Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.050790

REVIEW

A Comprehensive Survey of Recent Transformers in Image, Video and Diffusion Models

Dinh Phu Cuong Le[1,2], Dong Wang[1] and Viet-Tuan Le[3,*]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410082, China

[2]Faculty of Information Technology, Yersin University of Da Lat, Da Lat, 66100, Vietnam

[3]Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, 722000, Vietnam

*Corresponding Author: Viet-Tuan Le. Email: tuan.lv@ou.edu.vn

Received: 17 February 2024 Accepted: 17 May 2024 Published: 18 July 2024

ABSTRACT

Transformer models have emerged as dominant networks for various tasks in computer vision compared to Convolutional Neural Networks (CNNs). The transformers demonstrate the ability to model long-range dependencies by utilizing a self-attention mechanism. This study aims to provide a comprehensive survey of recent transformer-based approaches in image and video applications, as well as diffusion models. We begin by discussing existing surveys of vision transformers and comparing them to this work. Then, we review the main components of a vanilla transformer network, including the self-attention mechanism, feed-forward network, position encoding, etc. In the main part of this survey, we review recent transformer-based models in three categories: Transformer for downstream tasks, Vision Transformer for Generation, and Vision Transformer for Segmentation. We also provide a comprehensive overview of recent transformer models for video tasks and diffusion models. We compare the performance of various hierarchical transformer networks for multiple tasks on popular benchmark datasets. Finally, we explore some future research directions to further improve the field.

KEYWORDS

Transformer; vision transformer; self-attention; hierarchical transformer; diffusion models

1 Introduction

Transformer was designed for Natural Language Processing (NLP) tasks. Vaswani et al. [1] marked a milestone in the history of the transformer. Subsequently, BERT [2] achieved state-of-the-art performance across various tasks. The transformer has demonstrated its dominance in the field of NLP. Various versions of Generative Pre-trained Transformers (GPTs) [3,4] have been introduced for numerous NLP tasks. Moreover, articles generated by GPT-3 are often indistinguishable from those written by humans.

For many years, CNNs have been instrumental in solving a wide range of tasks in computer vision. AlexNet [5] is considered at the forefront of the CNNs when it outperformed the traditional handcraft methods on the ImageNet dataset. To further enhance CNN performance, numerous approaches have

incorporated self-attention in spatial [6], channel [7,8] or both spatial and channel [9]. However, self-attention is typically integrated as an additional layer within the convolutional network architecture.

The success of transformer-based approaches in NLP has sparked interest in applying similar techniques to computer vision. Many pure transformers have been proposed and utilized to replace the traditional CNNs since the transformers have achieved state-of-the-art performance across various computer vision tasks. In NLP, the original transformer model takes a 1D sequence of words as input. The Vision Transformer (ViT) [10] adapted the transformer architecture to handle 2D images by dividing them into a grid of patches, with each patch being flattened into a single vector. This work is known as the pioneer of using the transformer with visual data.

1.1 Review of Related Survey Articles

Recently, a multitude of transformer variants have been proposed, demonstrating that transformer-based models achieve state-of-the-art results across diverse tasks. To keep pace with the increase of transformer-based approaches, numerous surveys have been introduced to provide comprehensive overviews of the transformer landscape. Table 1 provides a comparison of recent survey works focusing on vision transformer models.

Table 1: Comparison of recent survey articles on vision transformer models

Survey	Year	Image	Video	Diffusion	Comparison
Khan et al. [11]	2020	✓	✓		✓
Liu et al. [12]	2021	✓			✓
Hafiz et al. [13]	2021	✓	✓		
Lin et al. [14]	2021	✓			
Liu et al. [15]	2022	✓			✓
Selva et al. [16]	2022		✓		✓
Min et al. [17]	2022				✓
Ruan et al. [18]	2022		✓		
Han et al. [19]	2022	✓	✓		
Yang et al. [20]	2022	✓	✓		✓
Islam [21]	2022	✓			
Ours	2023	✓	✓	✓	✓

Lin et al. [14] focused on the attention mechanism in their survey. They divided the improvement on attention into six categories including spare attention, linearized attention, prototype and memory compression, low-rank self-attention, attention with prior and improved multi-head mechanism. Then, they discussed position representations, layer normalization and position-wise feed-forward network which are three important parts of the transformer network. They also reviewed the transformer-based approach which modifies from the vanilla transformer to improve the computation of transformer networks.

Khan et al. [11] provided a survey of the transformer approaches in computer vision. Firstly, the methods using single-head self-attention are discussed. These methods are based on convolution operation and add a self-attention layer to exploit the long-range dependencies. In the second part, transformer (multi-head self-attention) methods are reviewed. In addition, the survey also discusses

six fields of computer vision that transformer have been applied, including object detection, segmentation, image and scene generation, low-level vision, multi-modal tasks, and video understanding. Han et al. [19] categorized the transformer-based methods into four main parts in their survey, including backbone network, high/mid-level vision, low-level vision, and video processing. In addition, they also discussed multi-modal tasks and the efficient transformer. Two kinds of backbone network were discussed, containing pure transformer and transformer with convolution. Yang et al. [20] reviewed methods using the transformer in image and video applications. In image tasks, the survey first reviews transformer networks as backbones. Then, they provide a detailed discussion about image classification, object detection, and image segmentation tasks in images. In the second part of the survey, the authors provide two aspects of video tasks, including object tracking and video classification.

Hafiz et al. [13] reviewed attention-based deep architectures for machine vision. A detailed discussion of five architectures which are based on attention is provided. Then, they discussed three combinations of CNNs and the transformer. The first kind is a convolutional neural network with extra attention layers [7,9]. CNNs are used to extract features that are input to the transformer. The third kind is the combination of CNN and transformer.

Liu et al. [12] reviewed three popular tasks of computer vision, containing classification, detection, and segmentation. The authors split classification methods into various categories, such as pure transformer, the combination of CNN and transformer, and deep transformer. Islam [21] reviewed recent transformer-based methods for image classification, segmentation, 3D point clouds, and person re-identification. This survey discussed semantic segmentation and medical image segmentation. Xu et al. [22] focused on transformer-based methods in low-level vision and generation in their survey. The authors also reviewed transformer methods for the backbone which are used for classification tasks. In addition, high-level vision and multi-model learning were discussed in this survey.

CNNs have obtained state-of-the-art performance in many fields of computer vision. Transformer has recently introduced and outperformed CNN-based methods in many tasks, such as classification, object detection, and segmentation. Liu et al. [15] reviewed recent deep Multi-layer Perceptron (MLP) approaches. The pioneering MLP methods [23–25] were discussed which obtained comparable performance to CNNs and the transformer. In the main part of the survey, they discuss three categories of MLP block variants. They also provide different architectures of MLP variants, such as single and pyramid architectures. A comparison of MLP, CNN, and transformer-based methods were provided on image classification, object detection, semantic segmentation, low-level vision, video analysis and point cloud.

In contrast, Selva et al. [16] focused on video transformers in their work. In the first main part, the survey discusses some pre-processing methods of video before feeding into the transformer network, such as embedding, tokenization, and positional embedding. Then, two main efficient designs were discussed for long sequences of video. The review provided three different approaches for multi-modality including multi-model fusion, multi-model translation, and multi-model alignment. Training a transformer and the performance of video classification using the transformer were compared in the last section of the survey.

Graphs have been used to represent structural information in many fields. In a graph, objects are represented by nodes/vertices while the relationships between objects are represented by the edges. Min et al. [17] provided an overview of transformers for graphs. The survey discussed three incorporations of transformer and graph, including Graph Neural Networks as auxiliary modules in the transformer, improved positional embedding from graphs, and improved attention matrices from

graphs. Moreover, the authors conducted an experiment to compare the effectiveness of methods in the three groups.

On the other hand, Ruan et al. [18] focused on transformer-based methods for video-language learning. A pre-training and fine-tuning strategy for video-language processing is discussed. Then, two types of model structures using the transformer are reviewed, including single-stream and multi-stream structures.

1.2 Contributions of this Survey Article

Recently, numerous methods based on transformers have been proposed for various tasks in computer vision. This review provides a comprehensive discussion of transformer-based approaches across different computer vision tasks. In summary, our main contributions are listed below:

- This paper comprehensively reviews recent visual transformers for image tasks, covering three fundamental areas: downstream, generation and segmentation.
- In addition, we delve into the state-of-the-art transformers for video tasks. Specifically, we comprehensively examine the success of transformers as backbones in a wide range of diffusion models.
- We present a detailed comparison of recent methods that utilize transformers as backbones.

