Intelligent Computing
A SCIENCE PARTNER JOURNAL

REVIEW ARTICLE

Understanding Video Transformers: A Review on Key Strategies for Feature Learning and Performance Optimization

Nan Chen[1,2,3], Tie Xu[2], Mingrui Sun[3], Chenggui Yao[3*], and Dongping Yang[2*]

[1]School of Mathematical Sciences, Zhejiang Normal University, Jinhua, Zhejiang, China. [2]Research Centre for Frontier Fundamental Studies, Zhejiang Lab, Hangzhou, Zhejiang, China. [3]School of Data Science, Jiaxing University, Jiaxing, Zhejiang, China.

*Address correspondence to: yaochenggui2006@126.com (C.Y.); dpyang@zhejianglab.com (D.Y.)

Citation: Chen N, Xu T, Sun M, Yao C, Yang D. Understanding Video Transformers: A Review on Key Strategies for Feature Learning and Performance Optimization. *Intell. Comput.* 2025;4:Article 0143. https://doi.org/10.34133/icomputing.0143

Submitted 11 February 2025
Revised 24 May 2025
Accepted 26 May 2025
Published 21 June 2025

The video transformer model, a deep learning tool relying on the self-attention mechanism, is capable of efficiently capturing and processing spatiotemporal information in videos through effective spatiotemporal modeling, thereby enabling deep analysis and precise understanding of video content. It has become a focal point of academic attention. This paper first reviews the classic model architectures and notable achievements of the transformer in the domains of natural language processing (NLP) and image processing. It then explores performance enhancement strategies and video feature learning methods for the video transformer, considering 4 key dimensions: input module optimization, internal structure innovation, overall framework design, and hybrid model construction. Finally, it summarizes the latest advancements of the video transformer in cutting-edge application areas such as video classification, action recognition, video object detection, and video object segmentation. A comprehensive outlook on the future research trends and potential challenges of the video transformer is also provided as a reference for subsequent studies.

Introduction

Since the rise of deep learning, neural network models have greatly promoted the progress of artificial intelligence (AI). Among them, the transformer model [1] has revolutionized traditional deep learning structures with its unique self-attention mechanism and excellent feature extraction ability, achieving remarkable results across various fields. In natural language processing (NLP), transformers have become the dominant architecture, effectively addressing the problem of long-term dependencies, improving training efficiency, and being widely used in tasks such as machine translation and text summarization. Through pre-training techniques, large language models such as bidirectional encoder representations from transformers (BERT) [2] and the generative pre-trained transformer (GPT) series [3] have rapidly developed, providing strong support for NLP tasks.

Inspired by the success of the transformer in NLP, Google applied it to computer vision (CV) and introduced the vision transformer (ViT) [4]. It has performed exceptionally well on specific visual recognition tasks, even surpassing human performance on datasets such as ImageNet [5,6]. By utilizing the self-attention mechanism to process images, ViT provides novel solutions for tasks such as image recognition and classification, exhibiting high robustness [7]. It challenges the dominance of traditional convolutional neural networks (CNNs), relaxing the constraint of translation invariance and possessing a weaker inductive bias. Research [8] has shown that image transformers are important for constructing visual models that are closer to human perception and for understanding human visual object recognition.

However, while image-based applications have received extensive attention, the domain of video—an inherently richer and more complex visual modality—remains comparatively underexplored. Videos present unique challenges due to their spatiotemporal nature: they capture not only spatial patterns across frames but also temporal dynamics over time. This increases data dimensionality and introduces motion-specific features such as object trajectories and temporal coherence, which require specialized modeling approaches. Despite the transformer's theoretical capacity to capture long-range dependencies, its lack of built-in inductive bias and high computational cost make it ill-suited for naively handling video data. As a result, substantial research has emerged proposing targeted adaptations, including spatiotemporal factorization, temporal attention schemes, token sparsification strategies, and hybrid CNN–transformer modules. However, existing surveys often overlook these innovations or fail to organize them into a comprehensive taxonomy [9–12].

Crucially, enhancing performance and feature learning capabilities is central to making video transformers viable for real-world deployment. Tasks such as action recognition,

video summarization, and event detection demand models that can simultaneously extract fine-grained motion cues and maintain global contextual understanding. Dual-branch architectures, for example, capture static (spatial) and dynamic (temporal) features in parallel, improve the model's representational richness. Furthermore, rapidly evolving application scenarios—such as real-time sports analytics, autonomous driving, and multi-view scene understanding—require lightweight, efficient, and high-performing video models. Yet, these use cases are often underrepresented in existing literature, creating a gap between current research and practical deployment needs [9–12].

Here, we bridge this gap by offering the following: (a) a comprehensive taxonomy of architectural innovations tailored to video data, clearly mapping to the dual challenges of dimensionality explosion and motion-dynamics modeling; (b) rigorous comparative analysis of how these modifications affect computational cost, feature-learning capacity, and end-to-end performance across diverse benchmark tasks; and (c) application-driven insights, spotlighting underexplored deployment scenarios—from safety-critical autonomous driving to on-device video generation—and identifying performance bottlenecks and promising avenues for lightweight, application-oriented, and robust model design. By synthesizing these perspectives and focusing on the dual goals of improving performance and enhancing feature learning, this survey not only charts the current landscape of video transformer research but also illuminates pathways toward more efficient, robust, and practically deployable spatiotemporal models.

Transformers

In 2017, Vaswani et al. introduced the transformer [1] in the NLP field. With its unique architecture and excellent performance, the model quickly became the focus of industry attention. The core architecture of the transformer includes an input module, an encoder–decoder module, and an output module. Among these, the crucial encoder–decoder module allows the model to mine and understand the underlying semantics of the text. The encoder involves 2 sub-layers: the multi-head self-attention (MHSA) layer and the feed-forward neural network (FFN) layer. Each integrates with residual connections and layer normalization techniques to improve the fluency of information transmission and the trainability of the model. The decoder uses a similar structure and incorporates masked multi-head attention, ensuring the model does not "peep" into future information during prediction.

Positional embedding

Transformer architecture relies entirely on the self-attention mechanism but lacks word position information. Positional embedding applies with the commonly used sine and cosine functions, given as the following formula:

$$PE(pos,i)=\begin{cases}\sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right), & i=2k,\\ \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right), & i=2k+1,\end{cases} \tag{1}$$

where pos is the absolute position of the word in the sentence, i is the index value of this vector, and d is the length of the encoding vector (the same as the embedding vector).

Self-attention

The attention mechanism is inspired by the way humans process external information, involving both autonomous and non-autonomous cues. Autonomous cues are guided by prior knowledge to direct attention, rather than being triggered by the salient features of the object. It is implemented using 3 key elements: query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$), where $\boldsymbol{Q}$ represents self-prompting, reflecting the information the model wants to focus on; $\boldsymbol{K}$ represents non-autonomous cues, reflecting the salient features of the object; and $\boldsymbol{V}$ is the specific feature information of the object.

The self-attention mechanism improves the attention mechanism by ensuring that $\boldsymbol{Q}$ and $\boldsymbol{K}$ originate from the same source. In the self-attention layer, 3 learnable weight matrices are defined: $\boldsymbol{W^Q}$, $\boldsymbol{W^K}$, and $\boldsymbol{W^V}$. The input sequence $\boldsymbol{X}$ generates the corresponding $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ through the projection of these weight matrices:

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W^Q}, \tag{2}$$

$$\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W^K}, \tag{3}$$

$$\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W^V}. \tag{4}$$

Next, the self-attention calculation is performed using the generated ($\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$) triples. The similarity score between $\boldsymbol{Q}$ and $\boldsymbol{K}$ is first calculated, reflecting the degree of correlation between different input elements. These similarity scores are then normalized, typically using the softmax function, ensuring that the sum of all scores is 1 and yielding the attention weight. Finally, the attention weight is applied on $\boldsymbol{V}$ and then summed to obtain the scaled dot-product self-attention:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}}}}\right)\boldsymbol{V}, \tag{5}$$

where $d_{\boldsymbol{K}}$ is the dimension size of the matrix $\boldsymbol{X}$.

The multi-head attention mechanism is an extension of the single-head mechanism. The input sequence $\boldsymbol{X}$ is divided into h groups along the channel dimension (i.e., h heads), with each group representing different patterns of attention. The results are then concatenated and transformed via a linear transformation to produce the output. This mechanism enables the model to capture a diverse set of features in the input sequence, thereby improving its representational ability and generalization performance, summarized as follows:

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}\left(head_1,head_2,\cdots,head_h\right)\boldsymbol{W^O}, \tag{6}$$

where h represents the number of heads, $head_i$ represents the output of the ith head, and $\boldsymbol{W^O}$ is the output transformation matrix. Each head is given by:

$$head_i=\text{Attention}\left(\boldsymbol{X}\boldsymbol{W_i^Q},\boldsymbol{X}\boldsymbol{W_i^K},\boldsymbol{X}\boldsymbol{W_i^V}\right), \tag{7}$$

where $\boldsymbol{Wi^Q}$, $\boldsymbol{Wi^K}$, and $\boldsymbol{Wi^V}$ are the $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ transformation matrices of the ith header.

Feed-forward neural network

The FFN consists of 2 linear layers and a rectified linear unit (ReLU) nonlinear activation layer. The first linear layer maps the input to a high-dimensional space, followed by the ReLU

layer, which processes the filtered information. The second linear layer converts the data to generate a rich feature representation. This structure allows the FFN to effectively extract and integrate key information from the input sequence, facilitating better understanding and processing of complex data, given as:

$$\mathrm{FFN}(x) = \max\left(0, x\boldsymbol{W}_1 + b_1\right)\boldsymbol{W}_2 + b_2. \quad (8)$$

Add and norm

Layer normalization (LN) is an important regularization technique that can improve the training efficiency and generalization ability of the model. In the transformer, LN before each layer reduces differences in data distribution, aids feature learning, and enhances model robustness. Combined with residual connections, LN also improves gradient stability, preventing issues like vanishing or exploding gradients.

Innovations in Image Processing

ViT [4], proposed by Dosovitskiy et al., represents a major breakthrough in the field of CV through the use of transformers, especially in the task of image classification. ViT primarily consists of embedding, encoder, and classification modules. Among these, the embedding module is the core innovation for image-based tasks. This design enables the model to process image data more effectively by reshaping the input 2-dimensional (2D) image data $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$ into a flattened image block sequence $\boldsymbol{X}_p \in \mathbb{R}^{N\times(p^2\times C)}$ with length $N = HW/p^2$, where C is the number of channels and p is the size of the image block.

Building on the success of BERT [2] in the field of NLP, the ViT [4] drew inspiration from BERT's architectural innovations by introducing a classification token and pre-training strategies. Specifically, in ViT, the first token of every input sequence is designated as the class token. The final hidden state corresponding to this token is then utilized as the aggregated sequence representation for downstream classification tasks. The key characteristics and distinctions between the ViT and BERT are summarized in Table 1.

The success of ViT lies not only in its compelling demonstration that standard transformer encoder architectures alone can achieve comparable or superior performance to CNNs in CV tasks, but also in its role as a bridge between the CV and NLP domains. Furthermore, a subsequent study [13] has further confirmed that the image transformer possesses stronger robustness and shape recognition capabilities, akin to human perception, when combined with distillation techniques.

Innovations in Video Processing

To fully harness the advantages of transformers in video processing, it is crucial to address the high-dimensional data challenges and the unique issues posed by video tasks when transitioning from NLP to image and, subsequently, video domains. As such, video transformers require innovations in their overall architecture.

Innovations in the input module

In video transformers, the input module serves as a critical bridge that connects raw data to the model's thematic structure. It not only converts video data into a format suitable for the model but also substantially improves its ability to capture and understand video information.

Tokenization and embedding

The input module first performs tokenization, dividing the continuous video stream into discrete tokens. Following tokenization, embedding extracts key features from each token and maps them into a high-dimensional space to enhance the model's understanding. Embedding is commonly implemented using linear layers [14,15], a few convolutional layers [16–18], or a complete CNN [19–22]. In specific tasks [23,24], convolutional layers or CNN embedding operations directly achieved tokenization, simplifying the process and reducing parameters while improving generalization due to convolution's local perception and weight sharing.

