu, "Transformers in computational visual media: A survey," Computational Visual Media, vol. 8, pp. 33–62, 2022.

[26] S. Jamil, M. Jalil Piran, and O.-J. Kwon, "A comprehensive survey of transformers for computer vision," Drones, vol. 7, no. 5, p. 287, 2023.

[27] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Shi, J. Fan, and Z. He, "A survey of visual transformers," IEEE Transactions on Neural Networks and Learning Systems, 2023.

[28] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, ´ "Video transformers: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[29] S. Khan, M. Naseer, M. Hayat, S. Zamir, F. Khan, and M. Shah, "Transformers in vision: A survey," pp. 1–28,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2101.01169.

[30] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," pp. 53 040–53 065,, 2019.

[31] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," 2021, available:. [Online]. Available: http://arxiv.org/abs/2106.04554.

[32] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, Xiao, C. Xu, Y. Xu et al., "A survey on vision transformer," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 1, pp. 87–110, 2022.

[33] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 357–366.

[34] M. Gehrig and D. Scaramuzza, "Recurrent vision transformers for object detection with event cameras," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13 884–13 893.

[35] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12 104–12 113.

[36] Y. Yang, L. Jiao, X. Liu, F. Liu, S. Yang, Z. Feng, and X. Tang, "Transformers meet visual learning understanding: A comprehensive review," arXiv preprint arXiv:2203.12944, 2022.

[37] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12 175–12 185.

[38] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, Zhang, and D. Shen, "Transformers in medical image analysis: A review," Intelligent Medicine, 2022.

[39] A. A. Aleissaee, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-Xia, and F. S. Khan, "Transformers in remote sensing: A survey," Remote Sensing, vol. 15, no. 7, p. 1860, 2023.

[40] "Vision transformer explained — papers with code," accessed Oct. 11, 2021). [Online]. Available: https://paperswithcode.com/method/vision-transformer

[41] J. Thickstun, "The transformer model equations," pp. 1–5,, 2019, available:. [Online]. Available: https://homes.cs.washington.edu/ thickstn/docs/transformers.pdf.

[42] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in 3rd International Conference on Learning Representations, ICLR 2015, 2015, p. 1–15.

[43] M. Luong, H. Pham, and C. Manning, "Effective approaches to attention-based neural machine translation," in Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process, 2015, pp. 1412–1421,.

[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit, 2018, pp. 7794–7803,.

[45] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus," pp. 1–9,, 2016, [Online]. Available: http://arxiv.org/abs/1606.08415.

[46] J. Ba, J. Kiros, and G. Hinton, "Layer normalization," 2016, available:. [Online]. Available: http://arxiv.org/abs/1607.06450.

[47] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019, available:. [Online]. Available: http://arxiv.org/abs/2007.07582.

[49] T. Brown, "Language models are few-shot learners," Adv. Neural Inf. Process. Syst, vol. 2020-Decem, 2020.

[50] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Le, "Xlnet: Generalized autoregressive pretraining for language understanding," Adv. Neural Inf. Process. Syst, vol. 32, no. NeurIPS, pp. 1–18,, 2019.

[51] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," 2019, available:. [Online]. Available: http://arxiv.org/abs/1907.11692.

[52] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and Soricut, "Albert: A lite bert for self-supervised learning of language representations," pp. 1–17,, 2019, available:. [Online]. Available: http://arxiv.org/abs/1909.11942.

[53] C. Rosset, "Turing-nlg: A 17-billion-parameter language model by microsoft," Microsoft Research, p. 17, 2020, accessed Sep. 05, 2021). [Online]. Available: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-

[54] H. Xu, "Pre-trained models: Past, present and future," 2021, available:.[Online]. Available: http://arxiv.org/abs/2106.07139.

[55] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, "Visformer: The vision-friendly transformer," 2021, available:. [Online]. Available: http://arxiv.org/abs/2104.12533.

[56] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and Vaswani, "Bottleneck transformers for visual recognition," 2021, available:. [Online]. Available: http://arxiv.org/abs/2101.11605.

[57] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," 2020, available:. [Online]. Available: http://arxiv.org/abs/2012.09164.

[58] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, available:. [Online]. Available: http://arxiv.org/abs/2010.11929.

[59] M. Chen, "Generative pretraining from pixels," in 37th Int. Conf. Mach. Learn. ICML 2020, 2020, vol. PartF16814, pp. 1669–1681,.

[60] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," LNCS, vol. 9908, pp. 646–661,, 2016.

[61] E. Hoffer, T. Ben-Nun, I. Hubara, N. Giladi, T. Hoefler, and D. Soudry, "Augment your batch: better training with larger batches," 2019, available:. [Online]. Available: http://arxiv.org/abs/1901.09335.

[62] B. Graham, "Levit: a vision transformer in convnet's clothing for faster inference," 2021, available:. [Online]. Available: http://arxiv.org/abs/2104.01136.