1.3 Roadmap of the Survey

The rest of the survey is organized as follows. Firstly, a discussion of the components of an original transformer network in Section 2. In Section 3, we discuss a wide range of vision transformers for image data. Next, we discuss recent transformers for video data in Section 4. Section 5 discusses recent transformer-based diffusion models. Then, Section 6 compares the performance of the recent methods based on the transformer network. Finally, we discuss some open research problems and give the conclusion of this survey in Sections 7 and 8, respectively.

2 Revisiting the Components of Transformer Network

Transformer was introduced by Vaswani et al. [1] for NLP. The transformer includes an encoder and a decoder which are used to encode the input and generate the output, respectively. Both the encoder and decoder have several transformer blocks. Each block contains a multi-head attention layer, a feed-forward neural network, and layer normalization as illustrated in Fig. 1.

2.1 Self-Attention Mechanism

The input vector x is transformed into query q, key k and value v vectors with dimension $d_q = d_k = d_v = d_{model}$:

$$k = x\boldsymbol{W}_k, v = x\boldsymbol{W}_v, q = x\boldsymbol{W}_q \tag{1}$$

where W_k, W_v, W_q are three matrices that are trained during the training phase. In practice, the queries, keys, and values are packed together into matrices Q, K, and V, respectively. Thus, the attention is computed with these matrices:

$$Attention\,(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

where the score is calculated by a dot product of the query and the key, and the score is normalized by a softmax operation *softmax*().

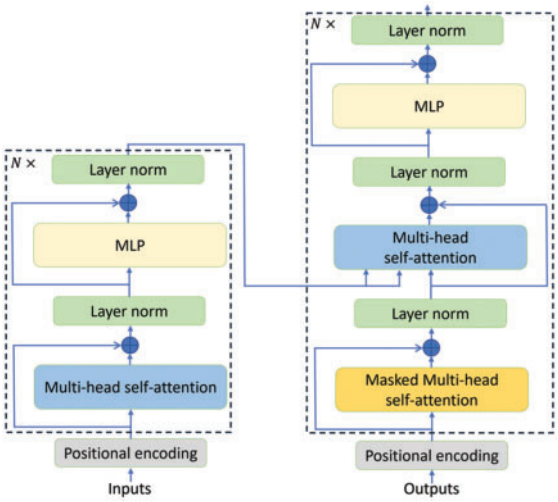

Figure 1: The vanilla transformer block, including an encoder (left) and a decoder (right). The encoder and decoder consist of several layers. Each layer of the encoder and the decoder contains multi-head self-attention mechanism and a multi-layer perceptron. In addition, the decoder has a masked multi-head self-attention

2.2 Multi-Head Attention

Multi-head attention is used to improve the performance of the attention mechanism by projecting the queries, keys and values into multiple subspaces. These projected outputs are processed parallel by attention heads. Then, the output matrices are concatenated and projected to the final output:

$$head_i = Attention\left(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\right)$$

$$MultiHead\left(\boldsymbol{Q}', \boldsymbol{K}', \boldsymbol{V}'\right) = Concat(head_i, \ldots head_h)\boldsymbol{W}^O \tag{3}$$

where $\boldsymbol{Q}'$, $\boldsymbol{K}'$, $\boldsymbol{V}'$ are the concatenation of $\left\{\boldsymbol{Q}_i\right\}_{i=1}^{h}$, $\{\boldsymbol{K}_i\}_{i=1}^{h}$, $\{\boldsymbol{V}_i\}_{i=1}^{h}$, respectively. $\boldsymbol{W}^0$ is the projection weight.

2.3 Feed-Forward Network

The second layer of a transformer block is a feed-forward network that contains two linear transformations and a nonlinear activation function in between:

$$FFN\left(\boldsymbol{X}\right) = \boldsymbol{W}_2\sigma\left(\boldsymbol{W}_1\boldsymbol{X}\right) \tag{4}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the two weight matrices of the two linear layers, σ is a nonlinear activation function, and $\boldsymbol{X}$ is used as the input of the feed-forward network.

2.4 Residual Connection and Layer Normalization

A residual connection [26] is added into each sub-layer, for example, the multi-head attention and the feed-forward network layer. In addition, a layer normalization [27] is followed each residual connection.

2.5 Positional Encoding

The positional information of words in a sentence is not encoded by the self-attention layer. To model the sequential information, the relative or absolute position of the tokens is added to the inputs. Sine and cosine functions are used for positional encoding.

$$PE\,(pos, i) = \begin{cases} \sin\,(pos \cdot w_k) & if\ i = 2k \\ \cos\,(pos \cdot w_k) & if\ i = 2k+1 \end{cases} \tag{5}$$

where $w_k = \dfrac{1}{\left(10000^{\frac{2k}{d}}\right)}$, pos denotes the position of the word, i denotes the current dimension of the positional encoding, and d denotes the dimension. Each positional encoding corresponds to a sinusoid. The transformer can learn the relative positions.

3 Vision Transformer for Image Data

3.1 Vision Transformer for Downstream Tasks

DINO [28] is a self-supervised approach, including student and teacher networks. Both student and teacher networks receive two transformations of input. Their outputs are normalized, and a cross-entropy loss is used to measure the similarity of them. To exchange visual information between regions, Fang et al. [29] introduced MSG-Transformer for image classification and object detection. Information in a local window is abstracted by a messenger token and is exchanged with other messenger tokens. Therefore, the information of local regions is exchanged by messenger tokens. To exchange information, groups of channels are obtained by splitting the channels of each messenger token. Then, obtained groups are shuffled with all other messenger tokens to exchange information. However, these transformers produce feature maps limited to a single scale while the CNN can output multi-scale feature maps suitable for various computer vision tasks.

A hierarchical transformer often includes four transformer stages in which different scales of feature maps are generated, as illustrated in Fig. 2. Each stage contains multiple transformer blocks which are composed of a multi-head attention layer and a feed-forward layer. The input is hierarchically reduced spatial size and expanded channel capacity through four stages of the transformer. PVT1 [30] introduced a pure transformer backbone that can be used as backbone for many downstream tasks. The output of the network is multi-scale feature maps which have a resolution of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$. The multi-scale feature maps are obtained by dividing the input features at each beginning stage. Moreover, a traditional multi-head attention is replaced by a Spatial-Reduction Attention (SRA) to reduce computational cost. In the SRA layer, the attention receives a key and a value which are reduced by the spatial scale. PVT2 [31] is improved from the previous version to address the computational complexity and arbitrary size of the input image. In PVT2, the spatial dimension is reduced by using average pooling instead of convolution and an overlapping patch embedding is introduced to capture the local continuity information. In addition, a zero-padding position encoding is introduced with a depth-wise convolution in feed-forward networks. Swin transformer [32] is one of the most novel transformer-based backbones that reduces the complexity of attention computation by proposing Shifted windows. To generate hierarchical features, a patch merging layer is applied at each stage of the network. Shift windows are proposed to compute self-attention within non-overlapping windows. Moreover, a shifted window partitioning was introduced to exploit the connection of the non-overlapping windows. Swin transformer 2 [33] is an improved version of Swin transformer 1. Swin transformer 2 introduced a residual post normalization approach by placing layer norm after

self-attention and MLP layers to resolve the increase of activation values at deeper layers. To solve the dominance of the attention map by a few pixel pairs, scaled cosine attention was introduced to compute the attention. In addition, a position bias method was proposed to transfer across windows. Given that these transformers generate multi-scale feature maps and possess a global receptive field, they can serve as a backbone for a variety of computer vision tasks, such as object detection, semantic segmentation, and video anomaly detection [34]. Furthermore, these hierarchical transformers can replace a CNN backbone and can be integrated into other networks.

