Neurocomputing 574 (2024) 127256

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

k-NN attention-based video vision transformer for action recognition

Weirong Sun, Yujun Ma *, Ruili Wang

School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

ARTICLE INFO

Keywords:
Action recognition
Vision transformer
Transformer
Attention mechanism

ABSTRACT

Action Recognition aims to understand human behavior and predict a label for each action. Recently, Vision Transformer (ViT) has achieved remarkable performance on action recognition, which models the long sequences token over spatial and temporal index in a video. The fully-connected self-attention layer is the fundamental key in the vanilla Transformer. However, the redundant architecture of the vision Transformer model ignores the locality of video frame patches, which involves non-informative tokens and potentially leads to increased computational complexity. To solve this problem, we propose a k-NN attention-based Video Vision Transformer (k-ViViT) network for action recognition. We adopt k-NN attention to Video Vision Transformer (ViViT) instead of original self-attention, which can optimize the training process and neglect the irrelevant or noisy tokens in the input sequence. We conduct experiments on the UCF101 and HMDB51 datasets to verify the effectiveness of our model. The experimental results illustrate that the proposed k-ViViT achieves superior accuracy compared to several state-of-the-art models on these action recognition datasets.

1. Introduction

Action recognition is a prevalent research task in computer vision, which aims to classify an action in a video and develops many real-world applications [1] such as virtual reality (VR), video retrieval, intelligent surveillance, and smart healthcare [2–6]. Many approaches based on convolutional neural networks (CNNs) for action recognition have been proposed in the last decades [7–9].

In recent years, Transformer [10] has been developed as a type of deep neural network that mainly depends on the self-attention mechanism. Inspired by the achievement of Transformer on natural language processing (NLP) [11,12], researchers are driving Transformer as a fresh paradigm shift in computer vision. Later, Dosovitskiy et al. [13] utilized the Transformer network in computer vision with the introduction of the Vision Transformer (ViT) model. ViT was proposed for image classification, which divided an image into multiple non-overlapping patches and then embedded the patches using a sequence of linear projection layers. The embedded patches were the input of the Transformer encoder.

Transformer based models [14–16] for action recognition resembled the patch embeddings of ViT, the individual video frame can be split into small patches and then tokenized. For example, Arnab et al. [17] proposed a Transformer-based model named the Video Vision Transformer (ViViT), an extension of the ViT model tailored action recognition, which was processing a regularized set of spatio-temporal tubelets from the input videos. Touvron et al. [18] introduced a convolution-free Transformer called DeiT for action recognition, which adopted a teacher–student strategy to Transformer and used the distillation tokens to reproduce the label. DeiT model split an image into several patches, and the positional embedding was utilized to the input tokens of the Transformer encoder. Ma et al. [14] introduced a relative positional embedding based spatially and temporally decoupled Transformer model for action recognition, which could mitigate the computational complexity associated with absolute positional embedding and the intertwined learning of spatial–temporal features in patches from input videos. Taking inspiration from this, we construct our model in a spatially and temporally decoupled manner.

It occurs to us that all the aforementioned works [17,19] share a commonality in their utilization of vanilla Transformers, which entails the use of self-attention. However, when Multi-Headed Self-Attention (MSA) modeling pairwise interaction, there has been a quadratic complexity according to the number of tokens. The complexity ratio is related to the augmentation of the input frames. Moreover, the redundant architecture of the vision Transformer model ignores the locality of video frame patches, which involve noisy and non-informative tokens.

We address the issues mentioned above by incorporating k-NN attention [20] into the video Transformer models by selecting top-k similar tokens from the input frames to alleviate the complexity of the attention mechanism. The k-NN attention mechanism extends the local bias of CNNs when the neighboring tokens have a similar tendency while establishing long-range dependencies by picking up the most relevant tokens. Consequently, this approach accelerates the training

* Corresponding author.
E-mail addresses: wsun@massey.ac.nz (W. Sun), yma1@massey.ac.nz (Y. Ma), ruili.wang@massey.ac.nz (R. Wang).

https://doi.org/10.1016/j.neucom.2024.127256
Received 14 December 2022; Received in revised form 23 December 2023; Accepted 8 January 2024
Available online 11 January 2024

process while effectively distilling noisy and non-informative tokens within videos.