Video transformers often draw inspiration from ViT [4] but are expanded to accommodate the high-dimensional nature of video data. Selva et al. [11] have identified that segmentation strategies can be categorized by the spatiotemporal coverage of tokens, ranging from fine-grained to coarse-grained, patch tokenization, instance tokenization, frame tokenization, and clip tokenization. As detailed in Table 2, each tokenization method is characterized by distinct implementation mechanisms and properties.

Table 1. Comparison of architectures and characteristics between ViT and BERT. In the ViT, an input image is partitioned into fixed-size patches (e.g., 16 × 16 pixels), which are then linearly projected into token embeddings. After concatenating 1D positional embeddings to these patch tokens, the resulting sequence serves as the input to the model. In contrast, BERT processes sequences of single or paired sentences tokenized using WordPiece embeddings. To these token sequences, BERT appends segment embeddings and positional embeddings. Segment embeddings distinguish whether a token belongs to sentence A or sentence B.

Model				Characteristics	
Model name	Input structure	Attention scope	Pre-training tasks	Core strengths	Application tasks
ViT [4]	Token embeddings + position embeddings	Global spatial attention	Supervised classification/ self-supervised	Long-range dependency modeling, multi-task scalability	Image classification, segmentation, object detection
BERT [2]	Token embeddings + segment embeddings + position embeddings	Bidirectional contextual attention	Masked language modeling	Context-sensitivity, cross-task generalization	Text understanding, generation, cross-modal alignment

Table 2. Comparison of four tokenization methods. When patch tokenization is employed, the model processes single-frame images by partitioning them into multiple 2D patches [21,25], which are then sequentially arranged in temporal order to form a frame-level input sequence. For entire video sequences, the model may instead segment the input into 3D patches [14,16,17] spanning multiple frames, enabling simultaneous capture of spatial and temporal information.

Tokenization			Characteristics		
Tokenization name	Division basis	Implementation approach	Advantages	Disadvantages	Typical application scenarios
Patch	2D/3D image patches	Direct splitting of raw pixels or generation via CNN feature maps	Fine-grained spatiotemporal modeling, high flexibility	High computational complexity, strong redundancy	Action localization, medical image analysis
Instance	Semantic instances (objects/regions)	Generation of instance embeddings via region proposal networks or hybridized with coarse-grained tokens	Reduces redundancy, semantically guided	Relies on detection models, ignores background	Visual question answering, action recognition
Frame	Single-frame global features	Encoding each frame as a single token	Efficient temporal modeling, low computational cost	Prone to losing spatial details, dependent on feature extractors	Video summarization, sentiment recognition
Clip	Multi-frame clips	Extraction of clip-level features using networks	Long-range dependency modeling, extremely low computational cost	Prone to information mixing, irreversibility	Long-video retrieval, global video classification

In patch tokenization strategies, 2 key segmentation details warrant attention: the size of patches and whether overlaps exist between adjacent patches. Generally, fixed-size segmentation of patches is adopted [14,21,25], which is simple and efficient, but may be limited in complex tasks such as video tracking and video super-resolution. To this end, the multi-size tokenization strategy emerged. By using methods such as direct segmentation [26,27] or convolution of different sizes [17,18,28,29] to dynamically adjust patch sizes, enabling more comprehensive capture of video features, the segmentation size is dynamically adjusted to capture video features more comprehensively. As for the issue of whether adjacent patches overlap, non-overlapping segmentation [14,21,25,30] is simple and straightforward, but it may ignore the boundary information. Studies [31] have shown that moderate overlap can retain important information. Thus, spatio-temporal pyramid transformer (STPT) [17], multiscale vision transformer (MViT) [32], and FuseFormer [33] adopt overlapping segmentation to make feature propagation between adjacent patches more effective.

Positional embedding

Positional embedding represents positional information by assigning a distinct vector to each position in the sequence, enabling the model to capture temporal dependencies between tokens, which are crucial for understanding context and dynamic changes within the sequence.

Inspired by NLP methodologies, several studies (e.g., Refs. [24,34]) encoded the absolute positions of input tokens as vectors, which are either summed or concatenated with input embeddings. While this strategy is intuitive and straightforward, its effectiveness in CV tasks remains empirically contested. To address this limitation, other studies (e.g., Refs. [35,36]) adopt relative positional concepts from image processing, learning pairwise relationships, and relative positions between input elements. These relative positional biases are typically incorporated into attention mechanisms as additive terms, critical for modeling structural dependencies within sequences.

Positional embedding can be categorized into pre-computed and dynamic types. Pre-computed positional embeddings [20,37] are derived and stored beforehand using mathematical formulas (e.g., sine and cosine functions); these embeddings offer high computational efficiency and perform well in resource-constrained scenarios. However, they lack adaptability to variable-length sequences. Dynamic positional embeddings [17,38] are generated on-the-fly during training or inference via deep convolutional networks, and they exhibit superior flexibility and adaptability. Despite their higher computational cost and reliance on large-scale training data, they excel in handling complex, dynamic sequences.

Innovations in the internal structure

The internal structure of video transformers primarily involves 2 key components: the attention mechanism and the FFN. Design regarding these components substantially impacts the model's ability to process data and capture complex dynamics. Additionally, improvements to these structures can reduce computational cost, increase efficiency, and enhance generalization.

Attention mechanism

Studies of video transformers [39,40] have designed diverse attention mechanisms to address the computational challenges

posed by high-dimensional data processing, aiming to reduce computational load, optimize performance, and enhance flexibility and efficiency.

Attention mechanisms can be categorized into local (Fig. 1A), axial (Fig. 1B), and sparse (Fig. 1C) attention based on attention regions. Local attention models [14,35,36] focus on local regions of the input sequence to reduce computational cost, where "local" refers to a neighborhood around the query [14] or a local window [35,36]. Axial attention models [25,40,41] restrict attention operations to specific axes, enabling independent processing of spatial or temporal dimensions. Sparse attention models [21,24,40] limit each query to compute only with a subset of keys, markedly reducing computational costs.

Depending on the order of spatial and temporal feature extraction, attention mechanisms in video transformers can be classified into 2 types: spatiotemporal sequential attention (Fig. 2A) and spatiotemporal parallel attention (Fig. 2B). The former extracts spatial and temporal features sequentially. Most studies [23,39,42,43] processed spatial features first, followed by temporal features, but others [25,44] reversed this order. The latter uses 2 independent attention modules to process spatial and temporal information separately. The information is then combined using methods such as weighted summation [45], concatenation [46,47], Hadamard product [47], or support vector machines [48] to form a unified spatiotemporal representation.

Furthermore, cross-attention and multi-scale attention have gained substantial attention in recent research. Cross-attention enables the fusion of features from different sources or modalities, thereby enhancing the model's ability to analyze complex video content. It can fuse multi-modal features [49–51] or local and global features [52,53] and can integrate query (current frame) and memory information [37,50,54,55]. Multi-scale attention captures features at different scales, which is discussed in detail in the "Multi-scale information extraction" section.

It is also worth noting that Liu et al. [16] and Ragini et al. [56] proposed multi-head convolutional self-attention mechanisms, where linear mappings are replaced with convolutional mappings. This allows for modeling long-range spatial and temporal dependencies in video frame sequences.

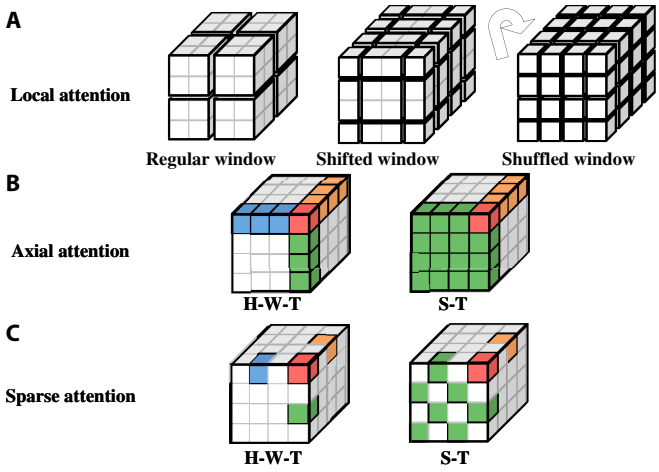

Fig. 1. Self-attention based on the scope of operation. (A) Regular window divides the input sequence into fixed-size, non-overlapping subsequences. Shifted window enables interaction between adjacent windows by shifting the window's position across different layers, while shuffled window randomly or in a specific manner shuffles the order or content of the windows to increase the model's focus on different parts of the sequence. In (B) and (C), H–W–T represents attention operations based on the height, width, and time length of the video data, while S–T represents attention operations based on the spatial and temporal dimensions of the video data.

Feed-forward neural network

The FFN is a crucial component of both encoder and decoder in video transformers. Building on the self-attention mechanism, the FFN refines the features and enhances the model's expressive power. Thus, improving the FFN is critical in the context of video transformers, as it substantially influences model performance.

Currently, one notable trend is the use of convolution to improve the FFN. For instance, Wang et al. [57] proposed a convolutional FFN consisting of one convolutional layer and 2 fully connected layers, along with a Gaussian error linear units (GELU) nonlinear activation function. Li et al. [58] replaced the fully connected layer with a step volume, while Zeng et al. [59] completely substituted the FFN with a 2-layer convolution. Additionally, DeTformer [56] and HiSViT [60] not only used convolution but also incorporated a gating mechanism to enhance the locality of the FFN. Specifically, FuseFormer [33] introduced soft composition and soft split into the FFN to enable one-dimensional linear layers to model 2D structures.

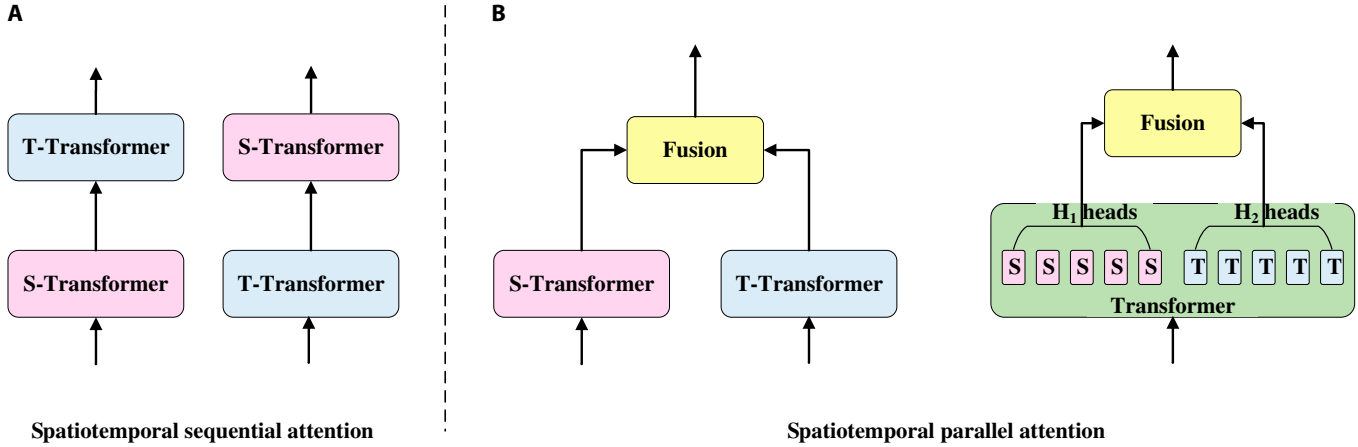

Fig. 2. Self-attention mechanism based on the order of operations. (A) Spatiotemporal sequential attention. (B) Spatiotemporal parallel attention. In (A) and (B), S represents the spatial dimension, and T represents the temporal dimension.

Innovations in model frameworks

The design of the overall model framework plays a critical role in enhancing the flexibility and efficiency of video transformers. These innovations help models process spatiotemporal information, enrich their ability to represent video content, and facilitate the effective fusion and generation of cross-modal information. The following subsections discuss the recent innovations in model frameworks and their limitations.

Encoder–decoder architecture

The encoder–decoder architecture stands as a cornerstone paradigm in deep learning models, designed to model complex mappings by segregating feature encoding from decoding-generation processes. Its exemplary deployment in transformers has enabled these models to exhibit robust modeling capabilities in NLP tasks, effectively addressing sequence-to-sequence challenges. DETR [61] and subsequent works [62] also achieved remarkable outcomes in image generation through this architecture. Furthermore, the encoder–decoder framework has been widely adopted in video segmentation [20,63] and multi-modal video tasks [64,65]. Particularly in vision-language tasks, encoders can independently process images and text, while decoders achieve semantic alignment through cross-modal attention, effectively capturing image-text associations. However, for specific tasks, transformers may employ encoder-only [21,32,39,40] or decoder-only [66,67] architectures.