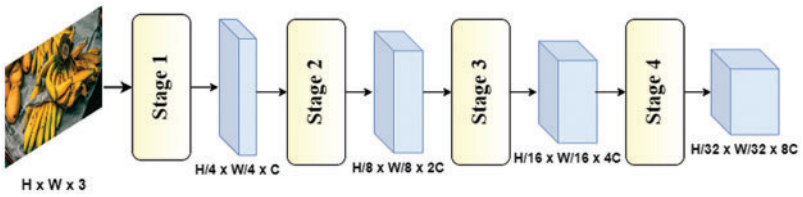

Figure 2: The architecture of a hierarchical transformer includes four stages for generating feature maps of different scales

Uformer [35] is a hierarchical transformer for image restoration. The network contains K encoder stages and K decoder stages. Each encoder stage includes a stack of locally enhanced window transformer blocks and one down-sampling layer. On the opposite stages, each has a stack of locally enhanced window transformer blocks and an up-sampling layer. A 2×2 transposed convolution with stride 2 is used to up-sample the features. The locally-enhanced window transformer block is introduced to capture long-range dependencies and local context by using convolution in the transformer as in [36,37]. Restormer [38] is a hierarchical transformer model for image restoration. Restormer replaces multi-head self-attention with a multi-Dconv head transposed attention to obtain linear complexity. Moreover, the proposed attention aims to compute the attention across channels instead of the channel dimension. A 1×1 convolution and a 3×3 depth-wise convolution are used to compute attention. In addition, two parallel paths of 1×1 and depth-wise convolutions are used in feed-forward network to improve representation.

Chu et al. [39] proposed Twins-PCPVT which is based on PVT [30] and Twins-SVT which is based on spatially separable self-attention. In Twins-PCPVT, conditional position encoding is used to replace absolute positional encoding. The spatially separable self-attention contains locally-grouped self-attention which is computed in each sub-window. To exchange the information between local windows, a global self-attention was proposed to communicate between sub-windows. Cswin transformer [40] computes self-attention in two directions by proposing cross-shaped window self-attention. The proposed attention obtains attention of a large area and global attention. In addition, locally-enhanced positional encoding was introduced for the downstream transformer network. Window-based transformers [32] have achieved promising results on multiple tasks of computer vision. Shuffle transformer [41] was proposed to improve the connection between non-overlapping local windows. A shuffle transformer block contains a shuffle multi-head self-attention to enhance the connection between the windows and neighbor-window connection to strengthen the information between windows by inserting a depth-wise convolution before the MLP module. Glance-and-Gaze Transformer [42] proposed Glance attention which computes self-attention with a global reception field. Since the feature maps are split into different dilated partitions, a partition contains information of the whole input feature instead of a local window. To capture the local connection between partitions, a Gaze branch was introduced using the depth-wise convolution.

Hassani et al. [43] introduced a Neighborhood Attention Transformer (NAT) which computes attention using proposed neighborhood attention. This attention has lower computational complexity and local inductive biases. Each point in features attends to its neighboring points. The NAT outputs pyramid features that are used for different downstream tasks in computer vision. DaViT [44] proposed a dual attention vision transformer that computes self-attention using both spatial tokens and channel tokens. Each stage of the transformer has dual attention blocks which include a spatial window attention block, a channel group attention block, and a feed-forward network. To obtain global information, self-attention is computed on the transpose of patch-level tokens instead of patch-level. Moreover, channels are grouped and compute attention to reduce the complexity. Zhang et al. [45] proposed a Multi-Scale Vision Longformer which is used for high-resolution image encoding. An efficient ViT was proposed by modifying the vanilla transformer. Multiple proposed ViT is stacked to construct a multi-scale vision transformer that generates different feature maps. In addition, the attention mechanism of vision longer is used to reduce the complexity. Both global and local tokens are used to access global and local information.

Convolutional Vision Transformer [46] is a hierarchical transformer that leverages convolution to the transformer. The convolution is applied to the Convolutional Token Embedding layer and convolutional transformer block to encode local spatial contexts. In the transformer block, a depth-wise convolution is used instead of the position-wise linear projection in the vanilla transformer. Li et al. [47] proposed a Multiscale Vision Transformer (MViTv2) for image and video classification. Moreover, the proposed method was evaluated with object detection and video recognition tasks. The relative positional embedding is used in the pooled self-attention to model the relative distance across tokens. A residual pooling connection is applied to enhance the representation. The Vitae [48] is a transformer network that contains two main cells, including a reduction cell and a normal cell. Reduction cells use convolutional layers with different dilation rates. The spatial dimension of features is reduced by using stride convolution. The normal cells have the same architecture as the reduction cell. However, the pyramid reduction module extracted multi-scale features are used only in the reduction cell. Chen et al. [49] transited a transformer-based model into a convolution-based model. There are eight steps, including replacing the token, replacing patch embedding, splitting the network into stages, replacing layer-norm, introducing 3×3 convolutions, removing position embedding, and adjusting the architecture of the network. The proposed network obtains better performance while having the same computational cost.

Tang et al. [50] proposed QuadTree Attention is computed from a rough to fine manner with lower computational complexity. Self-attention is computed with L-level pyramids. At the fine level, attention is calculated from subset tokens that are selected from the coarse level using attention score. Ding et al. [51] proposed a lightweight transformer that consists of a projector to reduce the size of the input feature, an encoder, and a decoder. Moreover, a multi-branch search space was proposed for dense prediction tasks. The search space models features with different scales and global contexts. Inception transformer [52] proposed a transformer-based network that captures both high and low-frequency features. The image tokens are passed through an inception token mixer which is composed of three branches to extract high and low frequency information. To extract high-frequency features, a combination of max-pooling and convolution operation is used while a self-attention is used to extract low-frequency features.

ConvMAE [53] is a hybrid convolution-transformer network that includes an encoder and a decoder. The encoder outputs multi-scale features of the input image. The self-attention of the transformer block is replaced by a 5×5 depthwise convolution. The random mask for stage-3 is generated by masking out $p\%$. Then, the mask of stage-2 and stage-1 are up-sampled from the

mask of the third stage. Li et al. [54] proposed masked auto-encoder pre-training for the hierarchical transformer. The proposed method contains uniform sampling and secondary masking stages. The input image with 25% visible image patches uses uniform constraint to ensure these patches as a compact image. A secondary masking was introduced to solve the degradation problem which is made by the uniform sampling. The secondary masking makes it more challenging for the recovery task to obtain a better representation of the network. Chen et al. [55] proposed an adapter that fine-tunes a transformer-based backbone on vision-specific tasks without changing the backbone network. The proposed network contains two parts, including the backbone network and the proposed adapter. The backbone network is an original transformer network that includes L transformer layers. The adapter has N blocks which composed of a spatial feature injector and a multi-scale feature extractor. A feature pyramid of the input is generated after passing through N blocks. VOLO [56] introduced an outlook attention mechanism which can encode fine-level features and contexts. The model is composed of a stack of Out-lookers and a stack of transformer blocks. The Out-looker has an outlook attention layer and a MLP layer. The former is used to extract fine-level features and the latter is used to aggregate global information. Although the performance of these transformers has improved significantly compared to previous transformers, the model sizes of these models have become bigger.

3.2 Vision Transformer for Generation

UNet [57] is a popular convolutional network architecture that was introduced for biomedical image segmentation. The network contains two branches. The left branch is a down-sampling of the feature map while the output features are up-sampled by the other branch. In this section, we discuss transformer networks that have a U-shaped architecture.

TransUNet [58] combines a Transformer and a CNN to extract both local and global context information. CNN is used to extract features of the input. Stacked transformer layers are applied to the extracted features and output the hidden features. A decoder up-samples the output features to the final segmentation mask using a 2×2 up-sampling operator. The input of each up-sampled stage includes the features of the previous stage and the corresponding features from the encoder. UNETR [59] was proposed for medical image segmentation. The network contains a transformer as the encoder and a CNN as the decoder. The encoder contains a stack of transformers to encode the features of a 3D input image. In the decoder, the combination of de-convolutional, convolutional, and normalization layers is applied to reshape the extracted features obtained from the encoder. Then, the reshaped features are concatenated with the previous feature stage. U-Net transformer [60] was proposed for image segmentation which has a U-shaped architecture. The network contained a self-attention module and a cross-attention module. The first module is used to exploit the global interactions of features while the second one keeps important information and discards irrelevant information from the skip connection features. UTNet [61] is a hybrid transformer network for medical image segmentation. This combination aims to capture local features by convolutional layer and long-range information by self-attention. The transformer block is applied after a residual convolutional block at each stage of the encoder except for the first resolution. To reduce the computational complexity, two projections are applied to the key and values. In addition, the relative position encoding is added to maintain the position information. UTNetV2 [62] is improved from UTNet [61] for medical image segmentation. A bidirectional multi-head attention was proposed to reduce the computational cost. The proposed attention maintains a semantic map through network stages. At each layer, the output from the previous layer and a semantic map projected by a depth-wise separable convolution and 1×1 convolution are used as input. The proposed attention encodes global context information with small computation. Mixed Transformer U-Net [63] introduced a mixed transformer module that includes

two types of attention. The first self-attention captures the short and long range dependencies by proposing a local-global strategy and Gaussian mask. The second attention captures inter-sample correlations. UCTransNet [64] introduced a transformer in U-Net architecture. The skip connections of U-Net are replaced by a channel transformer which includes a channel cross fusion module and channel wise cross-attention. Channel-wise cross fusion transformer fuses the multi-scale features that are extracted by the encoder. In addition, the channels-wise cross attention module fuses the output of the channel-wise cross fusion transformer module and the features of the previous decoder. Transfuse [65] combines CNN and transformer to capture both global and local information. The input image is processed by two parallel networks. The extracted features of the transformer and CNN are fused by a proposed BiFusion module which is composed of various mechanisms such as channel attention [7], spatial attention [9], and residual block. These models try to integrate the transformer model into an autoencoder. However, the main components of these models are still convolutional layers. For example, in models such as TransUNet and UNETR, a transformer functions as an encoder while a CNN serves as a decoder.