In this paper, we refer to this spatial and temporal factorized model, which utilizes k-NN attention, as Dk-ViViT. We have also conducted a comparison between Dk-ViViT and Uk-ViViT, where the latter solely employs k-NN attention without spatial–temporal decoupling.

To the best of our knowledge, we are pioneers in applying k-NN-based attention to the ViViT model, both in terms of development and efficiency analysis of k-NN attention for video-based action recognition. Instead of using the traditional self-attention mechanism, we adopt k-NN attention to the vision Transformer encoder to neglect the noisy tokens and increase the speed of training in Transformer. We conduct experiments on the UCF101 [21] and HMDB51 [22] datasets to verify the effectiveness of our model. The extensive experiments on two popular benchmarks (*i.e.*, UCF101, HMDB51) demonstrate that the proposed k-ViViT achieves higher accuracy and performs better than the several state-of-the-art models both qualitatively and quantitatively.

The rest of this paper is structured as follows. In Section 2, we review the related action recognition models and vision Transformer models. Then, we present our k-ViViT action recognition models in Section 3. Section 4 shows the experiment details of our k-ViViT models. Finally, we conclude this paper in Section 5.

2. Related works

2.1. Action recognition

Action recognition typically models the 2D or 3D [23–26] convolution layers on spatio-temporal data. Before Transformer showed its power in computer vision, CNNs were broadly used to classify complex images on high-dimensional datasets such as ImageNet [27]. Early action recognition models [28,29] started with hand-crafted features to identify the motion information. Since CNNs emerged as the superior recognition method, more and more extension models [30,31] were refined in action recognition. AlexNet [32] is the most outstanding achievement that began the 2D image convolutional networks in video datasets [33–35]. Ma et al. [15] proposed a convolutional transformer network (CTN) for fine-grained action recognition, which utilized 3D convolutions from raw input video clips for extracting the low-level spatial–temporal features. Liu et al. [33] proposed a multi-label learning network under infrared imaging for human facial expression recognition, which learned the expression multi-label through the Cauchy distribution function. Later, Liu et al. [8] proposed a human pose estimation model, which utilized joint direction cues and gaussian coordinate encoding to achieve accurate and flexible performance on different scenarios for infrared images. Meanwhile, Zhang et al. [34] proposed a CNN-based end-to-end learning model for facial expression recognition tasks, which learned the correlation emotion label distribution and associated multiple emotions with facial expressions based on their similarities.

Subsequently, the de-facto choices for action recognition backbones were CNNs [36,37] and RNNs [38]. With the introduction of large datasets like Kinetics [39], 3D CNNs [10,40] tended to be the predominant direction in computer vision, which led to an increase in the number of parameters. multi-cue based four-stream 3D ResNets (MF3D) model was proposed by Wang et al. [23] for action recognition, which consisted of four streams to extract more effective spatio-temporal features. Three connections were injected between these four streams and transferred different cues in the model. Later, Liu et al. [41] proposed a spatial–temporal interaction learning two-stream (STILT) model for action recognition, which developed a spatial–temporal learning module using an alternating co-attention mechanism to learn the interrelation between spatial features and temporal features in videos. Ma et al. [25] introduced a multi-stage factorized spatio-temporal model for RGB-D action recognition, which included a 3D central difference convolution stem module and multiple factorized spatio-temporal stages to capture the fine-grained motions from diverse modality. Zong et al. [42] proposed a spatial and temporal saliency-based four-stream model for action recognition, which adopted the multi-task learning-based LSTM to obtain long-term dependency correlation. Lu et al. [43] proposed a frequency-domain model for compressed action recognition, which utilized a frequency-domain partial decompression method to extract the salient spatial and motion features efficiently from the frequency-domain data. The spatial-to-frequency domain student–teacher network was achieved to reduce the computational cost with acceptable accuracy.