Encoder-only structures are frequently utilized in large language models and ViTs. By extracting global or local representations from input sequences, encoders generate fixed-dimensional features suitable for tasks like video classification and action recognition, which do not require generating new videos or texts. For instance, ViViT [39] and TimeSformer [40] leveraged encoder-only architectures with diverse spatiotemporal attention mechanisms to simultaneously capture spatial semantics and temporal dynamics, avoiding temporal redundancy introduced by decoders. Similarly, MViT [32] and STAR++ [21] relied on encoders to progressively fuse multi-scale spatiotemporal features, directly serving classification tasks without decoder-based feature reconstruction.

Decoder-only structures are commonly employed in generative tasks, accommodating outputs of unpredictable lengths such as video captioning and frame prediction. GPT [3] pioneered this architecture in NLP. Later, in video processing, to circumvent feature transfer issues between encoders and decoders, Kondratyuk et al. [68], Gupta et al. [69], and Miech et al. [66] adopted decoder-only architectures for multimodal generation tasks. Among these, the former 2 [68,69] integrated multimodal conditions (e.g., text descriptions and reference images) through cross-attention mechanisms within their decoders, enabling end-to-end generation. Conversely, Miech et al. [66] employed a dual-stream decoder architecture, enhancing cross-modal matching accuracy via a slow stream while ensuring real-time performance through a fast stream. Tan et al. [67] tailored DETR [61] for action proposal generation, utilizing a decoder to directly predict action boundaries and categories, eliminating the need of encoder-generated anchor boxes or candidate regions. For few-shot video segmentation, Siam et al. [29] employed a decoder to directly receive features from query and support frames, dynamically computing their similarity through self-attention mechanisms, obviating encoder preprocessing.

While decoder-only architectures excel in generative tasks with dynamic modeling advantages, they exhibit limitations in encoding contextual and positional information flexibly. Conversely, encoder-only structures demonstrate superiority in feature extraction and long-sequence processing but face constraints in generative tasks and scalability.

Dual-branch architecture

The Video Mobile-former [53] exemplified the trend toward efficiency-oriented dual-branch designs. Its mobile branch employed lightweight convolutional networks to extract local spatiotemporal features, while the former branch leveraged global self-attention to capture long-range dependencies. This architecture integrates lightweight networks with model compression techniques, substantially enhancing computational efficiency and scalability. Consequently, dual-branch transformer models are increasingly adopted in real-time video analytics and edge computing, where resource constraints demand both speed and precision. A core advantage of the dual-branch architecture is its ability to decompose tasks across 2 independent pathways. This division of labor helps mitigate feature entanglement, improves signal disentanglement, and allows for targeted resource allocation, making it highly adaptable for uni-modal and multi-modal video tasks. Depending on the information they process, uni-modal dual-branch architectures typically fall into 2 categories: static–dynamic (spatial–temporal) and local–global dual-branch architectures.

The static–dynamic dual-branch architecture is designed to disentangle static (spatial) and dynamic (temporal) information within videos. Inspired by the inherent properties of video data, this architecture employs a spatial branch to extract content from individual frames and a temporal branch to capture inter-frame motion dynamics, and thus has been widely applied to action segmentation [70] and action recognition [71–73]. Both Tu et al. [48] and Zhou et al. [74] applied this approach in face forgery detection, where subtle spatial–temporal inconsistencies must be identified. SlowFastFormer [75] applies a fast–slow branch design in skeleton-based action analysis, where the slow branch captures skeletal context and global motion trends from low-frame-rate inputs, and the fast branch captures fine-grained joint dynamics from high-frame-rate sequences. In essence, the fast–slow dual-branch architecture constitutes a specialized subtype of the static–dynamic dual-branch framework, wherein the fast branch aligns with dynamic information (characterized by high-frequency temporal variations) and the slow branch corresponds to static information (manifesting low-frequency spatial stability). Consequently, this architecture can be conceptualized as a temporal-domain instantiation of the static–dynamic paradigm, explicitly operationalizing the decoupling and hierarchical integration of motion-centric and context-aware features through frequency-modulated sampling strategies. This architecture effectively isolates static content from motion dynamics and is well-suited for tasks that rely on subtle temporal changes, e.g., forgery detection and action recognition. However, rigid separation of spatial and temporal processing may fail to model interdependencies effectively. It requires careful synchronization and fusion strategies to avoid information loss and redundancy during integration.

The local–global dual-branch architecture is designed to capture localized details and global contextual information simultaneously in video processing. The local branch focuses on fine-grained regions within video frames or clips to extract detailed features, while the global branch models the overall structural semantics and scene context. These branches typically

employ distinct neural network architectures. For example, ACTNet [76] utilized a CNN branch to capture local facial features from individual frames and a transformer branch to model long-term temporal dependencies across frames for global context. ViXNet [77] reversed this paradigm: a deep CNN generates global spatial features, while a transformer branch learns discriminative patterns in localized facial regions. Temporal motion and spatial enhanced appearance with transformer-based framework (T^2MEA) [78] introduced a content branch to extract the holistic video structure from a global perspective and a fovea branch to acquire localized fine-grained spatiotemporal features. This architecture balances detail sensitivity with holistic understanding, enabling scene-aware reasoning while preserving local discriminative power. However, dual processing can be computationally expensive, and fusing local and global features effectively remains a non-trivial design challenge, especially in long videos. Furthermore, it still lacks an effective strategy to keep the balance.

Furthermore, in video segmentation tasks, models often enhance the transformer's capability to process video data through external memory modules. For instance, Liang et al. [54] proposed a dual-branch framework: one branch employs a hierarchical transformer to extract high-level semantic features from key frames, improving segmentation accuracy, while the other branch utilizes a lightweight feature network to capture low-level features from non-key frames, enhancing segmentation efficiency. A dynamically updated memory matrix is introduced to store critical semantic information from historical frames, enabling cross-frame consistency. Cheng et al. [37] designed dual-branch memory mechanisms, pixel memory and object memory, that interact via bottom-up hierarchical attention. This design mitigated the impact of distractors or noise by decoupling fine-grained pixel details from object-centric semantics. These architectures enhance cross-frame consistency and long-term dependency modeling, and reduce redundant computation by reusing historical information. However, memory updates and maintenance add complexity and latency.

In multi-modal video tasks, several studies (e.g., Refs. [42,50,79]) processed data from different modalities (such as visual, audio, and textual) as inputs to separate branches. This approach prevents mutual interference among low-level features during fusion and enables parallel processing of heterogeneous data, thereby improving computational efficiency. After extracting the relevant features, the dual-branch architecture combines them through methods such as addition, concatenation, and cross-attention [53,54,70]. For example, Actor-T [34] proposed 2 fusion points: before entering the transformer and after classifier prediction. It preserves modality-specific cues while allowing flexible fusion strategies, promotes parallel processing, and reduces feature-level noise from early fusion. However, late fusion may miss early cross-modal interactions that are crucial for some tasks. Synchronizing temporal resolution across modalities requires alignment techniques, increasing model complexity with separate encoding pipelines.

Overall, dual-branch architectures represent a powerful design paradigm for video transformers, enabling efficient and modular modeling of complex video data. However, they are not without trade-offs: the design of fusion strategies, management of resource overhead, and handling of inter-branch dependencies are all critical to their success.

Multi-scale information extraction

For tasks involving complex scene analysis and video understanding, multi-scale information extraction is essential. It allows models to capture features at various levels of granularity, offering a more comprehensive understanding of video content. To achieve this, video transformer models often adopt strategies to form a feature pyramid structure [80], as shown in Fig. 3. In the input module (Fig. 4A), several studies (e.g., Refs. [28,29,63,79,81]) divided the input video into representations at different scales through the backbone or embedding network. Specifically, temporal pyramid transformer (TPT) [27] segmented the video into temporal patches of different lengths, as shown in Fig. 4B. In the intermediate module (Fig. 5), several studies (e.g., Refs. [17,18]) performed down-sampling through the patch embedding module, ensuring that each transformer layer learns information from different scales, thereby enhancing multi-scale processing.

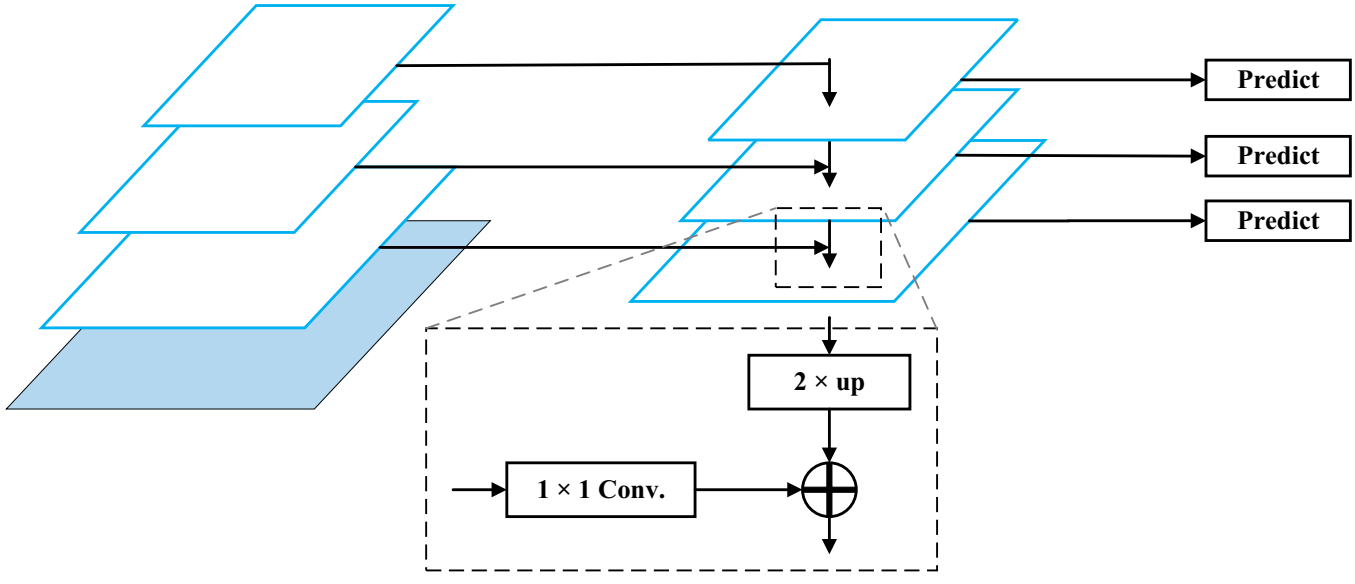

Fig. 3. Feature pyramid network (FPN). Reprinted with permission from Ref. [80]. Copyright by the Computer Vision Foundation.

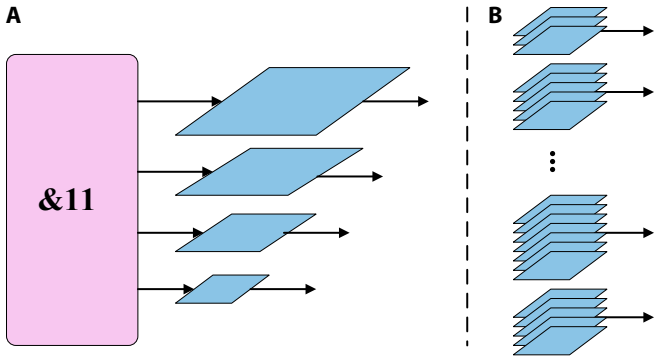

Fig. 4. Multi-scale partitioning of input modules. (A) The model utilizes a backbone network to segment video data into multi-scale subsets in terms of height, width, and temporal length. (B) The model does not use a backbone network and instead directly samples the data, segmenting it into subsets of different temporal lengths.

