

A survey of the Vision Transformers and their CNN-Transformer based Variants

Asifullah Khan^{1, 2, 3}, Zunaira Rauf^{1, 2}, Anabia Sohail^{1, 4}, Abdul Rehman Khan¹, Hifsa Asif^{1, 5}, Aqsa Asif^{1, 5}, and Umair Farooq¹*

¹Pattern Recognition Lab, Department of Computer & Information Sciences, Pakistan Institute of Engineering & Applied Sciences, Nilore, Islamabad 45650, Pakistan

²PIEAS Artificial Intelligence Center (PAIC), Pakistan Institute of Engineering & Applied Sciences, Nilore, Islamabad 45650, Pakistan

³Center for Mathematical Sciences, Pakistan Institute of Engineering & Applied Sciences, Nilore, Islamabad 45650, Pakistan

⁴Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi, UAE

⁵Air University, E-9, Islamabad 44230, Pakistan

Corresponding Authors: *Asifullah Khan, asif@pieas.edu.pk

Abstract

Vision transformers have become popular as a possible substitute to convolutional neural networks (CNNs) for a variety of computer vision applications. These transformers, with their ability to focus on global relationships in images, offer large learning capacity. However, they may suffer from limited generalization as they do not tend to model local correlation in images. Recently, in vision transformers hybridization of both the convolution operation and self-attention mechanism has emerged, to exploit both the local and global image representations. These hybrid vision transformers, also referred to as CNN-Transformer architectures, have demonstrated remarkable results in vision applications. Given the rapidly growing number of hybrid vision transformers, it has become necessary to provide a taxonomy and explanation of these hybrid architectures. This survey presents a taxonomy of the recent vision transformer architectures and more specifically that of the hybrid vision transformers. Additionally, the key features of these architectures such as the attention mechanisms, positional embeddings, multi-scale processing, and convolution are also discussed. In contrast to the previous survey papers that are primarily focused on individual vision transformer architectures or CNNs, this survey uniquely emphasizes the emerging trend of hybrid vision transformers. By showcasing the potential of hybrid vision transformers to deliver exceptional performance across a range of computer vision tasks, this survey sheds light on the future directions of this rapidly evolving architecture.

Key words: Auto Encoder, Channel Boosting, Computer Vision, Convolutional Neural Networks, Deep Learning, Hybrid Vision Transformers, Image Processing, Self-attention, and Transformer

1. Introduction

Digital images are complex in nature and exhibit high-level information, such as objects, scenes, and patterns (Khan et al. 2021a). This information can be analyzed and interpreted by computer vision algorithms to extract meaningful insights about the image content, such as recognizing objects, tracking movements, extracting features, etc. Computer vision has been an active area of research due to its applications in various fields (Bhatt et al. 2021). However, extracting high-level information from image data can be challenging due to variations in brightness, pose, background clutter, etc.

The emergence of convolutional neural networks (CNNs) has brought about a revolutionary transformation within the realm of computer vision. These networks have been successfully applied to a diverse range of computer vision tasks (Liu et al. 2018; Khan et al. 2020, 2022, 2023; Zahoor et al. 2022), especially image recognition (Sohail et al. 2021a; Zhang et al. 2023a), object detection (Rauf et al. 2023), and segmentation (Khan et al. 2021c). CNNs gained popularity due to their ability to automatically learn features and patterns from raw images (Simonyan and Zisserman 2014; Agbo-Ajala and Viriri 2021). Generally, local patterns, known as feature motifs are systematically distributed throughout the images. Different filters in the convolutional layers are specified to capture diverse feature motifs, while pooling layers in the CNNs are utilized for dimensionality reduction and to incorporate robustness against variations. This local-level processing of CNNs may result in a loss of spatial correlation, which can impact their performance when dealing with larger and more complex patterns.

Recently in computer vision, there has been some shift toward transformers, after they were first introduced by Vaswani et al. in 2017 for text processing applications (Vaswani et al. 2017a). In 2018, Parmer et al., exploited transformers for image recognition tasks, where they demonstrated

outstanding results (Parmar et al. 2018). Since then, there has been a growing interest in applying transformers to various vision-related applications (Liu et al. 2021b). In 2020, Dosovitskiy et al., introduced a transformer architecture, Vision Transformer (ViT), specifically designed for image analysis, which showed competitive results (Dosovitskiy et al. 2020). ViT models work by dividing an input image into a certain number of patches, each patch is subsequently flattened and fed to a sequence of transformer layers. The transformer layers enable the model to learn the relationships between the patches and their corresponding features, allowing it to identify feature motifs on a global scale in the image. Unlike CNNs that have a local receptive field, ViTs utilize its self-attention module to model long-range relationships, which enables them to capture the global view of an image (Ye et al. 2019; Guo et al. 2021). The global receptive field of ViTs helps them retain the global relationship and thus identify complex visual patterns distributed across the image (Bi et al. 2021; Wu et al. 2023b). In this context, Maurício et al. have reported that ViTs may show promising results as compared to CNNs in various applications (Zhang et al. 2021a; Maurício et al. 2023).

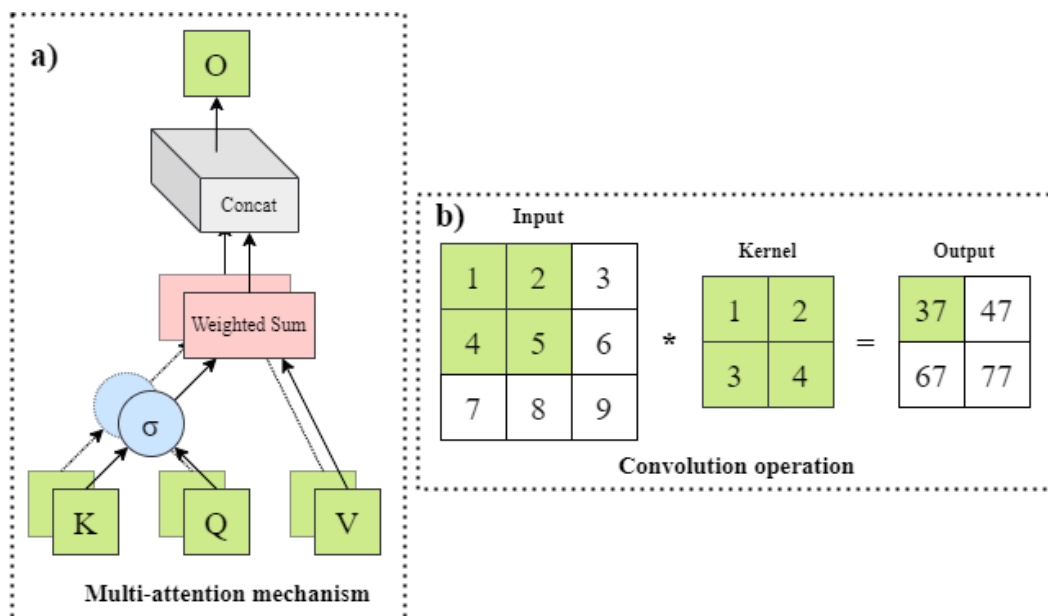


Figure 1: Depiction of the multi self-attention (MSA) mechanism and convolution operation. MSA tends to capture global relationships, whereas the convolution operation has a local receptive field to model pixel neighborhood information in the images.

In addition to the difference in their design and the way of capturing visual patterns, (shown in Fig 1) CNNs and ViTs also differ in their inductive biases. CNNs heavily rely on the correlation among the neighboring pixels, whereas ViTs assume minimal prior knowledge, making them significantly dependent on large datasets (Han et al. 2023). While ViT models have produced outstanding results on object recognition, classification, semantic segmentation, and other computer vision tasks (Kirillov et al. 2023; Dehghani et al. 2023), they are not a one-size-fits-all solution. In the case of small training data, despite the large learning capacity of ViTs, they may show limited performance as compared to CNNs (Morra et al. 2020; Jamali et al. 2023). In addition, their large receptive field demands significantly more computation. Therefore, the concept of Hybrid Vision Transformers (HVT) also known as CNN-Transformer, was introduced to combine the power of both CNNs and ViTs (Maaz et al. 2023). These hybrid models leverage the convolutional layers of CNNs to capture local features, which are then fed to ViTs to gain global context using the self-attention mechanism. The HVTs have shown improved performance in many image recognition tasks.

Recently, different interesting surveys have been conducted to discuss the recent architectural and implementational advancements in transformers (Liu et al. 2021b; Du et al. 2022; Islam 2022; Aleissae et al. 2022; Ulhaq et al. 2022; Shamshad et al. 2023). Most of these survey articles either focus on specific computer vision applications or delve into discussions on transformer models specifically developed for Natural Language Processing (NLP) applications. In contrast, this survey paper emphasizes recent developments in HVTs (CNN-Transformer), which combine concepts from both CNNs and transformers. It introduces a taxonomy for general ViTs and aims to thoroughly classify the emerging approaches based on their core architectural designs. Furthermore, this survey also conducts an in-depth analysis of HVTs, providing a

taxonomy grounded in their architectural composition as well as exploring their diverse applications.

The paper begins with an introduction to the essential components of the ViT networks and then discusses various recent ViT architectures. The reported ViT models are broadly classified into six categories based on their distinct features. Additionally, a detailed discussion on HVTs is included, highlighting their focus on leveraging the advantages of both convolutional operations and multi-attention mechanisms. The survey paper covers the recent architectures and applications of HVTs in various computer vision tasks. Moreover, a taxonomy is presented for HVTs, classifying them based on the way these architectures incorporate convolution operations in combination with self-attention mechanisms. This taxonomy divides HVTs into seven major groups, each of which reflects a different way of taking advantage of both the convolutional and multi-attention operations. Frequently used abbreviations are listed in Table 1.

The paper is structured as follows: (illustrated in Fig. 2) Section 1 presents a systematic understanding of the ViT architecture, highlighting its dissimilarities with CNNs and the advent of HVT architectures. Moving on, section 2 covers the fundamental concepts used in different ViT variants, while section 3 and section 4 provide a taxonomy of the recent ViTs and HVTs architectures, respectively. Section 5 focuses on the usage of HVTs, particularly in the area of computer vision, and section 6 presents current challenges and future directions. Finally, in section 7, the survey paper is concluded.

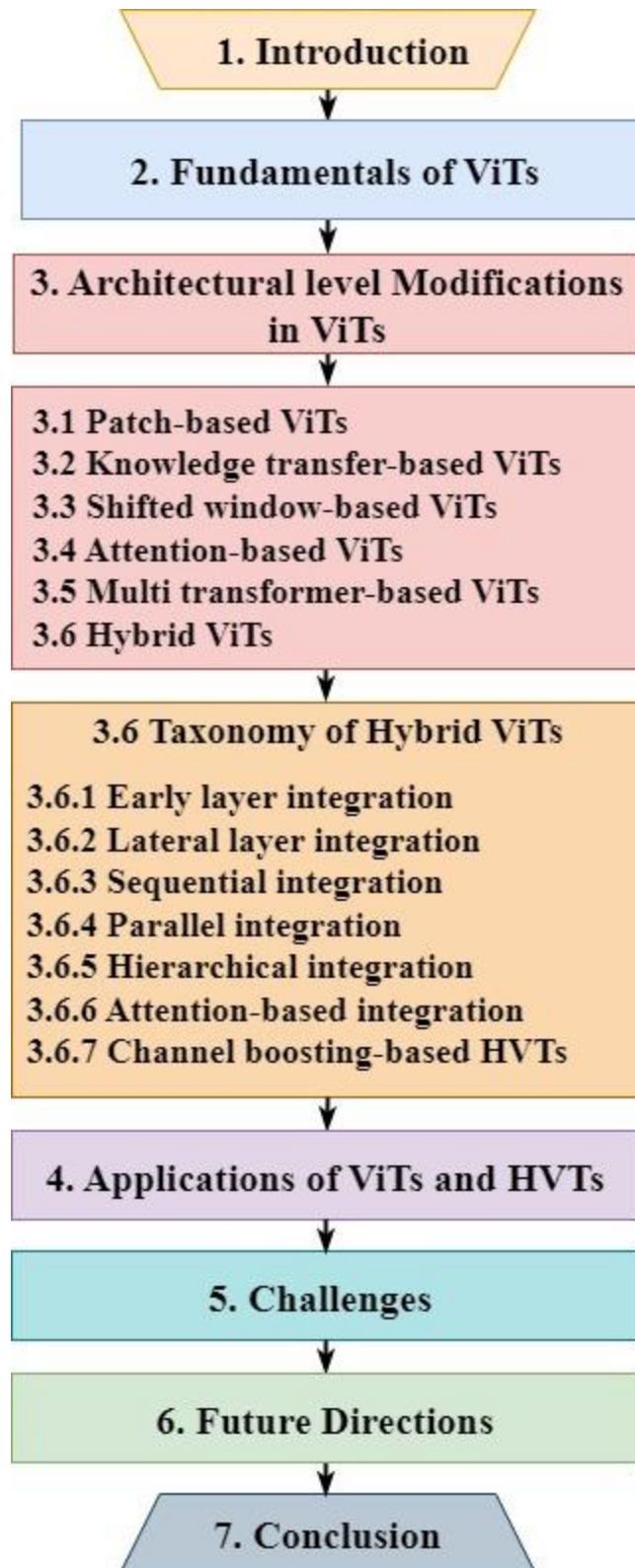


Figure 2: Layout of the different sections of the survey paper.

Table 1: Table of abbreviations.

Abbreviation	Definition
CNN	Convolutional Neural Network
ViT	Vision Transformer
NLP	Natural Language Processing
HVT	Hybrid Vision Transformer
DL	Deep Learning
MSA	Multi-Head Self-Attention
FFN	Feed Forward Network
MLP	Multi-Layer Perceptron
APE	Absolute Position Embedding
RPE	Relative Position Embedding
CPE	Convolution Position Embedding
Pre-LN	Pre-Layer Normalization
GELU	Gaussian Error Linear Unit
CB	Channel Boosting
CvT	Convolutional Vision Transformer
LeFF	Locally-enhanced Feed Forward
CeiT	Convolution Enhanced Image Transformer
I2T	Image To Transformer
MoFFN	Mobile Feed Forward Network
CCT	Compact Convolutional Transformer
Local ViT	Local Vision Transformer
LeViT	LeNet-Based Vision Transformer
PVT	Pyramid Vision Transformer
MaxViT	Multi-Axis Attention-based Vision Transformer
MBConv	Mobile inverted bottleneck convolution
DPT	Deformable Patch-based Transformer
TNT	Transformer iN Transformer
DeiT	Data-efficient Image Transformer
TaT	Target aware Transformer
CaiT	Class attention in image Transformer
IRFFN	Inverted Residual Feed Forward Network
LPU	Local Perceptron Unit
ResNet	Residual Network
STE	Standard Transformer Encoder
SE-CNN	Squeeze and Excitation CNN
FPN	Feature Pyramid Network
UAV	Unmanned Aerial Vehicle
EA	Evolving Attention
RC	Reduction Cells
NC	Normal Cells
ConTNet	Convolution Transformer Network
FCT	Fully Convolutional Transformer

2. Fundamental Concepts in ViTs

Fig. 3 illustrates the fundamental architectural layout of a transformer. Initially, the input image is divided, flattened, and transformed into lower dimensional linear embeddings known as Patch Embeddings. Then positional embeddings and class tokens are attached to these embeddings and fed into the encoder block of the transformer for generating class labels. In addition to the multi-head attention (MSA) layer, the encoder block contains a feed-forward neural (FFN) network, a normalization layer, and a residual connection. Finally, the last head (an MLP layer, or a decoder block) predicts the final output. Each of these components is discussed in detail in the following subsections.

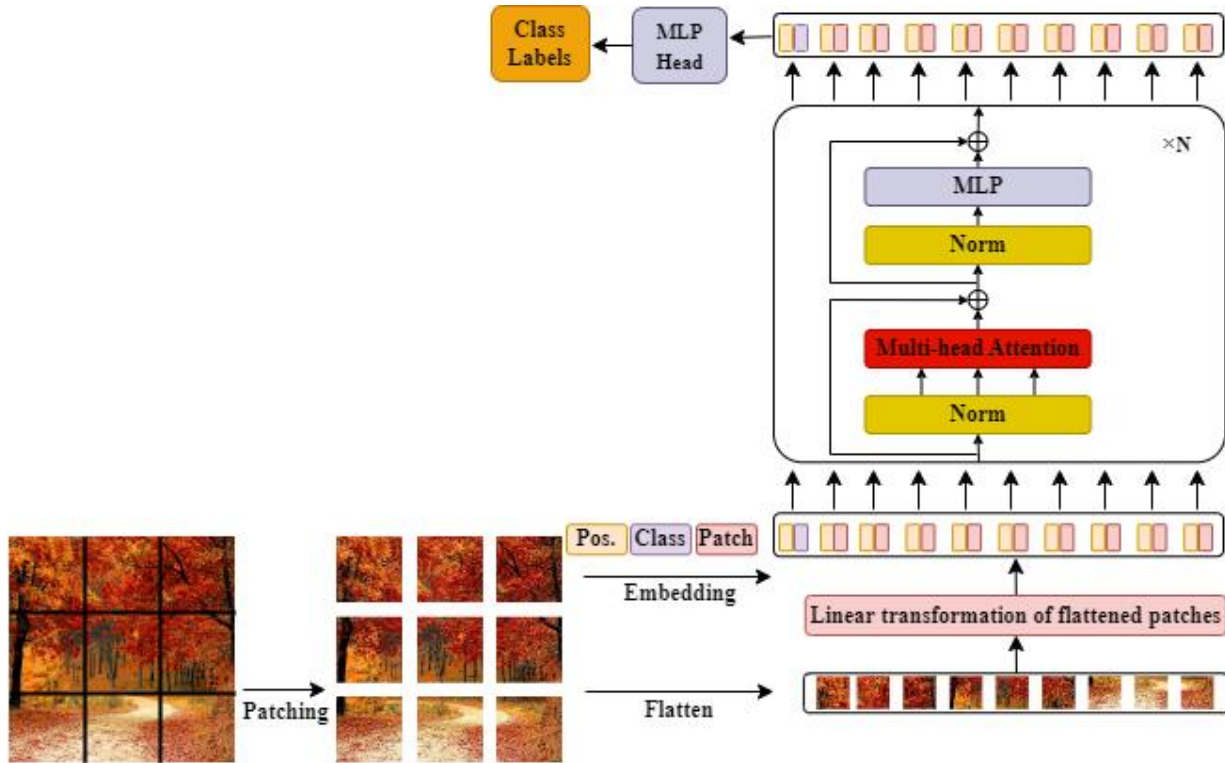


Figure 3: Detail architecture of ViT. An input image is at first divided into patches, then their linearly transformed embeddings are combined with positional information and processed through multiple encoder/decoder blocks for downstream tasks.

2.1. Patch embedding

Patch embedding is an important concept in ViT architecture. It involves converting the image patches into vector representations, which enables ViT to process images as sequences of tokens using a transformer-based approach (Dosovitskiy et al. 2020). The input image is partitioned into fixed-size non-overlapping parts, flattened into one-dimensional vectors, and projected to a higher-dimensional feature space using a linear layer with D embedding dimensions (Equation 1). This approach enables ViT to learn the long-range dependencies between different patches, allowing it to attain promising results on tasks that involve images.

$$\mathbf{X}_{patch}^{N \times D} = R(\mathbf{I}_{image}^{A \times B \times C}) \quad \text{Eq. 1}$$

The input image is \mathbf{I}_{image} with size $A \times B \times C$, $R()$ is the reshaping function to produce N number of patches “ \mathbf{X}_{patch} ” with size D , and $N = A/P \times B/P$, $D = P \times P \times C$, P = patch size and C = channels.

2.2. Positional embedding

ViTs utilize positional encoding to add positional information into the input sequence and retain it throughout the network. The sequential information between patches is captured through position embeddings, which is incorporated within the patch embeddings. Since the development of ViTs, numerous position embedding techniques have been suggested for learning sequential data (Jiang et al. 2022). These techniques fall into three categories:

2.2.1. Absolute Position Embedding (APE)

The positional embeddings are integrated into the patch embeddings by using APE before the encoder blocks.

$$\mathbf{X} = \mathbf{X}_{patch} + \mathbf{X}_{pos} \quad \text{Eq. 2}$$

where, the transformer's input is represented by \mathbf{X} , \mathbf{X}_{patch} represents patch embeddings, and \mathbf{X}_{pos} is the learnable position embeddings. Both \mathbf{X}_{patch} & \mathbf{X}_{pos} have dimensions $(N + 1) \times D$, where D represents the dimension of an embedding. It is possible to train \mathbf{X}_{pos} corresponding to positional embeddings of a single or two sets that can be learned (Carion et al. 2020).

2.2.2. Relative Position Embedding (RPE)

The Relative Position Embedding (RPE) technique is primarily used to incorporate information related to relative position into the attention module (Wu et al. 2021b). This technique is based on the idea that the spatial relationships between patches carry more weight than their absolute positions. To compute the RPE value, a lookup table is used, which is based on learnable parameters. The lookup process is determined by the relative distance between patches. Although the RPE technique is extendable to sequences of varying lengths, it may increase training and testing time (Chu et al. 2021b).

2.2.3. Convolution Position Embedding (CPE)

The Convolutional Position Embeddings (CPE) method takes into account the 2D nature of the input sequences. 2D convolution is employed to gather position information using zero-padding to take advantage of the 2D nature (Islam et al. 2021). Convolutional Position Embeddings (CPE) can be used to incorporate positional data at different stages of the ViT. The CPE can be introduced specifically to the self-attention modules (Wu et al. 2021a), the Feed-Forward Network (FFN) (Li et al. 2021c; Wang et al. 2021b), or in between two encoder layers (Chu et al. 2021a).