[63] C.-F. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," 2021, available:. [Online]. Available: http://arxiv.org/abs/2103.14899.

[64] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," pp. 1–12,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2103.00112.

[65] H. Chen, "Pre-trained image processing transformer," 2020, available:. [Online]. Available: http://arxiv.org/abs/2012.00364.

[66] A. Vaswani, "Tensor2tensor for neural machine translation," in AMTA 2018 - 13th Conference of the Association for Machine Translation in the Americas, Proceedings, 2018, vol. 1, p. 193–199.

[67] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," pp. 1–16,, 2020, available:. [Online]. Available: http://arxiv.org/abs/2010.04159.

[68] T. Lin, "Microsoft coco: Common objects in context," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics, no. PART 5, pp. 740–755,, 2014.

[69] M. Zheng, P. Gao, X. Wang, H. Li, and H. Dong, "End-to-end object detection with adaptive clustering transformer," 2020, available:. [Online]. Available: http://arxiv.org/abs/2011.09315.

[70] Z. Dai, B. Cai, Y. Lin, and J. Chen, "Up-detr: Unsupervised pre-training for object detection with transformers," 2020, available:. [Online]. Available: http://arxiv.org/abs/2011.09094.

[71] Z. Sun, S. Cao, Y. Yang, and K. Kitani, "Rethinking transformer-based set prediction for object detection," 2020, available:. [Online]. Available: http://arxiv.org/abs/2011.10881.

[72] J. Beal, E. Kim, E. Tzeng, D. Park, A. Zhai, and D. Kislyuk, "Toward transformer- based object detection," 2020, available:. [Online]. Available: http://arxiv.org/abs/2012.09958.

[73] J. Yang, "Focal self-attention for local-global interactions in vision transformers," pp. 1–21,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2107.00641.

[74] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," 2021, available:. [Online]. Available: http://arxiv.org/abs/2102.04378.

[75] Z. Liu, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, available:. [Online]. Available: http://arxiv.org/abs/2103.14030.

[76] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," pp. 248–255,, 2014-05.

[77] B. Zhou, "Semantic understanding of scenes through the ade20k dataset," Int. J. Comput. Vis, vol. 127, no. 3, pp. 302–321,, 2019.

[78] S. Zheng, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," pp. 6881–6890,, 2020, available:. [Online]. Available: http://arxiv.org/abs/2012.15840.

[79] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Max-deeplab: End-to-end panoptic segmentation with mask transformers," pp. 5463–5474,, 2020, available:. [Online]. Available: http://arxiv.org/abs/2012.00759.

[80] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," 2021, available:. [Online]. Available: http://arxiv.org/abs/2103.13413.

[81] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics, vol. 9351, pp. 234–241,, 2015.

[82] J. Chen, "Transunet: Transformers make strong encoders for medical image segmentation," pp. 1–13,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2102.04306.

[83] B. Yun, Y. Wang, J. Chen, H. Wang, W. Shen, and Q. Li, "Spectr: Spectral transformer for hyperspectral pathology image segmentation," 2021, available:. [Online]. Available: http://arxiv.org/abs/2103.03604.

[84] J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," pp. 1–18,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2102.10662.

[85] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L. Chen, "Axial-deeplab: Stand- alone axial-attention for panoptic segmentation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, vol. 12349, pp. 108–126,.

[86] M. Guo, J. Cai, Z. Liu, T. Mu, R. Martin, and S. Hu, "Pct: Point cloud transformer," Comput. Vis. Media, vol. 7, no. 2, pp. 187–199,, 2021.

[87] N. Parmar, "Image transformer," in 35th Int. Conf. Mach. Learn. ICML 2018, 2018, vol. 9, pp. 6453–6462,.

[88] Y. Jiang, S. Chang, and Z. Wang, "Transgan: Two pure transformers can make one strong gan, and that can scale up," 2021, available:. [Online]. Available: http://arxiv.org/abs/2102.07074.

[89] K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu, and C. Liu, "Vitgan: Training gans with vision transformers," pp. 1–13,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2107.04589.

[90] Z. Huang, J.-X. Du, and H.-B. Zhang, "A multi-stage vision transformer for fine-grained image classification," in 2021 11th Inter-national Conference on Information Technology in Medicine and Education (ITME). IEEE, 2021, pp. 191–195.

[91] D. Zhou, "Deepvit: Towards deeper vision transformer," 2021, available:. [Online]. Available: http://arxiv.org/abs/2103.11886.

[92] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jegou,´ "Going deeper with image transformers," 2021, available:. [Online]. Available: http://arxiv.org/abs/2103.17239.

[93] W. Wang, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," 2021, available:. [Online]. Available: http://arxiv.org/abs/2102.12122.

[94] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit, vol. 2016-Decem, 2016, pp. 770–778,.