Swin-Unet [66] proposed a pure transformer that has a shape like UNet for medical image segmentation. Both the encoder and decoder are composed of Swin transformer blocks [32]. Patch merging layer is used to down-sample and increase dimension while the patch expanding layer up-samples and restores the resolution. The extracted features from the encoder are fused with the features from the previous decoder layer via skip connections. Swin UNETR [67] combines Swin transformer [32] and CNN for 3D brain tumor semantic segmentation. A sequence of 3D tokens of the input is generated by a patch partition. The embedding tokens are extracted features by a Swin transformer-based encoder. A decoder is used to predict the final segmentation outputs. VT-UNet [68] proposed a transformer which has U-shaped architecture. The encoder includes three main stages including encoder block and patch merging. The encoder block is composed of two types of windows like a Swin transformer [32]. The decoder contains various decoder blocks, patch expanding and a classifier. Each decoder block has two self-attention encoders as regular and shifted window attentions. These models offer the advantage of proposing a pure transformer network, comprising both a transformer-based encoder and a transformer-based decoder.

3.3 Vision Transformer for Segmentation

Segmenter [69] is a transformer network for semantic segmentation. To exploit the global information, Segmenter is based on the vision transformer which does not use convolutions in the network. The network includes an encoder and decoder. The former is used to exploit the contextualized information while the latter up-samples the output of the encoder to pixel-level scores. In addition, two types of decoder were introduced, including a linear decoder and a mask transformer. In the mask transformer, a set of class embeddings was used to generate a class mask. This work was one of the pioneers in applying transformers to semantic segmentation. By introducing transformers into this domain, the study opened avenues for capturing a global receptive field in segmentation tasks. TopFormer [70] was proposed for semantic segmentation on mobile devices. The network uses stacked MobileNetV2 blocks [71] to create tokens at different scales. Semantic information is extracted by stacked transformer blocks with the generated tokens as input. The transformer block has the same architecture as the original transformer. However, linear layers are replaced by a 1×1 convolution layer, GELU is replaced by ReLU6, and a depth-wise convolution layer is used in the feed-forward network. Gu et al. [72] proposed a multi-scale transformer for semantic segmentation. The network contains four stages which include many parallel multi-scale transformer branches. An efficient self-attention was introduced to balance between efficiency and performance. Lawin Transformer

[73] solved the lack of contextual information by proposing a large window attention. To capture multi-scale representations, five parallel branches composed of three window attention branches, one shortcut connection, and one pooling branch were used. The proposed attention is inserted into a hierarchical vision transformer to exploit multi-scale representations.

4 Vision Transformer for Video

Space-Time Attention Model (STAM) [74] contains a spatial and temporal transformer that is used to extract both spatial and temporal information from video frames. The spatial attention is applied on patches of each frame while the temporal attention is applied on the output of the spatial attention to capture the temporal information of frames. Bain et al. [75] proposed a transformer-based model that includes two encoders for encoding image/video and a sequence of words. To process video input, the divided space-time attention is used with a modification of the residual connection of the temporal self-attention and spatial self-attention blocks. TimeSformer [76] is a transformer-based model for video classification. The model exploits both spatial and temporal information by obtaining temporal attention and spatial attention separately at each block of the transformer. The reduced computational complexity due to the temporal attention and spatial attention are computed once after the other. Zhang et al. [77] proposed a token shift module for modeling temporal information in the transformer. Several shift variants were introduced, including token shift, temporal shift, and patch shift. The token shift module can be inserted into various positions in a transformer-based encoder. Each position of the token shift will determine the degree of motion information. The shift module can be insert before the layer-norm layer, before the multi-head attention and feed-forward network, or post multi-head attention and feed-forward network. VidTr [78] is a video transformer for video classification. To reduce memory consumption, VidTr exploits spatiotemporal features by using spatial and temporal attention separately. In addition, a topK-based pooling was proposed to down-sample temporal since the video contains redundant information. Many works [76–79] have tried to reduce the complexity of the space-time attention. Multiple transformer-based architectures [79] were introduced for video classification. The interactions of all spatiotemporal tokens lead to quadratic complexity while computing multi-head self-attention. Model 2 solves the above limitation using two separate transformer encoders. However, this model increases transformer layers. Model 3 solves this disadvantage by computing temporal self-attention after spatial self-attention in a transformer block as in [76]. In model 4, the keys and values for each query are separated into spatial and temporal dimensions. XViT [80] tries to encode space-time attention which has linear complexity $O(TS^2)$ with the number of frames. The time attention is computed from a local temporal window and the temporal of the whole video is obtained through the depth of the transformer. To reduce the complexity, the computation of space-time attention has been used shift module [81]. The complexity of a model that computes both space and time attention is $O(T^2S^2)$. Since the space-time transformers require high computational cost, devided attention computes spatial attention and temporal attention separately. This approach not only proves to be more efficient but also improves accuracy.

ConvTransformer [82] was introduced for video frame synthesis. The input frames are extracted features by a feature embedding module. The extracted features with positional maps are used as the input of an encoder-decoder. The generated frames are decoded by a synthesis feed-forward network. Both the encoder and decoder contain a multi-head convolutional self-attention layer and a 2D convolutional feed-forward network. VisTR [83] is an end-to-end transformer-based model for video instance segmentation. The extracted features of a backbone network are passed through an encoder-decoder transformer to output a sequence of object prediction. The instance sequence matching strategy and instance sequence segmentation module are proposed to match the same

instance in different images and predict the mask sequence for each instance. TeViT [84] proposed a transformer backbone that exploits temporal features efficiently. To exploit the temporal information, messenger tokens leaned embedding are shifted along the temporal axis. The temporal information is exploited at each stage of the network and the shift mechanism has no extra parameter. In addition, a spatiotemporal query interaction head network is introduced to exploit the temporal information at the instance level. Hwang et al. [85] introduced a transformer-based model for video instance segmentation. The proposed model reduces the cost of the space-time attention by proposing an Inter-Frame Communication transformer (IFC) that solves the heavy computation and memory usage of previous per-frame methods. The information between frames is exchanged when the feature maps of input video are passed through an inter-frame communication encoder. The encoder is composed of transformer-based encoder-receive and gather-communicate.

Yan et al. [86] introduced a multi-view transformer for video recognition. A multi-view transformer contains separate transformer encoders which are used to process tokens of different views. To fuse information from different views, three fusion methods were introduced, including cross-view attention, bottleneck tokens, and MLP fusion. The output is produced by a global encoder. Neimark et al. [87] proposed a video transformer network for video recognition. The entire video is processed using Longformer [88] which has a linear computation complexity. Girdhar et al. [89] proposed an anticipative architecture instead of aggregation of features over the temporal axis. Vision transformer [10] is used as a backbone network to extract features of individual video frames. Then, the extracted features are processed by a causal transformer decoder to predict future features. Fan et al. [90] proposed a multi-scale vision transformer that generates a multi-scale pyramid of features of the input. To generate multi-scale features, a multi-head pooling attention was proposed. The queries Q, keys K, and values V are pooled before computing attention. The network contains multi-stages. At each stage, the channel dimension is increased while the spatiotemporal resolution is reduced. Weng et al. [91] proposed a combination of CNN and transformer network for video reconstruction. A multi-scale feature pyramid is generated by a recurrent convolution backbone including several ConvLSTM layers. The generated features are used as input for token pyramid aggregation which models the internal and intersected dependency of the input features. An up-sampler is used to reconstruct the intensity image.