2.3. Attention Mechanism

The core component of the ViT architecture is the self-attention mechanism, which plays a crucial role in explicitly representing the relationships between entities within a sequence. It calculates the significance of one item to others by representing each entity in terms of the global contextual information and capturing the interaction between them (Vaswani et al. 2017b). The self-attention module transforms the input sequence into three different embedding spaces namely query, key, and value. The sets of key-value pairs with query vectors are taken as inputs. The output vector is calculated by taking a weighted sum of the values followed by the softmax operator, where the weights are calculated by a scoring function (Equation 3).

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}}\right) \cdot \mathbf{V} \quad \text{Eq. 3}$$

where, \mathbf{Q} , \mathbf{V} , and \mathbf{K}^T are query, value, and transposed key matrix, respectively. $\frac{1}{\sqrt{d_k}}$ is the scaling factor, and d_k is dimensions of the key matrix.

2.3.1. Multi-Head Self-Attention (MSA)

The limited capacity of a single-head self-attention module often leads to its focus on only a few positions, potentially overlooking other important positions. To address this limitation, MSA is employed. MSA utilizes parallel stacking of self-attention blocks to increase the effectiveness of the self-attention layer (Vaswani et al. 2017b). It captures a diverse range of complex interactions among the sequence elements by assigning various representation subspaces (query, key, and value) to the attention layers. The MSA constitutes multiple self-attention blocks. Each is equipped with learnable weight matrices for query, key, and value sub-spaces (Equation 4). The outputs of these blocks are then concatenated and projected to the output space using the learnable parameter

W^O (Equation 5) This enables the MSA to focus on multiple portions and to effectively capture the relationships in all areas. The mathematical representation of the attention process is given below:

$$head_i = Attention(Q_i, K_i, V_i), \text{ and } i = 1, 2, \dots, h \quad \text{Eq. 4}$$

$$MSA(Q, K, V) = Concat(head_1, head_2, \dots, head_h) \cdot W^O \quad \text{Eq. 5}$$

In Eq. 4, $head_i$ is the output of each self-attention block, whereas W^O in Eq.5 represents the learnable parameters.

Self-attention's capability to dynamically compute filters for every input sequence is a significant advantage over convolutional processes. Unlike convolutional filters, which are often static, self-attention can adjust to the particular context of the input data. Self-attention is also robust to changes in the number of input points or their permutations, which makes it a good choice for handling irregular inputs. Traditional convolutional procedures, on the other hand, are less adaptable to handling inputs with variable objects and require a grid-like structure, like 2D images. Self-attention is a powerful tool for modeling sequential data and has been effective in various tasks including NLP (Khan et al. 2021b).

2.4. Transformer layers

A ViT encoder consists of several layers to process the input sequence. These layers comprise the MSA mechanism, feed-forward neural network (FFN), residual connection, and normalization layer. These layers are arranged to create a unified block that is repeated several times to learn the complex representation of the input sequence.

2.4.1. Feed-forward network

A transformer-specific feed-forward network (FFN) is employed in models to obtain more complex attributes from the input data. It contains multiple fully connected layers and a nonlinear activation function, such as GELU in between the layers (Equation 6). FFN is utilized in every encoder block after the self-attention module. The hidden layer of the FFN usually has a dimensionality of 2048. These FFNs or MLP layers are local and translationally equivalent to global self-attention layers (Dosovitskiy et al. 2020).

$$FFN(X) = b_2 + W_2 * \sigma(b_1 + W_1 * X) \quad \text{Eq. 6}$$

In Eq. 6, the non-linear activation function GELU is represented by σ . Weights of the network are represented as W_1 , and W_2 , whereas b_1 , and b_2 correspond to layer-specific bias

2.4.2. Residual connection

Sub-layers in the encoder/decoder block (MSA and FFN) utilize a residual link to improve performance and strengthen the information flow. Original input positional embedding is added to the output vector of MSA, as additional information. The residual connection is then followed by a layer-normalization operation (Equation 7).

$$X_{output} = LayerNorm(X \oplus O_{SL}) \quad \text{Eq. 7}$$

Where X is the original input and O_{SL} is the output of each sub-layer, \oplus representing the residual connection.

2.4.3. Normalization layer

There are various methods for layer normalization, such as pre-layer normalization (Pre-LN) (Kim et al. 2023), which is utilized frequently. The normalization layer is placed prior to the MSA or FFN and inside the residual connection. Other normalization procedures, including batch

normalization, have been suggested to enhance the training of transformer models, however, they might not be as efficient due to changes in the feature values (Jiang et al. 2022).

2.5. Hybrid Vision Transformers (CNN-Transformer Architectures)

In the realm of computer vision tasks, ViTs have gained popularity, but compared to CNNs, they still lack image-specific inductive bias often referred to as prior knowledge (Seydi and Sadegh 2023). This inductive bias includes characteristics like translation and scale invariance due to the shared weights across different spatial locations (Moutik et al. 2023). In CNNs, the locality, translational equivariance, and two-dimensional neighborhood structure are ingrained in every layer throughout the whole model. Additionally, the kernel leverages the correlation between neighboring pixels, which facilitates the extraction of good features quickly (Woo et al. 2023). On the other hand, in ViT, the image is split into linear patches (tokens) that are fed into encoder blocks through linear layers to model global relationships in the images. However, linear layers lack effectiveness in extracting local correlation (Woo et al. 2023).

Many HVT designs have focused on the efficiency of convolutions in capturing local features in images, especially at the start of the image processing workflow for patching and tokenization (Guo et al. 2023). The Convolutional Vision Transformer (CvT), for instance, uses a convolutional projection to learn the spatial and low-level information in image patches. It also utilizes a hierarchical layout with a progressive decrease in token numbers and an increase in token width to mimic the spatial downsampling effect in CNNs (Wu et al. 2021a). Similarly, Convolution-enhanced Image Transformers (CeiT) leverage convolutional operations to extract low-level features via an image-to-token module (Yuan et al. 2021a). A novel sequence pooling technique is presented by the Compact Convolutional Transformer (CCT), which also integrates

conv-pool-reshape blocks to carry out tokenization (Hassani et al. 2021). It also showed an accuracy of about 95% on smaller datasets like CIFAR10 when trained from scratch, which is generally difficult for other traditional ViTs to achieve.

Several recent studies have investigated ways to enhance the local feature modeling capabilities of ViTs. LocalViT employs depthwise convolutions to improve the ability to model local features (Li et al. 2021c). LeViT uses a CNN block with four layers at the beginning of the ViT architecture to gradually increase channels and improve efficiency at inference time (Graham et al. 2021). Similar methods are employed by ResT, however, to manage fluctuating image sizes, depth-wise convolutions, and adaptive position encoding are used (Zhang and Yang 2021).

Without additional data, CoAtNets' unique architecture of depthwise convolutions and relative self-attention achieves outstanding ImageNet top-1 accuracy (Dai et al. 2021). In order to create stronger cross-patch connections, Shuffle Transformer provides a shuffle operation (Huang et al. 2021b) and CoaT is a hybrid approach that incorporates depthwise convolutions and cross-attention to encode relationships between tokens at various scales (Xu et al. 2021a). Another method “Twins” builds upon PVT by incorporating separable depthwise convolutions and relative conditional position embedding (Chu et al. 2021a). Recently, MaxVit, hybrid architecture, introduced the idea of multi-axis attention. Their hybrid block consists of MBConv-based convolution followed by block-wise self-attention and grid-wise self-attention, and when repeated multiple times this block creates a hierarchical representation and is capable of tasks like image generation and segmentation (Tu et al. 2022b). The block-wise and grid-wise attention layers are capable of extracting local and global features respectively. Convolution and transformer model strengths are intended to be combined in these hybrid designs.

3. Architectural level modifications in ViTs

In recent years, different modifications have been carried out in ViT architectures (Zhou et al. 2021). These modifications can be categorized based on their attention mechanism, positional encoding, pre-training strategies, architectural changes, scalability, etc. ViT architectures can be broadly classified into five main classes based on the type of architectural modification, namely, (i) patch-based approaches, (ii) knowledge transfer-based approaches, (iii) shifted window-based approaches, (iv) attention-based approaches, and (v) multi-transformer-based approaches. However, it is observed that with the introduction of CNN's inductive bias to ViTs there came a boost in its performance. In this regard, we also classified the HVTs into seven categories based on their structural design. The taxonomy of ViT architectures is shown in Fig. 4. In addition a comprehensive overview of various online resources relevant to ViTs including libraries, lecture series, datasets, and computing platforms are provided in Table 2.

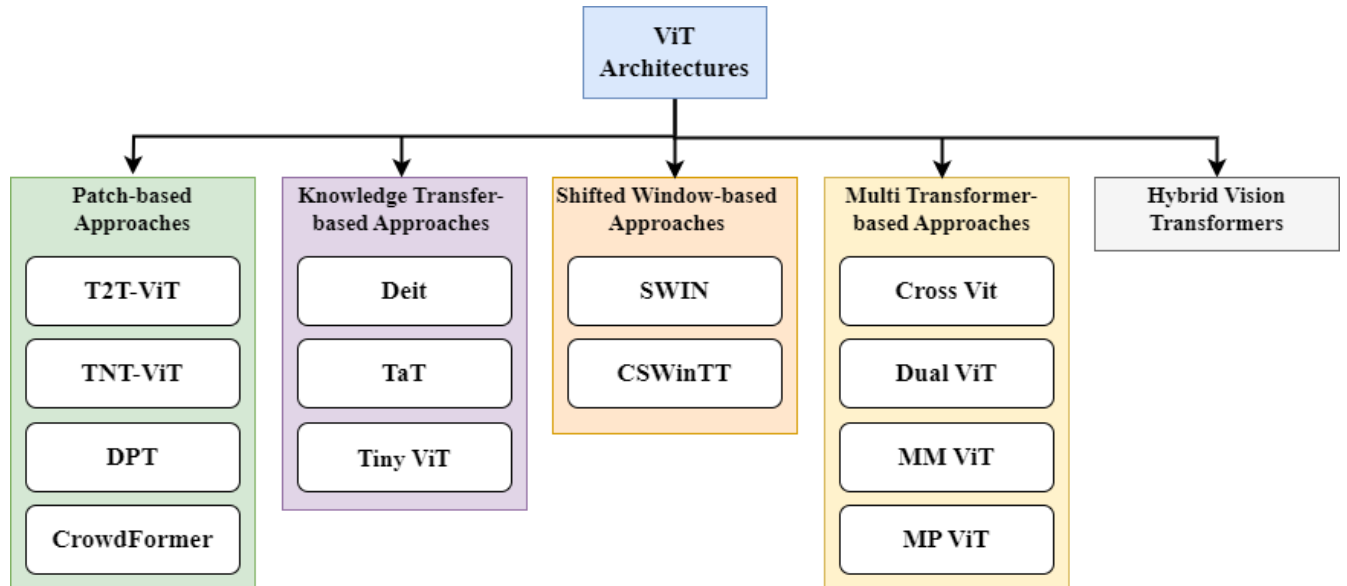


Figure 4: Taxonomy of Vision ViT architectures.

Table 2: Distinct online resources relevant to DL and ViT.

Category	Description	Source
Cloud Computing Solutions	Online available datasets and computational resources	Google Colab: https://colab.research.google.com/notebooks/welcome.ipynb
		Kaggle: https://www.kaggle.com/
	DL commercially available services	FloydHub: https://www.floydhub.com/
		Amazon SageMaker: https://aws.amazon.com/deep-learning/
		Microsoft Azure ML Services: https://azure.microsoft.com/en-gb/services/machine-learning/
		Google Cloud: https://cloud.google.com/deep-learning-vm/
		IBM Watson Studio: https://www.ibm.com/cloud/deep-learning
DL Libraries	DL libraries with integrated neural network classes and optimization for CPU and GPU	Pytorch: https://pytorch.org/
		Tensorflow: https://www.tensorflow.org/
		MatConvNet: http://www.vlfeat.org/matconvnet/
		Keras: https://keras.io/
		Theano: http://deeplearning.net/software/theano/
		Caffe: https://caffe.berkeleyvision.org/
		Detectron2: https://github.com/facebookresearch/detectron2
		OpenMMLAB: https://openmmlab.com/
Lecture Series	Free online courses on DL	Stanford Lecture Series: https://web.stanford.edu/class/cs224n/
		Youtube: https://www.youtube.com/watch?v=SZorAJ4I-sA
		Youtube: https://www.youtube.com/watch?v=iDulhoQ2pro
		Coursera: https://www.coursera.org/learn/nlp-sequence-models
Datasets	Diverse categories of annotated image datasets available online for free access	ImageNet: http://image-net.org/
		COCO: http://cocodataset.org/#home

		Visual Genome: http://visualgenome.org/
		Open images: https://ai.googleblog.com/2016/09/introducing-open-images-dataset.html
		Places: https://places.csail.mit.edu/index.html
		Youtube-8M: https://research.google.com/youtube8m/index.html
		CelebA: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
		Wiki Links: https://code.google.com/archive/p/wiki-links/downloads
		EXCITEMENT dataset: https://github.com/hltfbk/EOP-1.2.1/wiki/Data-Sets
		Ubuntu Dialogue Corpus: https://www.kaggle.com/datasets/rtatman/ubuntu-dialogue-corpus
		ConvAI3: https://github.com/DeepPavlovAdmin/convai
		Large Movie Review Dataset: https://ai.stanford.edu/~amaas/data/sentiment/
		CIFAR10: https://www.cs.toronto.edu/~kriz/cifar.html
		Indoor Scene Recognition: http://web.mit.edu/torralba/www/indoor.html
		Computer Vision Datasets: https://computervisiononline.com/datasets
		MonuSeg: https://monuseg.grand-challenge.org/Data/
		Oxford-IIIT Pets: https://www.robots.ox.ac.uk/~vgg/data/pets/
		Fashion MNIST: https://research.zalando.com/welcome/mission/research-projects/fashion-mnist
Blogs/ Repositories	High-quality, free online articles and blogs	Github.io: http://jalammar.github.io/illustrated-transformer/
		Github: https://github.com/huggingface/pytorch-image-models
		Viso Ai: https://viso.ai/deep-learning/vision-transformer-vit/
		Github: https://github.com/google-research/vision_transformer
		HuggingFace: https://huggingface.co/docs/transformers/model_doc/vit

Hardware Resources	Energy-efficient and computationally optimized hardware solutions for DL processing	NVIDIA: http://nvidia.org/
		FPGA: https://www.intel.com/content/www/us/en/artificial-intelligence/programmable/overview.html
		Eyeriss: http://eyeriss.mit.edu/
		AMD: https://www.amd.com/en/graphics/instinct-server-accelerators
		Google's TPU: https://cloud.google.com/tpu/

3.1. Patch-based approaches

In ViT, an image is first divided into a grid of patches, which are subsequently flattened to generate linear embedding, treated as a sequence of tokens (Dosovitskiy et al. 2020). Positional embedding and class tokens are added to these embeddings, which are then given to the encoder for feature learning. Several studies exploited different ways of patch extraction mechanisms to improve the performance of ViTs. These mechanisms include fixed-size patching (Wang et al. 2021c), dynamic patching (Ren et al. 2022; Zhang et al. 2022c), and overlapping patching (Wang et al. 2021b). In this regard, we discuss several architectures and their patching criteria.

3.1.1. Tokens-to-Token Vision Transformer (T2T-ViT)

Tokens-to-Token Vision Transformer (T2T-ViT) utilizes a fixed size and iterative approach to generate patches (Yuan et al. 2021b). It utilizes the proposed Token-to-Token module iteratively to generate patches from the images. The generated patches are then fed to the T2T-ViT network to obtain final predictions.

3.1.2. Transformer in Transformer (TNT-ViT)

Transformer in Transformer ViT (TNT-ViT) presented a multi-level patching mechanism to learn representations from objects with different sizes and locations (Han et al. 2021). It first divides the input image into patches then each patch is further divided into sub-patches. Later, the architecture utilizes different transformer blocks to model the relationship between the patches and sub-patches. Extensive experiments showed the efficiency of TNT-ViT in terms of image classification on the ImageNet dataset.

3.1.3. Deformable Patch-based Transformer (DPT)

Deformable Patch-based Transformer (DPT) presented an adaptive patch embedding module named DePatch (Chen et al. 2021e). Fixed-size patching in transformers results in a loss of semantic information, which affects the system's performance. In this regard, the proposed DePatch module in DPT splits the images in an adaptive way to obtain patches with variable sizes and strong semantic information.

3.1.4. CrowdFormer

Yang and co-authors developed a ViT architecture, CrowdFormer for crowd counting (Yang et al. 2022b). The proposed architecture utilizes its overlap patching transformer block to capture the crowd's global contextual information. To consider images at different scales and in a top-down manner, the overlap patching layer is exploited, where instead of fixed-sized patches, a sliding window is used to extract overlapping patches. These overlapping patches tend to retain the relative contextual information for effective crowd counting.

3.2. Knowledge transfer-based approaches

This category enlists those ViT architectures that utilize a knowledge transfer (knowledge distillation) approach. It involves conveying knowledge from a larger network to a smaller network, much like a teacher imparting knowledge to a student (Kanwal et al. 2023; Habib et al. 2023). The teacher model is usually a complex model with ample learning capability, while the student model is simpler. The basic idea behind knowledge distillation is to facilitate the student model in acquiring and incorporating the distinctive features of the teacher model. This can be particularly useful for tasks where computational resources are limited, as the smaller ViT model can be deployed more efficiently than the larger one.

3.2.1. Data-efficient Image Transformers (DeiT)

DeiT is a smaller and more efficient version of ViT, which has shown competitive performance on various tasks (Touvron et al. 2020). It uses a pre-trained ViT model for the teacher and a smaller version for the student. Usually, supervised and unsupervised learning is used in combination, with the teacher network supervising the student network to produce similar results. In addition to the fast inference time and limited computational resources of DeiT, it also has an improved generalization performance because the student model has learned to capture the most important features and patterns in the data, rather than just memorizing the training data.

3.2.2. Target-aware Transformer (TaT)

Target-aware Transformer (TaT) (Lin et al. 2022) utilized one-to-many relation to exchange information from the teacher to the student network. The feature maps were first divided into a number of patches, then for each patch, all the teacher's features were transferred to all the student features rather than employing correlation between all spatial regions. All the features inside a

patch were then averaged into a single vector to make the knowledge transfer computationally efficient.

3.2.3. Tiny Vision Transformer (TinyViT)

Wu et al. suggested a fast distillation methodology along with a novel architecture, known as TinyViT (Wu et al. 2022a). Their main concept was to convey the learned features of the large pre-trained models to the tiny ones during pre-training (Fig. 5). The output logits of the instructor models were reduced and stored in addition to the encoded data augmentations on the disc beforehand to save memory and computational resource. The student model then employs a decoder to re-construct the saved data augmentations and knowledge is transferred via the output logits with both the models trained independently. Results demonstrated TinyViT's effectiveness on large-scale test sets.

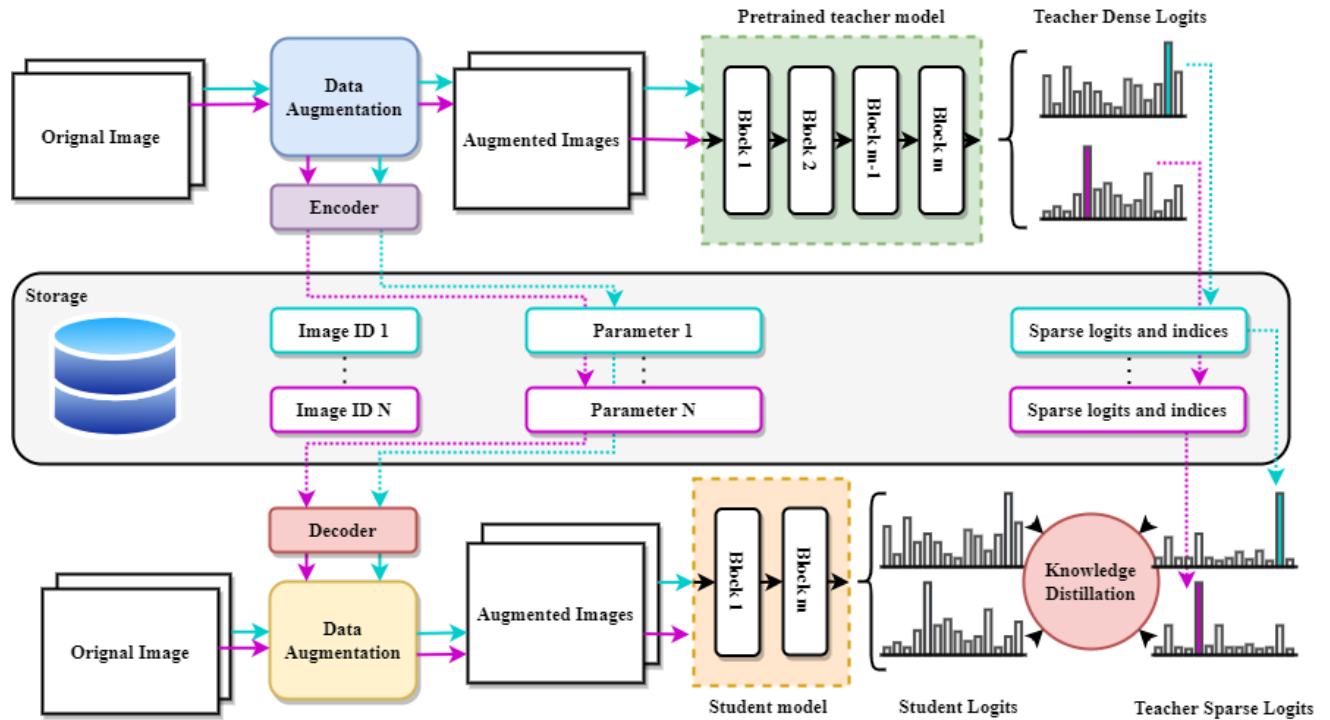


Figure 5: Detailed workflow of knowledge transfer-based approach (TinyViT).