Inspired by the advancements of Transformers in natural language processing [44], action recognition models whether based on a combination of CNNs and Transformers or purely Transformer models gradually gained popularity. Transformers have been proven to be effective in addressing the challenge of capturing long-range dependencies in both spatial and temporal dimensions. Therefore, in this paper, we have adopted the Vision Transformer (ViT) model [13] as our baseline model.

2.2. Vision transformer

Transformer is a novel deep neural network primarily based on the self-attention mechanism and is used in NLP. Transformer consisted of encoder and decoder layers. The encoder was composed of several self-attention blocks and a feed-forward layer. The decoder had a similar architecture to the encoder except for an extra encoder–decoder attention mechanism. The residual connections were used at each layer, followed by layer normalization. According to the strong representation capabilities and promising prospects, numerous studies [45,46] employed Transformer for computer vision. ViT was first proposed by Dosovitskiy et al. [13], which applied a pure-Transformer to the sequences of split image patches.

There are a lot of models integrated Transformer with convolution. For example, Girdhar et al. [47] proposed a convolution-based Transformer model for human action recognition, aggregating features extracted from the relevant regions of the object. Liu et al. [48] proposed an inductive bias of the locality model in video Transformer for action recognition, which computed self-attention globally with spatial–temporal factorization. Another representative CNN-enhanced Transformer-based action recognition model is [45]. Spatio-temporal attention network (STAN) introduced inductive bias vias convolution to achieve computational efficiency. The model extended two-stream Transformer model to learn temporal dependencies for classifying long videos. The model replaced the tokenization method with spatial and temporal information extracted by pre-trained CNNs.

Although most CNN-based models can be integrated into any design and extend their strength, they still suffer in dealing with positional and temporal information in videos. Pure-Transformer model abolished convolution-integrated architecture and was applied to the issue of positional and temporal information. For example, TimeSformer [49] was proposed for action recognition, which allowed spatio-temporal features to be learned from a sequence of frame-level patches. TimeSformer was a convolution-free model built on specific self-attention in space and time dimensions and developed a “divided attention” to separate spatial and temporal attention in each Transformer block. Without using 3D convolution layers in the vision Transformer models, Sharir et al. [50] proposed a global attention temporal Transformer model for action recognition, subsampled from spatial and temporal input, then training to learn spatial and temporal representation from the fraction of input video frames. The model reduced the computational and maintained the accuracy when reviewing a small number of video frames instead of learning the entire video. Arnab et al. [17] developed several variant models to factorize spatial- and temporal-dimensions in the Transformer encoders, such as factorized encoder, factorized self-attention, and factorized dot-Product. The

result of the paper showed that several datasets performed well on unfactorised models. Yan et al. [19] proposed Multiview Transformers for Video Recognition (MTV), which split the encoders for diverse video views and used the lateral connections for information fusion. MTV used multi-head self-attention on the tokens that capture the temporal features.

However, the redundant architecture of ViT ignores the locality of video frame patches, which involves non-informative and noisy tokens and is inefficient during the training procedure. To solve the problems mentioned above, we propose a pure-Transformer model based on the ViViT model for action recognition. The model uses k-NN attention to select the top-k similar tokens and filter out the noisy tokens in the computation of self-attention, further reducing overfitting in relatively small datasets.

3. k-NN Video Vision Transformers

We begin by introducing the ViT model and then extend our discussion to ViViT in Section 3.1. ViViT serves as the foundational architecture for our model. Then k-NN attention will be discussed in Section 3.2, which differs from vanilla self-attention in ViviT. Finally, we present our proposed k-NN attention-based Transformer model for action recognition in Section 3.3.