Various attention mechanisms have been specifically designed to enhance multi-scale feature extraction within transformer-based video models. These mechanisms aim to effectively capture both fine-grained local patterns and broader contextual dependencies across spatial and temporal dimensions. For example, a transformer network with a set of constrained self-attention operations in pyramid structures (PCSA) [19] and Video Swin [36] introduced multi-scale window attention, capturing information at different scales by adjusting the size of local windows. This approach enhances the model's ability to process hierarchical spatial information while maintaining computational efficiency through localized attention. A joint spatial-temporal transformer network (STTN) [26] restricted attention to constrain spatial dimensions to focus on multi-scale feature learning, improving performance in spatially dense scenes, while MViT [32] and subsequent studies [82,83] employed multi-head pooling attention, further enhancing feature diversity and translation invariance while reducing computational costs overhead by replacing token-heavy self-attention with hierarchically pooled representations. MViTv2 [82] and MeMViT [83] utilized multi-scale representations on the encoder side, enabling efficient hierarchical abstraction. A meta-learned multiscale memory comparator (MMC) [29] integrated multi-scale attention into the decoder, enabling more precise reconstruction and feature refinement at multiple resolutions. MED-VT [63] leveraged multi-scale attention throughout the entire encoder–decoder pipeline, allowing consistent feature resolution adaptation across both encoding and decoding stages, thereby supporting more effective representation learning across tasks like segmentation or temporal localization.

However, selecting optimal window sizes or pooling strategies often requires empirical tuning and task-specific adaptation. In models with independently scaled branches (e.g., PCSA and MViT), inconsistent feature alignment across scales can lead to fusion bottlenecks or degraded accuracy. While these approaches effectively address spatial multi-scale modeling, they often underexploit temporal scale diversity, limiting performance on fast vs. slow motion dynamics. Furthermore, when applied to the decoder (e.g., MMC), multi-scale processing may induce latency and memory overhead, particularly in real-time applications.

Innovations in hybrid models

Hybrid models offer a balanced technical pathway for various video understanding tasks, e.g., by synergizing CNN's localized feature extraction with transformer's global contextual awareness. They typically demonstrate superior computational efficiency, improved generalizability, and multi-task adaptability, enabling flexible deployment across heterogeneous scenarios. Nevertheless, ongoing research continues to explore transformer integrations with other deep learning methodologies. In the following subsections, we systematically analyze the integration of transformers with CNNs, U-Nets, and graph neural networks (GNNs), highlighting their architectural evolution, technical benefits, and inherent limitations.

CNNs and transformers

The fusion of CNNs with video transformers has become a cornerstone in video understanding. CNNs excel at extracting localized and hierarchical features, while transformers offer superior capabilities in modeling global spatiotemporal dependencies. The hybridzation of the two, as highlighted by Djenouri and Belbachir [84], has led to a proliferation of architectures [53,76,85] that benefit from both computational efficiency and enhanced representational richness. Three main architectural paradigms exist: (a) CNN-as-backbone: CNNs extract low- and mid-level features that are passed to transformer modules for global reasoning [20,23,28,29,81,86]; (b) parallel branches: CNN and transformer

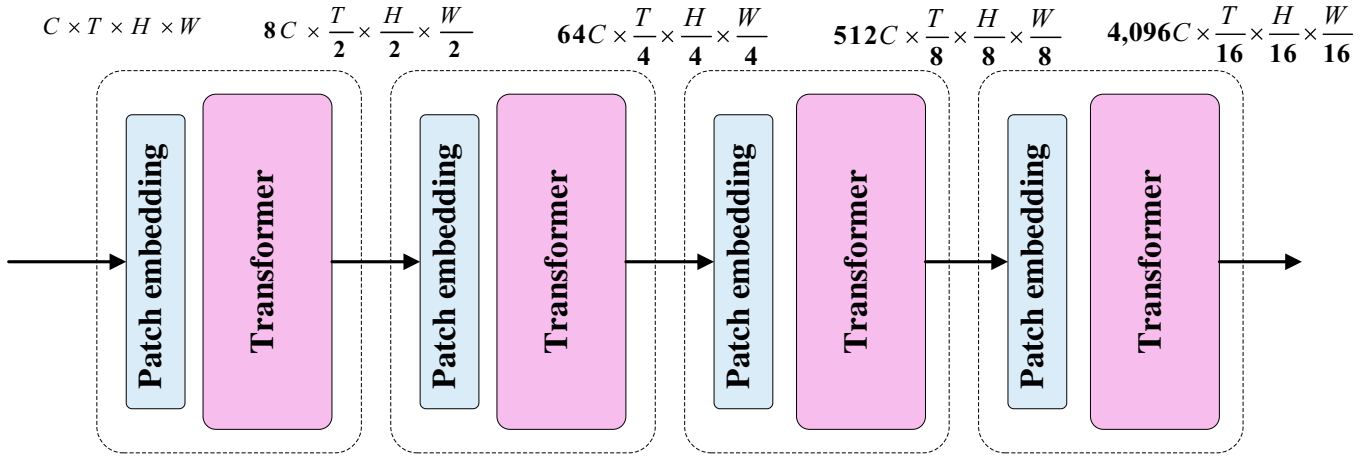

Fig. 5. Multi-scale partitioning of intermediate modules. *C* represents the channel dimension, *T* represents the time dimension, while *H* represents height, and *W* represents width.

modules operate concurrently to capture complementary spatiotemporal features [53,54,76]; and (c) encoder–decoder hybrids: works like HSTforU [87], Swin-VEC [88], Rangevit [89], and Fast-BEV [90] utilize CNNs either for encoding fine-grained spatial information or for decoding high-level transformer representations. Moreover, a recent work by Ahmadabadi et al. [91] applied knowledge distillation techniques to optimize CNN–transformer hybrids, facilitating efficient knowledge transfer from complex models to lightweight variants while enhancing operational efficiency.

However, integration complexity can lead to suboptimal information fusion: (a) transformers may overshadow CNN outputs in late-fusion setups, leading to underutilization of fine spatial cues; (b) inconsistent feature scales between CNN and transformer branches can hinder alignment and joint optimization; and (c) dual-pipeline systems may increase inference latency and memory usage, particularly in real-time systems.

U-Nets and transformers

Originally designed for medical image segmentation, U-Nets offer strong localization capabilities through encoder–decoder symmetry and skip connections. When integrated with transformers, the resulting hybrids maintain spatial precision while enhancing global context modeling, thereby improving performance in complex video processing tasks. This approach has gained traction in medical imaging and anomaly detection tasks. In medical processing, TT U-Net [92] employed temporal transformer layers to capture dynamic cardiac motion patterns while leveraging U-Net's spatial segmentation capability. This effectively reduces motion artifacts in cardiac CT images, enhancing segmentation accuracy and reliability. In video anomaly detection, models such as TransAnomaly [44] and CViT [93] combined U-Net with transformer models to improve anomaly localization and precision. In comparison with TransAnomaly, CViT [93] extracted richer features from RGB frames by stacking convolutional layers and transformer modules, demonstrating superior performance in appearance anomaly detection. In other fields, SeTHPose [94] extracted visual features from hand images, employed a transformer to learn contextual relationships for 2D joint estimation, and then refined them into 3D poses via a U-Net-based graph convolutional network (GCN). Transframer [95] leveraged generative modeling for arbitrary frame prediction in video sequences, while U-Transformer [96] designed a pure transformer model with U-Net-inspired architecture, improving action segmentation efficiency and precision while reducing model complexity.

Although this hybrid brings excellent spatial detail preservation and flexibility across tasks such as segmentation, anomaly detection, and prediction, limitations persist in behavioral anomaly detection, suggesting future development of real-time anomaly activity detection systems for multi-camera complex scenarios to enhance model robustness. Berroukham et al. [97] proposed leveraging the attention mechanism unique to ViT to effectively capture spatial features in video frames for frame-level anomaly classification, followed by U-Net integration for precise anomaly localization. Nevertheless, these models lack explicit temporal information modeling modules to improve dynamic anomaly detection capabilities.

GNNs and transformers

GNNs offer a natural representation for relational structures such as skeleton joints or object interactions, making them highly suitable for video understanding when fused with transformers. The integration aims to unify GNNs' structural reasoning with transformers' sequential modeling capabilities. For skeleton-based action recognition, a novel hybrid dual-branch network (HDBN) [72] innovatively integrated both transformer and GCNs to model 2D and 3D skeletal data, respectively. Compared to conventional single-backbone architectures, HDBN substantially enhanced robustness for action recognition, improved discriminative capability for complex actions, and maintained stable performance under challenging scenarios such as occlusions. For sign language and gesture recognition, Tunga et al. [98] proposed combining GCNs with BERT [2], utilizing the former for spatial modeling and the latter for temporal modeling. However, in terms of spatiotemporal feature fusion, the study still has room for further optimization. The relation-enhanced spatial–temporal hierarchical transformer (RESTHT) [99] integrated transformers with GCNs for spatial modeling while relying solely on transformers for temporal modeling, capturing key video information from multiple dimensions. Other advanced applications, for example, TransMOT [100], employed graph-structured representations to model video objects and their motion trajectories, using graph attention mechanisms to capture dynamic interactions between objects. ViGAT [101] applied graph attention networks (GANs) to video event recognition and explanation, achieving comprehensive coverage from low-level features to high-level interpretation. This framework demonstrates superior event recognition performance while providing complete explanations for classifier decisions. However, compared to efficient top-down approaches, ViGAT incurs substantially higher computational costs in terms of memory consumption and inference time due to imperfections in its object detector.

Applications of Transformers in Video

As illustrated in Fig. 6, video understanding tasks can be categorized into video level, frame level, and pixel level based on their granularity of reasoning. The transformer architecture has demonstrated remarkable advancements across video tasks at all granularity levels. This section focuses on video classification and action recognition, object detection and tracking, and video object and semantic segmentation, introducing the state-of-the-art techniques in these domains. Additionally, we emphasize the latest developments in emerging research areas such as autonomous driving, deepfake detection, and video generation, while analyzing their technical innovations and practical implications.

Video action recognition and classification

The application of transformers to video action recognition and classification builds upon foundational architectures originally

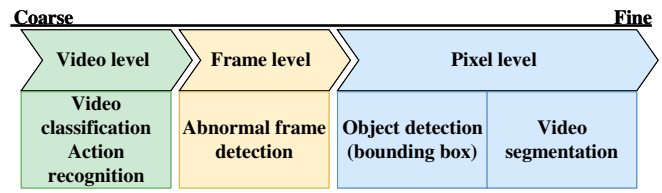

Fig. 6. Video understanding tasks. Video-level inference focuses on global semantic comprehension of the entire video sequence. Frame-level inference involves independent analysis of individual frames, treating each as a discrete unit for localized interpretation. Pixel-level inference addresses fine-grained spatial parsing, comprising 2 hierarchical sub-levels: bounding-box-level detection for region-of-interest localization and per-pixel-level segmentation for dense semantic annotation.

developed for image tasks. Early adaptations, such as TimeSformer [40], ViViT [39], and MViT [32], extended the self-attention mechanism to capture spatiotemporal dynamics, thus aligning with the unique requirements of video data. Later models, including Video Swin [36] and follow-ups [102], introduced window-based attention schemes derived from the Swin Transformer [103], enhancing computational efficiency and scalability.

For video classification, TimeSformer [40] adopted a pure attention mechanism with a "divided attention" strategy, applying temporal and spatial attention separately. It achieved 82.2% and 80.7% top 1 accuracy on Kinetics-400 and Kinetics-600 [104], respectively, demonstrating the effectiveness of decoupled spatiotemporal attention. However, it lacks hierarchical feature learning, limiting its representation of complex spatiotemporal patterns, and thus may lead to less efficiency on long video sequences or fine-grained actions. ViViT [39] introduced efficient model variants to handle long video sequences by factorizing input dimensions and leveraging pretrained image models, enabling effective application on smaller datasets. Although it demonstrates superior performance over conventional 3D CNNs, achieving 80.0% top 1 accuracy on Kinetics-400 [104], it still suffers from high computational overhead and heavy dependence on pre-trained image models. To address this, the improved k-ViViT [14] replaced the original dense self-attention with k-NN attention, optimizing the training process by ignoring irrelevant or noisy tokens. However, heavy reliance on pretrained 2D image models may reduce training flexibility. MViT [32] pioneered the integration of multi-scale feature hierarchies with transformers, constructing a multi-scale feature pyramid. MViT-B achieved 81.2% video classification accuracy on Kinetics-400 [104], enabling synergistic perception of fine-grained local features and global semantics. Subsequently, MViTv2 [82] enhanced the model's ability to model translation invariance in visual content by introducing decomposed relative positional embeddings and residual pooling connections, improving feature propagation and mitigating information loss in deep networks. MeMViT [83] further proposed processing videos in an online fashion while caching "memory" at each iteration, allowing the model to reference prior context for long-term modeling with only marginal computational overhead. However, increased architectural complexity may lead to optimization challenges and multi-scale fusion may introduce redundancy or inconsistencies between resolution levels.