[95] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015, 2015, p. 1–14.

[96] L. Yuan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," 2021, available:. [Online]. Available: http://arxiv.org/abs/2101.11986.

[97] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021, available:. [Online]. Available: http://arxiv.org/abs/2104.05704.

[98] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. Oh, "Rethinking spatial dimensions of vision transformers," 2021, available:. [Online]. Available: http://arxiv.org/abs/2103.16302.

[99] Z. Lu, H. Liu, J. Li, and L. Zhang, "Efficient transformer for single image super resolution," pp. 1–13,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2108.11084.

[100] N. Kitaev, Kaiser, and A. Levskaya, "Reformer: The efficient transformer," pp. 1–12,, 2020, available:. [Online]. Available: http://arxiv.org/abs/2001.04451.

[101] H. Wu, "Cvt: Introducing convolutions to vision transformers," 2021, available:. [Online]. Available: http://arxiv.org/abs/2103.15808.

[102] X. Chu, "Twins: Revisiting the design of spatial attention in vision transformers," pp. 1–14,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2104.13840.

[103] "Conditional positional encodings for vision transformers," 2021, available:. [Online]. Available: http://arxiv.org/abs/2102.10882.

[104] Z. Zhang, H. Zhang, L. Zhao, T. Chen, and T. Pfister, "Aggregating nested transformers," pp. 1–18,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2105.12723.

[105] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in 36th Int. Conf. Mach. Learn. ICML 2019, 2019, vol. 2019-June, pp. 10 691–10 700,.

[106] A. Krizhevsky, "Learning multiple layers of features from tiny im-ages," 2009.

[107] B. Wu, "Visual transformers: Token-based image representation and processing for computer vision," 2020, available:. [Online]. Available: http://arxiv.org/abs/2006.03677.

[108] M. Caron, "Emerging properties in self-supervised vision transformers," 2021, [Online]. Available: http://arxiv.org/abs/2104.14294.

[109] M. Patacchiola and A. Storkey, "Self-supervised relational reasoning for representation learning," Adv. Neural Inf. Process. Syst, vol. 2020-Decem, no. NeurIPS, 2020.

[110] J. Donahue and K. Simonyan, "Large scale adversarial representation learning," Adv. Neural Inf. Process. Syst, vol. 32, no. NeurIPS, pp. 1–32,, 2019.

[111] S. Jenni and P. Favaro, "Self-supervised feature learning by learning to spot artifacts," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit, 2018, pp. 2733–2742,.

[112] S. Atito, M. Awais, and J. Kittler, "Sit: Self-supervised vision transformer," pp. 1–10,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2104.03602.

[113]H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," pp. 1– 16,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2106.08254.

[114]A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Luciˇc,´ and C. Schmid, "Vivit: A video vision transformer," 2021, available:. [Online]. Available: http://arxiv.org/abs/2103.15691.

[115]Y. Wang, "End-to-end video instance segmentation with transformers," p. 8741–8750, 2020.

[116]H. Fan, "Multiscale vision transformers," 2021. [Online]. Available: http://arxiv.org/abs/2104.11227.

[117]R. Liu, Z. Yuan, T. Liu, and Z. Xiong, "End-to-end lane shape prediction with transformers," 2020.

[118]J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," Adv. Neural Inf. Process. Syst, vol. 32, pp. 1–11,, 2019.

[119]J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit, 2020, pp. 10 434–10 443,.

[120]H. Akbari, "Vatt: Transformers for multimodal self-supervised learning from raw video," Audio and Text,", 2021, available:. [Online]. Available: http://arxiv.org/abs/2104.11178.

[121]L. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," pp. 1–14,, 2019, available:. [Online]. Available: http://arxiv.org/abs/1908.03557.

[122]J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," IEEE Trans. Circuits Syst. Video Technol, vol. 30, no. 12, pp. 4467–4480,, 2020.

[123]W. Liu, S. Chen, L. Guo, X. Zhu, and J. Liu, "Cptr: Full transformer network for image captioning," pp. 1–5,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2101.10804.

[124]R. Hu and A. Singh, "Unit: Multimodal multitask learning with a unified transformer," 2021, available:. [Online]. Available: http://arxiv.org/abs/2102.10772.

[125]K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," 2020, available:. [Online]. Available: http://arxiv.org/abs/2006.06666.

[126]L. Chen, "Decision transformer: Reinforcement learning via sequence modeling," pp. 1–21,, 2021, available:. [Online]. Available: http://arxiv.org/abs/2106.01345.

[127]Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 Inter, 2015, pp. 3730–3738,.

[128]F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," 2015, available:. [Online]. Available: http://arxiv.org/abs/1506.03365.

[129]K. Han, "A survey on vision transformer," pp. 1–25,, 2020, available:. [Online]. Available: http://arxiv.org/abs/2012.12556.