3.1. Originations: ViT and ViViT

ViT converts the Transformer model from NLP to classify 2D images. Especially, ViT splits up an image, $x_i \in \mathbb{R}^{h \times w \times c}$, into N non-overlapping image patches, $x_p \in \mathbb{R}^{n \times (p^2 \cdot c)}$, c is the number of channels, $h \times w$ is the resolution of the initial image, p^2 is the resolution of the partition image patch. The patches are then linearly projected and rasterized into 1D tokens $z_i \in \mathbb{R}^d$. This tokenization will be forwarded through the following Transformer encoder:

$$z = [z_{cls}, E_{x_1}, E_{x_2}, \ldots, E_{x_N}] + p, \tag{1}$$

where the linear projection E can be seen as a 2D convolution. All the tokens are concatenated to form a sequence with z_{cls}, which is prepended as a learned token [51] at the final layer of the encoder serves. Furthermore, $p \in \mathbb{R}^{N \times d}$ refers to a learned positional embedding, which is also added to the token to preserve the position information.

The differential between ViT and ViViT occurs from embedding video clips. The extension of ViT's embedding to video is ViViT, which extracts a non-overlapping spatio-temporal "tubelet" from the input video and performs a linear projection. Compared to uniformly sample n_t frames from the video, tubelet embedding dimension $t \times h \times w$, $n_t = \left[\frac{T}{t}\right]$, $n_h = \left[\frac{H}{h}\right]$ and $n_w = \left[\frac{W}{w}\right]$, the tokens gain from the temporal, height, and width dimensions separately. Then a total of $n_t \cdot n_h \cdot n_w$ tokens will be passed through into the Transformer encoder. In other words, the tubelet embedding may be considered as constructing 3D blocks, the spatio-temporal information is fused during tokenization, as shown in the left side of Fig. 1. Multi-Headed Self-Attention(MSA) is included in the transformed embedding of vanilla ViT and ViViT models after layer normalization(LN) blocks. Multilayer perceptron (MLP) blocks are also requested in the Transformer encoder after LN blocks and consist of two linear projections separated by a GELU non-linearity [52]. The MSA blocks and MLP blocks are as follows:

$$y^{\ell} = MSA\left(LN\left(z^{\ell}\right)\right) + z^{\ell}, \tag{2}$$

$$z^{\ell+1} = MLP\left(LN\left(y^{\ell}\right)\right) + y^{\ell}, \tag{3}$$

where ℓ denotes each layer in L Transformer encoders.

3.2. Vanilla self-attention and k-NN attention

As the embedding tokens pass through in the Transformer encoder, MSA [10] and LN [53] process the tokens in L times, as shown on the right side of Fig. 1. The residual connections are performed after self-attention blocks. The vanilla attention is dot product attention in Transformer. The attention matrix A is computed based on the pairwise similarity between two elements of the sequence and their query Q and key K representations [13], which is defined as:

$$A = softmax\left(QK^{\top}/\sqrt{d}\right). \tag{4}$$

As shown in Eq. (4), $Q \in \mathbb{R}^{n \times d}$ represents the queries and $K \in \mathbb{R}^{n \times d}$ denominate the keys, the d denotes the dimension. The new value $\hat{V}$ is calculated by multiplying the value V with the attention matrix A, as shown below:

$$\hat{V} = AV. \tag{5}$$

Intuitively, Eq. (5) denotes that the weighted average over the old value, the attention matrix defines the weights in the Transformer. According to the knowledge of vanilla Transformer, the Q, K and V are computed by the linear projection of the input token X:

$$Q = XW_Q,\ K = XW_K,\ V = XW_V, \tag{6}$$

where W_Q, W_K, W_V are learnable weights. The ViViT model achieves success for video action recognition for now. However, the drawback of vanilla self-attention rejects a certain efficiency and accuracy when calculating irrelevant and noisy tokens, even though smaller weights are distributed for the digressive tokens. Fully-connected self-attention takes every token to compute the attention map, including the noisy tokens about cluttered backgrounds and occlusion situations. The issue led to a protracted training process.