In the application of action recognition, 3D human pose estimation, particularly from monocular or multi-view video, represents a challenging subdomain (Table 3). For monocular videos, SlowFastFormer [75] introduced parallel encoding modules to extract temporal–contextual features from both slow and fast branches. To address performance degradation caused by error accumulation in diffusion processes, HSTDenoiser [43] integrated hierarchical spatiotemporal denoising modules, forming a disentangled diffusion-based 3D human pose estimation framework. The reverse diffusion process is enhanced by reinforcing attention weights between adjacent joints, leading to a marked reduction in prediction errors for key joints such as the pelvis. Meanwhile, HDFormer [105] incorporated GCNs into a transformer-based framework, leveraging multi-head graph attention to capture cross-hierarchy anatomical semantics while balancing performance and efficiency. However, performance declines with long-sequence inputs, likely due to limited model scale and ineffective handling of temporal redundancy and noise in dense sequences. For multi-view videos, FusionFormer [106] tackled depth uncertainty by first encoding 2D pose estimates into pose features, then using a transformer encoder to jointly aggregate multi-view and multi-frame features into a unified global representation.

Transformer-based models have shown promising advances, but notable limitations persist. As is widely recognized, deep learning techniques require substantial amounts of data to achieve optimal performance. However, the scarcity of real-world 3D data often leads to poor generalization in models trained on such datasets, presenting substantial challenges for 3D human pose estimation. In particular, for monocular videos, current approaches primarily focus on frame-level feature correlations while frequently neglecting inter-frame node relationships, making effective spatiotemporal information integration remain challenging. Furthermore, existing methods often exhibit reduced robustness when the target subject undergoes substantial scale variations (either shrinking or enlarging) or exhibits extreme motion speeds (either too fast or too slow). In multi-view approaches for 3D human pose estimation, effective fusion of multi-view and multi-frame features is essential for efficient information aggregation. However, interactions among multiple subjects further complicate the fusion process, rendering real-time multi-person 3D pose estimation particularly challenging. Additionally, the use of multiple cameras for

Table 3. Video transformers for action recognition using 3D human pose estimation. Human 3.6M is widely used in the 3D human pose estimation task. It contains 3.6 million 3D human poses and corresponding images with 11 professional actors and collected in 17 scenarios.

Model				Performance		
Model name	Encoder-Decoder	Backbone	Architecture	Dataset	Evaluation metric	Value (%)
SlowFastFormer ($T = 243$) [75]	E	–	Dual-branch	Human 3.6M	MPJPE	42.6
				Human 3.6M	P-MPJPE	34.2
HSTDenoiser ($T = 243$) [43]	–	–	–	Human 3.6M	MPJPE	39.0
HDformer ($T = 96$) [105]	–	–	–	Human 3.6M	MPJPE	42.6
				Human 3.6M	P-MPJPE	33.1
FusionFormer ($T = 27$) [106]	E–D	–	–	Human 3.6M	MPJPE	7.9

T, the number of input frames; MPJPE, the mean per joint position error; P-MPJPE, procrustes mean per joint position error

data acquisition introduces system complexity through requirements for camera synchronization and calibration, consequently increasing the computational overhead [107].

Video object detection and tracking

Video transformers have brought substantial improvements to object detection and tracking by leveraging self-attention mechanisms and global context modeling. They excel at identifying and localizing objects under challenging conditions, such as occlusions, deformations, and complex motion. However, despite their accuracy gains, they face practical challenges in real-time deployment, small object recognition, and adverse environmental conditions.

Building upon Sparse R-CNN [108], SparseVOD [109] adds a temporal feature extractor and an attention-guided semantic proposal module to enhance spatiotemporal object detection. Using ResNet50 backbone, it achieved 80.3% mean average precision (mAP) on the ImageNet VID [110] dataset. However, it had limited performance on fast-moving or small objects due to coarse feature granularity. TransVOD [23] was proposed as an end-to-end model that simplifies video object detection by employing a temporal transformer to aggregate single-frame spatial object queries and feature memory. This approach attained 90% mAP on ImageNet VID [110], setting a new benchmark for the task. For small object detection, both PSCA [19] and STPT [17] employed feature pyramid structures to detect targets at varying scales and velocities. FAQ [111] designed a fundamental query aggregation module and extends it into a dynamic version. When integrated with state-of-the-art transformer-based object detectors, this module achieves over 2.4% mAP improvement on the ImageNet VID [110] benchmark. However, these models are susceptible to temporal noise when object appearance changes drastically, and aggregation may blur distinctions between similar objects.

Furthermore, due to the temporal dimension inherent in videos, practical applications require not only object detection but also continuous tracking. In practical applications, video object detection is critical for the safety of autonomous vehicles, ensuring precise identification of vehicles and obstacles. To address this, TransTrack [86] was proposed as a concise yet efficient multi-object tracking solution. It innovatively employed an attention-based query-key mechanism, achieving unified detection and tracking through dual-path queries. On MOT17 [112], it attained 74.5% multiple object tracking accuracy (MOTA). However, it struggles with re-identification after long occlusion gaps. Similarly, TrackFormer [113] was developed by integrating DETR [61], establishing an encoder–decoder-based end-to-end trainable framework. This approach utilized track queries to enable seamless transition from detection to tracking tasks. However, it requires large-scale training data for robust tracking across scenarios and real-time performance is limited by attention computation over long sequences.

To address the scarcity of annotated vehicle datasets and overcome the limitations of existing techniques in adapting to unstructured traffic environments, STVD [114] proposed a Swin Transformer-based [103] vehicle detection framework. By enabling global information interaction both within and between image patches, it generates hierarchical feature maps, effectively alleviating multi-scale feature extraction challenges. Chen et al. [115] achieved dynamic feature alignment for LiDAR–camera fusion using transformers. However, there remains considerable room for optimization in terms of runtime performance. Existing bird's-eye view (BEV)-based perception systems for autonomous driving either demand substantial computational resources or exhibit modest performance. To resolve these issues, Fast-BEV [90] adopted lightweight view transformation, multi-scale image encoding, and an efficient BEV encoder. Accurate perception and decision-making in autonomous driving systems rely heavily on long-range detection of small traffic signs. Thus, TSD-DETR [81] incorporated a dedicated shallow-feature detection head to preserve and enhance fine-grained details of small objects through high-resolution feature maps. However, the model's generalization capability requires further improvement. In practical applications, enhanced robustness is needed to enable earlier decision-making for autonomous driving under nighttime and adverse weather conditions, thereby effectively reducing accident rates (Table 4).

Table 4. Video transformers-based object detection for autonomous driving and deepfake detection

Model				Performance		
Model name	Encoder–Decoder	Backbone	Architecture	Dataset	Evaluation metric	Value (%)
STVD [114]	–	–	–	KITTI	mAP	88.45
Fast-BEV [90]	E	ResNet50	Multi-scale	NuScenes	mAP	33.4
		ResNet101	Multi-scale	NuScenes	mAP	40.2
TSD-DETR [81]	–	–	Multi-scale	TT-100K	mAP	96.8
DF-TransFusion [49]	E	–	–	DFDC/FakeAVCeleb/ DF-TIMIT	AUC	97.9/74.8/100
AVTENet [42]	E	ViViT	Dual-branch	FakeAVCeleb Testset-II	Acc	99
HCiT [85]	E	Xception	–	DeepFake/ FaceSwap/ Face2Face	Acc	96.0/97.82/95.85
Swin-Fake [15]	E	Swin	–	DFDC	Acc	93.7

mAP, mean average precision; AUC, area under the curve; Acc, accuracy

Additionally, deepfake has recently emerged as a new technology impacting cybersecurity, making transformer-based deepfake video detection models a trending research focus. DF-TransFusion [49] is a multimodal deepfake detection framework that integrates lip-audio cross-attention with facial self-attention mechanisms. This architecture achieves state-of-the-art performance on multimodal deepfake detection datasets, outperforming 11 existing detection methods. However, several failure cases persist, particularly when speakers' lip and facial regions deviate from camera alignment or experience occlusion. AVTENet [42] simultaneously considered acoustic–visual tampering. Future research directions can include integrating advanced self-supervised learning models to further enhance detection performance. In terms of foundational transformer architectures, Ramadhani et al. [116] utilized ViViT [39], HCiT [85] employed ViT [4], and Swin-Fake [15] integrated the Swin Transformer [103]. The proposed Swin-Fake [15] innovatively employed the Swin Transformer [103] as the feature extractor and utilizes average cosine distance as the consistency loss metric, demonstrating superior generalization capability across multiple deepfake detection benchmarks. However, limitations remain in temporal feature extraction and utilization. Future improvements should focus on temporal modeling and spatio-temporal fusion (Table 4).

Despite the widespread applications of object detection technology, maintaining stable recognition performance under varying lighting conditions, occlusions, viewpoint changes, and object scale variations [17,19,49,111] remains a formidable challenge. Meanwhile, the stringent requirements of real-time detection impose demanding computational efficiency standards, which are critical for applications requiring instant decision-making [117]. Furthermore, in autonomous driving scenarios, models typically require adaptation to high-performance computing chips and GPU platforms for deployment. However, the continued widespread use of low-computing-power chips leads to degraded model performance in practical applications [90].

Video object and semantic segmentation

Object-based video segmentation is an essential task in CV, enabling fine-grained scene understanding by assigning pixel-level labels across frames. While transformers have markedly advanced the field by modeling long-range dependencies and temporal consistency, challenges remain, particularly in computational efficiency, occlusion handling, and generalization across conditions.

Temporal context enhanced referring video object segmentation network (TCE-RVOS) [50] proposed a frame label fusion encoder and instance query decoder, maximizing the potential information gain of the video relative to a single image. It utilizes the joint visual-textual encoding to improve instance-level distinction and is thus suitable for weakly supervised video segmentation. However, aggregation between video and text can be inconsistent under real-world conditions. Sstvos [24] introduced a sparse attention mechanism to enhance processing efficiency in long video sequences, while TransVOS [20] designed a unified dual-path transformer-based feature extractor, simplifying the dual-encoder pipeline. Both models achieved competitive performance on the YouTube-VOS [118] and DAVIS [119] benchmarks. They maintain competitive segmentation performance on long videos and greatly reduce computational overhead. Such lightweight architecture is suitable for deployment on resource-limited devices. However, sparse attention may lead to incomplete object representation in scenes with complex interactions and reduced flexibility in handling notable domain variation across video inputs. In recent advancements (Table 5), the fully transformer-equipped architecture (FTEA) [79] formulated referring video object segmentation as mask sequence learning in a pure transformer-based end-to-end framework, while DCT [51] designed a language-guided visual enhancement module for the reference video object segmentation task and adopted a cross-layer feature pyramid network as the spatial decoder to better utilize multimodal and multiscale information to generate high-quality object boundaries. However, they are

Table 5. Video transformers for video object and semantic segmentation

Model				Performance		
Model name	Encoder–Decoder	Backbone	Architecture	Dataset	Evaluation metric	Value (%)
Cutie [37]	E-D	ResNet	Dual-branch	DAVIS-17 test	$\mathcal{J}\&\mathcal{F}/\mathcal{J}/\mathcal{F}$	85.3/81.4/89.3
	E-D	ResNet	Dual-branch	YouTube-VOS-2019 val	$\mathcal{J}_s/\mathcal{F}_s/\mathcal{J}_u/\mathcal{F}_u$	85.4/90.0/81.3/89.3
TCE-RVOS [50]	E-D	ResNet	Dual-branch	YouTube-VOS	$\mathcal{J}\&\mathcal{F}/\mathcal{J}/\mathcal{F}$	60.8/59.4/62.2
	E-D	Video Swin	Dual-branch	YouTube-VOS	$\mathcal{J}\&\mathcal{F}/\mathcal{J}/\mathcal{F}$	61.3/59.8/62.7
DCT [51]	E-D	Video Swin-T	–	Ref-YouTube-VOS	$\mathcal{J}\&\mathcal{F}/\mathcal{J}/\mathcal{F}$	56.6/55.4/57.8
MAVOS [55]	E-D	ResNet	–	LVOS	$\mathcal{J}\&\mathcal{F}/\mathcal{J}/\mathcal{F}$	63.6/57.6/69.0
	E-D	Swin-B	–	LVOS	$\mathcal{J}\&\mathcal{F}/\mathcal{J}/\mathcal{F}$	64.8/58.7/70.9
FIEA [79]	E-D	Swin-T	Dual-branch/ Multi-scale	Ref-YouTube-VOS val	$\mathcal{J}\&\mathcal{F}/\mathcal{J}/\mathcal{F}$	56.5/55.0/58.0
RangeViT	E-D	ViT	–	NuScenes	mIoU	75.21
TPVFormer	–	–	–	NuScenes	mIoU	69.4