Zhang et al. [92] proposed a cross-frame transformer for video super-resolution network. The similarity and similarity coefficient matrixes of the input frames are obtained using self-attention computation. The obtained matrixes are used to reconstruct the super resolution frame using a multi-level reconstruction. Geng et al. [93] proposed a transformer network that has UNet architecture for video super resolution tasks. The proposed network contains an encoder to extract features and a decoder to reconstruct output frames. Both the encoder and decoder have four stages that include many Swin transformer blocks [32]. In addition, the extracted features of each stage of the encoder and a single frame query are used as input for the corresponding decoder. Liu et al. [94] proposed a transformer-based network that aims to exploit both object movements and background textures for video in-painting. A sequence of input frames is down-sampled and up-sampled by a CNN encoder and decoder, respectively. In addition, a decoupled spatial-temporal transformer is placed between the encoder and decoder to exploit spatial and temporal information effectively. By disentangling the spatial and temporal attention computation, the computational complexity is reduced significantly. VDTR [95] is a transformer-based model for video de-blurring. The features of the input frames are extracted by a transformer-based auto-encoder. The extracted spatial features are used as the input of a temporal transformer to exploit information from neighboring frames. The attention between the

frames is computed by using a temporal cross-attention module which the queries are calculated from the reference feature maps. The output frame is reconstructed by several transformer blocks.

5 Transformer for Diffusion Models

5.1 Diffusion Models

The forward process of the Gaussian diffusion models [96] gradually injects noise into real data:

$$q\left(x_t|x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \boldsymbol{I}\right) \tag{6}$$

We can sample x_t at any timestep t by using:

$$q\left(x_t|x_0\right) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha_t}}x_0, \left(1-\overline{\alpha_t}\right)\boldsymbol{I}\right), \tag{7}$$

where $\alpha_t = 1 - \beta_t$ and $\overline{\alpha_t} = \prod_{s=1}^{t} \alpha_s$.

The reverse process inverts the forward process:

$$p_\theta\left(x_{t-1}|x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta\left(x_t, t\right)\right) \tag{8}$$

The reverse process model is trained to optimize the ELBO on the log-likelihood:

$$L = \mathbb{E}\left[-logp_\theta\left(x_0\right)\right] \leq \mathbb{E}_q\left[-log\frac{p_\theta\left(x_{0:T}\right)}{q\left(x_{1:T}|x_0\right)}\right] \tag{9}$$

Reparameterizing μ_θ with a model to predict the noise ϵ:

$$\mu_\theta\left(x_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha_t}}}\epsilon_t\left(x_t, t\right)\right), \tag{10}$$

where ϵ_θ is a learned function.

5.2 Transformer-Based Diffusion Models

Diffusion models often leverage a convolutional U-Net to learn the reverse process to construct the output from the noise. DiTs [97] replace the U-Net with a transformer for operating on latent patches and achieve state-of-the-art performance on the class conditional generation tasks. Swinv2-Imagen [98] introduces a diffusion model for text-to-image task, which is based on the Swinv2 transformer and Scene Graph generator. The scene graph generator enhances the text understanding by generating a scene graph and extracting the relational embeddings for generating image. UniDiffuser [99] uses a transformer to process all input types of various modalities, which performs text-to-image, image-to-text, and image-text pair generation. To generate high-quality and realistic outputs from textual descriptions, ET-DM [100] combines the advantages of the diffusion model and transformer model for text-to-image generation. The transformer model exploits the mapping relationship between textual descriptions and image representation. However, the text-to-image (T2I) models require high training costs. PIXART-α [101] solves this issue by introducing three advance designs, including training strategy decomposition, efficient T2I transformer, and high-informative data. PIXART-δ [102] achieves a 7 $\times$ improvement over the previous version PIXART-α by combining the Latent Consistency Model and ControlNet. The ControlNet is integrated with the transformer, which achieves effectiveness in controlling information and generating high-quality output.

Diffusion models have been applied to various fields. LayoutDM [103] uses a pure transformer to generate a layout, which captures relationship information between elements effectively. DiffiT [104] proposes a diffusion vision transformer with a hierarchical encoder and decoder, consisting of novel time-dependent self-attention modules. To speed up the learning process of the diffusion probabilistic model, Gao et al. [105] introduced a Masked Diffusion Transformer (MDT), which masks the input image in the latent space and generates images from masked input by an asymmetric masking diffusion transformer. MDT [106] introduces a multimodal diffusion transformer, which encodes the image observation using two vision-language models. In addition, a CLIP model is used to encode the goal images or language annotations. For medical image segmentation, a diffusion transformer U-Net [107] introduces a transformer-based U-Net for extracting various scales of contextual information. Moreover, a cross-attention module fuses the embeddings of the source image and noise map to enhance the relationship from source images. Zhao et al. [108] proposed a spatio-temporal transformer-based diffusion model for realistic precipitation nowcasting. The past observations are used as a condition for the diffusion model to generate the target image sequence from noise. Sora [109] is a large-scale training of generative models, which generates a minute of high-fidelity video or images. A raw input video is compressed into a latent spacetime representation. Then, a sequence of latent spacetime patches is extracted to capture both the appearance and motion information. A diffusion transformer model is used to construct videos from these patches and work tokens.

6 A Comparison of Methods

Table 2 summarizes the popular transformer-based architectures on the ImageNet-1K classification task. This dataset consists of 1.28 M training images and 50 K validation images for 1000 classes. In addition, different configurations are compared to evaluate the efficiency of proposed methods, including model size, number of parameters, FLOPs, and Top-1 accuracy with a single 224 × 224 pixels.

Table 2: Comparison of different transformer models on ImageNet-1K classification

Method	Size	Year	# Params	FLOPs	Top-1 acc
Glance-and-gaze [42]	Tiny	2021	28 M	4.5 G	82.0
	Small	2021	50 M	8.7 G	83.4
Shuffle transformer [41]	Tiny	2021	29M	4.6 G	82.5
	Small	2021	50 M	8.9 G	83.5
	Base	2021	88 M	15.6	84.0
HR-NAS [51]	HR-NAS-A	2021	5.5 M	267 M	76.6
	HR-NAS-B	2021	6.4 M	325 M	77.3
CVT [46]	CvT-13	2021	20 M	4.5 G	81.6
	CvT-21	2021	30 M	7.1 G	82.5
Vision longformer [45]	Tiny	2021	6.7 M	1.3 G	76.7
	Small	2021	24.6 M	4.9 G	82.4
	Medium	2021	39.7 M	8.7 G	83.5
	Base	2021	55.7 M	13.4 G	83.7

(Continued)

Table 2 (continued)

Method	Size	Year	# Params	FLOPs	Top-1 acc
MViTv2 [47]	Tiny	2021	24 M	1.3 G	82.3
	Small	2021	35 M	7 G	83.6
	Base	2021	52 M	10.2 G	84.4
	Large	2021	218 M	42.1 G	85.3
ViTAE [48]	ViTAE-T	2022	4.5 M	1.5 G	75.3
	ViTAE-6M	2022	6.5 M	2 G	77.9
	ViTAE-13M	2022	13.2 M	3.4 G	81.0
	ViTAE-S	2022	23.6 M	5.6 G	82.0
Visformer [49]	Tiny	2021	10.3 M	1.3 G	78.6
	Small	2021	40.2 M	4.9 G	82.2
Swin transformer 1 [32]	Tiny	2021	29 M	4.5 G	81.3
	Small	2021	50 M	8.7 G	83.0
	Base	2021	88 M	15.4 G	83.5
Swin transformer 2 [33]	SwinV2-B	2022	88 M	–	78.08
	SwinV2-L	2022	197 M	–	78.31
	SwinV2-G	2022	3.0 B	–	84.0
PVTv1 [30]	Tiny	2021	13.2 M	1.9 G	75.1
	Small	2021	24.5 M	3.8 G	79.8
	Medium	2021	44.2 M	6.7 G	81.2
	Large	2021	61.4 M	9.8 G	81.7
PVTv2 [31]	PVTv2-B1	2022	13.1 M	2.1 G	78.7
	PVTv2-B2	2022	25.4 M	4 G	82.0
	PVTv2-B3	2022	45.2 M	6.9 G	83.2
	PVTv2-B4	2022	62.6 M	10.1 G	83.6
	PVTv2-B5	2022	82.0 M	11.8 G	83.8
Neighborhood attention [43]	Mini	2022	20 M	20 G	81.8
	Tiny	2022	28 M	4.3 G	83.2
	Small	2022	51 M	7.8 G	83.7
	Base	2022	90 M	13.7 G	84.3
QuadTree [50]	QuadTree-B-b0	2022	3.5 M	0.7 G	72.0
	QuadTree-B-b1	2022	13.6 M	2.3 G	80.0
	QuadTree-B-b2	2022	24.2 M	4.5 G	82.7
	QuadTree-B-b3	2022	46.3 M	7.8 G	83.7
	QuadTree-B-b4	2022	64.2 M	11.5 G	84.0
CSWin transformer [40]	Tiny	2022	23 M	4.3 G	82.7
	Small	2022	35 M	6.9 G	83.6
	Base	2022	78 M	47 G	84.2