The k-NN attention was developed to select the top-k most related keys and values for every query within the self-attention mechanism [20]. Initially, two versions of k-NN attention were introduced: the slow version and the fast version. In the initial method, Euclidean distance was used to calculate the top-k most related keys and values for each query. However, the computational speed is particularly slow due to the requirement to calculate distances between different keys. According to [20], the fast version of k-NN attention leverages the advantage of matrix multiplication operations. It has been empirically validated for its effectiveness across eleven different vision Transformer models. Wang et al. [20] provided empirical evidence showcasing that the proposed k-NN attention mechanism enhanced the classification performance by 0.8% for global and local vision Transformers. Therefore, in this paper, we exclusively discuss the fast version of k-NN attention. The detail of the fast version for k-NN is to select row-wise top-k elements, $\mathcal{T}_k(\cdot)$, in softmax computing, which is shown below:

$$\hat{V}^{knn} = softmax\left(\mathcal{T}_k\left(QK^{\top}/\sqrt{d}\right)\right) \cdot V, \tag{7}$$

$$[\mathcal{T}_k(A)]_{ij} = \begin{cases} A_{ij} & A_{ij} \in top-k(row\ j) \\ -\infty & otherwise. \end{cases} \tag{8}$$

Our proposed model marks the initial achievement of excellent performance in the Video Vision Transformer model using k-NN attention mechanism. The original MSA mechanism in the ViViT model calculates all the tokens embedded by the linearly enhanced input frames. The computational of MSA becomes quadratic complexity. Meanwhile, there are some redundancy input frames in the video datasets, such as cluttered backgrounds, irrelated objects, and input frames in which no action occurs. Thus, we adopt k-NN attention to replace the fully-connected attention in the ViViT model for the first time. According to the details of k-NN attention, we find that k-NN attention collects the top-k similar tokens from the sequence to calculate through the attention map instead of using all the input tokens. Our model is robust in increasing training speed and mitigating the influence of noise from input tokens.

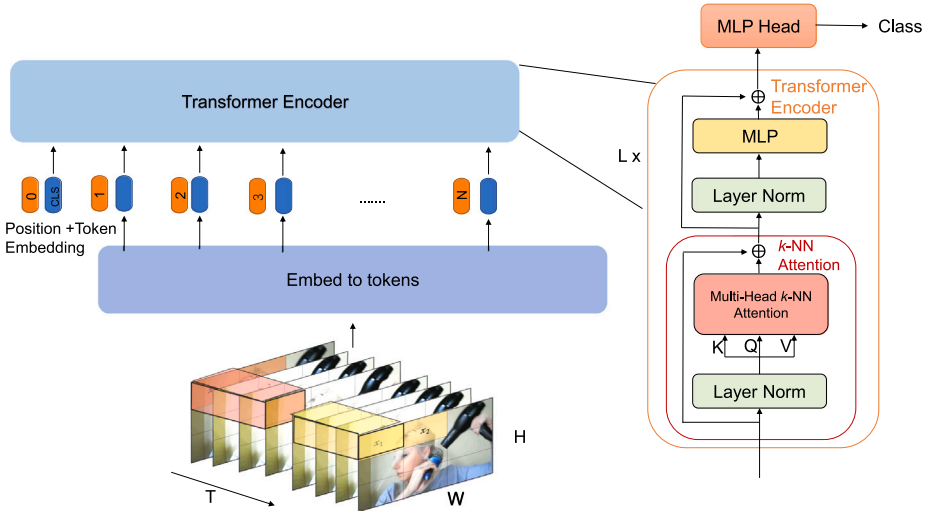

Fig. 1. The framework of the proposed k-NN attention-based pure-Transformer model for action recognition, named Undecomposed k-NN Video Transformer (Uk-ViViT).

3.3. k-NN attention-based Video Vision Transformer

As shown in Fig. 1, we propose a k-NN attention-based pure-Transformer model for action recognition. Our model is inspired by ViViT to extend the spatio-temporal Transformer attention more efficiently. The attention mechanism is also used on the decomposed encoder, as shown in Fig. 2, which enhances network training speed and reduces the noise tokens in videos for action recognition.