$\mathcal{J}$, Jaccard index; $\mathcal{F}$, contour accuracy; $\mathcal{J}\&\mathcal{F}$: the average of $\mathcal{J}$ and $\mathcal{F}$; s and u, the calculation of $\mathcal{J}$ and $\mathcal{F}$ for the "seen" and "unseen" categories in the YouTube-VOS dataset, respectively; mIoU, mean intersection over union; E-D, the encoder–decoder architecture

extremely resource-intensive during training and inference, which hinders its widespread adoption. They are susceptible to tracking drift in long-term sequences and show poor temporal coherence due to the lack of a memory mechanism to store the historical features of the referenced objects, resulting in insufficient utilization of temporal information. Furthermore, Cutie [37] reintegrated object-level reasoning into the segmentation pipeline by combining foreground–background mask attention mechanisms, while MAVOS [55] introduced an optimized and dynamic long-term modulation cross-attention to model temporal smoothness. They enhance reasoning about object boundaries and foreground continuity and are thus robust in structured environments, showing excellent segmentation performance in many cases. However, they struggle with close-range movement of visually similar objects or sudden occlusion or target disappearance in mutual occlusion or fast motion scenarios.

Video semantic segmentation extends object segmentation by assigning class labels to every pixel and enabling vehicles to understand and interact with their visual environment, making it particularly critical for autonomous driving and robotic perception. RangeViT [89] adapted pre-trained ViTs for LiDAR-based 3D semantic segmentation through customized tokenization and preprocessing pipelines for ViT encoders, coupled with an efficient convolutional decoder. It handled 3D point clouds effectively and leveraged pre-trained image models for better generalization. However, it struggles with sparsely populated or occluded LiDAR returns and requires complex preprocessing and tuning for different sensor setups. TPVFormer [120] projected image features into an enhanced 3D space via its encoder architecture, effectively aggregating multi-plane features for comprehensive 3D scene understanding. Thus, it performed well in the extended tasks of semantic segmentation and semantic possession prediction, successfully capturing the positions and sizes of near and far objects with high precision. However, it still exhibits invalid predictions, including failure to distinguish closely adjacent pedestrians and distant pedestrian misprojection as strip-shaped artifacts (Table 5).

Despite promising results, both video object and semantic segmentation approaches face shared limitations. Illumination and weather variations, as well as object size and occlusion factors, can also affect segmentation accuracy. For example, while Cutie [37] demonstrates superior robustness compared to other state-of-the-art methods, it tends to fail under 2 challenging scenarios: close-range movement of highly similar objects and mutual occlusion cases. Similarly, FIEA [79] shows limitations in complete contour prediction (e.g., appearance similarity between target and background, poor illumination, and object overlap). Moreover, high computational demands required for processing high-detail images further hinder real-time applications [117].

Video generation

Video generation refers to the automatic creation of continuous, realistic, and temporally coherent video content using AI technologies. Recent breakthroughs—particularly the fusion of transformer architecture and diffusion models—have substantially pushed the boundaries of this field, enabling the development of tools for "one-click video creation" that reduce the creative and technical barrier for users. Despite these advances, the field still faces notable challenges in realism, temporal consistency, multi-modal control, and computational scalability.

Inspired by the success of diffusion models in image generation [121], transformer-based video generation models have quickly integrated with diffusion techniques. VDT [122] pioneered this integration, leveraging the inherent sequence modeling capability of transformers to seamlessly extend to video prediction tasks through a straightforward token concatenation strategy. To effectively model the substantial number of tokens extracted from videos, Latte [123] introduced 4 efficient variants by decomposing the spatial and temporal dimensions of input videos. Through systematic experimentation, including video clip patch embedding, temporal positional embedding, and learning strategies, it achieved state-of-the-art video generation quality. While considerable progress has been made in human motion video generation, existing methods still struggle to accurately render detail-rich body parts such as hands and facial features, particularly in long sequences and complex motions. To address this, HumanDiT [124] proposed a pose-guided diffusion transformer that supports multiple video resolutions and variable sequence lengths, facilitating long-sequence video generation learning. However, it is susceptible to temporal artifacts or motion jitter in high-speed movements.

Despite rapid progress, current transformer-based video generation models face several persistent limitations: detail preservation and temporal coherence. Even advanced models like HumanDiT [124] struggle with rendering fine-grained details, particularly in high-motion regions such as limbs, facial features, and clothing dynamics. Most models are constrained in sequence length, with quality often degrading in long videos due to accumulation of temporal drift or frame incoherence. More critically, the rapid advancement and widespread application of video generation technologies have raised increasingly stringent demands for video security, which should become a key focus area for subsequent research.

Future Technologies and Prospects

Video transformers excel in spatiotemporal modeling and feature extraction for video processing, but face challenges like high computational complexity, substantial data dependence, and sensitivity to hyperparameter tuning. To address these limitations and inspire the next generation of models, this section outlines several forward-looking research directions, each grounded in emerging trends and recent theoretical and empirical findings.

Improving efficiency of transformer architectures

A central concern in transformer design is architectural redundancy. Future work should ask whether all transformer layers contribute to performance or if some can be streamlined. Two approaches to investigate this are "freezing" and "skipping":

Freezing layers: One technique involves selectively freezing layers during training. For example, when our experiment froze the FFN layers in ViViT [39] on the DVS dataset [125], the accuracy dropped only marginally (from 94.31% to 93.97%) while saving 14.29% of computation time per epoch. This suggests that static substructures in deep networks can retain representational power while reducing computational overhead.

Skipping layers: Skip-based architectures [126] create direct connections between non-adjacent layers, effectively skipping intermediate layers between them. This increases model flexibility and reduces redundancy without a substantial drop in performance. Such dynamic skipping strategies may inspire new families

of lightweight and flexible video transformers with more computational efficiency for real-time applications and edge devices.

Thus, by incorporating freezing or skipping techniques, future models can become more computationally efficient while maintaining high accuracy, making video processing tasks more feasible on resource-constrained systems.

Hybrid models

Hybrid models offer a promising avenue for combining the strengths of multiple paradigms. While transformers provide powerful global modeling, combining them with domain-specific architectures may yield more adaptable and efficient systems.

Reservoir computing and transformers [127]: Reservoir computing offers fast training and inherent temporal dynamics, making it suitable for tasks like streaming video analysis. When integrated with the representational depth of Transformers, these hybrid models could reduce training costs and improve robustness in low-data regimes.

Wavelet transforms and transformers [128]: Wavelet decomposition enhances temporal frequency analysis, helping to capture both fine and coarse features. Its integration with transformers could improve multi-scale representation for tasks such as video synthesis or fine-grained action recognition.

You only look once (YOLO) and transformers [129]: YOLO excels in fast object detection. Merging its real-time capabilities with transformer's global contextual understanding creates a powerful architecture for high-speed video understanding, especially in real-time video processing applications. However, balancing computational performance and transfer learning capabilities remains a challenge for such hybrid models, necessitating deeper exploration of optimization techniques.

Despite their potential, hybrid models often struggle with modular compatibility, increased model complexity, and domain-specific transferability—challenges that must be addressed to ensure practical deployment.

Multi-modal data integration and adaptive models

As real-world video understanding tasks increasingly involve multi-modal inputs (e.g., vision, sound, and language), future video transformers should evolve to handle heterogeneous data efficiently. Recent works [42,49,79] demonstrate how multi-modal transformers can process synchronized audiovisual signals for tasks like lip-reading or deepfake detection. Future models should extend this capability by adapting inference strategies dynamically based on input modality or context to handle a broader range of applications, from video understanding to complex decision-making tasks.

Adaptive computation time: Dynamically halting processing when sufficient confidence is reached to save computation.

Context-aware modules: Adjusting attention mechanisms or embedding strategies based on modality or environment.

Moreover, incorporating reinforcement learning and online learning [130,131] could enable transformers to operate effectively in real time, continuously evolving settings, such as autonomous driving or surveillance.

Brain-inspired models: Neural dynamics and memory

A novel and promising direction lies in drawing inspiration from cognitive neuroscience, particularly mechanisms in the entorhinal–hippocampal system that govern memory formation, retrieval, and spatial navigation in the brain.

Brain-inspired video transformers could emulate core neural processes including the following:

Dynamic information encoding: Mimicking grid and place cell functionality to spatially encode visual content over time.

Temporal context modeling: Using recurrent neural dynamics and the corresponding high-dimensional manifolds of neural dynamics to simulate working memory and temporal trace formation.

Memory retrieval and updating: Inspired by hippocampal replay and attention gating, enabling selective memory retrieval and real-time adaptation.

For instance, the integration of spiking neural networks or continuous attractor dynamics with transformer layers could lead to neuromorphic video models capable of lifelong learning, few-shot generalization, and interpretability. Such models may reduce dependence on large-scale data by leveraging more structured internal dynamics, akin to how the brain processes sequences efficiently and robustly.

While still largely theoretical, this direction represents a critical interface between computational neuroscience and AI, with transformative implications for video processing.

Conclusions

Video transformers have proven to be powerful tools for video processing, capable of handling multi-dimensional features in parallel and capturing the temporal and spatial dependencies inherent in video data. Their flexibility, adjustability, and universal applicability make them ideal for a wide range of video-based tasks, and their ability to integrate with other models enhances their robustness.

This paper has reviewed the current state of research on transformer models in video processing, covering input modules, architectures, hybrid models, and recent applications. However, the understanding of the internal mechanisms of video transformers remains limited, and their complexity often leads to poor interpretability, which hinders their use in certain applications. Therefore, future research should aim to deepen the understanding of these models' internal workings, which will help improve their performance and broaden their applicability in real-world scenarios.

Acknowledgments

Funding: This work was supported by the Research Initiation Project of Zhejiang lab (no. K2022Ki0Pi01), the Natural Science Foundation of Zhejiang Province (grant nos. IZ24A050007 and LY24A050003), the National Natural Science Foundation of China (grant no. 12175242), and the Public Welfare Research Plan of Jiaxing (grant nos. 2025CGZ037 and 2023AY31029).
Author contributions: C.Y., N.C., and D.Y. wrote the abstract and introduction. N.C. and D.Y. wrote the theory section. M.S., N.C., and T.X. wrote the model performance analysis section. D.Y. and T.X. wrote the outlook section. The manuscript was revised by all authors.
Competing interests: The authors declare that they have no competing interests.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In: *Proceedings of the Conference and Workshop on Neural Information Processing Systems.* Long Beach (USA); 2017.