3.3. Shifted window-based approaches

Several ViT architectures have adopted the shifted window-based approach to enhance their performance. This approach was first introduced by Liu et al. in their Swin Transformer (Liu et al. 2021c). The Swin Transformer has a similar architecture to ViT but with a shifted windowing scheme, as shown in Fig. 6. It controls the self-attention computation by computing it within each non-overlapping local window, while still providing cross-window connections to improve the efficacy. This is achieved by implementing shifted window-based self-attention as two consecutive Swin Transformer blocks. The first block uses regular window-based self-attention, and the second block shifts those windows and applies regular window-based self-attention again. The idea behind shifting the windows is to enable cross-window connections, which can help the network in improving its capability to model global relationships.

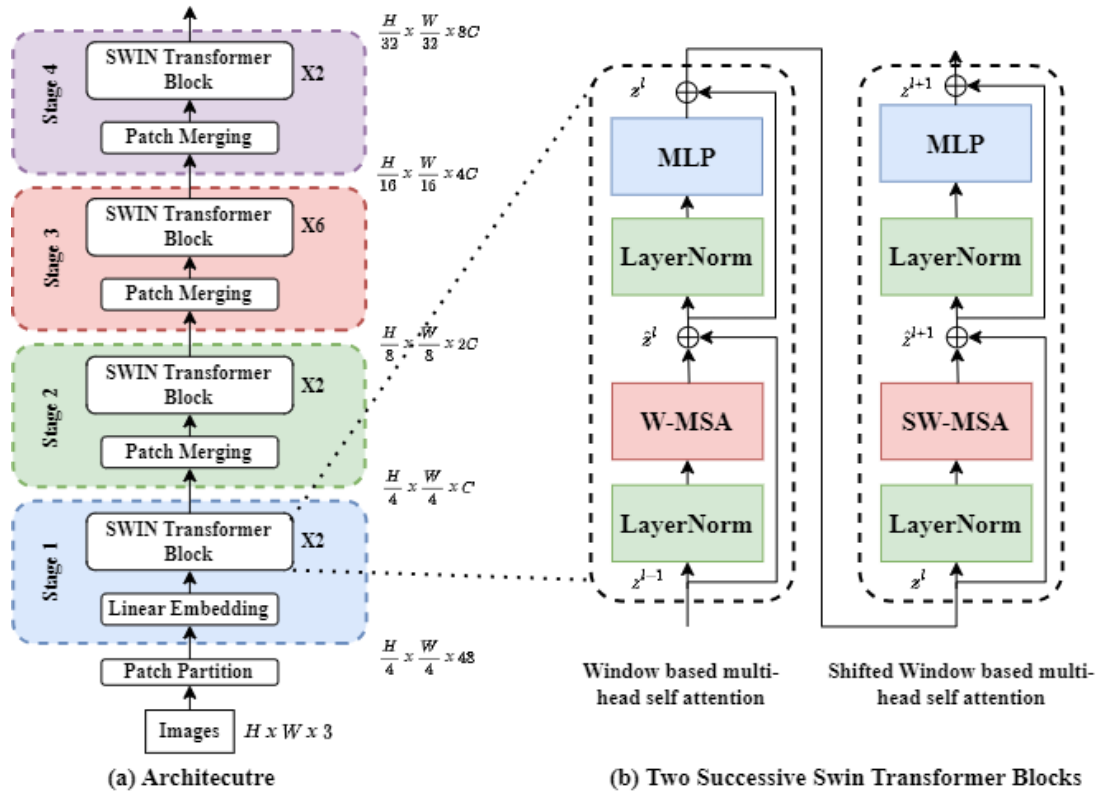


Figure 6: Architectural diagram of Swin Transformer (shifted window-based approach).

Song et al. proposed a novel ViT architecture for visual object tracking, named CSWinTT which utilizes cyclic shifting window-based attention at multi-scales (Song et al. 2022b). This approach enhances pixel attention to window attention and enables cross-window multi-scale attention to aggregate attention at different scales. This ensures the integrity of the tracking object and generates the best fine-scale match for the target object. Moreover, the cyclic shifting technique expands the window samples with positional information, which leads to greater accuracy and computational efficiency. By incorporating positional information into the attention mechanism, the model is better equipped to handle changes in the object's position over time and can track the object more effectively. Overall, the proposed architecture has shown promising results in improving the accuracy and efficiency of visual object tracking using ViT-based models.

3.4. Attention-based approaches

Numerous ViT architectures have been proposed that modify the self-attention module to enhance their performance. Some of these models utilize dense global attention mechanisms (Vaswani et al. 2017a; Dosovitskiy et al. 2020), while others utilize sparse attention mechanisms (Jiang et al. 2021; Liu et al. 2021c; Dai et al. 2021) to capture global-level dependencies in the images with no spatial correlation. These types of attention mechanisms are known to be computationally expensive. A number of works have been done to improve the attention modules in terms of performance and computational complexity (Tu et al. 2022b).

3.4.1. Class attention layer (CaiT)

Touvron et al. introduced a new approach to improve the performance of deep transformers (Touvron et al. 2021). Their architecture, named, CaiT contains a self-attention module and a class attention module. The self-attention module is just like a normal ViT architecture, but the class token (class information) is not added in the initial layers. The class embeddings are added in the

class attention module, later in the architecture. Their approach showed good results with a few numbers of parameters.

3.4.2. Deformable attention transformer (DAT)

Xia and co-authors proposed a data-dependent attention mechanism to focus on the more reliable regions (Xia et al. 2022). Their architecture has a modular design with each stage having a local attention layer followed by a deformable attention layer in each stage. The proposed DAT architecture showed exemplary performance on benchmark datasets.

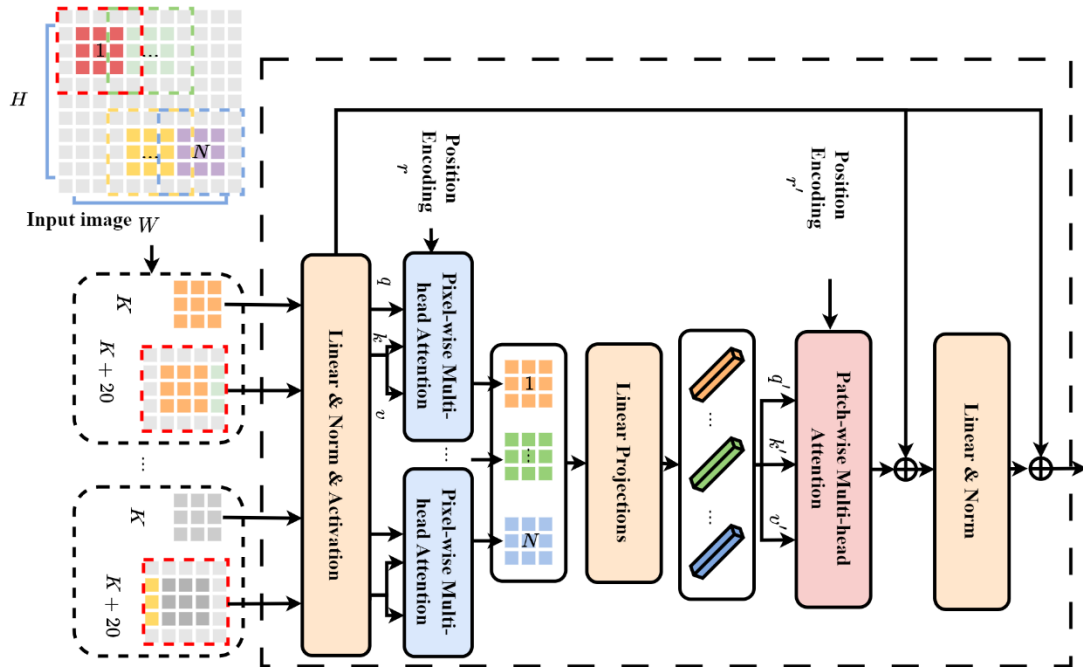


Figure 7: Architecture of patch-based Separable Transformer (SeT), which modified its MSA layer by introducing two diverse attention blocks.

3.4.3. Patch-based Separable Transformer (SeT)

Sun et al. used two different attention modules in their ViT architecture to fully capture global relationships in images (Sun et al. 2022) (Fig. 7). They proposed a pixel-wise attention module to learn local interactions in the initial layers. Later they utilized a patch-wise attention module to

extract global-level information. SeT showed superior results than other methods on several datasets, including ImageNet and MS COCO datasets.

3.5. Multi-transformer-based approaches

Many approaches utilized multiple ViTs in their architecture to improve their performance on various tasks that require multi-scale features. This section discusses such multi-transformer-based ViT architectures.

3.5.1. Cross Vision Transformer (CrossViT)

Chen and co-authors proposed a ViT architecture having dual branches which they named as CrossViT (Chen et al. 2021a). The key innovation in the proposed model is the combination of image patches of different sizes, which enables CrossViT to generate highly domain-relevant features. The smaller and larger patch tokens are processed using two separate branches with varying computational complexities. The two branches are fused multiple times using an efficient cross-attention module. This module enables knowledge transfer between the branches by creating a non-patch token. The attention map generation is achieved linearly, rather than quadratically, through this process. This makes CrossViT more computationally efficient than other models that use quadratic attention.

3.5.2. Dual Vision Transformer (Dual-ViT)

The Dual Vision Transformer (Dual-ViT) is a new ViT architecture that reduces the computational cost of self-attention mechanisms (Yao et al.). This architecture utilizes two individual pathways to capture global and local-level information. The semantic branch learns the coarse details, whereas the pixel pathway captures more fine details in the images. both of these branches are

integrated and trained in parallel. The proposed dualViT showed good results on the ImageNet dataset with fewer parameters as compared to other existing models.

3.5.3. Multiscale Multiview Vision Transformer (MMViT)

Multiscale Multiview Vision Transformer (MMViT) incorporates multiscale feature maps and multiview encodings into transformer models. The MMViT model utilizes several feature extraction stages to process multiple views of the input at various resolutions in parallel. At each scale stage, a cross-attention block is exploited to merge data across various perspectives. This approach enables the MMViT model to obtain high-dimensional representations of the input at multiple resolutions, leading to complex and robust feature representations.

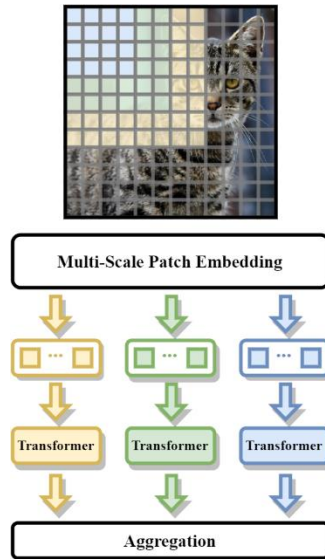


Figure 8: Architecture of Multi-transformer-based MPViT, which utilize multiple transformers in its architecture.

3.5.4. Multi-Path Vision Transformer (MPViT)

MPViT utilizes the multi-scale patching technique and multi-path-based ViT architecture to learn feature representations at different scales (Lee et al. 2021b). Their proposed multi-scale patching technique utilizes CNNs to create feature maps at different scales (Fig. 8). Later they utilize

multiple transformer encoders to process multi-scale patch embeddings. Lastly, they aggregate the outputs from each encoder to generate an aggregated output. The proposed MPViT demonstrated superior results as compared to existing approaches on the ImageNet dataset.

3.6. Taxonomy of HVTs (CNN-Transformer architectures)

Despite their successful performance, ViTs face three main issues, a) inability to capture low-level features by considering correlation in the local neighborhood, b) expensive in terms of computation and memory consumption due to their MSA mechanism, c) and fixed-sized input tokens, embedding. To overcome these issues, there comes the boom of hybridization of CNNs and ViTs after 2021. HVTs combine the strengths of both CNNs and transformer architectures to create models for capturing both the local patterns and global context in images (Yuan et al. 2023b). They have gained valuable focus in the research community due to their promising results in several image-related tasks (Li et al. 2022). Researchers have proposed a variety of architectures in this field by exploiting different approaches to merge CNNs and transformers (Heo et al. 2021; Si et al. 2022). These approaches include but are not limited to, adding some CNN layers within transformer blocks (Liu et al. 2021a; He et al. 2023; Wei et al. 2023), introducing a multi-attention mechanism in CNNs (Zhang et al. 2021b; Ma et al. 2023b), or using CNNs to extract local features and transformers to capture long-range dependencies (Yuan et al. 2021a, 2023a; Zhang et al. 2023c). In this regard, we define some subcategories based on the pattern of integration of the convolution operation within ViT architectures. These include (1) early-layer integration, (2) lateral-layer integration, (3) sequential integration, (4) parallel integration, (5) block integration, (6) hierarchical integration, (7) attention-based integration, and (8) channel boosting integration, depicted in Fig. 9.

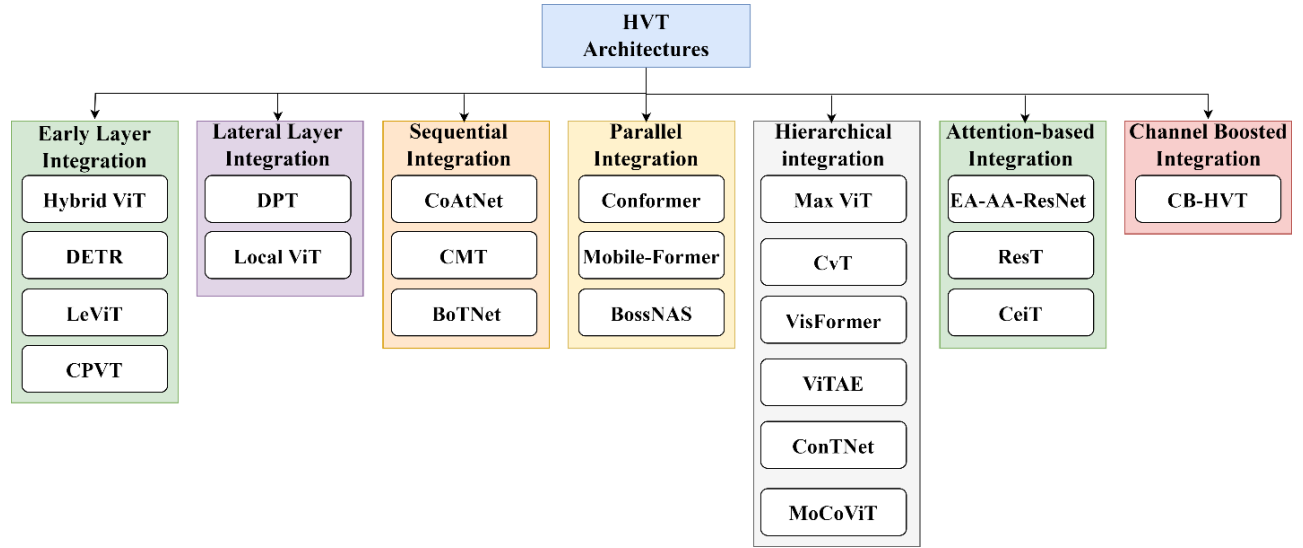


Figure 9: Taxonomy of Hybrid ViTs.

3.6.1. Early-layer integration

Long-range dependencies in the images are well-captured by ViTs, but since there is no inductive bias, training them needs a lot of data. On the other hand, CNNs inherent image-related inductive bias and capture high-level correlation present in the images locally. Therefore, researchers are focusing on designing HVTs, to merge the benefits of both CNNs and transformers (Pan et al. 2022). A lot of work is done to find out the most optimal way to fuse the convolution and attention in the transformer architectures. CNNs can be utilized at different levels to incorporate the locality in the architectures. Various studies have suggested the idea that it is beneficial to first capture local patterns and then learn the long-range dependencies to have a more optimized local and global perspective of an image (Peng et al. 2023).

Hybrid ViT

The first ViT architecture was proposed by Dosovitskiy *et al.* in 2020 (Dosovitskiy et al. 2020). In their work, they suggested the idea of considering image patches as sequences of tokens and feeding them into a transformer-based network to perform image recognition tasks. In their paper,

they laid the foundation for HVTs by presenting a hybrid version of ViT. In the hybrid architecture, the input sequences were obtained from CNN feature maps instead of raw image patches (LeCun et al. 1989). The input sequence was created by flattening the feature maps spatially, and the patches were produced using a 1x1 filter. They utilized ResNet50 architecture to obtain the feature maps as input to ViT (Wu et al. 2019). In addition, they carried out extensive experiments to identify the optimal intermediate block for feature map extraction.

Detection Transformer (DETR)

Carion et al. proposed a Detection Transformer (DETR) for performing object detection in natural images in 2020 (Carion et al. 2020). In their end-to-end proposed approach, they initially utilized a CNN to process the input before feeding it to the ViT architecture. The feature maps from the CNN backbone were combined with fixed-sized positional embeddings to create input for the ViT encoder. The outputs from the ViT decoder were then fed to a feed-forward network to make final predictions. DETR showed better performance when compared to other revolutionary detection models like Faster R-CNN. Their detailed idea is depicted in Fig. 10.

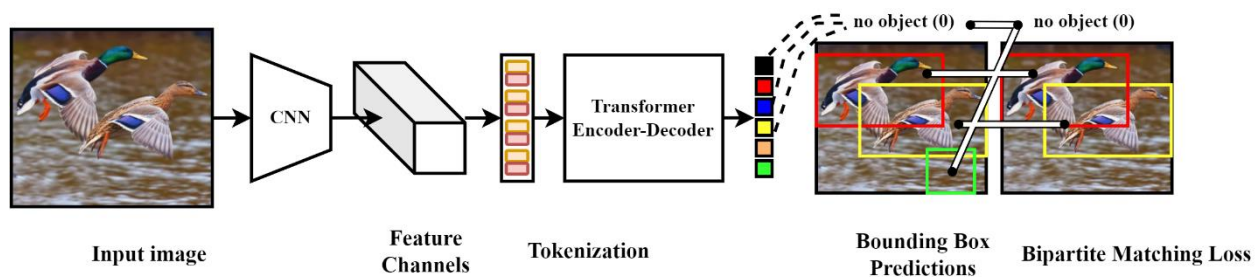


Figure 10: Architecture of DETR, with CNN integration as an initial stem block.

LeNet-based Vision Transformer (LeViT)

Graham et al. proposed a hybrid ViT “LeViT” in 2021 (Graham et al. 2021). In their model, they utilized convolution layers initially for processing the input. The proposed architecture combined

a CNN with the MSA of ViT architecture to extract local as well as global features from the input image. The LeViT architecture at first utilized a four-layered CNN model to reduce the image resolution and to obtain local feature representations. These representations were then fed to a ViT-inspired multi-stage architecture with MLP and attention layers to generate output.

Conditional Positional Encodings for Vision Transformers (CPVT)

CPVT was proposed by Chu et al. in 2023 (Chu et al. 2021b). In their work, they devised a new scheme of conditional positional embeddings to improve the performance of ViTs (Fig. 11). In this regard, they proposed Positional Encoding Generators (PEGs) which utilized depth-wise convolutions to make positional embeddings more local and translational equivalent. They also developed a ViT based on the proposed scheme that utilized their PEGs to incorporate more positional information into their architecture and showed good results. In addition, they also showed that instead of the class token, the global average pooling layer above the final MLP layer resulted in boosted performance. Xiao et al. in their study estimated that utilizing CNN layers at early layers in the ViTs can increase their performance (Xiao et al. 2021). For comparison, they replaced the conventional ViT patching with a convolutional stem and reported more generalized and enhanced performance.

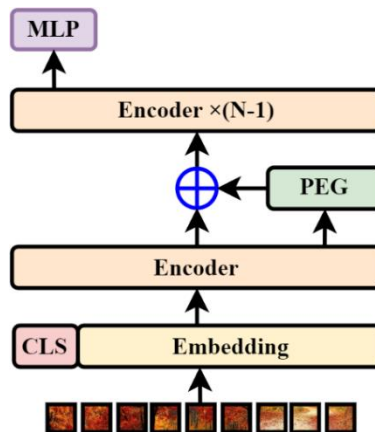


Figure 11: Architecture of CPVT, which incorporated CNN in their PEG block.

3.6.2. Lateral-layer integration

Models that use a CNN layer or block at the end of the transformer network, such as in place of the last linear layer, or as a post-processing layer fall under this category.

Dense Prediction Transformer (DPT)

Ranftl et al. proposed a dense prediction transformer “DPT” for segmentation in natural images. DPT has an encoder-decoder-based design, with a ViT as the encoder and a CNN as the decoder. It captured the global perspective and long-range dependencies by the backbone architecture. The learned global representations were then decoded into image-based embeddings taken by utilizing a CNN. Outputs from the ViT-based encoder were decoded at different levels to carry out dense predictions (Ranftl et al. 2021).

Local Vision Transformer (LocalViT)

Li et al, in their research, also incorporated locality into ViT architecture for image classification. The architecture of LocalViT is just like a conventional ViT, with its MSA module specialized to capture global-level features of images. The feed-forward network in ViT encoders performs final predictions by taking input from the learned encodings from the attention module. LocalViT modifies its FFN to incorporate local information into its architecture by employing depth-wise convolution (Li et al. 2021c).

3.6.3. Sequential integration

This category describes some of the popular hybrid ViTs that leveraged the benefits of CNN in their ViT architectures by following some sequential integration (Wang et al. 2023c).

Convolution and Attention Networks (CoAtNet)

Dai et al. carried out an extensive study to find out the most optimal and efficient way of merging convolutions and attention mechanisms in a single architecture to increase its generalization and capacity (Dai et al. 2021). In this regard, they introduced CoAtNet, by vertically stacking several convolutional and transformer blocks. For the convolutional blocks, they employed MBConv blocks which are based on depth-wise convolutions. Their findings suggested that stacking two convolutional blocks followed by two transformers blocks, sequentially showed efficient results.