(Continued)

Table 2 (continued)

Method	Size	Year	# Params	FLOPs	Top-1 acc
VOLO [56]	VOLO-D1	2021	27 M	6.8 B	84.2
	VOLO-D2	2021	59 M	14.1 G	85.2
	VOLO-D3	2021	86 M	20.6 G	85.2
	VOLO-D4	2021	193 M	43.8 G	85.7
	VOLO-D5	2021	296 M	69 G	86.1
Twins [39]	Small	2022	24 M	2.9 G	81.7
	Base	2022	56 M	8.6 G	83.2
	Large	2022	99.2 M	15.1 G	83.7
Cswin transformer [40]	Tiny	2022	23 M	4.3 G	82.7
	Small	2022	35 M	6.9 G	83.6
	Base	2022	78 M	15 G	84.2
Inception transformer [52]	Small	2022	20 M	4.8 G	83.4
	Base	2022	48 M	9.4 G	84.6
	Large	2022	87 M	14 G	84.8
Dual AVT [44]	Tiny	2022	28.3 M	4.5 G	82.8
	Small	2022	49.7 M	8.8 G	84.2
	Base	2022	87.9 M	15.5 G	84.6

ADE20K is a challenging dataset, including 20 K images for training and 2 K images for validation. Table 3 compares mIoU results on the ADE20K dataset with different transformer models.

Table 3: Performance comparison of different transformers on ADE20K

Method	Size	Setting	# Params	mIoU
VOLO [56]	VOLO-D1	VOLO	ImgNet-1k	50.5
	VOLO-D3	VOLO	ImgNet-1k	52.9
	VOLO-D5	VOLO	ImgNet-1k	54.3
Twins [39]	Small	PVT	ImgNet-1k	43.2
	Base	PVT	ImgNet-1k	45.3
	Large	PVT	ImgNet-1k	46.7
Cswin transformer [40]	Tiny	FPN	ImgNet-1k	48.2
	Small	FPN	ImgNet-1k	49.2
	Base	FPN	ImgNet-1k	49.9
Inception transformer [52]	Small	FPN	ImgNet-1k	48.6
Dual AVT [44]	Tiny	UperNet	ImgNet-1k	46.3
	Small	UperNet	ImgNet-1k	48.8
	Base	UperNet	ImgNet-1k	49.4

Swin transformer 1 [32] and Swin transformer 2 [33] are two popular window-based transformers. Pyramid Vision Transformer (PVT) 1 [30] and Pyramid Vision Transformer 2 [31] are two transformer architectures that are motion for other hierarchical transformers.

7 Open Research Problems

Transformer-based methods have achieved remarkable successes in natural language processing as well as computer vision. Transformers have a strong capability of capturing global context information (long-range dependencies). However, self-attention requires a huge computation cost to compute the attention map. In addition, convolutional neural networks can capture local context that is not modeled well by the transformer.

7.1 Decreasing the Computational Cost

The transformer shows the capability of modeling the long-range dependencies using self-attention mechanism. However, the computation of the full-attention mechanism [10,46–110] is inefficient because the complexity is quadratic to the size of the image. Many proposed methods have been introduced to solve the issues. For example, window-based methods [32,33–40] have linear complexity with the image size. To reduce the computational complexity to linear, many works proposed spatial reduction attention [30,31] by reducing the spatial scale of the key K and value V before the computation of self-attention. To reduce spatial dimension, the key K and value V are applied by a convolution operator or average pooling.

Recently, many studies [43,52] still try to decrease the computational cost of self-attention and compute attention more efficiently. This is an open research direction that many researchers aim to solve.

7.2 Capturing Both Local and Global Contexts

Transformers can capture the global context however it shows limitations in modeling the local context. Many studies try to capture local information by proposing a conv-attention mechanism [111] which introduces convolution in attention mechanism. Reference [46] introduced convolution to token embedding and convolutional projection for attention.

On the other hand, TransUNet [58] extracts local features by using a CNN and a transformer to aggregate global features from extracted local features. TransFuse [65] used two parallel networks including a CNN and a transformer network to capture both local and global features. STransFuse [112] combines transformer and CNN to exploit the benefits of both networks.

Transformer-based models can model global information using the self-attention mechanism. However, recent approaches combine CNN and transformer to exploit local features for the transformer. A pure transformer network that can model both local and global information is an open research direction.

8 Conclusion

Transformers have demonstrated remarkable performance across various computer vision tasks. In this survey, we have comprehensively reviewed recent transformer-based methods for image, video tasks, and diffusion models. We first categorize the methods for image tasks into three fundamental categories, including downstream, segmentation, and generation tasks. We discuss state-of-the-art transformer-based methods for video tasks and the complexity of these models. Specifically, we

provide an overview of the diffusion model and discuss recent diffusion models using a transformer as a backbone network. In addition, we provide a detailed comparison of recent transformer-based models on ImageNet and ADE20K datasets.

Acknowledgement: None.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grants 61502162, 61702175, and 61772184, in part by the Fund of the State Key Laboratory of Geo-information Engineering under Grant SKLGIE2016-M-4-2, in part by the Hunan Natural Science Foundation of China under Grant 2018JJ2059, in part by the Key R&D Project of Hunan Province of China under Grant 2018GK2014, and in part by the Open Fund of the State Key Laboratory of Integrated Services Networks under Grant ISN17-14. Chinese Scholarship Council (CSC) through College of Computer Science and Electronic Engineering, Changsha, 410082, Hunan University with Grant CSC No. 2018GXZ020784.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Dinh Phu Cuong Le, Viet-Tuan Le; analysis and interpretation of results: Dinh Phu Cuong Le, Dong Wang; draft manuscript preparation: Dinh Phu Cuong Le, Dong Wang, Viet-Tuan Le. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Vaswani *et al.*, "Attention is all you need," in *31st Int. Conf. Neural Inf. Process. Syst. (NIPS'17)*, NY, USA, 2017, vol. 30, pp. 6000–6010.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *2019 Conf. North American Chapter Assoc. Comput. Linguist.: Human Lang. Technol.*, Minneapolis, Minnesota, 2019, vol. 1, pp. 4171–4186.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, pp. 9, 2019.

[4] T. Brown *et al.*, "Language models are few-shot learners," in *34th Int. Conf. Neural Inf. Process. Syst.*, NY, USA, 2020, vol. 33, pp. 1877–1901.

[5] A. Krizhevsky, I. Sutskever, and E. G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.*, Lake Tahoe, Nevada, USA, 2012, vol. 25, pp. 1097–1105.

[6] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7794–7803. doi: 10.1109/CVPR.2018.00813.

[7] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7132–7141. doi: 10.1109/CVPR.2018.00745.

[8] J. Wang, Y. Chen, R. Chakraborty, and S. X. Yu, "Orthogonal convolutional neural networks," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 11505–11515. doi: 10.1109/CVPR42600.2020.01152.

[9] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *European Conf. Comput. Vis. (ECCV)*, Cham, Munich, Germany, Springer, 2018, vol. 11211, pp. 3–19.

[10] A. Dosovitskiy *et al.*, "An image is worth 16 ×16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent.*, Austria, 2021.