2. Kenton J D M-W C, Toutanova L K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Minneapolis (MN): Association for Computational Linguistics; 2019. p. 4171–4186.
3. Radford A. Improving language understanding by generative pre-training. 2018. https://openai.com/research/language_unsupervised
4. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. 2020. https://doi.org/10.48550/arXiv.2010.11929
5. Ridnik T, Ben-Baruch E, Noy A, Zelnik-Manor L. Imagenet-21k pretraining for the masses. arXiv. 2021. https://doi.org/10.48550/arXiv.2104.10972
6. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition.* Miami (FL): IEEE; 2009. p. 248–255.
7. Naseer MM, Ranasinghe K, Khan SH, Hayat M, Shahbaz Khan F, Yang M-H. Intriguing properties of vision transformers. In: *Proceedings of the Conference and Workshop on Neural Information Processing Systems.* Virtual; 2021. p. 23296–23308.
8. Tuli S, Dasgupta I, Grant E, Griffiths T L. Are convolutional neural networks or transformers more like human vision? arXiv. 2021. https://doi.org/10.48550/arXiv.2105.07197
9. Yang Y, Jiao L, Liu X, Liu F, Yang S, Feng Z, Tang X. Transformers meet visual learning understanding: A comprehensive review. arXiv. 2022. https://doi.org/10.48550/arXiv.2203.12944
10. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. *ACM Comput Surv*. 2022;54(10s):1–41.
11. Selva J, Johansen AS, Escalera S, Nasrollahi K, Moeslund TB, Clapés A. Video transformers: A survey. *IEEE Trans Pattern Anal Mach Intell*. 2023;45(11):12922–12943.
12. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, et al. A survey on vision transformer. *IEEE Trans Pattern Anal Mach Intell*. 2022;45(1):87–110.
13. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: *Proceedings of the International Conference on Machine Learning.* PMLR. Virtual; 2021. p. 10347–10357.
14. Sun W, Ma Y, Wang R. *k*-NN attention-based video vision transformer for action recognition. *Neurocomputing*. 2024;574:127256.
15. Gong LY, Li XJ, Chong PHJ. Swin-fake: A consistency learning transformer-based deepfake video detector. *Electronics*. 2024;13(15):3045.
16. Liu Z, Luo S, Li W, Lu J, Wu Y, Sun S, Li C, Yang L. Convtransformer: A convolutional Transformer network for video frame synthesis. arXiv. 2020. https://doi.org/10.48550/arXiv.2011.10185
17. Weng Y, Pan Z, Han M, Chang X, Zhuang B. An efficient spatio-temporal pyramid Transformer for action detection. In: *Proceedings of the European Conference on Computer Vision.* Tel Aviv (Israel): Springer; 2022. p. 358–375.
18. Yu L, Huang L, Zhou C, Zhang H, Ma Z, Zhou H, Tian Y. SVFormer: A direct training spiking transformer for efficient video action recognition. arXiv. 2024. https://doi.org/10.48550/arXiv.2406.15034
19. Gu Y, Wang L, Wang Z, Liu Y, Cheng M-M, Lu S-P. Pyramid constrained self-attention network for fast video salient object detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* New York (USA): AAAI; 2020. p. 10869–10876.
20. Mei J, Wang M, Lin Y, Yuan Y, Liu Y. TransVOS: Video object segmentation with transformers. arXiv. 2021. https://doi.org/10.48550/arXiv.2106.00588
21. Ahn D, Kim S, Ko BC. STAR++: Rethinking spatio-temporal cross attention transformer for video action recognition. *Appl Intell*. 2023;53(23):28446–28459.
22. Lin K-Y, Zhou J, Zheng W-S. Human-centric transformer for domain adaptive action recognition. *IEEE Trans Pattern Anal Mach Intell*. 2024;4(2):1–18.
23. He L, Zhou Q, Li X, Niu L, Cheng G, Li X, Liu W, Tong Y, Ma L, Zhang L. End-to-end video object detection with spatial-temporal Transformers. In: *Proceedings of the 29th ACM International Conference on Multimedia.* Chengdu (China): ACM; 2021. p. 1507–1516.
24. Duke B, Ahmed A, Wolf C, Aarabi P, Taylor GW. Sstvos: Sparse spatiotemporal Transformers for video object segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* IEEE. Virtual; 2021. p. 5912–5921.
25. Truong T-D, Bui Q-H, Duong C N, Seo H-S, Phung S L, Li X, Luu K. Direcformer: A directed attention in transformer approach to robust action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* New Orleans (LA): IEEE; 2022. p. 20030–20040.
26. Zeng Y, Fu J, Chao H. Learning joint spatial-temporal transformations for video inpainting. In: *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference.* Glasgow (UK): Springer; 2020. p. 528–543.
27. Peng M, Wang C, Gao Y, Shi Y, Zhou X-D. Temporal pyramid Transformer with multimodal interaction for video question answering. arXiv. 2021. https://doi.org/10.48550/arXiv.2109.04735
28. Yan S, Xiong X, Arnab A, Lu Z, Zhang M, Sun C, Schmid C. Multiview Transformers for video recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* New Orleans (LA): IEEE; 2022. p. 3333–3343.
29. Siam M, Karim R, Zhao H, Wildes R. Multiscale memory comparator transformer for few-shot video segmentation. arXiv. 2023. https://doi.org/10.48550/arXiv.2307.07812
30. Pizarro R, Valle R, Bergasa LM, Buenaposada JM, Baumela L. Pose-guided multi-task video transformer for driver action recognition. arXiv. 2024. https://doi.org/10.48550/arXiv.2407.13750
31. Jelassi S, Sander M, Li Y. Vision Transformers provably learn spatial structure. In: *Proceedings of the Conference and Workshop on Neural Information Processing Systems.* New Orleans (LA): MIT Press; 2022. p. 37822–37836.
32. Fan H, Xiong B, Mangalam K, Li Y, Yan Z, Malik J, Feichtenhofer C. Multiscale vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* Virtual: IEEE; 2021. p. 6824–6835.
33. Liu R, Deng H, Huang Y, Shi X, Lu L, Sun W, Wang X, Dai J, Li H. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* Virtual: IEEE; 2021. p. 14040–14049.
34. Gavrilyuk K, Sanford R, Javan M, Snoek C G. Actor-transformers for group activity recognition. In: *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle (WA): IEEE; 2020. p. 839–848.
35. Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, Ning J, Cao Y, Zhang Z, Dong L. Swin Transformer v2: Scaling up capacity and resolution. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans (LA): IEEE; 2022. p. 12009–12019.
36. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H. Video Swin Transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans (LA): IEEE; 2022. p. 3202–3211.
37. Cheng H K, Oh S W, Price B, Lee J-Y, Schwing A. Putting the object back into video object segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle (WA): IEEE; 2024. p. 3151–3161.
38. Shou Z, Yuan X, Li D, Mo J, Zhang H, Zhang J, Wu Z. A dynamic position embedding-based model for student classroom complete meta-action recognition. *Sensors*. 2024;24(16):5371.
39. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. ViViT: A video vision transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal (Canada): IEEE; 2021. p. 6836–6846.
40. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? In: *Proceedings of the International Conference on Machine Learning*. Virtual; 2021. p. 4.
41. Tan Y, Qiu Z, Hao Y, Yao T, He X, Mei T. Selective volume mixup for video action recognition. arXiv. 2023. https://doi.org/10.48550/arXiv.2309.09534
42. Hashmi A, Shahzad S A, Lin C-W, Tsao Y, Wang H-M. AVTENet: Audio-visual transformer-based ensemble network exploiting multiple experts for video deepfake detection. arXiv. 2023. https://doi.org/10.48550/arXiv.2310.13103
43. Cai Q, Hu X, Hou S, Yao L, Huang Y. Disentangled diffusion-based 3D human pose estimation with hierarchical spatial and temporal denoiser. In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Vancouver (Canada): AAAI; 2024. p. 882–890.
44. Yuan H, Cai Z, Zhou H, Wang Y, Chen X. Transanomaly: Video anomaly detection using video vision transformer. *IEEE Access*. 2021;9:123977–123986.
45. Ouyang Y, Zhang T, Gu W, Wang H. Adaptive perception transformer for temporal action localization. arXiv. 2022. https://doi.org/10.48550/arXiv.2208.11908
46. Zhang T, Yang J. Transformer with hybrid attention mechanism for stereo endoscopic video super resolution. *Symmetry*. 2023;15(10):1947.
47. Ning X, Cai F, Li Y, Ding Y. Parallel spatio-temporal attention transformer for video frame interpolation. *Electronics*. 2024;13(10):1981.
48. Tu Y, Wu J, Lu L, Gao S, Li M. Face forgery video detection based on expression key sequences. *J King Saud Univ-Com*. 2024;36(7):102142.
49. Kharel A, Paranjape M, Bera A. DF-TransFusion: Multimodal deepfake detection via lip-audio cross-attention and facial self-attention. arXiv. 2023. https://doi.org/10.48550/arXiv.2309.06511
50. Hu X, Hampiholi B, Neumann H, Lang J. Temporal context enhanced referring video object segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa (HI): IEEE; 2024. p. 5574–5583.
51. Wu A, Wang R, Tan Q, Song Z. Decoupled cross-modal transformer for referring video object segmentation. *Sensors*. 2024;24(16):5375.
52. Wang X, Rong Y, Wang S, Chen Y, Wu Z, Jiang B, Tian Y, Tang J. Unleashing the power of CNN and Transformer for balanced RGB-event video recognition. arXiv. 2023. https://doi.org/10.48550/arXiv.2312.11128
53. Wang R, Wu Z, Chen D, Chen Y, Dai X, Liu M, Zhou L, Yuan L, Jiang Y-G. Video Mobile-former: Video recognition with efficient global spatial-temporal modeling. arXiv. 2022. https://doi.org/10.48550/arXiv.2208.12257
54. Liang Z, Dong W, Zhang B. A dual-branch hybrid network of CNN and transformer with adaptive keyframe scheduling for video semantic segmentation. *Multimedia Systems*. 2024;30(2):67.
55. Shaker A, Wasim ST, Danelljan M, Khan S, Yang M-H, Khan F S. Efficient video object segmentation via modulated cross-attention memory. arXiv. 2024. https://doi.org/10.48550/arXiv.2403.17937
56. Ragini T, Prakash K, Cheruku R. DeTformer: A novel efficient transformer framework for image deraining. *Circ Syst Signal Pr*. 2024;43(2):1030–1052.
57. Wang H, Zhao B, Zhang W, Liu G. LGANet: Local and global attention are both you need for action recognition. *IET Image Process*. 2023;17(12):3453–3463.
58. Li W, Liu H, Ding R, Liu M, Wang P, Yang W. Exploiting temporal contexts with strided transformer for 3D human pose estimation. *IEEE Trans Multimed*. 2022;25:1282–1293.
59. Zeng Y, Zeng B, Hu H, Zhang H. PRAT: Accurate object tracking based on progressive attention. *Eng Appl Artif Intell*. 2023;126:106988.
60. Wang P, Zhang Y, Wang L, Yuan X. Hierarchical separable video transformer for snapshot compressive imaging. arXiv. 2024. https://doi.org/10.48550/arXiv.2407.11946
61. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with Transformers. In: *Proceedings of the European Conference on Computer Vision*. Glasgow (UK): Springer; 2020. p. 213–229.
62. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: Deformable transformers for end-to-end object detection. arXiv. 2020. https://doi.org/10.48550/arXiv.2010.04159
63. Karim R, Zhao H, Wildes R P, Siam M. MED-VT: Multiscale encoder-decoder video transformer with application to object segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, BC (Canada): IEEE; 2023. p. 6323–6333.
64. Wang X, Li P, Wang R. CEDT2M: Text-driven human motion generation via cross-modal mixture of encoder-decoder. *Neural Comput Appl*. 2025;1–19.
65. Li Y, Wu S, Zhu Y, Sun W, Zhang Z, Song S. SAMR: Symmetric masked multimodal modeling for general multi-modal 3D motion retrieval. *Displays*. 2025;87:102987.
66. Miech A, Alayrac J-B, Laptev I, Sivic J, Zisserman A. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Virtual: IEEE; 2021. p. 9826–9836.
67. Tan J, Tang J, Wang L, Wu G. Relaxed transformer decoders for direct action proposal generation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal (Canada): IEEE; 2021. p. 13526–13535.
68. Kondratyuk D, Yu L, Gu X, Lezama J, Huang J, Schindler G, Hornung R, Birodkar V, Yan J, Chiu M-C. Videopoet: A large