CNNs Meet Transformers (CMT)

Despite their successful performance, ViTs face three main issues, a) inability to capture low-level features by considering correlation in the local neighborhood, b) expensive in terms of computation and memory consumption due to their MSA mechanism, c) and fixed sized input tokens, embedding. To overcome these issues, there comes the boom of hybridization of CNNs and ViTs after 2021. Guo et al. in 2021 also proposed a hybrid ViT, named CMT (CNNs Meet Transformers) (Guo et al. 2021). Inspired by CNNs (Tan and Le 2019), CMT also consists of an initial stem block followed by the sequential stacking of the CNN layer and CMT block. The designed CMT block was inspired by the ViT architecture, therefore contained a lightweight MSA block in place of conventional MSA, and the MLP layer was replaced with an inverted residual feed-forward network (IRFFN). In addition, a local perception unit (LPU) is added to the CMT block to increase the representation capacity of the network. The architecture is shown in Fig. 12.

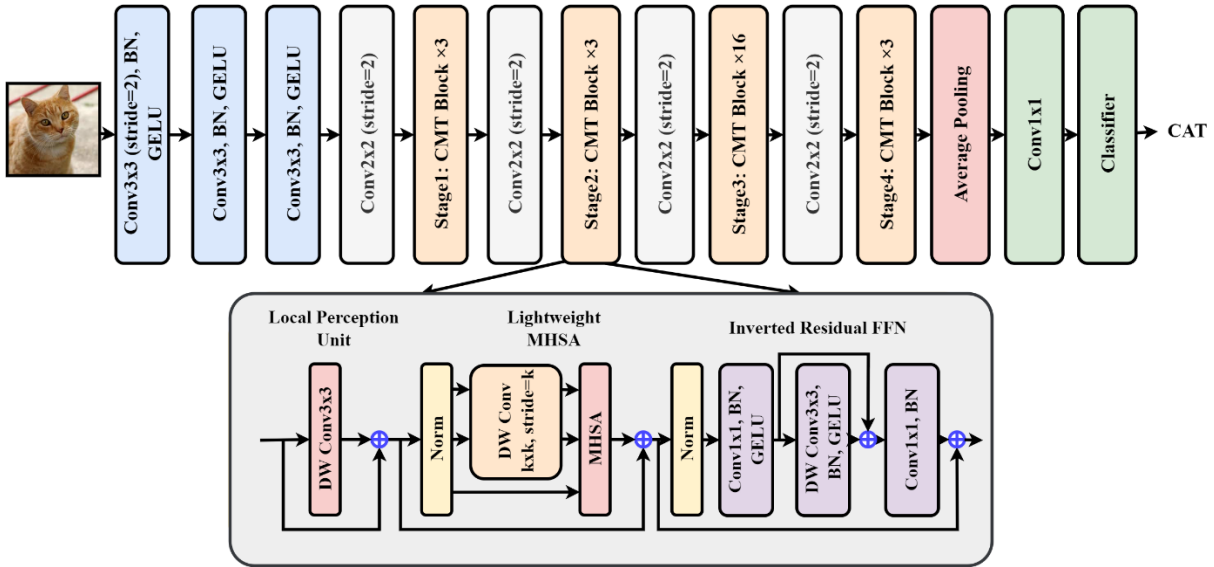


Figure 12: Architecture of CMT, with the integration of CNN in sequential order

Bottleneck Transformers (BoTNet)

Since the convolutional layer captures low-level features that are the main building blocks of many structuring elements in the images, Srinivas et al, introduced a hybrid ViT, BoTNet (Bottleneck Transformers for Visual Recognition) to benefit both from the CNN and ViT (Srinivas et al. 2021). The architecture of BoTNet is simply a sequential combination of ResNet blocks where the attention mechanism is incorporated in the last three blocks. ResNet block contains two 1x1 convolutions and a 3x3 convolution. The MSA is added in place of 3x3 convolution to capture long-term dependencies in addition to local features.

3.6.4. Parallel integration

This category includes those HVT architectures that use both CNNs and transformer architectures in parallel and their predictions are then combined in the end (Wang et al. 2021a).

Convolution-augmented Transformer (Conformer)

In 2021, Peng et al. conducted a study to perform visual recognition in natural images. In this regard, they proposed an architecture named, Conformer (Peng et al. 2021). Due to the popularity

of ViTs the architecture of Conformer was also based on ViTs. To improve the perception capacity of the network, they integrated the benefits of CNN and to multi-head self-attention mechanism. Conformer, a hybrid ViT, contained two separate branches, a CNN branch to capture local perceptions and a transformer branch to capture global-level features. Subsequent connections were built from the CNN branch to the transformer branch to make each branch local-global context-aware. Final predictions were obtained from a CNN classifier and a transformer classifier. Cross-entropy loss function was used to train each classifier. Conformer showed better performance than other outperforming ViT architectures such as DeiT, and ViT.

MobileNet-based Transformer (MobileFormer)

Chen et al. proposed a concurrent hybrid ViT architecture with two different pathways for a CNN and transformer (Chen et al. 2022e). Like other hybrid ViTs, MobileFormer employed the CNN model to learn spatial correlation and used a transformer to capture long-term dependencies in the images, thus fusing both the local correlation and global representations. The CNN architecture was based on MobileNet, which uses inverted residual blocks with a reduced number of parameters. Information among both the branches was synchronized using connections, which kept the CNN pathway aware of global information and the transformer aware of local information. Concatenated output from both branches followed by a pooling layer was then fed to a two-layer classifier for final predictions. Fig. 13 shows their detailed architecture.

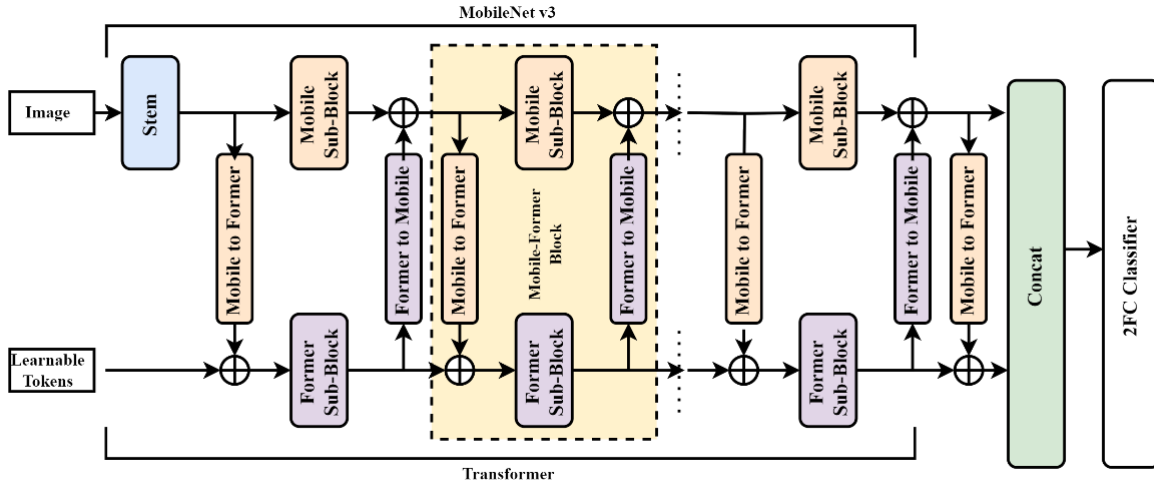


Figure 13: Architecture of Mobile-former (CNN and transformer with parallel integration)

Block-wisely Self-supervised Neural Architecture Search (BossNAS)

Li et al. developed a search space (HyTra) to evaluate hybrid architectures and advised that each block should be trained separately (Li et al. 2021a). In every layer within the HyTra search space, they utilized CNN and transformer blocks with various resolutions in parallel and freely selectable form. This broad search area includes conventional CNNs with progressively smaller spatial scales and pure transformers with fixed content lengths.

3.6.5. Hierarchical integration

Those HVT architectures that adopt a hierarchical design, similar to CNNs, fall under this category. Many of these models have designed a unified block for integrating CNN and ViT, which is then repeated throughout the architecture (Tu et al. 2022b).

Multi-Axis Attention-based Vision Transformer (MaxViT)

MaxViT is a variant of the ViT architecture that was introduced by Tu et al., in their paper "Multi-Axis Attention Based Vision Transformer" (Tu et al. 2022b). It introduced the Multi-Axis attention mechanism consisting of blocked local and dilated global attention. It proved to be an efficient and

scalable attention mechanism as compared to previous architectures. A new hybrid block was introduced as a basic element, which consists of MBConv-based convolution and Multi-Axis based attention. The basic hybrid block was repeated over multiple stages to obtain a hierarchical backbone, similar to CNN-based backbones that can be used for classification, object detection, segmentation, and generative modeling. MaxViT can see locally and globally across the whole network, including the earlier stages.

Convolutional Vision Transformer (CvT)

CvT was introduced by Wu et al. in 2021 (Wu et al. 2021a). The architecture of CvT contained several stages like CNNs to make up a hierarchical framework. They added convolution in their architecture in two ways. At first, they used a convolutional token embedding to extract token sequences, which not only incorporated locality in the network but also shortened the sequence length gradually. Secondly, they proposed a convolutional projection that used depth-wise separable convolution to replace the linear projection before each self-attention block in the encoder block. CvT outperformed other approaches for image recognition.

Vision-Friendly Transformer (Visformer)

Visformer was introduced as a vision-friendly transformer in 2020 (Chen et al. 2021d) presenting a modular design for efficient performance. The architecture had several modifications to a conventional ViT network. In Visformer, global average pooling was employed in place of classification token, and layer normalization was replaced with batch normalization. In addition, they utilized convolutional blocks inspired by ResNeXt (Xie et al.) instead of self-attention in each stage to efficiently capture both spatial and local features. However, to model the global dependencies they adopted self-attention in the last two stages. Another notable modification in Visformer's architecture was the addition of 3x3 convolutions in the MLP block.

Vision Transformer Advanced by Exploring intrinsic Inductive Bias (ViTAE)

The authors suggested a novel ViT architecture called ViTAE, that combines two different basic cell types (shown in Fig. 14): reduction cells (RC) and normal cells (NC) (Xu et al. 2021b). RCs are used to downscale input images and embed them into enriched multi-scale contextual tokens, while NCs are used to model local and long-term dependencies concurrently within the token sequence. The underlying structure of these two types of cells is also similar, consisting of parallel attention modules, convolutional layers, and an FFN. RC includes contextual information in tokens by utilizing several convolutions with different dilation rates in the pyramid reduction module. The authors also presented a more optimized version, ViTAEv2, that showed better performance than earlier method (Zhang et al. 2022d).

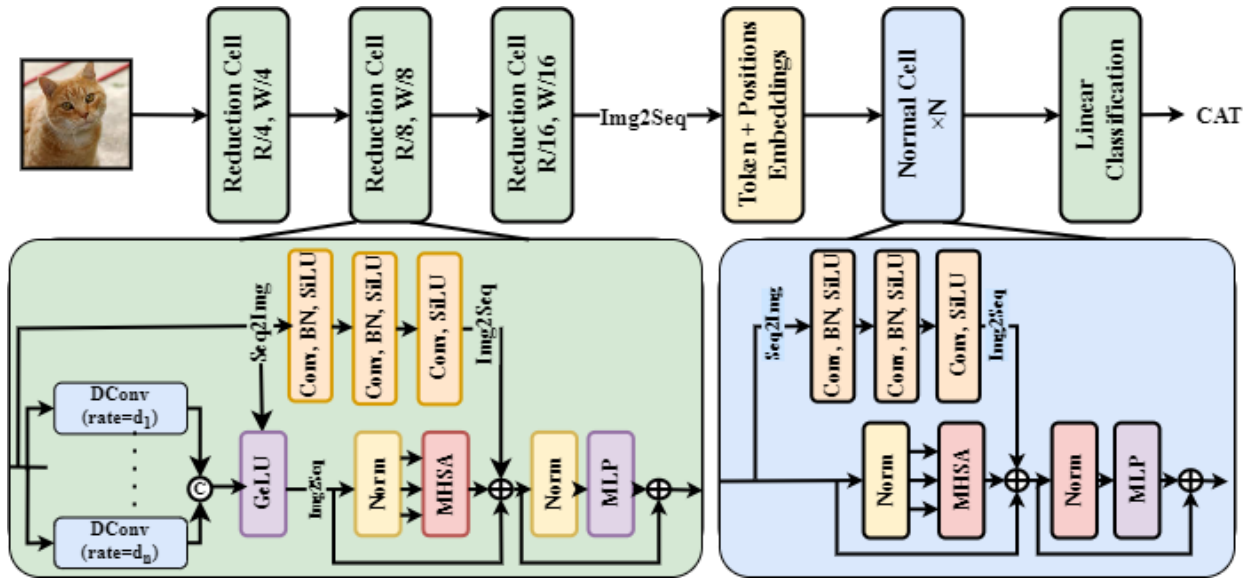


Figure 14: Architectural diagram of ViTAE

Convolution-Transformer Network (ConTNet)

A novel Convolution-Transformer Network (ConTNet) is proposed for computer vision tasks to address the challenges faced in this area. The ConTNet is implemented by stacking multiple ConT blocks (Yan et al.) (shown in Fig. 15). The ConT block treats the standard transformer encoder (STE) as an independent component similar to a convolutional layer. Specifically, a feature map is divided into several patches of equal size and each patch is flattened to a (super) pixel sequence, which is then input to the STE. After reshaping the patch embeddings, the resulting feature maps are then passed on to the next convolutional layer or to the STE module.

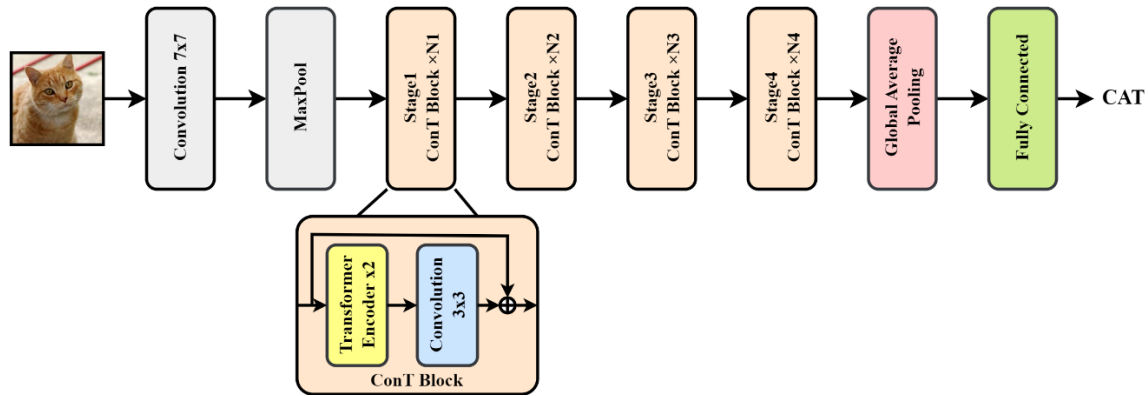


Figure 15: Architecture of ConTNet, which integrated CNN and ViT in its ConT block to form a hierarchical architecture.

3.6.6. Attention-based integration

This section discusses those HVT architectures, which have utilized CNNs in their attention mechanism to incorporate locality.

Evolving Attention with Residual Convolutions (EA-AA-ResNet)

Due to the limited generalizability of independent self-attention layers in capturing underlying dependencies between tokens, Wang et al. extended the attention mechanism by adding convolutional modules (Wang et al.). Specifically, they adopted a convolutional unit with residual

connections to generalize the attention maps in each layer by exploiting the knowledge inherited from previous layers, named as Evolving Attention (EA). The proposed EA-AA-ResNet architecture extends attention mechanisms by bridging attention maps across different layers and learning general patterns of attention using convolutional modules.

ResNet Transformer (ResT)

A hybrid architecture that integrates convolution operation in its attention mechanism, allowing it to capture both global and local features effectively (Zhang and Yang 2021). The authors utilized a new efficient transformer block in their architecture where they replaced the conventional MSA block with its efficient variant. In the proposed efficient multi-head self-attention, they employed depth-wise convolution to reduce the spatial dimensions of the input token map before computing the attention function.

Convolution-Enhanced Image Transformer (CeiT)

CeiT was proposed by Yuan et al. in 2021 in their paper “Incorporating Convolution Designs into Visual Transformers” (Yuan et al. 2021a). The proposed CeiT combined the benefits of CNNs and ViTs in extracting low-level features, capturing locality, and learning long-range dependencies. In their CeiT, they made three main advancements in conventional ViT architecture. They modified the patch extraction scheme, the MLP layer and added a last layer above the ViT architecture. For patch extraction, they proposed an Image-to-Tokens (I2T) module in which they utilized CNN-based blocks to process the inputs. Instead of utilizing raw input images, they used low-level features learned from the initial convolutional blocks to extract patches. I2T contained convolutional, max pooling, and batch normalization layers in its architecture to fully leverage the benefits of CNNs in ViTs. They utilized a Locally-enhanced Feed-Forward (LeFF) layer in place of the conventional MLP layer in the ViT encoder, in which depth-wise convolutions were utilized

to capture more spatial correlation. In addition, a last class token attention (LCA) layer was devised to systematically combine the outputs from different layers of ViT. CeiT not only showed promising results on several image and scene recognition datasets (including ImageNet, CIFAR, and Oxford-102) but is also computationally efficient as compared to ViT.

3.6.7. Channel boosting-based integration

Channel boosting (CB) is an idea used in DL to increase the representation learning ability of CNN models. In CB, besides the original channels, boosted channels are generated using transfer learning-based auxiliary learners to capture diverse and complex patterns from images. CB-based CNNs (CB-CNN) have shown outstanding performance in various vision-related tasks. In a study by Ali et al., they proposed a CB-based HVT architecture (Ali et al. 2023b). In CB-HVT they utilized CNNs and ViT-based auxiliary learners to generate boosted channels. The CNN-based channels captured local-level diversity in the image patterns, whereas Pyramid Vision Transformer (PVT)-based channels learned global-level contextual information. The authors evaluated CB-HVT on the lymphocyte assessment dataset, where it showed reasonable performance. An overview of their architecture is shown in Fig. 16.

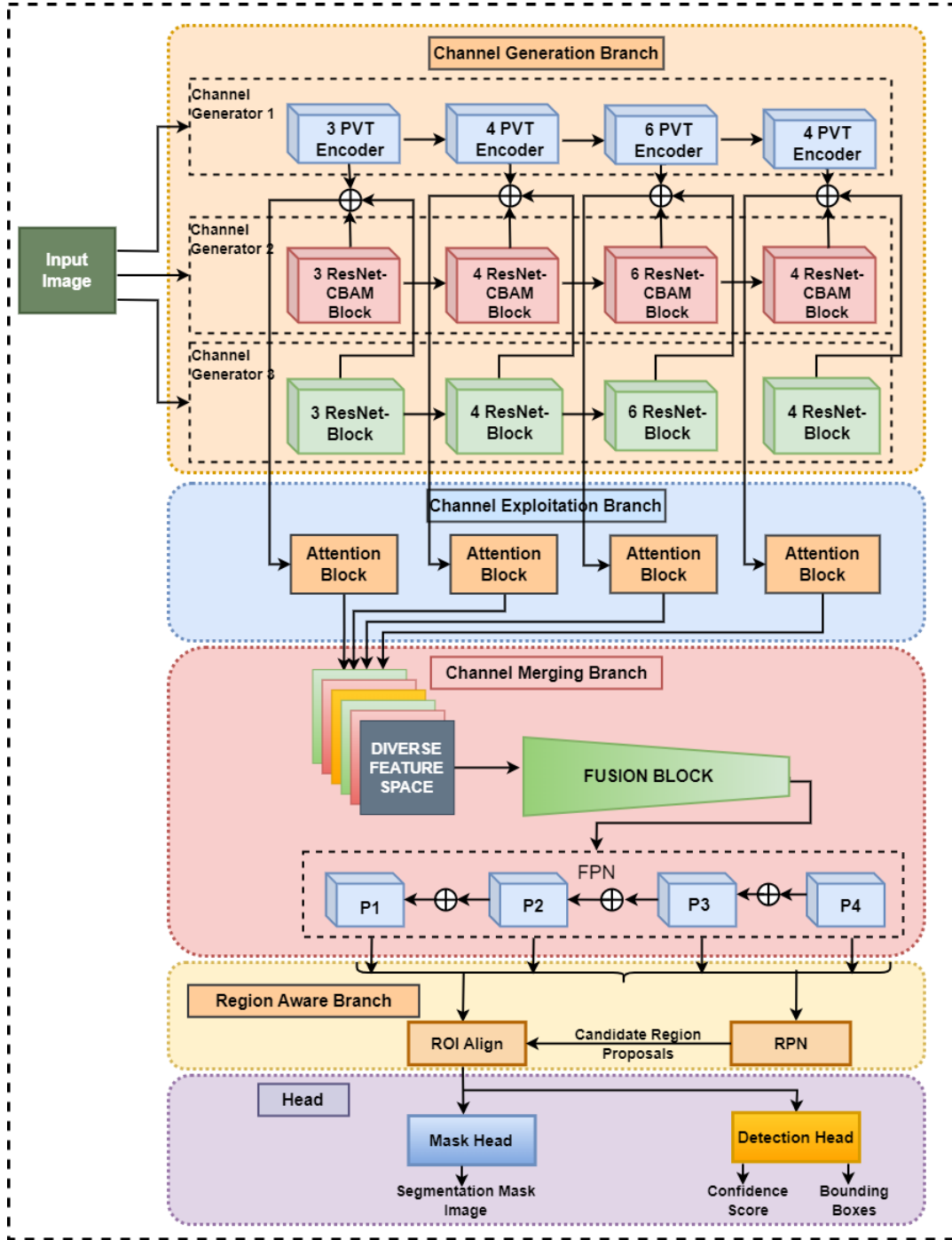


Figure 16: Overview of CB-HVT, where PVT (a VIT) is combined within CNN architecture using channel boosting.