[11] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022. doi: 10.1145/3505244.
[12] Y. Liu *et al.*, "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–21, 2023. doi: 10.1109/TNNLS.2022.3227717.
[13] A. M. Hafiz, S. A. Parah, and R. U. A. Bhat, "Attention mechanisms and deep learning for machine vision: A survey of the state of the art," arXiv preprint arXiv:2106.07550, 2021.
[14] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, no. 120, pp. 111–132, 2022. doi: 10.1016/j.aiopen.2022.10.001.
[15] R. Liu, Y. Li, L. Tao, D. Liang, and H. T. Zheng, "Are we ready for a new paradigm shift? A survey on visual deep MLP," *Patterns*, vol. 3, no. 7, pp. 100520, 2022. doi: 10.1016/j.patter.2022.100520.
[16] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12922–12943, 2023. doi: 10.1109/TPAMI.2023.3243465.
[17] E. Min *et al.*, "Transformer for graphs: An overview from architecture perspective," arXiv preprint arXiv:2202.08455, 2022.
[18] L. Ruan and Q. Jin, "Survey: Transformer based video-language pre-training," *AI Open*, vol. 3, pp. 1–13, 2022. doi: 10.1016/j.aiopen.2022.01.001.
[19] K. Han *et al.*, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, 2022. doi: 10.1109/TPAMI.2022.3152247.
[20] Y. Yang *et al.*, "Transformers meet visual learning understanding: A comprehensive review," arXiv preprint arXiv:2203.12944, 2022.
[21] K. Islam, "Recent advances in vision transformer: A survey and outlook of recent work," arXiv preprint arXiv:2203.01536, 2022.
[22] Y. Xu *et al.*, "Transformers in computational visual media: A survey," *Comput. Vis. Media*, vol. 8, no. 1, pp. 33–62, 2022. doi: 10.1007/s41095-021-0247-3.
[23] I. Tolstikhin *et al.*, "MLP-Mixer: An all-MLP architecture for vision," in *Adv. Neural Inf. Process. Syst.*, Virtual, 2021, vol. 34, pp. 24261–24272.
[24] H. Touvron *et al.*, "ResMLP: Feedforward networks for image classification with data-efficient training," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5314–5321, 2023. doi: 10.1109/TPAMI.2022.3206148.
[25] L. Melas-Kyriazi, "Do you even need attention? A stack of feed-forward layers does surprisingly well on imagenet," arXiv preprint arXiv:2105.02723, 2021.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv preprint arXiv:1607.06450, 2016.
[28] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 9650–9660. doi: 10.1109/ICCV48922.2021.00951.
[29] J. Fang, L. Xie, X. Wang, X. Zhang, W. Liu, and Q. Tian, "MSG-Transformer: Exchanging local spatial information by manipulating messenger tokens," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 12063–12072. doi: 10.1109/CVPR52688.2022.01175.
[30] W. Wang *et al.*, "Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 568–578. doi: 10.1109/ICCV48922.2021.00061.
[31] W. Wang *et al.*, "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–418, 2022. doi: 10.1007/s41095-022-0274-8.
[32] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[33] Z. Liu *et al.*, "Swin Transformer V2: Scaling up capacity and resolution," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 11999–12009. doi: 10.1109/CVPR52688.2022.01170.

[34] V. T. Le and Y. G. Kim, "Attention-based residual autoencoder for video anomaly detection," *Appl. Intell.*, vol. 53, no. 3, pp. 3240–3254, 2023. doi: 10.1007/s10489-022-03613-1.

[35] Z. Wang, X. Cun, J. Bao, and J. Liu, "Uformer: A general U-shaped transformer for image restoration," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 17662– 17672. doi: 10.1109/CVPR52688.2022.01716.

[36] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. van Gool, "LocalViT: Bringing locality to vision transformers," arXiv preprint arXiv:2104.05707, 2021.

[37] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu and W. Wu, "Incorporating convolution designs into visual transformers," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 559–568. doi: 10.1109/ICCV48922.2021.00062.

[38] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan and M. H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 5718–5729. doi: 10.1109/CVPR52688.2022.00564.

[39] X. Chu *et al.*, "Twins: Revisiting the design of spatial attention in vision transformers," in *Adv. Neural Inf. Process. Syst.*, Virtual, 2021, vol. 34, pp. 9355–9366.

[40] X. Dong *et al.*, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 12114–12124. doi: 10.1109/CVPR52688.2022.01181.

[41] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu and B. Fu, "Shuffle transformer: Rethinking spatial shuffle for vision transformer," arXiv preprint arXiv:2106.03650, 2021.

[42] Q. Yu, Y. Xia, Y. Bai, Y. Lu, A. L. Yuille and W. Shen, "Glance-and-Gaze vision transformer," in *Adv. Neural Inf. Proce. Syst.*, Virtual, 2021, vol. 34, pp. 12992–13003.

[43] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 6185–6194. doi: 10.1109/CVPR52729.2023.00599.

[44] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang and L. Yuan, "DaViT: Dual attention vision transformers," in *Eur. Conf. Comput. Vis. (ECCV)*, Cham, Springer Nature Switzerland, Tel Aviv, Israel, 2022, pp. 74–92. doi: 10.1007/978-3-031-20053-3_5.

[45] P. Zhang *et al.*, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 2978–2988. doi: 10.1109/ICCV48922.2021.00299.

[46] H. Wu *et al.*, "CvT: Introducing convolutions to vision transformers," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 22–31. doi: 10.1109/ICCV48922.2021.00009.

[47] Y. Li *et al.*, "MViTv2: Improved multiscale vision transformers for classification and detection," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 4804–4814. doi: 10.1109/CVPR52688.2022.00476.

[48] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "ViTAE: Vision transformer advanced by exploring intrinsic inductive bias," in *Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 28522–28535.

[49] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei and Q. Tian, "Visformer: The vision-friendly transformer," in *EEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 589–598. doi: 10.1109/ICCV48922.2021.00063.

[50] S. Tang, J. Zhang, S. Zhu, and P. Tan, "Quadtree attention for vision transformers," in *Int. Conf. Learn. Represent.*, 2022.

[51] M. Ding *et al.*, "HR-NAS: Searching efficient high-resolution neural architectures with lightweight transformers," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 2981–2991. doi: 10.1109/CVPR46437.2021.00300.

[52] C. Si, W. Yu, P. Zhou, Y. Zhou, X. Wang and S. Yan, "Inception transformer," in *Adv. Neural Inf. Process. Syst.*, New Orleans, LA, USA, 2022, vol. 35, pp. 23495–23509.
[53] P. Gao, T. Ma, H. Li, Z. Lin, J. Dai and Y. Qiao, "MCMAE: Masked convolution meets masked autoencoders," in *Adv. Neural Inf. Process. Syst.*, New Orleans, LA, USA, 2022, vol. 35, pp. 35632–35644.
[54] X. Li, W. Wang, L. Yang, and J. Yang, "Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality," arXiv preprint arXiv:2205.10063, 2022.
[55] Z. Chen *et al.*, "Vision transformer adapter for dense predictions," in *The Eleventh Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, 2023.
[56] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, "VOLO: Vision outlooker for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6575–6586, 2023. doi: 10.1109/TPAMI.2022.3206108.
[57] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Comput. Computer-Assisted Interven.–MICCAI 2015: 18th Int. Conf.*, Munich, Germany, Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.
[58] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.
[59] A. Hatamizadeh *et al.*, "UNETR: Transformers for 3D medical image segmentation," in *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, 2022, pp. 1748–1758. doi: 10.1109/WACV51458.2022.00181.
[60] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins and L. Soler, "U-Net Transformer: Self and cross attention for medical image segmentation," in *Mach. Learn. Med. Imaging: 12th Int. Workshop*, Strasbourg, France, Springer, 2021, pp. 267–276. doi: 10.1007/978-3-030-87589-3_28.
[61] Y. Gao, M. Zhou, and D. N. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Medical Image Comput. Computer Assisted Interven.–MICCAI 2021: 24th Int. Conf.*, Strasbourg, France, Springer, 2021, pp. 61–71. doi: 10.1007/978-3-030-87199-4_6.
[62] Y. Gao, M. Zhou, D. Liu, and D. Metaxas, "A multi-scale transformer for medical image segmentation: Architectures, model efficiency, and benchmarks," arXiv preprint arXiv:2203.00131, 2022.
[63] H. Wang *et al.*, "Mixed transformer U-Net for medical image segmentation," in *ICASSP 2022-2022 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Singapore, 2022, pp. 2390–2394. doi: 10.1109/ICASSP43922.2022.9746172.
[64] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer," in *AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 2441–2449. doi: 10.1609/aaai.v36i3.20144.
[65] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Medical Image Comput. Comput. Assisted Interven.-MICCAI 2021: 24th Int. Conf., Proc., Part I 24*, Strasbourg, France, Springer, 2021, pp. 14–24. doi: 10.1007/978-3-030-87193-2_2.
[66] H. Cao *et al.*, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Eur. Conf. Comput. Vis. (ECCV)*, Cham, Springer Nature Switzerland, Tel Aviv, Israel, 2023, pp. 205–218. doi: 10.1007/978-3-031-25066-8_9.
[67] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Int. MICCAI Brain. Workshop*, Virtual Event, Springer International Publishing, 2022, pp. 272–284. doi: 10.1007/978-3-031-08999-2_22.
[68] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A robust volumetric transformer for accurate 3D tumor segmentation," in *Medical Image Comput. Computer-Assisted Interven.–MICCAI 2022*, Cham, Singapore, Springer Nature Switzerland, 2022, pp. 162–172. doi: 10.1007/978-3-031-16443-9_16.
[69] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 7242–7252. doi: 10.1109/ICCV48922.2021.00717.
[70] W. Zhang *et al.*, "TopFormer: Token pyramid transformer for mobile semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 12073–12083. doi: 10.1109/CVPR52688.2022.01177.