language model for zero-shot video generation. arXiv. 2023. https://doi.org/10.48550/arXiv.2312.14125
69. Gupta A, Likhomanenko T, Yang K D, Bai R H, Aldeneh Z, Jaitly N. Visatronic: A multimodal decoder-only model for speech synthesis. arXiv. 2024. https://doi.org/10.48550/arXiv.2411.17690
70. Lu Z, Elhamifar E. BIT: Bi-level temporal modeling for efficient supervised action segmentation. arXiv. 2023. https://doi.org/10.48550/arXiv.2308.14900
71. Herzig R, Ben-Avraham E, Mangalam K, Bar A, Chechik G, Rohrbach A, Darrell T, Globerson A. Object-region video transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* New Orleans (LA): IEEE; 2022. p. 3148–3159.
72. Liu J, Yin B, Lin J, Wen J, Li Y, Liu M. HDBN: A novel hybrid dual-branch network for robust skeleton-based action recognition. arXiv. 2024. https://doi.org/10.48550/arXiv.2404.15719
73. Hussain A, Khan SU, Khan N, Ullah W, Alkhayyat A, Alharbi M, Baik SW. Shots segmentation-based optimized dual-stream framework for robust human activity recognition in surveillance video. *Displays*. 2024;91:632–647.
74. Zhou C, Li FW, Song C, Zheng D, Yang B. 3D data augmentation and dual-branch model for robust face forgery detection. *Graphic Models*. 2025;138:101255.
75. Zhou L, Chen Y, Wang J. SlowFastFormer for 3D human pose estimation. *Comput Vis Image Und*. 2024;243:103992.
76. Chen H, Zhang X, Guo Z, Ying N, Yang M, Guo C. ACTNet: Attention based CNN and transformer network for respiratory rate estimation. *Biomed Signal Process Control*. 2024;96:106497.
77. Ganguly S, Ganguly A, Mohiuddin S, Malakar S, Sarkar R. ViXNet: Vision transformer with Xception Network for deepfakes based video and image forgery detection. *Expert Syst Appl*. 2022;210:118423.
78. Ma H, Zhang C, Ning E, Chuah CW. Temporal motion and spatial enhanced appearance with Transformer for video-based person ReID. *Knowl Based Systems*. 2025;317:113461.
79. Li P, Zhang Y, Yuan L, Xu X. Fully transformer-equipped architecture for end-to-end referring video object segmentation. *Inform Process Manag*. 2024;61(1):103566.
80. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Honolulu (HI): IEEE; 2017. p. 2117–2125.
81. Zhang L, Yang K, Han Y, Li J, Wei W, Tan H, Yu P, Zhang K, Yang X. TSD-DETR: A lightweight real-time detection transformer of traffic sign detection for long-range perception of autonomous driving. *Eng Appl Artif Intell*. 2025;139:109536.
82. Li Y, Wu C-Y, Fan H, Mangalam K, Xiong B, Malik J, Feichtenhofer C. Mvitv2: Improved multiscale vision transformers for classification and detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* New Orleans (LA): IEEE; 2022. p. 4804–4814.
83. Wu C-Y, Li Y, Mangalam K, Fan H, Xiong B, Malik J, Feichtenhofer C. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* New Orleans (LA): IEEE; 2022. p. 13587–13597.
84. Djenouri Y, Belbachir AN. A hybrid visual transformer for efficient deep human activity recognition. In: *Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision.* Paris (France): IEEE; 2023. p. 721–730.
85. Kaddar B, Fezza SA, Akhtar Z, Hamidouche W, Hadid A, Serra-Sagristà J. Deepfake detection using spatiotemporal transformer. *ACM T Multim Comput*. 2024;20(11):1–21.
86. Sun P, Cao J, Jiang Y, Zhang R, Xie E, Yuan Z, Wang C, Luo P. TransTrack: Multiple object tracking with transformer. arXiv. 2020. https://doi.org/10.48550/arXiv.2012.15460
87. Le V-T, Jin H, Kim Y-G. HSTforU: Anomaly detection in aerial and ground-based videos with hierarchical spatio-temporal transformer for U-net. *Appl Intell*. 2025;55(4):261.
88. Zhang B, Ma R, Cao Y, An P. Swin-VEC: Video Swin Transformer-based GAN for video error concealment of VVC. *Vis Comput*. 2024;40(10):7335–7347.
89. Ando A, Gidaris S, Bursuc A, Puy G, Boulch A, Marlet R. Rangevit: Towards vision transformers for 3D semantic segmentation in autonomous driving. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Vancouver (Canada): IEEE; 2023. p. 5240–5250.
90. Li Y, Huang B, Chen Z, Cui Y, Liang F, Shen M, Liu F, Xie E, Sheng L, Ouyang W. Fast-BEV: A fast and strong bird's-eye view perception baseline. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE; 2024. p. 8665–8679.
91. Ahmadabadi H, Manzari O N, Ayatollahi A. Distilling knowledge from CNN-transformer models for enhanced human action recognition. In: *Proceedings of the 13th International Conference on Computer and Knowledge Engineering (ICCKE).* Mashhad (Iran): IEEE; 2023. p. 180–184.
92. Deng Z, Zhang W, Chen K, Zhou Y, Tian J, Quan G, Zhao J. TT U-Net: Temporal transformer U-Net for motion artifact reduction using PAD (pseudo all-phase clinical-dataset) in cardiac CT. *IEEE Trans Med Imaging*. 2023;42(12):3805–3816.
93. Roka S, Diwakar M. CViT: A convolution vision transformer for video abnormal behavior detection and localization. *SN Comput Sci*. 2023;4(6):829.
94. Khaleghi L, Marshall J, Etemad A. Learning sequential contexts using transformer for 3D hand pose estimation. In: *Proceedings of the 26th International Conference on Pattern Recognition (ICPR).* Montreal, QC (Canada): IEEE; 2022. p. 535–541.
95. Nash C, Carreira J, Walker J, Barr I, Jaegle A, Malinowski M, Battaglia P. Transframer: Arbitrary frame prediction with generative models. arXiv. 2022. https://doi.org/10.48550/arXiv.2203.09494
96. Du D, Su B, Li Y, Qi Z, Si L, Shan Y. Do we really need temporal convolutions in action segmentation? In: *Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME).* Brisbane (Australia): IEEE; 2023. p. 1014–1019.
97. Berroukham A, Housni K, Lahraichi M. Detection and localization of anomalous objects in video sequences using vision transformers and U-net model. *Signal Image Video P*. 2024;18(8–9):1–12.
98. Tunga A, Nuthalapati S V, Wachs J. Pose-based sign language recognition using GCN and BERT. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa (HI): IEEE; 2021. p. 31–40.
99. Zheng L, Xu W, Miao Z, Qiu X, Gong S. RESTHT: Relation-enhanced spatial–temporal hierarchical transformer for video captioning. *Vis Comput*. 2024;41(1):591–604.

100. Chu P, Wang J, You Q, Ling H, Liu Z. Transmot: Spatial-temporal graph transformer for multiple object tracking. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* Waikoloa (HI): IEEE; 2023. p. 4870–4880.
101. Gkalelis N, Daskalakis D, Mezaris V. ViGAT: Bottom-up event recognition and explanation in video using factorized graph attention network. *IEEE Access*. 2022;10:108797–108816.
102. Chen T, Mo L. Swin-fusion: Swin-transformer with feature fusion for human action recognition. 2023;55(8):11109–11130.
103. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin Transformer: Hierarchical vision Transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* Virtual: IEEE; 2021. p. 10012–10022.
104. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P. The kinetics human action video dataset. arXiv. 2017. https://doi.org/10.48550/arXiv.1705.06950
105. Chen H, He J-Y, Xiang W, Cheng Z-Q, Liu W, Liu H, Luo B, Geng Y, Xie X. HDFormer: High-order directed transformer for 3D human pose estimation. arXiv. 2023. https://doi.org/10.48550/arXiv.2302.01825
106. Cai Y, Zhang W, Wu Y, Jin C. Fusionformer: A concise unified feature fusion transformer for 3D pose estimation. *Proc AAAI Conf Artif Intell*. 2024;38(2):900–908.
107. Neupane RB, Li K, Boka TF. A survey on deep 3D human pose estimation. *Artif Intell Rev*. 2024;58(1):24.
108. Sun P, Zhang R, Jiang Y, Kong T, Xu C, Zhan W, Tomizuka M, Li L, Yuan Z, Wang C. Sparse R-CNN: End-to-end object detection with learnable proposals. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Virtual: IEEE; 2021. p. 14454–14463.
109. Hashmi K A, Stricker D, Afzal M Z. Spatio-temporal learnable proposals for end-to-end video object detection. arXiv. 2022. https://doi.org/10.48550/arXiv.2210.02368
110. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. *Imagenet large scale visual recognition challenge*. 2015;115:211–252.
111. Cui Y. Feature aggregated queries for Transformer-based video object detectors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Vancouver (Canada): IEEE; 2023. p. 6365–6376.
112. Milan A, Leal-Taixé L, Reid I, Roth S, Schindler K. MOT16: A benchmark for multi-object tracking. arXiv. 2016. https://doi.org/10.48550/arXiv.1603.00831
113. Meinhardt T, Kirillov A, Leal-Taixe L, Feichtenhofer C. Trackformer: Multi-object tracking with transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* New Orleans (LA): IEEE; 2022. p. 8844–8854.
114. Deshmukh P, Satyanarayana G, Majhi S, Sahoo UK, Das SK. Swin transformer based vehicle detection in undisciplined traffic environment. *Expert Syst Appl*. 2023;213(Part B):118992.
115. Chen M, Liu P, Zhao H. LiDAR-camera fusion: Dual transformer enhancement for 3D object detection. 2023;120:105815.
116. Ramadhani KN, Munir R, Utama NP. Improving video vision transformer for deepfake video detection using facial landmark, depthwise separable convolution and self attention. *IEEE Access*. 2024;12:8932–8939.
117. Zhao J, Wu Y, Deng R, Xu S, Gao J, Burke A. A survey of autonomous driving from a deep learning perspective. *ACM J*. 57(10):1–60.
118. Xu N, Yang L, Fan Y, Yue D, Liang Y, Yang J, Huang T. YouTube-VOS: A large-scale video object segmentation benchmark. arXiv. 2018. https://doi.org/10.48550/arXiv.1809.03327
119. Pont-Tuset J, Perazzi F, Caelles S, Arbeláez P, Sorkine-Hornung A, Van Gool L. The 2017 DAVIS challenge on video object segmentation. arXiv. 2017. https://doi.org/10.48550/arXiv.1704.00675
120. Huang Y, Zheng W, Zhang Y, Zhou J, Lu J. Tri-perspective view for vision-based 3D semantic occupancy prediction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Vancouver (Canada): IEEE; 2023. p. 9223–9232.
121. Peebles W, Xie S. Scalable diffusion models with transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* Paris (France): IEEE; 2023. p. 4195–4205.
122. Lu H, Yang G, Fei N, Huo Y, Lu Z, Luo P, Ding M. VDT: General-purpose video diffusion transformers via mask modeling. arXiv. 2023. https://doi.org/10.48550/arXiv.2305.13311
123. Ma X, Wang Y, Jia G, Chen X, Liu Z, Li Y-F, Chen C, Qiao Y. Latte: Latent diffusion transformer for video generation. arXiv. 2024. https://doi.org/10.48550/arXiv.2401.03048
124. Gan Q, Ren Y, Zhang C, Ye Z, Xie P, Yin X, Yuan Z, Peng B, Zhu J. HumanDiT: Pose-guided diffusion transformer for long-form human motion video generation. arXiv. 2025. https://doi.org/10.48550/arXiv.2502.04847
125. Vasudevan A, Negri P, Di Ielsi C, Linares-Barranco B, Serrano-Gotarredona T. SL-Animals-DVS: Event-driven sign language animals dataset. *Pattern Anal Appl*. 2022;25(3):1–16.
126. Sun Q, Pickett M, Nain A K, Jones L. Transformer layers as painters. arXiv. 2024. https://doi.org/10.48550/arXiv.2407.09298
127. Yang D, Xu T, Lin F. Reservoir computing in rehabilitation video analyses. In: *Proceedings of the 2023 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM).* Penang (Malaysia): IEEE; 2023. p. 210–214.
128. Zhang K, Li G, Su Y, Wang J. WTVI: A wavelet-based transformer network for video inpainting. *IEEE Signal Proc Let*. 2024;31:616–620.
129. Liu D, Wang Z, Meng X. Fast intensive crowd counting model of internet of things based on multi-scale attention mechanism. *IET Image Process*. 2022;19(1):Article e12686.
130. Suzuki T, Aoki Y. RetinaViT: Efficient visual backbone for online video streams. *Sensors*. 2024;24(17):5457.
131. Bajgoti A, Gupta R, Balaji P, Dwivedi R, Siwach M, Gupta D. SwinAnomaly: Real-time video anomaly detection using video Swin transformer and SORT. *IEEE Access*. 2023;24(17):5457.