3.7. Empirical comparison of different methods

In this section, we present a brief yet comprehensive empirical comparison of several ViT and HVT architectures that have demonstrated exceptional performance across various computer vision tasks. To get insights into their strengths and weaknesses, we have provided a detailed overview in Table 3 and Table 4. In addition, we have also highlighted the primary modifications made in each model, along with the underlying rationale, as per their taxonomy.

Table 3: Empirical comparison of various ViT architectures, based on their strengths, weaknesses, rationale, and performance on benchmark datasets (For comparison, we have reported the results of the best performing variants of the mentioned architecture).

Architecture	Strength (Architectural modification)	Weakness	Rationale	Metric
Patch-Based Approaches				
T2T-ViT (Yuan et al. 2021b) CODE	Utilized Token-to-Token module for iterative patching.	May have higher computational requirements due to increased tokenization.	To improve the representation by focusing on tokens instead of patches.	ImageNet 83.3% Top-1 Acc @ 384x384
TNT-ViT (Han et al. 2021) CODE	Utilizes multi-level patching, to capture objects with spatial and size variations.	May require more parameters, leading to increased model size.	To capture the attention inside the local patches.	ImageNet 82.9% Top-1 Acc @ 224x224
DPT (Chen et al. 2021e) CODE	Used DePatch, to have patches of variable sizes.	Could be sensitive to the selection of deformable patches, impacting performance.	For better handling of irregular-shaped objects in the image.	ImageNet 81.9% Top-1 Acc @ 224x224
CrowdFormer (Yang et al. 2022b)	Models global context by learning features at different scales, effective for crowd counting.	Could be computationally expensive for real-world applications.	The global context is incorporated to deal with the uneven distribution of crowds.	NWPU 67.1 MAE, 301.6 MSE @ 512x512
Knowledge Transfer-Based Approaches				

DeiT (Touvron et al. 2020) CODE	Based on a teacher-student model, where knowledge from teacher is transferred to student model	Performance is highly dependent upon the transferred knowledge.	Enables efficient image representation, suitable for resource-constrained scenarios.	ImageNet 85.2% Top-1 Acc @ 384x384
TaT (Lin et al. 2022) CODE	Utilized one-to-all spatial. matching knowledge distillation approach	Relatively computationally expensive due to one-to-many mapping	For an effective knowledge transfer one-to-many mapping was done between the teacher and the student models.	ImageNet 72.41% Top-1 Acc @ 513x513
TinyViT (Wu et al. 2022a) CODE	Compact model with reduced output logits of teacher model to have a light model.	May not achieve the same level of accuracy as larger, more complex models.	Provides an efficient and lightweight model, efficient for deployment.	ImageNet 86.5% Top-1 Acc @ 512x512
Shifted Window-Based Approaches				
Swin Transformer (Liu et al. 2021c) CODE	Shifted windowing scheme allows to efficiently compute self-attention over large images.	Relatively computationally expensive, which may limit its applicability to real-time applications.	To enable cross-window connections, which can help in enhancing the learning capacity of the model.	ImageNet 84.5% Top-1 Acc @ 384x384
Attention-Based Approaches				
CaiT (Touvron et al. 2021) CODE	LayerScale and class-attention stage significantly improve the accuracy of deep transformers.	Performance may be influenced by class attention accuracy and distribution.	To give more focus to important classes, leading to better classification.	ImageNet 86.5% Top-1 Acc @ 448x448
DAT (Xia et al. 2022) CODE	Utilizes deformable attention mechanisms, improving feature representation.	The deformable attention may have increased computational cost.	To enable better modeling of local image structures and deformable objects.	ImageNet 84.8% Top-1 Acc @ 384x384
SeT (Sun et al. 2022)	Factorizes the spatial attention into pixel-	May have limited representation	To effectively capture both fine-grained and	ImageNet 83.3%

	wise and patch-wise attention, which reduces the computational cost.	capacity compared to non-separable approaches.	coarse-grained features.	Top-1 Acc @ 224x224
Multi-Transformer-Based Approaches				
CrossViT (Chen et al. 2021a) CODE	Capable of modeling multi-scale feature representations in linear time.	Does not explore the use of different token fusion strategies.	To explore multi-scale tokens and demonstrate the effect of dual-branch feature fusion	ImageNet 82.8% Top-1 Acc @ 224x224
Dual-ViT (Yao et al.) CODE	The semantic pathway captures global semantics and a pixel pathway learns finer local details.	Complex architecture as compared to other SOTA Transformers which could make it difficult to train.	To capture global and local features simultaneously and effectively.	ImageNet 85.7% Top-1 Acc @ 384x384
MMViT (Liu et al. 2023b) CODE	Combines multi-scale and multi-view images, improving performance and robustness.	May require additional pre-processing steps to handle multi-view images.	To enable comprehensive image understanding by leveraging multiple scales and views.	ImageNet 83.2% Top-1 Acc @ 224x224
MPViT (Lee et al. 2021b) CODE	Multi-scale patch embedding and multi-path structure helps in learning fine and coarse feature representations.	Increased depth and complexity may lead to higher training time and resource usage.	To demonstrate that multi-scale and multi-path structures are beneficial for dense prediction tasks.	ImageNet 83.4% Top-1 Acc @ 384x384

Table 4: Empirical comparison of several HVT architectures, based on their strengths, weaknesses, rationale and performance on benchmark datasets (For comparison, we have reported the results of the best performing variants of the mentioned architecture)

Architecture	Strength	Weakness	Rationale	Metric
Early-Layer Integration				

DETR (Carion et al. 2020) CODE	End-to-end object detection with transformers, eliminating the need for separate region proposal networks.	The approach maybe computationally expensive, especially for large images.	To demonstrate the effectiveness of transformers in dense prediction tasks.	COCO 44.9% Box mAP @ 384x384
LeViT (Graham et al. 2021) CODE	Utilized convolution layers in initial layers to capture local features and reduce image resolutions.	May require more memory than some of the best CNNs.	To combine the strengths of both CNNs and ViTs, while avoiding some of their weaknesses.	ImageNet 82.6% Top-1 Acc @ 384x384
CPVT (Chu et al. 2021b) CODE	Utilized a new scheme for conditional position encodings for boosted performance	May have increased complexity and computational requirements.	To make positional embeddings translational invariant, they used depth-wise convolutions	ImageNet 82.7% Top-1 Acc @ 224x224
Lateral-Layer Integration				
DPT (Ranftl et al. 2021) CODE	Utilized a ViT-based and a CNN-based decoder	Higher memory requirements compared to some traditional dense prediction models.	To demonstrate the suitability of transformers for dense prediction tasks by capturing long-range dependencies between pixels.	ADE20K 49.02% IoU @ 520x520
LocalViT (Li et al. 2021c) CODE	Introduces depth-wise convolutions into its FFN.	May have limitations in capturing global context information effectively.	To explicitly incorporate local dependencies.	ImageNet 80.8% Top-1 Acc @ 224x224
Sequential Integration				
CoAtNet (Dai et al. 2021) CODE	Integrates convolutional layers within MSA.	Not as accurate as other ViTs on large datasets.	To effectively combine the strengths of both convolutional and MSA mechanism	ImageNet 86.0% Top-1 Acc @ 512x512
CMT (Guo et al. 2021) CODE	Incorporated a lightweight MSA	Complex than traditional CNNs, difficult to implement and train.	To integrate the strengths of both CNNs and Transformers for vision tasks while avoiding some weaknesses.	ImageNet 84.8% Top-1 Acc @ 288x288

BoTNet (Srinivas et al. 2021)	Employs bottleneck structures to improve memory efficiency and computational speed.	Performance may be influenced by the choice of bottleneck design and model depth.	To enhance the efficiency and speed of ViTs through bottleneck structures.	ImageNet 84.7% Top-1 Acc @ 224x224
Parallel Integration				
Conformer (Peng et al. 2021) CODE	Feature Coupling Unit (FCU) allows for efficient fusion of local features and global representations.	As compared to traditional CNNs, it may have complex to train and deploy.	To combine the strengths of both CNNs and self-attention mechanisms, while mitigating their weaknesses.	ImageNet 84.1% Top-1 Acc @ 224x224
Mobile-Former (Chen et al. 2022e) CODE	The bridge between MobileNet and transformer enables bidirectional fusion of local and global features.	The light-weight cross attention in the bridge may not be able to fully capture the interactions between local and global features.	To provide parallel interaction of MobileNet and transformer, allowing the model to achieve a good balance between efficiency and representation power.	ImageNet 79.3% Top-1 Acc @ 224x224
BossNAS (Li et al. 2021a) CODE	Can effectively search for hybrid CNN-transformer architectures.	Can be computationally expensive to train, especially for large search spaces.	Large and diverse search space of hybrid architectures makes it difficult for traditional NAS methods to be effective.	ImageNet 82.5% Top-1 Acc @ 512x512
Hierarchical Integration				
MaxViT (Tu et al. 2022b) CODE	Introduces a number of novel ideas, including multi-axis attention, hierarchical stacking, and linear-complexity global attention.	Can be more difficult to train because of the complex attention mechanism and may require more data to achieve good results.	To enable local and global feature extraction through self-attention in linear time.	ImageNet 86.70% Top-1 Acc @ 512x512
CvT (Wu et al. 2021a) CODE	Combines convolutional and MSA blocks, striking a balance between efficiency and accuracy.	Performance may be influenced by the specific configuration of the CvT architecture.	Integrates convolutional and ViT elements for effective vision tasks.	ImageNet 87.7% Top-1 Acc @ 384x384
Visformer (Chen et al. 2021d) CODE	Optimizes transformer architecture for vision tasks, considering image-specific challenges.	May require further architectural advancements to achieve state-of-the-art performance.	To tailor the transformer architecture for vision-specific challenges.	ImageNet 82.19% Top-1 Acc @ 224x224

ViTAE (Xu et al. 2021b) CODE	Introduces an inductive bias-aware architecture, improving generalization.	May not be as effective for tasks that require fine-grained visual reasoning	To incorporate the inductive biases to enhance model generalization and adaptability.	ImageNet 83.0% Top-1 Acc @ 384x384
ConTNet (Yan et al.) CODE	More robust to changes in the input data than transformer-based models.	Model complexity may increase due to the combination of convolution and transformers.	To obtain hierarchical features using both convolution and transformers for various vision-related tasks.	ImageNet 81.8% Top-1 Acc @ 224x224
Attention-Based Integration				
EA-AA-ResNet (Wang et al.) CODE	Evolves attention mechanisms with residual convolutions, enhancing feature representation.	May have higher computational cost compared to standard convolutional models.	To improve feature representation through evolving attention with residual convolutions.	ImageNet 79.63% Top-1 Acc @ 224x224
ResT (Zhang and Yang 2021) CODE	The introduction of Memory-Efficient MSA, Spatial Attention for positional encoding and Stack of conv. layers for patch embedding.	May be computationally more expensive than traditional CNN-based models.	To provide novel and innovative techniques that make transformer models efficient and versatile for visual recognition tasks.	ImageNet 83.6% Top-1 Acc @ 224x224
CeiT (Yuan et al. 2021a) CODE	Enhances ViTs with convolutional operations, improving efficiency and performance.	Model complexity may increase with the addition of convolutional operations.	To improve local features extraction of ViTs with convolutional components.	ImageNet 83.3% Top-1 Acc @ 384x384
Channel Boosting-Based Integration				
CB-HVT (Ali et al. 2023b)	Employs channel boosting for better feature representation, improving model accuracy.	Increased computational cost due to additional channel boosting computations.	To enhance feature representation through channel boosting in a hybrid architecture.	LYSTO 88.0% F-Score @ 256x256 NuClick 82.0% F-Score @ 256x256

4. Applications of ViTs and HVTs

In recent years, both ViTs and HVTs have gained prominence across a range of vision-based applications (Lee et al. 2021b; Deng et al. 2023; Xue and Ma 2023; Yan et al. 2023; Cheng et al. 2023; Lian et al. 2023; Chen et al. 2023d; Dehghani et al. 2023; Liu et al. 2023a; Ye et al. 2023a), including image and video recognition (Pecoraro et al. 2022; Zhang and Zhang 2022; Chen et al. 2023a; Xia et al. 2023; Mogan et al. 2023; Liang et al. 2023), object detection (Wang et al. 2023c) (Dehghani-Dehcheshmeh et al. 2023; Huang et al. 2023a; Yu and Zhou 2023), segmentation (Chen et al. 2021b; Li et al. 2023e; Quan et al. 2023), image restoration (Zhou et al. 2023a), and medical image analysis (Gao et al. 2021; An et al. 2022; Chen et al. 2022a; Song et al. 2022a; Zhang et al. 2022b; Yang and Yang 2023; Nafisah et al. 2023; Wu et al. 2023c). Transformer-based modules and CNNs are combined to create HVTs, an effective approach that can interpret intricate visual patterns (Li et al. 2023a). Some notable applications for ViTs and HVTs are discussed below.

4.1. Image/video recognition

CNNs have been extensively utilized for image and video processing due to their capability to automatically extract complex information from visual data (Fan et al. 2016; Fang et al. 2018; Yao et al. 2019; Kaur et al. 2022; Rafiq et al. 2023). Nevertheless, ViTs have revolutionized the field of computer vision by achieving outstanding performance on various challenging tasks, including image and video recognition (Chen and Ho 2022; Jing and Wang 2022; Chen et al. 2022c, b; Wensel et al. 2022; Ulhaq et al. 2022). The success of ViTs can be attributed to their self-attention mechanism, which enables them to capture long-range dependence in images (Ji et al. 2023). In recent years, HVTs have gained popularity, as they combine the power of both CNNs and transformers (Zhang et al. 2022a; Li et al. 2023g; Ma et al. 2023a). Various methods have been

proposed based on HVTs for recognition in both images and videos (Jiang et al. 2019; Li et al. 2021b; Huang et al. 2021a; GE et al. 2021; Yang et al. 2022a; Leong et al. 2022; Zhao et al. 2022a; Raghavendra et al. 2023; Zhu et al. 2023b). Xiong et al. proposed a hybrid multi-modal approach based on ViT and CNN to enhance fine-grained 3D object recognition (Xiong and Kasaei 2022). Their approach encodes the global information of the object using the ViT network and the local representation of the object using a CNN network through both RGB and depth views of the object. Their technique outperforms both CNN-only and ViT-only baselines. In another technique, Tiong et al. presented a novel hybrid attention vision transformer (HA-ViT) to carry out face-periocular cross identification (Tiong et al. 2023). HA-ViT utilizes depth-wise convolution and convolution-based MSA concurrently in its hybrid attention module in parallel to integrate local and global features. The proposed methodology outperforms three benchmark datasets in terms of Face Periocular Cross Identification (FPCI) accuracy. Wang et al. proposed a novel approach for visual place recognition using an HVT-based architecture (Wang et al. 2022d). Their method aims to improve the robustness of the visual place recognition system by combining both CNN and ViT to capture local details, spatial context, and high-level semantic information. To recognize vehicles Shi et al. developed a fused network that used SE-CNN architecture for feature extraction followed by the ViT architecture to capture global contextual information (Shi et al. 2023). Their proposed approach demonstrated good accuracy values for the road recognition task.

4.2. Image generation

Image generation is an interesting task in computer vision and can serve as a baseline for many downstream tasks (Frolov et al. 2021). Generative adversarial networks (GANs) are widely used for image generation in various domains (Arjovsky et al. 2017; Karras et al. 2019). Additionally, ViT-based GANs have shown promising performance in this task (Lee et al. 2021a; Naveen et al.

2021; Rao et al. 2022; Gao et al. 2022b). Recently, researchers have also utilized HVT-based GANs and demonstrated outstanding performance on various benchmark datasets (Tu et al. 2022a; Lyu et al. 2023). Torbunov, et al. reported UVCGAN, a hybrid GAN model, for image generation (Torbunov et al. 2022). The architecture of the UVCGAN model is based on the original CycleGAN model (Zhu et al. 2017) with some modifications. The generator of UVCGAN is a hybrid architecture based on a UNet (Weng and Zhu 2015) and a ViT bottleneck (Devlin et al. 2018). Experimental results demonstrated its superior performance compared to earlier best performing models while retaining a strong correlation between the original and generated images. In another work, SwinGAN was introduced for MRI reconstruction by Zhao, et al (Zhao et al. 2023). They utilized Swin Transformer U-Net-based generator network and CNN-based discriminator network. The generated MRI images by SwinGAN showed good reconstruction quality due to its ability to capture more effective information. Tu et al. combined the Swin transformer and CNN layers in their proposed SWCGAN (Tu et al. 2022a). In their architecture they utilized CNN layers initially to capture local level features and then in later layers utilized Residual Dense Swin Transformer Blocks “RDST” to capture global level features. The developed method showed good reconstruction performance compared to existing approaches in remote sensing images. Recently, Bao et al. proposed a spatial attention-guided CNN-Transformer aggregation network (SCTANet) to reconstruct facial images (Bao et al. 2023b). They utilized both CNN and transformer in their Hybrid Attention Aggregation (HAA) block for deep feature extraction. Their experimental results demonstrated better performance than other techniques. Zheng et al. in their approach presented a HVT-based GAN network for medical image generation (Zheng et al. 2023). In their approach, named L-former they utilize transformers in the shallow

layers and CNNs in the deeper layers. Their approach demonstrated outperformance as compared to conventional GAN architectures.

4.3. Image segmentation

Although CNNs and ViT-based approaches have shown exceptional performance in complex image-related tasks such as image segmentation, there is currently an emphasis on combining the strengths of both approaches to achieve boosted performance (Dolz et al. 2019; Wang et al. 2020, 2022c, b; Jing et al. 2023; Shafri et al. 2023; Yang et al. 2023c). In this regard, Wang et al. presented a new semantic segmentation method called DualSeg for grapes segmentation (Wang et al. 2023a). Their method combines Swin Transformer and CNN to leverage the advantages of both global and local features. In another work, Zhou and co-authors proposed a hybrid approach named SCDeepLab to segment tunnel cracks (Zhou et al. 2023b). Their approach outperformed other CNN-only and transformer-only-based models in segmenting cracks in tunnel lining. Feng et al. carried out segmentation recognition in metal couplers to detect fracture surfaces (Feng et al. 2023). To this end, they proposed an end-to-end HVT-based approach by utilizing a CNN for automatic feature extraction and a hybrid convolution and transformer (HCT) module for feature fusion and global modeling. Recently, Xia and Kim developed Mask2Former, an HVT approach, to address the limitations of ViT or CNN-based systems (Xia and Kim 2023). The developed approach achieved better results as compared to other techniques on both the ADE20K and Cityscapes datasets. Li et al. proposed an HVT-based method called MCAFNet for semantic segmentation of remote sensing images (Li et al. 2023d).

4.4. Image Restoration

A crucial task in computer vision is image restoration, which tends to restore the original image from its corrupted version. Image restoration-based systems have generally shifted from the use of CNNs to ViT models (Ali et al. 2023a; Song et al. 2023) and more recently to HVTs that combine the strengths of both CNNs and transformers (Gao et al. 2022a; Wu et al. 2023c). Yi et al. proposed an auto-encoder based hybrid method to carry out single infrared image blind deblurring (Yi et al. 2023). Their approach utilizes hybrid convolution-transformer blocks for extracting context-related information between the objects and their backgrounds. To hasten the convergence of the training process and achieve superior image deblurring outcomes, the study also incorporated a multi-stage training technique and mixed error function. In another technique, Chen et al. developed an efficient image restoration architecture called Dual-former, which combines the local modeling ability of convolutions and the global modeling ability of self-attention modules (Chen et al. 2022d). The proposed architecture achieves superior performance on multiple image restoration tasks while consuming significantly fewer GFLOPs than previously presented methods. To address the issue of high computational complexity Fang et al. utilized a hybrid network, HNCT, for lightweight image super-resolution (Fang et al. 2022). HNCT leverages the advantages of both CNN and ViT and extracts features that consider both local and non-local priors, resulting in a lightweight yet effective model for super-resolution. Experimental results demonstrate that HNCT's improved results as compared to existing approaches with fewer parameters. Zhao et al. developed a hybrid denoising model, called Transformer Encoder and Convolutional Decoder Network (TECDNet), for efficient and effective real image denoising (Zhao et al. 2022b). TECDNet attained outstanding denoising results while maintaining relatively a low computational cost. Recently, Chen et al. presented an end-to-end HVT-based image fusion

approach for infrared and visible image fusion (Chen et al. 2023b). The proposed technique consists of a CNN module with two branches to extract coarse features, and a ViT module to obtain global and spatial relationships in the image. Their method was able to focus on global information and overcome the flaws of CNN-based methods. In addition, to retain the textural and spatial information a specialized loss function is designed.

4.5. Feature extraction

Feature extraction is essential in computer vision to identify and extract relevant visual information from images. Initially, CNNs were generally used for feature extraction in computer vision applications, but now ViTs have gained attention due to their impressive results in image classification as well as other applications like pose estimation, and face recognition (Wang et al. 2023d; Zhu et al. 2023a; Su et al. 2023).

Li and Li, in their work, presented a hybrid approach, ConVit, to merge the advantages of both CNNs and transformers for effective feature extraction to identify crop disease (Li and Li 2022). The experimental results of the developed approach showed good performance in the plant disease identification task. A cascaded approach was proposed by Li et al. for recaptured scene image identification (Li et al. 2023b). In their approach, they initially employed CNN layers to extract local features and later in the deeper layers, they utilized transformer blocks to learn global-level image representations. The high accuracy value of their proposed approach demonstrated its effectiveness in identifying recaptured images. Li and co-authors developed HVT architecture to detect defects in strip steel surfaces. Their approach utilized a CNN module, followed by a patch embedding block and two transformer blocks to extract high-domain relevant features. Their experiments showed good classification performance as compared to existing methods. Recently, Rajani et al. in their approach, proposed an encoder-decoder approach for categorizing different

seafloor types. Their developed method is a ViT-based architecture with its MLP block replaced with CNN-based feature extraction module. The modified architecture achieves outstanding results while meeting real-time computational requirements.