[71] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[72] J. Gu *et al.*, "Multi-scale high-resolution vision transformer for semantic segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 12084–12093. doi: 10.1109/CVPR52688.2022.01178.

[73] H. Yan, C. Zhang, and M. Wu, "Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention," arXiv preprint arXiv:2201.01615, 2022.

[74] G. Sharir, A. Noy, and L. Zelnik-Manor, "An image is worth 16 × 16 words, what is a video worth?," arXiv preprint arXiv:2103.13915, 2021.

[75] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 1728–1738. doi: 10.1109/ICCV48922.2021.00175.

[76] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual, 2021, vol. 2.

[77] H. Zhang, Y. Hao, and C. W. Ngo, "Token shift transformer for video classification," in *29th ACM Int. Conf. Multimed.*, China, Virtual Event, 2021, pp. 917–925. doi: 10.1145/3474085.3475272.

[78] Y. Zhang *et al.*, "VidTr: Video transformer without convolutions," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, vol. 696, pp. 13557–13567. doi: 10.1109/ICCV48922.2021.01332.

[79] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić and C. Schmid, "ViViT: A video vision transformer," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 6816–6826. doi: 10.1109/ICCV48922.2021.00676.

[80] A. Bulat, J. M. Perez Rua, S. Sudhakaran, B. Martinez, and G. Tzimiropoulos, "Space-time mixing attention for video transformer," in *Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 19594–19607.

[81] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *IEEE/CVF Int. Conf. Comput. Visi. (ICCV)*, Seoul, Republic of Korea, 2019, pp. 7082–7092. doi: 10.1109/ICCV.2019.00718.

[82] Z. Liu *et al.*, "ConvTransformer: A convolutional transformer network for video frame synthesis," arXiv preprint arXiv:2011.10185, 2011.

[83] Y. Wang *et al.*, "End-to-End video instance segmentation with transformers," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 8737–8746. doi: 10.1109/CVPR46437.2021.00863.

[84] S. Yang *et al.*, "Temporally efficient vision transformer for video instance segmentation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 2885–2895. doi: 10.1109/CVPR52688.2022.00290.

[85] S. Hwang, M. Heo, S. W. Oh, and S. J. Kim, "Video instance segmentation using inter-frame communication transformers," in *Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 13352–13363.

[86] S. Yan *et al.*, "Multiview transformers for video recognition," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 3323–3333. doi: 10.1109/CVPR52688.2022.00333.

[87] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *2021 IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, BC, Canada, 2021, pp. 3156–3165. doi: 10.1109/ICCVW54120.2021.00355.

[88] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," arXiv preprint arXiv:2004.05150, 2004.

[89] R. Girdhar and K. Grauman, "Anticipative video transformer," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 13485–13495. doi: 10.1109/ICCV48922.2021.01325.

[90] H. Fan *et al.*, "Multiscale vision transformers," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 6804–6815. doi: 10.1109/ICCV48922.2021.00675.

[91] W. Weng, Y. Zhang, and Z. Xiong, "Event-based video reconstruction using transformer," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 2563–2572. doi: 10.1109/ICCV48922.2021.00256.

[92] W. Zhang, M. Zhou, C. Ji, X. Sui, and J. Bai, "Cross-frame transformer-based spatiotemporal video super-resolution," *IEEE Trans. Broadcast.*, vol. 68, no. 2, pp. 359–369, 2022. doi: 10.1109/TBC.2022.3147145.

[93] Z. Geng, L. Liang, T. Ding, and I. Zharkov, "RSTT: Real-time spatial temporal transformer for space-time video super-resolution," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 17420–17430. doi: 10.1109/CVPR52688.2022.01692.

[94] R. Liu *et al.*, "Decoupled spatial-temporal transformer for video inpainting," arXiv preprint arXiv:2104.06637, 2021.

[95] M. Cao, Y. Fan, Y. Zhang, J. Wang, and Y. Yang, "VDTR: Video deblurring with transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 160–171, 2023. doi: 10.1109/TCSVT.2022.3201045.

[96] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Adv. Neural Inf Process. Syst.*, Virtual, 2020, vol. 33, pp. 6840–6851.

[97] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, 2023, pp. 4172–4182. doi: 10.1109/ICCV51070.2023.00387.

[98] R. Li, W. Li, Y. Yang, H. Wei, J. Jiang and Q. Bai, "Swinv2-Imagen: Hierarchical vision transformer diffusion models for text-to-image generation," *Neural Comput. Appl.*, vol. 8, no. 12, pp. 153113, 2023. doi: 10.1007/s00521-023-09021-x.

[99] F. Bao *et al.*, "One transformer fits all distributions in multi-modal diffusion at scale," in *Int. Conf. Mach. Learn.*, Honolulu, HI, USA, 2023, pp. 1692–1717.

[100] H. Li, F. Xu, and Z. Lin, "ET-DM: Text to image via diffusion model with efficient Transformer," *Displays*, vol. 80, no. 1, pp. 102568, 2023. doi: 10.1016/j.displa.2023.102568.

[101] J. Chen *et al.*, "PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," in *The twelfth Int. Conf. Learn. Represent.*, Vienna, Austria, 2024.

[102] J. Chen *et al.*, "PIXART-δ: Fast and controllable image generation with latent consistency models," arXiv preprint arXiv:2401.05252, 2024.

[103] S. Chai, L. Zhuang, and F. Yan, "LayoutDM: Transformer-based diffusion model for layout generation," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 18349–18358. doi: 10.1109/CVPR52729.2023.01760.

[104] H. Ali, S. Jiaming, L. Guilin, K. Jan, and V. Arash, "DiffiT: Diffusion vision transformers for image generation," arXiv preprint arXiv:2312.02139, 2023.

[105] S. Gao, P. Zhou, M. M. Cheng, and S. Yan, "Masked diffusion transformer is a strong image synthesizer," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, 2023, pp. 23107–23116. doi: 10.1109/ICCV51070.2023.02117.

[106] M. Reuss and R. Lioutikov, "Multimodal diffusion transformer for learning from play," in *2nd Workshop on Lang. Robot Learn.: Lang. Ground.*, Atlanta, Georgia, USA, 2023.

[107] G. J. Chowdary and Z. Yin, "Diffusion transformer U-Net for medical image segmentation," in *Medical Image Comput. Comput. Assisted Interven.–MICCAI 2023*, Vancouver, BC, Canada, 2023, pp. 622–631. doi: 10.1007/978-3-031-43901-8_59.

[108] Z. Zhao, X. Dong, Y. Wang, and C. Hu, "Advancing realistic precipitation nowcasting with a spatiotemporal transformer-based denoising diffusion model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024. doi: 10.1109/TGRS.2024.3355755.

[109] OpenAI, "Sora: Creating video from text," 2024. Accessed: Apr. 29, 2024. [Online]. Available: https://openai.com/sora.

[110] L. Yuan *et al.*, "Tokens-to-Token ViT: Training vision transformers from scratch on imagenet," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 538–547. doi: 10.1109/ICCV48922.2021.00060.

[111] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," in *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 9961–9970. doi: 10.1109/ICCV48922.2021.00983.

[112] L. Gao *et al.*, "STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 10990–11003, 2021. doi: 10.1109/JSTARS.2021.3119654.