4.6. Medical image analysis

CNN-based approaches have been frequently employed for analyzing medical images due to their capability to capture diverse and complex patterns (Zafar et al. 2021; Sohail et al. 2021b; Rauf et al. 2023). Nevertheless, driven by the requirement to model image representations at a global level, researchers have found inspiration in employing ViTs within the field of medical image analysis (Obeid et al. 2022; Cao et al. 2023; Zou and Wu 2023; Li et al. 2023c; Zidan et al. 2023; Xiao et al. 2023). Recently, several studies have proposed integrating CNNs and ViTs to capture both local and global image features in medical images, allowing for more comprehensive analysis (Tragakis et al.; Springenberg et al. 2022; Wu et al. 2022c, 2023a; Jiang and Li 2022; Bao et al. 2023a; Dhamija et al. 2023; Huang et al. 2023b; Ke et al. 2023; Yuan et al. 2023a). These hybrid architectures (CNN-transformer) have shown tremendous performance in a number of medical images-related applications (Zhang et al. 2021c; Shen et al. 2022; Rehman and Khan 2023; Li et al. 2023f; Wang et al. 2023b). Tragakis, et al. proposed a novel Fully Convolutional Transformer (FCT) approach to segment medical images (Tragakis et al.). FCT adapted both ViT and CNN in its architecture by combining the ability of CNNs in learning effective image representations with the ability of Transformers to capture long-term dependencies. The developed approach showed outstanding performance on various medical challenge datasets as compared to other existing architectures. In another work, Heidari, et al. proposed HiFormer, an HVT to capture multi-scale feature representations by utilizing a Swin Transformer module and a CNN-based encoder (Heidari et al. 2022). Experimental results demonstrated the effectiveness of HiFormer in

segmenting medical images in various benchmark datasets. In their paper, Yang and colleagues presented a novel hybrid approach called TSEDeepLab, which combines convolutional operations with transformer blocks to analyze medical images (Yang et al. 2023a). Specifically, the approach utilizes convolutional layers in the early stages for learning local features, which are then processed by transformer blocks to extract global patterns. Their approach demonstrated exceptional segmentation accuracy and strong generalization performance on multiple medical image segmentation datasets.

4.7. Object Detection

Object detection is a crucial computer vision task with a wide range of real-world applications such as surveillance, robotics, crowd counting, and autonomous driving (Liu et al. 2023a). The progress of DL has significantly contributed to the advancements in object detection over the years (Er et al. 2023). ViTs have also shown impressive performance in object detection due to their self-attention mechanism that allows them to capture long-range dependencies between image pixels and identify complex object patterns across the entire image (Carion et al. 2020; Chen et al. 2021c; Wang and Tien 2023; Heo et al. 2023). Recently, there has been a lot of interest in HVTs for combining CNNs with self-attention mechanisms to improve object detection performance (Jin et al. 2021; Maaz et al. 2022; Mathian et al. 2022; Ye et al. 2023b; Zhang et al. 2023b; Lu et al. 2023a; Ullah et al. 2023). Beal et al. proposed an HVT approach named ViT-FRCNN for object detection in natural images. In their approach, they utilized a ViT-based backbone for a Faster R-CNN object detector. ViT-FRCNN showed improved detection results with a better generalization ability (Beal et al. 2020). Chen et al. introduced a single-stage hybrid detector for detection in remote sensing images. their proposed approach, MDCT leveraged both the CNNs and transformers in its architecture and showed better performance as compared to other single-stage

detectors (Chen et al. 2023c). Lu et al. developed an HVT-based approach for object detection in unmanned aerial vehicle (UAV) images (Lu et al. 2023b). The proposed approach utilized a transformer-based backbone to extract features with global-level information, which were then fed to FPN for multi-scale feature learning. The proposed method demonstrated good performance as compared to earlier approaches. Yao and his colleagues proposed a fusion network that utilizes individual transformer and CNN-based branches to learn global and local-level features (Yao et al. 2023). Experimental results showed satisfactory performance of the developed method as compared to other methods.

4.8. Pose Estimation

Human pose estimation tends to identify important points in various scenarios. Both CNNs and transformers have shown exemplary performance in pose estimation tasks (Sun et al. 2019; Huang et al. 2019; Cao et al. 2022). In a recent study, a method for multi-person pose estimation has been described, utilizing the multi-head attention mechanism of ViT (Yang et al. 2023b). Currently, researchers are focusing to combine CNNs and transformers in a unified method to incorporate both local and global level information for accurate pose estimation (Stoffl et al. 2021; Mao et al. 2021; Li et al. 2021d; Wu et al. 2022b). Zhao et al. presented a new Dual-Pipeline Integrated Transformer “DPIT” for human pose estimation (Zhao et al. 2022c). In Zhao’s approach initially, two CNN-based branches are employed to extract local features followed by the transformer encoder blocks to capture long-range dependencies in the image (Wang et al. 2022a). In another technique, Wang and coauthors used a CNN and a transformer branch to learn local and global image representations, which were then integrated to generate the final output. Their approach demonstrated significant improvement as compared to other existing approaches. Hampali and coauthors developed a hybrid pose estimation method, named Keypoint Transformer (Hampali et al.

2021). In the proposed method they utilized both CNN and transformer-based modules to efficiently estimate human joints as 2D keypoints. Experimental results showed exemplary results of this approach on datasets including InterHand2.6M.

5. Challenges

ViTs and HVTs have demonstrated exceptional performance not only in computer vision, but also in various other domains. Nonetheless, integrating convolutional operations effectively into the transformer architecture poses several challenges for HVTs. Some of the ViT and HVT challenges include:

- Humans have a capacity for leveraging rotation invariance and equivariance to comprehend the world within its context. However, deep learning models encounter difficulties in upholding equivariance/invariance as required. The challenge to equip ViTs with the appropriate inductive biases to effectively represent objects and video frames in an equivariant fashion is undeniably complex.
- The MSA mechanism in ViTs and the convolution operation in CNNs both rely on dense matrix multiplication to capture data dependencies. Therefore, both ViT and HVT architectures (CNN-Transformers) may face high computational complexity and memory overhead. As a result, they may encounter challenges when attempting to model dense applications such as volumetric analysis and segmentation.
- Training ViTs and HVTs generally require powerful hardware resources like GPUs due to their computational complexity. This can limit their deployment in real-world applications, especially on edge devices, due to the hardware constraints and associated costs.

- The intricate architecture of ViT poses challenges for its interpretability, particularly in the context of multi-head attention, where attention weights from each layer are intricately blended. As a result, comprehending which segments of the input tokens are actively involved in the decision-making process proves to be challenging.
- A major challenge faced by HVT architectures is the efficient merging of learned features from both transformer and convolutional layers. While the transformer layers learn global features that are independent of spatial location, convolutional layers learn local features that are spatially correlated. In architectural terms, the efficient unification of MSA and CNN layers can potentially result in improved performance in various vision tasks.
- ViTs and HVTs can process complex image data accurately due to their high learning capacity. However, this also means that they require large training datasets to effectively learn and generalize from the data. This poses a challenge, particularly in the medical image domain, where obtaining a large amount of annotated data is often difficult and time-consuming. The need for obtaining extensive labeled data can be a significant obstacle, consuming valuable resources and time, and impeding the development and application of HVTs in medical imaging.

6. Future directions

Understanding how ViTs make decisions is crucial due to their significant capabilities. As ViTs evolve, comprehending their inner workings becomes even more important. Knowing how the model functions is essential for understanding its decision-making instead of treating it as a black box. To improve ViTs, future research can explore different ways of interpreting the model. These methods include visualizing features, generating activation maps, and using techniques like Grad-

CAM. Such approaches are vital for understanding the model's behavior and improving its transparency. By revealing the decision-making process, biases and errors can be identified and addressed more effectively. This involves focusing on specific regions rather than arbitrary ones, improving model development and deployment.

ViTs and HVTs are large models with billions of parameters, which necessitates the need for lightweight architectures. Their high complexity may lead to latency in inference and significant overhead on energy consumption. There is a need to explore new and innovative design principles for efficient ViTs and HVTs with significant inference rates to enable their practical deployment in real-world applications, edge devices, and computationally limited systems, such as satellites. Knowledge distillation emerges as a promising approach to generate data-efficient and compact models by transferring knowledge from high-capacity models to simpler ones.

HVTs combine the strengths of CNNs and transformers, making significant advancements in image analysis and computer vision. However, to fully utilize their potential, it is important to explore suitable ways in which the convolution and self-attention mechanisms can be integrated for specific vision applications. This involves an in-depth analysis of integration methods based on their suitability for various contexts, such as early layer integration, lateral layer integration, sequential integration, parallel integration, hierarchical integration, attention-based integration, and attention-based integration.

The HVT's local and global processing capabilities make them quite promising for a wide range of vision applications, with potential benefits beyond vision-related tasks. To further enhance the performance of HVTs, it is important to gain a deeper understanding of image content and associated operations, which can help in devising better hybrid and deep architectures. The

investigation of the potential utilization of hand-crafted operators in combination with the hybrid and dynamic feature extraction mechanisms of CNN-Transformer architectures may be particularly important in the near future. Developing new and effective blocks using both convolution and self-attention mechanisms is also a promising area for research.

In summary, the future of HVTs looks bright, with immense potential for various applications in the field of image analysis, computer vision, etc. In our opinion, it is better to also focus on possible integration methods that merge self-attention and convolution layers within HVT architectures for specific vision tasks. This focus should also extend to understanding image content and operations, developing effective blocks that combine convolution and self-attention, and utilizing multimodality and multitasking in ViT and HVT architectures.

7. Conclusion

The ViT has gained considerable attention in research due to its promising performance in specific image-related tasks. This success can be attributed to the MSA module integrated into ViT architectures, enabling the modeling of global interactions within images. To enhance their performance, various architectural improvements have been introduced. These improvements can be categorized as patch-based, knowledge distillation-based, attention-based, multi-transformer-based, and hybrid approaches. This paper not only examines the architectural taxonomy of ViTs but also explores the fundamental concepts underlying ViT architectures.

While ViTs have impressive learning capacities, they may suffer from limited generalization in some applications due to their lack of inductive bias that can capture local relations in images. To address this, researchers have developed HVTs, also known as CNN-Transformers, which

leverage both self-attention and convolution mechanisms to learn both local and global information.

Several studies have proposed ways to integrate convolution-specific inductive bias into transformers to improve their generalization and capacity. Integration methodologies include early-layer integration, lateral-layer integration, sequential integration, parallel integration, hierarchical integration, and channel boosting-based integration. In addition to introducing taxonomy for HVT architectures based on their integration methodology, we also provide an overview of how they are used in various real-world computer vision applications. Despite current challenges, we believe that HVTs have enormous potential due to their capability to perform learning at both local and global levels.

Acknowledgments

This work has been conducted at the pattern recognition lab, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan. We extend our sincere gratitude to Dr. Abdul Majid and Dr. Naeem Akhter of DCIS, PIEAS for their invaluable assistance in improving the manuscript. Additionally, we acknowledge Pakistan Institute of Engineering and Applied Sciences (PIEAS) for a healthy research environment which led to the work presented in this article.

Competing interests

The authors declare no competing financial and/or non-financial interests about the described work.

References

- Agbo-Ajala O, Viriri S (2021) Deep learning approach for facial age classification: a survey of the state-of-the-art. *Artif Intell Rev* 54:179–213. <https://doi.org/10.1007/S10462-020-09855-0/TABLES/4>
- Aleissae AA, Kumar A, Anwer RM, et al (2022) Transformers in Remote Sensing: A Survey. <https://doi.org/10.3390/rs15071860>
- Ali AM, Benjdira B, Koubaa A, et al (2023a) Vision Transformers in Image Restoration: A Survey. *Sensors* 23:. <https://doi.org/10.3390/s23052385>
- Ali ML, Rauf Z, Khan A, et al (2023b) CB-HVTNet: A channel-boosted hybrid vision transformer network for lymphocyte assessment in histopathological images
- An L, Wang L, Li Y (2022) HEA-Net: Attention and MLP Hybrid Encoder Architecture for Medical Image Segmentation. *Sensors* 2022, Vol 22, Page 7024 22:7024. <https://doi.org/10.3390/S22187024>
- Arjovsky M, Chintala S, Bottou L (2017) Wasserstein GAN
- Bao H, Zhu Y, Li Q (2023a) Hybrid-scale contextual fusion network for medical image segmentation. *Comput Biol Med* 152:106439. <https://doi.org/10.1016/J.COMPBIOMED.2022.106439>
- Bao Q, Liu Y, Gang B, et al (2023b) SCTANet: A Spatial Attention-Guided CNN-Transformer Aggregation Network for Deep Face Image Super-Resolution. *IEEE Trans Multimed* 1–12. <https://doi.org/10.1109/TMM.2023.3238522>
- Beal J, Kim E, Tzeng E, et al (2020) Toward Transformer-Based Object Detection
- Bhatt D, Patel C, Talsania H, et al (2021) CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope. *Electron* 2021, Vol 10, Page 2470 10:2470. <https://doi.org/10.3390/ELECTRONICS10202470>
- Bi J, Zhu Z, Meng Q (2021) Transformer in Computer Vision. 2021 IEEE Int Conf Comput Sci Electron Inf Eng Intell Control Technol CEI 2021 178–188. <https://doi.org/10.1109/CEI52496.2021.9574462>
- Cao H, Wang Y, Chen J, et al (2023) Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. 205–218. https://doi.org/10.1007/978-3-031-25066-8_9
- Cao X, Li X, Ma L, et al (2022) AggPose: Deep Aggregation Vision Transformer for Infant Pose Estimation. *IJCAI Int Jt Conf Artif Intell* 5045–5051. <https://doi.org/10.24963/ijcai.2022/700>

- Carion N, Massa F, Synnaeve G, et al (2020) End-to-End Object Detection with Transformers. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 12346 LNCS:213–229. https://doi.org/10.1007/978-3-030-58452-8_13
- Chen CF, Fan Q, Panda R (2021a) CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. *Proc IEEE Int Conf Comput Vis* 347–356. <https://doi.org/10.48550/arxiv.2103.14899>
- Chen H, Li C, Wang G, et al (2022a) GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recognit* 130:108827. <https://doi.org/10.1016/J.PATCOG.2022.108827>
- Chen J, Chen X, Chen S, et al (2023a) ShapeFormer: Bridging CNN and Transformer via ShapeConv for multimodal image matching. *Inf Fusion* 91:445–457. <https://doi.org/10.1016/J.INFFUS.2022.10.030>
- Chen J, Ding J, Yu Y, Gong W (2023b) THFuse: An infrared and visible image fusion network using transformer and hybrid feature extractor. *Neurocomputing* 527:71–82. <https://doi.org/10.1016/J.NEUCOM.2023.01.033>
- Chen J, Ho CM (2022) MM-ViT: Multi-Modal Video Transformer for Compressed Video Action Recognition. 1910–1921
- Chen J, Hong H, Song B, et al (2023c) MDCT: Multi-Kernel Dilated Convolution and Transformer for One-Stage Object Detection of Remote Sensing Images. *Remote Sens* 2023, Vol 15, Page 371 15:371. <https://doi.org/10.3390/RS15020371>
- Chen J, Lu Y, Yu Q, et al (2021b) TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation
- Chen J, Zhang Y, Pan Y, et al (2023d) A transformer-based deep neural network model for SSVEP classification. *Neural Networks* 164:521–534. <https://doi.org/10.1016/J.NEUNET.2023.04.045>
- Chen S, Ge C, Tong Z, et al (2022b) Token Merging: Your ViT But Faster
- Chen S, Ge C, Tong Z, et al (2022c) AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition
- Chen S, Ye T, Liu Y, Chen E (2022d) Dual-former: Hybrid Self-attention Transformer for Efficient Image Restoration
- Chen S, Yu T, Li P (2021c) MVT: Multi-view Vision Transformer for 3D Object Recognition
- Chen Y, Dai X, Chen D, et al (2022e) MobileFormer: Bridging MobileNet and

- Transformer. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2022-June:5260–5269. <https://doi.org/10.1109/CVPR52688.2022.00520>
- Chen Z, Xie L, Niu J, et al (2021d) Visformer: The Vision-friendly Transformer. Proc IEEE Int Conf Comput Vis 569–578. <https://doi.org/10.1109/ICCV48922.2021.00063>
- Chen Z, Zhu Y, Zhao C, et al (2021e) DPT: Deformable Patch-based Transformer for Visual Recognition. MM 2021 - Proc 29th ACM Int Conf Multimed 2899–2907. <https://doi.org/10.1145/3474085.3475467>
- Cheng M, Ma H, Ma Q, et al (2023) Hybrid Transformer and CNN Attention Network for Stereo Image Super-resolution
- Chu X, Tian Z, Wang Y, et al (2021a) Twins: Revisiting the Design of Spatial Attention in Vision Transformers. Adv Neural Inf Process Syst 12:9355–9366
- Chu X, Tian Z, Zhang B, et al (2021b) Conditional Positional Encodings for Vision Transformers
- Dai Z, Liu H, Le Q V., Tan M (2021) CoAtNet: Marrying Convolution and Attention for All Data Sizes. Adv Neural Inf Process Syst 5:3965–3977. <https://doi.org/10.48550/arxiv.2106.04803>
- Dehghani-Dehcheshmeh S, Akhoondzadeh M, Homayouni S (2023) Oil spills detection from SAR Earth observations based on a hybrid CNN transformer networks. Mar Pollut Bull 190:114834. <https://doi.org/10.1016/J.MARPOLBUL.2023.114834>
- Dehghani M, Mustafa B, Djolonga J, et al (2023) Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution
- Deng Y, Meng Y, Chen J, et al (2023) TChange: A Hybrid Transformer-CNN Change Detection Network. Remote Sens 15:. <https://doi.org/10.3390/rs15051219>
- Devlin J, Chang MW, Lee K, Toutanova K (2018) 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 1:4171–4186
- Dhamija T, Gupta A, Gupta S, et al (2023) Semantic segmentation in medical images through transfused convolution and transformer networks. Appl Intell 53:1132–1148. <https://doi.org/10.1007/S10489-022-03642-W/FIGURES/9>
- Dolz J, Gopinath K, Yuan J, et al (2019) HyperDense-Net: A Hyper-Densely Connected CNN for Multi-Modal Image Segmentation. IEEE Trans Med Imaging 38:1116–1126. <https://doi.org/10.1109/TMI.2018.2878669>

- Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. <https://doi.org/10.48550/arxiv.2010.11929>
- Du Y, Liu Z, Li J, Zhao WX (2022) A Survey of Vision-Language Pre-Trained Models. *IJCAI Int Jt Conf Artif Intell* 5436–5443. <https://doi.org/10.24963/ijcai.2022/762>
- Er MJ, Zhang Y, Chen J, Gao W (2023) Ship detection with deep learning: a survey. *Artif Intell Rev* 1–41. <https://doi.org/10.1007/S10462-023-10455-X/TABLES/3>
- Fan Y, Lu X, Li D, Liu Y (2016) Video-Based emotion recognition using CNN-RNN and C3D hybrid networks. *ICMI 2016 - Proc 18th ACM Int Conf Multimodal Interact* 445–450. <https://doi.org/10.1145/2993148.2997632>
- Fang J, Lin H, Chen X, Zeng K (2022) A Hybrid Network of CNN and Transformer for Lightweight Image Super-Resolution. *IEEE Comput Soc Conf Comput Vis Pattern Recognit Work 2022-June*:1102–1111. <https://doi.org/10.1109/CVPRW56347.2022.00119>
- Fang W, Zhang F, Sheng VS, Ding Y (2018) A method for improving CNN-based image recognition using DCGAN. *Comput Mater Contin* 57:167–178. <https://doi.org/10.32604/CMC.2018.02356>
- Feng Q, Li F, Li H, et al (2023) Hybrid convolution and transformer network for coupler fracture failure pattern segmentation recognition in heavy-haul trains. *Eng Fail Anal* 145:107039. <https://doi.org/10.1016/J.ENGFAILANAL.2022.107039>
- Frolov S, Hinz T, Raue F, et al (2021) Adversarial text-to-image synthesis: A review. *Neural Networks* 144:187–209. <https://doi.org/10.1016/J.NEUNET.2021.07.019>
- Gao G, Xu Z, Li J, et al (2022a) CTCNet: A CNN-Transformer Cooperation Network for Face Image Super-Resolution. <https://doi.org/10.1109/TIP.2023.3261747>
- Gao P, Yang X, Zhang R, et al (2022b) Generalised Image Outpainting with U-Transformer. *Neural Networks* 162:1–10. <https://doi.org/10.1016/j.neunet.2023.02.021>
- Gao Y, Zhou M, Metaxas DN (2021) UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 12903 LNCS:61–71. https://doi.org/10.1007/978-3-030-87199-4_6/COVER
- GE C, Liang Y, SONG Y, et al (2021) Revitalizing CNN Attention via Transformers in Self-Supervised Visual Representation Learning. *Adv Neural Inf Process Syst* 34:4193–4206
- Graham B, El-Nouby A, Touvron H, et al (2021) LeViT: a Vision Transformer in

- ConvNet's Clothing for Faster Inference. *Proc IEEE Int Conf Comput Vis* 12239–12249
- Guo H, Song M, Ding Z, et al (2023) Vision-Based Efficient Robotic Manipulation with a Dual-Streaming Compact Convolutional Transformer. *Sensors* 2023, Vol 23, Page 515 23:515. <https://doi.org/10.3390/S23010515>
- Guo J, Han K, Wu H, et al (2021) CMT: Convolutional Neural Networks Meet Vision Transformers. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2022-June*:12165–12175. <https://doi.org/10.1109/CVPR52688.2022.01186>
- Habib G, Saleem TJ, Lall B (2023) Knowledge Distillation in Vision Transformers: A Critical Review
- Hampali S, Sarkar SD, Rad M, Lepetit V (2021) Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2022-June*:11080–11090. <https://doi.org/10.1109/CVPR52688.2022.01081>
- Han K, Wang Y, Chen H, et al (2023) A Survey on Vision Transformer. *IEEE Trans Pattern Anal Mach Intell* 45:87–110. <https://doi.org/10.1109/TPAMI.2022.3152247>
- Han K, Xiao A, Wu E, et al (2021) Transformer in Transformer. *Adv Neural Inf Process Syst* 19:15908–15919
- Hassani A, Walton S, Shah N, et al (2021) Escaping the Big Data Paradigm with Compact Transformers
- He Q, Yang Q, Xie M (2023) HCTNet: A hybrid CNN-transformer network for breast ultrasound image segmentation. *Comput Biol Med* 155:106629. <https://doi.org/10.1016/J.COMPBIOMED.2023.106629>
- Heidari M, Kazerouni A, Soltany M, et al (2022) HiFormer: Hierarchical Multi-scale Representations Using Transformers for Medical Image Segmentation. *Proc - 2023 IEEE Winter Conf Appl Comput Vision, WACV 2023* 6191–6201. <https://doi.org/10.1109/WACV56688.2023.00614>
- Heo B, Yun S, Han D, et al (2021) Rethinking Spatial Dimensions of Vision Transformers. *Proc IEEE Int Conf Comput Vis* 11916–11925. <https://doi.org/10.48550/arxiv.2103.16302>
- Heo YJ, Yeo WH, Kim BG (2023) DeepFake detection algorithm based on improved vision transformer. *Appl Intell* 53:7512–7527. <https://doi.org/10.1007/S10489-022-03867-9/TABLES/4>
- Huang J, Zhu Z, Huang G (2019) Multi-Stage HRNet: Multiple Stage High-Resolution Network for Human Pose Estimation

- Huang K, Wen M, Wang C, Ling L (2023a) FPDT: a multi-scale feature pyramidal object detection transformer. <https://doi.org/10.1117/1JRS.17026510>.
<https://doi.org/10.1117/1JRS.17.026510>
- Huang Q, Huang C, Wang X, Jiang F (2021a) Facial expression recognition with grid-wise attention and visual transformer. *Inf Sci (Ny)* 580:35–54.
<https://doi.org/10.1016/J.INS.2021.08.043>
- Huang X, Chen J, Chen M, et al (2023b) FRE-Net: Full-region enhanced network for nuclei segmentation in histopathology images. *Biocybern Biomed Eng* 43:386–401.
<https://doi.org/10.1016/J.BBE.2023.02.002>
- Huang Z, Ben Y, Luo G, et al (2021b) Shuffle Transformer: Rethinking Spatial Shuffle for Vision Transformer
- Islam K (2022) Recent Advances in Vision Transformer: A Survey and Outlook of Recent Work
- Islam MA, Kowal M, Jia S, et al (2021) Position, Padding and Predictions: A Deeper Look at Position Information in CNNs. *ArXiv*
- Jamali A, Roy SK, Ghamisi P (2023) WetMapFormer: A unified deep CNN and vision transformer for complex wetland mapping. *Int J Appl Earth Obs Geoinf* 120:103333.
<https://doi.org/10.1016/J.JAG.2023.103333>
- Ji GP, Zhuge M, Gao D, et al (2023) Masked Vision-language Transformer in Fashion. *Mach Intell Res* 20:421–434. <https://doi.org/10.1007/S11633-022-1394-4/METRICS>
- Jiang A, Yan N, Wang F, et al (2019) Visible Image Recognition of Power Transformer Equipment Based on Mask R-CNN. *iSPEC 2019 - 2019 IEEE Sustain Power Energy Conf Grid Mod Energy Revolution, Proc* 657–661.
<https://doi.org/10.1109/ISPEC48194.2019.8975213>
- Jiang K, Peng P, Lian Y, Xu W (2022) The encoding method of position embeddings in vision transformer. *J Vis Commun Image Represent* 89:103664.
<https://doi.org/10.1016/J.JVCIR.2022.103664>
- Jiang S, Li J (2022) TransCUNet: UNet cross fused transformer for medical image segmentation. *Comput Biol Med* 150:106207.
<https://doi.org/10.1016/J.COMPBIOMED.2022.106207>
- Jiang Y, Chang S, Wang Z (2021) TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. *Adv Neural Inf Process Syst* 18:14745–14758
- Jin W, Yu H, Luo X (2021) CvT-ASSD: Convolutional vision-Transformer Based Attentive Single Shot MultiBox Detector. *Proc - Int Conf Tools with Artif Intell ICTAI 2021-Novem*:736–744. <https://doi.org/10.1109/ICTAI52525.2021.00117>

- Jing T, Meng Q-H, Hou H-R (2023) SmokeSegger: A Transformer-CNN coupled model for urban scene smoke segmentation. *IEEE Trans Ind Informatics* 1–12.
<https://doi.org/10.1109/TII.2023.3271441>
- Jing Y, Wang F (2022) TP-VIT: A TWO-PATHWAY VISION TRANSFORMER FOR VIDEO ACTION RECOGNITION. *ICASSP, IEEE Int Conf Acoust Speech Signal Process - Proc 2022-May*:2185–2189.
<https://doi.org/10.1109/ICASSP43922.2022.9747276>
- Kanwal N, Eftestøl T, Khoraminia F, et al (2023) Vision Transformers for Small Histological Datasets Learned Through Knowledge Distillation. 167–179.
https://doi.org/10.1007/978-3-031-33380-4_13
- Karras T, Laine S, Aittala M, et al (2019) Analyzing and Improving the Image Quality of StyleGAN. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 8107–8116.
<https://doi.org/10.1109/CVPR42600.2020.00813>
- Kaur G, Sinha R, Tiwari PK, et al (2022) Face mask recognition system using CNN model. *Neurosci Informatics* 2:100035.
<https://doi.org/10.1016/J.NEURI.2021.100035>
- Ke J, Lu Y, Shen Y, et al (2023) ClusterSeg: A crowd cluster pinpointed nucleus segmentation framework with cross-modality datasets. *Med Image Anal* 85:102758.
<https://doi.org/10.1016/J.MEDIA.2023.102758>
- Khan A, Khan SH, Saif M, et al (2023) A Survey of Deep Learning Techniques for the Analysis of COVID-19 and their usability for Detecting Omicron.
<https://doi.org/10.1080/0952813X20232165724>.
<https://doi.org/10.1080/0952813X.2023.2165724>
- Khan A, Qureshi AS, Wahab N, et al (2021a) A recent survey on the applications of genetic programming in image processing. *Comput Intell* 37:1745–1778.
<https://doi.org/10.1111/coin.12459>
- Khan A, Sohail A, Zahoora U, Qureshi AS (2020) A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev* 53:5455–5516.
<https://doi.org/10.1007/s10462-020-09825-6>
- Khan S, Naseer M, Hayat M, et al (2021b) Transformers in Vision: A Survey. *ACM Comput Surv* 54:. <https://doi.org/10.1145/3505244>
- Khan SH, Khan A, Lee YS, et al (2021c) Segmentation of Shoulder Muscle MRI Using a New Region and Edge based Deep Auto-Encoder
- Khan SH, Shah NS, Nuzhat R, et al (2022) Malaria parasite classification framework using a novel channel squeezed and boosted CNN. *Microscopy*.

<https://doi.org/10.1093/JMICRO/DFAC027>

Kim BJ, Choi H, Jang H, et al (2023) Improved robustness of vision transformers via prelayernorm in patch embedding. *Pattern Recognit* 141:109659.
<https://doi.org/10.1016/J.PATCOG.2023.109659>

Kirillov A, Mintun E, Ravi N, et al (2023) Segment Anything

LeCun Y, Boser B, Denker JS, et al (1989) Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput* 1:541–551.
<https://doi.org/10.1162/NECO.1989.1.4.541>

Lee K, Chang H, Jiang L, et al (2021a) ViTGAN: Training GANs with Vision Transformers

Lee Y, Kim J, Willette J, Hwang SJ (2021b) MPViT: Multi-Path Vision Transformer for Dense Prediction. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2022-June*:7277–7286. <https://doi.org/10.1109/CVPR52688.2022.00714>

Leong MC, Zhang H, Tan HL, et al (2022) Combined CNN Transformer Encoder for Enhanced Fine-grained Human Action Recognition

Li C, Tang T, Wang G, et al (2021a) BossNAS: Exploring Hybrid CNN-transformers with Block-wisely Self-supervised Neural Architecture Search. *Proc IEEE Int Conf Comput Vis* 12261–12271. <https://doi.org/10.48550/arxiv.2103.12424>

Li G, Chen R, Zhang J, et al (2023a) Fusing enhanced Transformer and large kernel CNN for malignant thyroid nodule segmentation. *Biomed Signal Process Control* 83:104636. <https://doi.org/10.1016/J.BSPC.2023.104636>

Li G, Yao H, Le Y, Qin C (2023b) Recaptured screen image identification based on vision transformer. *J Vis Commun Image Represent* 90:103692.
<https://doi.org/10.1016/J.JVCIR.2022.103692>

Li J, Chen J, Tang Y, et al (2023c) Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Med Image Anal* 85:102762. <https://doi.org/10.1016/J.MEDIA.2023.102762>

Li J, Du Q, Li W, et al (2023d) MCAFNet: A Multiscale Channel Attention Fusion Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens* 2023, Vol 15, Page 361 15:361. <https://doi.org/10.3390/RS15020361>

Li R, Mai Z, Zhang Z, et al (2023e) TransCAM: Transformer attention-based CAM refinement for Weakly supervised semantic segmentation. *J Vis Commun Image Represent* 92:103800. <https://doi.org/10.1016/J.JVCIR.2023.103800>

Li X, Li S (2022) Transformer Help CNN See Better: A Lightweight Hybrid Apple

- Disease Identification Model Based on Transformers. *Agric* 2022, Vol 12, Page 884 12:884. <https://doi.org/10.3390/AGRICULTURE12060884>
- Li X, Li X, Zhang S, et al (2023f) SLViT: Shuffle-convolution-based lightweight Vision transformer for effective diagnosis of sugarcane leaf diseases. *J King Saud Univ - Comput Inf Sci* 35:101401. <https://doi.org/10.1016/J.JKSUCI.2022.09.013>
- Li X, Xiang Y, Li S (2023g) Combining convolutional and vision transformer structures for sheep face recognition. *Comput Electron Agric* 205:107651. <https://doi.org/10.1016/J.COMPAG.2023.107651>
- Li Y, Yao T, Pan Y, Mei T (2021b) Contextual Transformer Networks for Visual Recognition. *IEEE Trans Pattern Anal Mach Intell*. <https://doi.org/10.1109/TPAMI.2022.3164083>
- Li Y, Zhang K, Cao J, et al (2021c) LocalViT: Bringing Locality to Vision Transformers. <https://doi.org/10.48550/arxiv.2104.05707>
- Li Y, Zhang S, Wang Z, et al (2021d) TokenPose: Learning Keypoint Tokens for Human Pose Estimation. *Proc IEEE Int Conf Comput Vis* 11293–11302. <https://doi.org/10.1109/ICCV48922.2021.01112>
- Li Z, Li D, Xu C, et al (2022) TFCNs: A CNN-Transformer Hybrid Network for Medical Image Segmentation. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 13532 LNCS:781–792. https://doi.org/10.1007/978-3-031-15937-4_65/COVER
- Lian J, Liu T, Zhou Y, et al (2023) Aurora Classification in All-Sky Images via CNN-Transformer. *Universe* 2023, Vol 9, Page 230 9:230. <https://doi.org/10.3390/UNIVERSE9050230>
- Liang S, Hua Z, Li J (2023) Hybrid transformer-CNN networks using superpixel segmentation for remote sensing building change detection. <https://doi.org/101080/0143116120232208711> 44:2754–2780. <https://doi.org/10.1080/01431161.2023.2208711>
- Lin S, Xie H, Wang B, et al (2022) Knowledge Distillation via the Target-aware Transformer. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2022-June:10905–10914. <https://doi.org/10.1109/CVPR52688.2022.01064>
- Liu J, Li H, Kong W (2023a) Multi-level learning counting via pyramid vision transformer and CNN. *Eng Appl Artif Intell* 123:106184. <https://doi.org/10.1016/J.ENGAPPAI.2023.106184>
- Liu X, Deng Z, Yang Y (2018) Recent progress in semantic image segmentation. *Artif Intell Rev* 52:1089–1106. <https://doi.org/10.1007/s10462-018-9641-3>

- Liu Y, Ong N, Peng K, et al (2023b) MMViT: Multiscale Multiview Vision Transformers
- Liu Y, Wu Y-H, Sun G, et al (2021a) Vision Transformers with Hierarchical Attention
- Liu Y, Zhang YY, Wang Y, et al (2021b) A Survey of Visual Transformers.
<https://doi.org/10.1109/TNNLS.2022.3227717>
- Liu Z, Lin Y, Cao Y, et al (2021c) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proc IEEE Int Conf Comput Vis 9992–10002.
<https://doi.org/10.48550/arxiv.2103.14030>
- Lu T, Wan L, Qi S, Gao M (2023a) Land Cover Classification of UAV Remote Sensing Based on Transformer–CNN Hybrid Architecture. Sensors 2023, Vol 23, Page 5288 23:5288. <https://doi.org/10.3390/S23115288>
- Lu W, Lan C, Niu C, et al (2023b) A CNN-Transformer Hybrid Model Based on CSWin Transformer for UAV Image Object Detection. IEEE J Sel Top Appl Earth Obs Remote Sens 16:1211–1231. <https://doi.org/10.1109/JSTARS.2023.3234161>
- Lyu J, Li G, Wang C, et al (2023) Region-focused multi-view transformer-based generative adversarial network for cardiac cine MRI reconstruction. Med Image Anal 85:102760. <https://doi.org/10.1016/J.MEDIA.2023.102760>
- Ma F, Sun B, Li S (2023a) Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion. IEEE Trans Affect Comput 14:1236–1248.
<https://doi.org/10.1109/TAFFC.2021.3122146>
- Ma Z, Qi Y, Xu C, et al (2023b) ATFE-Net: Axial Transformer and Feature Enhancement-based CNN for ultrasound breast mass segmentation. Comput Biol Med 153:106533. <https://doi.org/10.1016/J.COMPBIOMED.2022.106533>
- Maaz M, Shaker A, Cholakkal H, et al (2023) EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 13807 LNCS:3–20. https://doi.org/10.1007/978-3-031-25082-8_1/COVER
- Maaz M, Shaker A, Cholakkal H, et al (2022) EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications.
<https://doi.org/10.48550/arxiv.2206.10589>
- Mao W, Ge Y, Shen C, et al (2021) TFPose: Direct Human Pose Estimation with Transformers
- Mathian E, Liu H, Fernandez-Cuesta L, et al (2022) HaloAE: An HaloNet based Local Transformer Auto-Encoder for Anomaly Detection and Localization
- Maurício J, Domingues I, Bernardino J (2023) Comparing Vision Transformers and

- Convolutional Neural Networks for Image Classification: A Literature Review. *Appl Sci* 2023, Vol 13, Page 5521 13:5521. <https://doi.org/10.3390/APP13095521>
- Mogan JN, Lee CP, Lim KM, et al (2023) Gait-CNN-ViT: Multi-Model Gait Recognition with Convolutional Neural Networks and Vision Transformer. *Sensors* 2023, Vol 23, Page 3809 23:3809. <https://doi.org/10.3390/S23083809>
- Morra L, Piano L, Lamberti F, Tommasi T (2020) Bridging the gap between natural and medical images through deep colorization. In: *Proceedings - International Conference on Pattern Recognition*
- Moutik O, Sekkat H, Tigani S, et al (2023) Convolutional Neural Networks or Vision Transformers: Who Will Win the Race for Action Recognitions in Visual Data? *Sensors* 2023, Vol 23, Page 734 23:734. <https://doi.org/10.3390/S23020734>
- Nafisah SI, Muhammad G, Hossain MS, AlQahtani SA (2023) A Comparative Evaluation between Convolutional Neural Networks and Vision Transformers for COVID-19 Detection. *Math* 2023, Vol 11, Page 1489 11:1489. <https://doi.org/10.3390/MATH11061489>
- Naveen S, Ram Kiran MSS, Indupriya M, et al (2021) Transformer models for enhancing AttnGAN based text to image generation. *Image Vis Comput* 115:104284. <https://doi.org/10.1016/J.IMAVIS.2021.104284>
- Obeid A, Mahbub T, Javed S, et al (2022) NucDETR: End-to-End Transformer for Nucleus Detection in Histopathology Images. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 13574 LNCS:47–57. https://doi.org/10.1007/978-3-031-17266-3_5/COVER
- Pan X, Ge C, Lu R, et al (2022) On the Integration of Self-Attention and Convolution. 815–825
- Parmar N, Vaswani A, Uszkoreit J, et al (2018) Image transformer. *35th Int Conf Mach Learn ICML 2018* 9:6453–6462. <https://doi.org/10.48550/arxiv.1802.05751>
- Pecoraro R, Basile V, Bono V (2022) Local Multi-Head Channel Self-Attention for Facial Expression Recognition. *Inf* 2022, Vol 13, Page 419 13:419. <https://doi.org/10.3390/INFO13090419>
- Peng Z, Guo Z, Huang W, et al (2023) Conformer: Local Features Coupling Global Representations for Recognition and Detection. *IEEE Trans Pattern Anal Mach Intell* 1–15. <https://doi.org/10.1109/TPAMI.2023.3243048>
- Peng Z, Huang W, Gu S, et al (2021) Conformer: Local Features Coupling Global Representations for Visual Recognition. *Proc IEEE Int Conf Comput Vis* 357–366. <https://doi.org/10.1109/ICCV48922.2021.00042>

- Quan J, Ge B, Wang M (2023) CrackViT: a unified CNN-transformer model for pixel-level crack extraction. *Neural Comput Appl* 35:10957–10973. <https://doi.org/10.1007/S00521-023-08277-7/TABLES/7>
- Rafiq G, Rafiq · Muhammad, Gyu ·, et al (2023) Video description: A comprehensive survey of deep learning approaches. *Artif Intell Rev* 2023 1–80. <https://doi.org/10.1007/S10462-023-10414-6>
- Raghavendra S, Ramyashree, Abhilash SK, et al (2023) Efficient Deep Learning Approach to Recognize Person Attributes by Using Hybrid Transformers for Surveillance Scenarios. *IEEE Access* 11:10881–10893. <https://doi.org/10.1109/ACCESS.2023.3241334>
- Ranftl R, Bochkovskiy A, Koltun V (2021) Vision Transformers for Dense Prediction. *Proc IEEE Int Conf Comput Vis* 12159–12168
- Rao D, Wu X-J, Xu T (2022) TGFuse: An Infrared and Visible Image Fusion Approach Based on Transformer and Generative Adversarial Network. <https://doi.org/10.1109/TIP.2023.3273451>
- Rauf Z, Sohail A, Khan SH, et al (2023) Attention-guided multi-scale deep object detection framework for lymphocyte analysis in IHC histological images. *Reprod Syst Sex Disord* 72:27–42. <https://doi.org/10.1093/jmicro/dfac051>
- Rehman A, Khan A (2023) MaxViT-UNet: Multi-Axis Attention for Medical Image Segmentation. *arXiv Prepr arXiv230508396*
- Ren P, Li C, Wang G, et al (2022) Beyond Fixation: Dynamic Window Visual Transformer. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2022-June:11977–11987. <https://doi.org/10.1109/CVPR52688.2022.01168>
- Seydi ST, Sadegh M (2023) Improved burned area mapping using monotemporal Landsat-9 imagery and convolutional shift-transformer. *Measurement* 216:112961. <https://doi.org/10.1016/J.MEASUREMENT.2023.112961>
- Shafri MBA;, Al-Ruzouq HZM;, Shanableh R;, et al (2023) Large-Scale Date Palm Tree Segmentation from Multiscale UAV-Based and Aerial Images Using Deep Vision Transformers. *Drones* 2023, Vol 7, Page 93 7:93. <https://doi.org/10.3390/DRONES7020093>
- Shamshad F, Khan S, Zamir SW, et al (2023) Transformers in medical imaging: A survey. *Med Image Anal* 102802. <https://doi.org/10.1016/j.media.2023.102802>
- Shen X, Xu J, Jia H, et al (2022) Self-attentional microvessel segmentation via squeeze-excitation transformer Unet. *Comput Med Imaging Graph* 97:102055. <https://doi.org/10.1016/J.COMPMEDIMAG.2022.102055>

- Shi R, Yang S, Chen Y, et al (2023) CNN-Transformer for visual-tactile fusion applied in road recognition of autonomous vehicles. *Pattern Recognit Lett* 166:200–208. <https://doi.org/10.1016/J.PATREC.2022.11.023>
- Si C, Yu W, Zhou P, et al (2022) Inception Transformer
- Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc
- Sohail A, Khan A, Nisar H, et al (2021a) Mitotic Nuclei Analysis in Breast Cancer Histopathology Images using Deep Ensemble Classifier Mitotic Nuclei Analysis in Breast Cancer Histopathology Images using Deep Ensemble Classifier. <https://doi.org/10.1016/j.media.2021.102121>
- Sohail A, Khan A, Nisar H, et al (2021b) Mitotic nuclei analysis in breast cancer histopathology images using deep ensemble classifier. *Med Image Anal* 72:102121
- Song L, Liu G, Ma M (2022a) TD-Net:unsupervised medical image registration network based on Transformer and CNN. *Appl Intell* 52:18201–18209. <https://doi.org/10.1007/S10489-022-03472-W/TABLES/3>
- Song Y, He Z, Qian H, Du X (2023) Vision Transformers for Single Image Dehazing. *IEEE Trans Image Process* 1–1. <https://doi.org/10.1109/TIP.2023.3256763>
- Song Z, Yu J, Chen YPP, Yang W (2022b) Transformer Tracking with Cyclic Shifting Window Attention. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2022-June:8781–8790. <https://doi.org/10.1109/CVPR52688.2022.00859>
- Springenberg M, Frommholz A, Wenzel M, et al (2022) From CNNs to Vision Transformers -- A Comprehensive Evaluation of Deep Learning Models for Histopathology
- Srinivas A, Lin TY, Parmar N, et al (2021) Bottleneck Transformers for Visual Recognition. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 16514–16524. <https://doi.org/10.1109/CVPR46437.2021.01625>
- Stoffl L, Vidal M, Mathis A (2021) End-to-End Trainable Multi-Instance Pose Estimation with Transformers
- Su W, Wang Y, Li K, et al (2023) Hybrid token transformer for deep face recognition. *Pattern Recognit* 139:109443. <https://doi.org/10.1016/J.PATCOG.2023.109443>
- Sun K, Xiao B, Liu D, Wang J (2019) Deep High-Resolution Representation Learning for Human Pose Estimation. 5693–5703
- Sun S, Yue X, Zhao H, et al (2022) Patch-based Separable Transformer for Visual Recognition. *IEEE Trans Pattern Anal Mach Intell*.

<https://doi.org/10.1109/TPAMI.2022.3231725>

Tan M, Le Q V. (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 36th Int Conf Mach Learn ICML 2019 2019-June:10691–10700

Tiong LCO, Sigmund D, Teoh ABJ (2023) Face-Periocular Cross-Identification via Contrastive Hybrid Attention Vision Transformer. IEEE Signal Process Lett 1–5. <https://doi.org/10.1109/LSP.2023.3256320>

Torbunov D, Huang Y, Yu H, et al (2022) UVCAN: UNet Vision Transformer cycle-consistent GAN for unpaired image-to-image translation. Proc - 2023 IEEE Winter Conf Appl Comput Vision, WACV 2023 702–712. <https://doi.org/10.1109/WACV56688.2023.00077>

Touvron H, Cord M, Douze M, et al (2020) Training data-efficient image transformers & distillation through attention. <https://doi.org/10.48550/arxiv.2012.12877>

Touvron H, Cord M, Sablayrolles A, et al (2021) Going deeper with Image Transformers. Proc IEEE Int Conf Comput Vis 32–42. <https://doi.org/10.48550/arxiv.2103.17239>

Tragakis A, Kaul C, Murray-Smith R, Husmeier D The Fully Convolutional Transformer for Medical Image Segmentation. Institute of Electrical and Electronics Engineers Inc.

Tu J, Mei G, Ma Z, Piccialli F (2022a) SWCGAN: Generative Adversarial Network Combining Swin Transformer and CNN for Remote Sensing Image Super-Resolution. IEEE J Sel Top Appl Earth Obs Remote Sens 15:5662–5673. <https://doi.org/10.1109/JSTARS.2022.3190322>

Tu Z, Talebi H, Zhang H, et al (2022b) MaxViT: Multi-Axis Vision Transformer. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 13684 LNCS:459–479. <https://doi.org/10.48550/arxiv.2204.01697>

Ulhaq A, Akhtar N, Pogrebna G, Mian A (2022) Vision Transformers for Action Recognition: A Survey

Ullah W, Hussain T, Ullah FUM, et al (2023) TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection. Eng Appl Artif Intell 123:106173. <https://doi.org/10.1016/J.ENGAPPAI.2023.106173>

Vaswani A, Brain G, Shazeer N, et al (2017a) Attention is All you Need. Adv Neural Inf Process Syst 30:

Vaswani A, Shazeer N, Parmar N, et al (2017b) Attention Is All You Need. Adv Neural Inf Process Syst 2017-Decem:5999–6009. <https://doi.org/10.48550/arxiv.1706.03762>

Wang H, Zhu Y, Adam H, et al (2021a) Max-DeepLab: End-to-End Panoptic

- Segmentation with Mask Transformers. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 5459–5470. <https://doi.org/10.1109/CVPR46437.2021.00542>
- Wang J, Zhang Z, Luo L, et al (2023a) DualSeg: Fusing transformer and CNN structure for image segmentation in complex vineyard environment. *Comput Electron Agric* 206:107682. <https://doi.org/10.1016/J.COMPAG.2023.107682>
- Wang L, Pan L, Wang H, et al (2023b) DHUnet: Dual-branch hierarchical global–local fusion network for whole slide image segmentation. *Biomed Signal Process Control* 85:104976. <https://doi.org/10.1016/J.BSPC.2023.104976>
- Wang L, Tien A (2023) Aerial Image Object Detection With Vision Transformer Detector (ViTDet)
- Wang R, Geng F, Wang X (2022a) MTPose: Human Pose Estimation with High-Resolution Multi-scale Transformers. *Neural Process Lett* 54:3941–3964. <https://doi.org/10.1007/S11063-022-10794-W/TABLES/8>
- Wang W, Chen W, Qiu Q, et al (2023c) CrossFormer++: A Versatile Vision Transformer Hinging on Cross-scale Attention
- Wang W, Dai J, Chen Z, et al (2022b) InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. 14408–14419
- Wang W, Tang C, Wang X, Zheng B (2022c) A ViT-Based Multiscale Feature Fusion Approach for Remote Sensing Image Segmentation. *IEEE Geosci Remote Sens Lett* 19:. <https://doi.org/10.1109/LGRS.2022.3187135>
- Wang W, Wang J, Lu B, et al (2023d) MCPT: Mixed Convolutional Parallel Transformer for Polarimetric SAR Image Classification. *Remote Sens* 2023, Vol 15, Page 2936 15:2936. <https://doi.org/10.3390/RS15112936>
- Wang W, Xie E, Li X, et al (2021b) PVT v2: Improved Baselines with Pyramid Vision Transformer. *Comput Vis Media* 8:415–424. <https://doi.org/10.1007/s41095-022-0274-8>
- Wang W, Xie E, Li X, et al (2021c) Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. *Proc IEEE Int Conf Comput Vis* 548–558. <https://doi.org/10.48550/arxiv.2102.12122>
- Wang Y, Qiu Y, Cheng P, Zhang J (2022d) Hybrid CNN-Transformer Features for Visual Place Recognition. *IEEE Trans Circuits Syst Video Technol*. <https://doi.org/10.1109/TCSVT.2022.3212434>
- Wang Y, Xu Z, Wang X, et al (2020) End-to-End Video Instance Segmentation with Transformers. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 8737–8746. <https://doi.org/10.1109/CVPR46437.2021.00863>

- Wang Y, Yang Y, Bai J, Zhang M Evolving Attention with Residual Convolutions
- Wei Z, Pan H, Li L, et al (2023) DMFormer: Closing the gap Between CNN and Vision Transformers. ICASSP 2023 - 2023 IEEE Int Conf Acoust Speech Signal Process 1–5. <https://doi.org/10.1109/ICASSP49357.2023.10097256>
- Weng W, Zhu X (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. IEEE Access 9:16591–16603. <https://doi.org/10.1109/ACCESS.2021.3053408>
- Wensel J, Ullah H, Member SS, et al (2022) ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos
- Woo S, Debnath S, Hu R, et al (2023) ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders
- Wu H, Xiao B, Codella N, et al (2021a) CvT: Introducing Convolutions to Vision Transformers. Proc IEEE Int Conf Comput Vis 22–31. <https://doi.org/10.48550/arxiv.2103.15808>
- Wu J, Fu R, Fang H, et al (2023a) MedSegDiff-V2: Diffusion based Medical Image Segmentation with Transformer
- Wu K, Peng H, Chen M, et al (2021b) Rethinking and Improving Relative Position Encoding for Vision Transformer. Proc IEEE Int Conf Comput Vis 10013–10021. <https://doi.org/10.1109/ICCV48922.2021.00988>
- Wu K, Zhang J, Peng H, et al (2022a) TinyViT: Fast Pretraining Distillation for Small Vision Transformers. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 13681 LNCS:68–85. https://doi.org/10.1007/978-3-031-19803-8_5
- Wu Q, Wu Y, Zhang Y, Zhang L (2022b) A Local-Global Estimator Based on Large Kernel CNN and Transformer for Human Pose Estimation and Running Pose Measurement. IEEE Trans Instrum Meas 71:. <https://doi.org/10.1109/TIM.2022.3200438>
- Wu Y, Lian C, Zeng Z, et al (2023b) An Aggregated Convolutional Transformer Based on Slices and Channels for Multivariate Time Series Classification. IEEE Trans Emerg Top Comput Intell 7:768–779. <https://doi.org/10.1109/TETCI.2022.3210992>
- Wu Y, Wang G, Wang Z, et al (2022c) DI-Unet: Dimensional interaction self-attention for medical image segmentation. Biomed Signal Process Control 78:103896. <https://doi.org/10.1016/J.BSPC.2022.103896>
- Wu Z, Liao W, Yan C, et al (2023c) Deep learning based MRI reconstruction with transformer. Comput Methods Programs Biomed 233:107452.

<https://doi.org/10.1016/J.CMPB.2023.107452>

Wu Z, Shen C, van den Hengel A (2019) Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. *Pattern Recognit* 90:119–133.

<https://doi.org/10.1016/J.PATCOG.2019.01.006>

Xia W, Han D, Li D, et al (2023) An ensemble learning integration of multiple CNN with improved vision transformer models for pest classification. *Ann Appl Biol* 182:144–158. <https://doi.org/10.1111/AAB.12804>

Xia Z, Kim J (2023) Enhancing Mask Transformer with Auxiliary Convolution Layers for Semantic Segmentation. *Sensors* 2023, Vol 23, Page 581 23:581.

<https://doi.org/10.3390/S23020581>

Xia Z, Pan X, Song S, et al (2022) Vision Transformer with Deformable Attention. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2022-June:4784–4793.

<https://doi.org/10.1109/CVPR52688.2022.00475>

Xiao H, Li L, Liu Q, et al (2023) Transformers in medical image segmentation: A review. *Biomed Signal Process Control* 84:104791.

<https://doi.org/10.1016/J.BSPC.2023.104791>

Xiao T, Singh M, Mintun E, et al (2021) Early Convolutions Help Transformers See Better. *Adv Neural Inf Process Syst* 36:30392–30400

Xie S, Girshick R, Dollár P, et al Aggregated residual transformations for deep neural networks. openaccess.thecvf.com

Xiong S, Kasaei H (2022) Fine-grained Object Categorization for Service Robots

Xu W, Xu Y, Chang T, Tu Z (2021a) Co-Scale Conv-Attentional Image Transformers. *Proc IEEE Int Conf Comput Vis* 9961–9970.

<https://doi.org/10.1109/ICCV48922.2021.00983>

Xu Y, Zhang Q, Zhang J, Tao D (2021b) ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias. *Adv Neural Inf Process Syst* 34:28522–28535

Xue T, Ma P (2023) TC-net: transformer combined with cnn for image denoising. *Appl Intell* 53:6753–6762. <https://doi.org/10.1007/s10489-022-03785-w>

Yan C, Fan X, Fan J, et al (2023) HyFormer: Hybrid Transformer and CNN for Pixel-Level Multispectral Image Land Cover Classification. *Int J Environ Res Public Heal* 2023, Vol 20, Page 3059 20:3059. <https://doi.org/10.3390/IJERPH20043059>

Yan H, Li Z, Li W, et al ConTNet : Why not use convolution and transformer at the same time ?

- Yang H, Yang D (2023) CSwin-PNet: A CNN-Swin Transformer combined pyramid network for breast lesion segmentation in ultrasound images. *Expert Syst Appl* 213:119024. <https://doi.org/10.1016/J.ESWA.2022.119024>
- Yang J, Du B, Wu C (2022a) Hybrid Vision Transformer Model for Hyperspectral Image Classification. *Int Geosci Remote Sens Symp 2022-July*:1388–1391. <https://doi.org/10.1109/IGARSS46834.2022.9884262>
- Yang J, Tu J, Zhang X, et al (2023a) TSE DeepLab: An efficient visual transformer for medical image segmentation. *Biomed Signal Process Control* 80:104376. <https://doi.org/10.1016/J.BSPC.2022.104376>
- Yang S, Feng Z, Wang Z, et al (2023b) Detecting and grouping keypoints for multi-person pose estimation using instance-aware attention. *Pattern Recognit* 136:109232. <https://doi.org/https://doi.org/10.1016/j.patcog.2022.109232>
- Yang S, Guo W, Ren Y (2022b) CrowdFormer: An Overlap Patching Vision Transformer for Top-Down Crowd Counting. *IJCAI Int Jt Conf Artif Intell* 2:1545–1551. <https://doi.org/10.24963/IJCAI.2022/215>
- Yang Y, Zhang L, Ren L, Wang X (2023c) MMViT-Seg: A lightweight transformer and CNN fusion network for COVID-19 segmentation. *Comput Methods Programs Biomed* 230:107348. <https://doi.org/10.1016/J.CMPB.2023.107348>
- Yao C, Feng L, Kong Y, et al (2023) Transformers and CNNs fusion network for salient object detection. *Neurocomputing* 520:342–355. <https://doi.org/10.1016/J.NEUCOM.2022.10.081>
- Yao G, Lei T, Zhong J (2019) A review of Convolutional-Neural-Network-based action recognition. *Pattern Recognit Lett* 118:14–22. <https://doi.org/10.1016/J.PATREC.2018.05.018>
- Yao T, Li Y, Pan Y, et al Dual Vision Transformer. *IEEE Trans Pattern Anal Mach Intell*. <https://doi.org/10.1109/TPAMI.2023.3268446>
- Ye D, Ni Z, Wang H, et al (2023a) CSformer: Bridging Convolution and Transformer for Compressive Sensing. *IEEE Trans Image Process* 32:2827–2842. <https://doi.org/10.1109/TIP.2023.3274988>
- Ye L, Rochan M, Liu Z, Wang Y (2019) Cross-Modal Self-Attention Network for Referring Image Segmentation. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June*:10494–10503. <https://doi.org/10.1109/CVPR.2019.01075>
- Ye T, Qin W, Zhao Z, et al (2023b) Real-Time Object Detection Network in UAV-Vision Based on CNN and Transformer. *IEEE Trans Instrum Meas* 72:. <https://doi.org/10.1109/TIM.2023.3241825>

- Yi S, Li L, Liu X, et al (2023) HCTIRdeblur: A hybrid convolution-transformer network for single infrared image deblurring. *Infrared Phys Technol* 131:104640. <https://doi.org/10.1016/J.INFRARED.2023.104640>
- Yu G, Zhou X (2023) An Improved YOLOv5 Crack Detection Method Combined with a Bottleneck Transformer. *Math* 2023, Vol 11, Page 2377 11:2377. <https://doi.org/10.3390/MATH11102377>
- Yuan F, Zhang Z, Fang Z (2023a) An effective CNN and Transformer complementary network for medical image segmentation. *Pattern Recognit* 136:109228. <https://doi.org/10.1016/J.PATCOG.2022.109228>
- Yuan J, Zhou F, Guo Z, et al (2023b) HCformer: Hybrid CNN-Transformer for LDCT Image Denoising. *J Digit Imaging* 1–16. <https://doi.org/10.1007/S10278-023-00842-9/TABLES/8>
- Yuan K, Guo S, Liu Z, et al (2021a) Incorporating Convolution Designs into Visual Transformers. *Proc IEEE Int Conf Comput Vis* 559–568. <https://doi.org/10.1109/ICCV48922.2021.00062>
- Yuan L, Chen Y, Wang T, et al (2021b) Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. *Proc IEEE Int Conf Comput Vis* 538–547
- Zafar MM, Rauf Z, Sohail A, et al (2021) Detection of Tumour Infiltrating Lymphocytes in CD3 and CD8 Stained Histopathological Images using a Two-Phase Deep CNN. *Photodiagnosis Photodyn Ther* 37:102676. <https://doi.org/10.1016/j.pdpdt.2021.102676>
- Zahoor MM, Qureshi SA, Bibi S, et al (2022) A New Deep Hybrid Boosted and Ensemble Learning-Based Brain Tumor Analysis Using MRI. *Sensors* 2022, Vol 22, Page 2726 22:2726. <https://doi.org/10.3390/S22072726>
- Zhang C, Zhang M, Zhang S, et al (2021a) Delving Deep into the Generalization of Vision Transformers under Distribution Shifts. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2022-June:7267–7276. <https://doi.org/10.1109/CVPR52688.2022.00713>
- Zhang J, Li C, Yin Y, et al (2023a) Applications of artificial neural networks in microorganism image analysis: a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer. *Artif Intell Rev* 56:1013–1070. <https://doi.org/10.1007/S10462-022-10192-7/FIGURES/2>
- Zhang K, Su Y, Guo X, et al (2021b) MU-GAN: Facial Attribute Editing Based on Multi-Attention Mechanism. *IEEE/CAA J Autom Sin* 8:1614–1626. <https://doi.org/10.1109/JAS.2020.1003390>

- Zhang N, Nex F, Vosselman G, Kerle N (2022a) Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation
- Zhang N, Yu L, Zhang D, et al (2022b) APT-Net: Adaptive encoding and parallel decoding transformer for medical image segmentation. *Comput Biol Med* 151:106292. <https://doi.org/10.1016/J.COMPBIOMED.2022.106292>
- Zhang Q, Xu Y, Zhang J, Tao D (2022c) VSA: Learning Varied-Size Window Attention in Vision Transformers. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 13685 LNCS:466–483. https://doi.org/10.1007/978-3-031-19806-9_27
- Zhang Q, Xu Y, Zhang J, Tao D (2022d) ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and Beyond. *Int J Comput Vis* 131:1141–1162. <https://doi.org/10.1007/s11263-022-01739-w>
- Zhang QL, Yang Y Bin (2021) ResT: An Efficient Transformer for Visual Recognition. *Adv Neural Inf Process Syst* 19:15475–15485. <https://doi.org/10.48550/arxiv.2105.13677>
- Zhang X, Cheng S, Wang L, Li H (2023b) Asymmetric Cross-Attention Hierarchical Network Based on CNN and Transformer for Bitemporal Remote Sensing Images Change Detection. *IEEE Trans Geosci Remote Sens* 61:. <https://doi.org/10.1109/TGRS.2023.3245674>
- Zhang X, Zhang Y (2022) Conv-PVT: a fusion architecture of convolution and pyramid vision transformer. *Int J Mach Learn Cybern* 14:2127–2136. <https://doi.org/10.1007/S13042-022-01750-0/TABLES/8>
- Zhang Y, Liu H, Hu Q (2021c) TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 12901 LNCS:14–24. https://doi.org/10.1007/978-3-030-87193-2_2/COVER
- Zhang Z, Sun G, Zheng K, et al (2023c) TC-Net: A joint learning framework based on CNN and vision transformer for multi-lesion medical images segmentation. *Comput Biol Med* 161:106967. <https://doi.org/10.1016/J.COMPBIOMED.2023.106967>
- Zhao L, Yu Q, Yang Y (2022a) Video Person Re-identification Based on Transformer-CNN Model. *2022 4th Int Conf Artif Intell Adv Manuf* 445–450. <https://doi.org/10.1109/AIAM57466.2022.00091>
- Zhao M, Cao G, Huang X, Yang L (2022b) Hybrid Transformer-CNN for Real Image Denoising. *IEEE Signal Process Lett* 29:1252–1256. <https://doi.org/10.1109/LSP.2022.3176486>

- Zhao S, Liu K, Huang Y, et al (2022c) DPIT: Dual-Pipeline Integrated Transformer for Human Pose Estimation. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 13605 LNAI:559–576.
https://doi.org/10.1007/978-3-031-20500-2_46/COVER
- Zhao X, Yang T, Li B, Zhang X (2023) SwinGAN: A dual-domain Swin Transformer-based generative adversarial network for MRI reconstruction. *Comput Biol Med* 153:106513. <https://doi.org/10.1016/J.COMPBIOMED.2022.106513>
- Zheng T, Oda H, Hayashi Y, et al (2023) L-former : a lightweight transformer for realistic medical image generation and its application to super-resolution.
<https://doi.org/10.1117/122653776> 12464:245–250.
<https://doi.org/10.1117/12.2653776>
- Zhou D, Kang B, Jin X, et al (2021) DeepViT: Towards Deeper Vision Transformer
- Zhou Z, Li G, Wang G (2023a) A hybrid of transformer and CNN for efficient single image super-resolution via multi-level distillation. *Displays* 76:102352.
<https://doi.org/10.1016/J.DISPLA.2022.102352>
- Zhou Z, Zhang J, Gong C (2023b) Hybrid semantic segmentation for tunnel lining cracks based on Swin Transformer and convolutional neural network. *Comput Civ Infrastruct Eng*. <https://doi.org/10.1111/MICE.13003>
- Zhu D, Tan J, Wu C, et al (2023a) Crop Disease Identification by Fusing Multiscale Convolution and Vision Transformer. *Sensors* 2023, Vol 23, Page 6015 23:6015.
<https://doi.org/10.3390/S23136015>
- Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Proc IEEE Int Conf Comput Vis* 2017-Octob:2242–2251. <https://doi.org/10.1109/ICCV.2017.244>
- Zhu X, Li Z, Sun J, et al (2023b) Expression recognition method combining convolutional features and Transformer. *Math Found Comput* 6:203–217.
<https://doi.org/10.3934/MFC.2022018>
- Zidan U, Gaber MM, Abdelsamea MM (2023) SwinCup: Cascaded swin transformer for histopathological structures segmentation in colorectal cancer. *Expert Syst Appl* 216:119452. <https://doi.org/10.1016/J.ESWA.2022.119452>
- Zou P, Wu JS (2023) SwinE-UNet3+: swin transformer encoder network for medical image segmentation. *Prog Artif Intell* 1–7. <https://doi.org/10.1007/S13748-023-00300-1/FIGURES/4>