3.3.1. Pre-trained models

As shown in Fig. 1, we propose an efficient video vision Transformer for action recognition. According to the massive success of the two datasets in CNN models, we start with data augmentation of the small video dataset, which will be presented in Section 4.

Meanwhile, we initialize the pure-Transformer model from the pre-trained image model for the small video datasets. The dimension of image position embedding is $n_w \cdot n_h \times d$ in the ViT model. However, ViT model is different from our video model with three dimensions, our models have n_t times more tokens than original ViT model. Thus, the position embedding is initialized from $\mathbb{R}^{n_w \cdot n_h \times d}$ to $\mathbb{R}^{n_t \cdot n_w \cdot n_h \times d}$, which provides the same spatial index with the same embedding. According to the 3D tensor in our model, we propose using "central frame initialisation" [17] to initialize the embedding weight with zeros for all temporal positions, except at the center,

$$E = [0, \ldots, E_{image}, \ldots, 0]. \tag{9}$$

Therefore, the model learns temporal information from previous video frames.

3.3.2. Undecomposed k-NN Video Vision Transformer (Uk-ViViT)

We also propose a k-NN attention-based Transformer model, which extends from the first undecomposed model of [17]. We use tubelets embedding to embed the input frames to tokens, which include temporal, height, and width dimensions, then feed into the Transformer encoder. All pairwise interaction is modeled between all spatio-temporal tokens in each Transformer layer so that the undecomposed video Transformer models the long-range interaction through video.

In addition, the fast version of k-NN attention is added after LN blocks in each Transformer layer. Our model selects the row-wise top-k elements for softmax computing to reduce the noisy tokens and speed up the training process. The mechanism only considers the top-k most similar patches, which means choosing a proper k value with high similarity of patches is significant. Though the fully-connected self-attention has the ability to capture long-range dependency in ViViT, the defect is mixing the irrelevant patches during the dot-product attention, which makes a slow training process. Therefore, the proposed Uk-ViViT model influences the convergence speed of the vital visual patches relevant to the target class.

3.3.3. Decomposed k-NN encoder (Dk-ViViT)

This model comprises two Transformer encoders, which are decomposed into the spatial encoder and the temporal encoder. Firstly, the tokens extracted from the same temporal index input the spatial encoder, as shown in Fig. 2. The dot product attention is replaced by k-NN attention in the spatial encoder, which only models the interaction among tokens derived from identical temporal indices. Then, each temporal index representation, $h_i \in \mathbb{R}^d$, is provided after L_s spatial layer. $z_{cls}^{L_s}$ is the prepend of the cls token to the input of the temporal encoder. The frame-level representations, h_i, are concatenated into $H \in \mathbb{R}^{n_t \times d}$, then pass through k-NN attention after layer normalization. The temporal encoder consists of k-NN attention and MLP in L_t Transformer layers to model interaction between tokens in different temporal indices. Then, the output is classified at the end of our model. Though the Dk-ViViT model could have more Transformer layers than Uk-ViViT, Dk-ViViT achieves fewer floating point operations and reduces the mix of irrelevant image patches by adding k-NN attention mechanism, *e.g.,* the background information or the other objects excluded the main objects.

4. Experimental setup

4.1. Data augmentation

We adopt the random crop, flip, and color jitter as the data augmentation strategy to increase the number of training samples. Random crops can especially enhance the accuracy and stability of models. We adopt Mixup [54] as the optimization, which is a learning principle to relieve the problem of large memorizing corrupt labels. Our augmentation strategy consists of random crop, flip, color jitter, and other regularizations, such as Kinetics-400 initialisation, stochastic depth [55], random augment [56], label smoothing [57], and Mixup [54], to enhance the robustness and accuracy of Transformer in small video datasets.

4.2. Model configuration and training

Our backbone network follows ViViT, and the pre-trained modal is the same as ViT. We set $L = 12$ as the number of Transformer layers

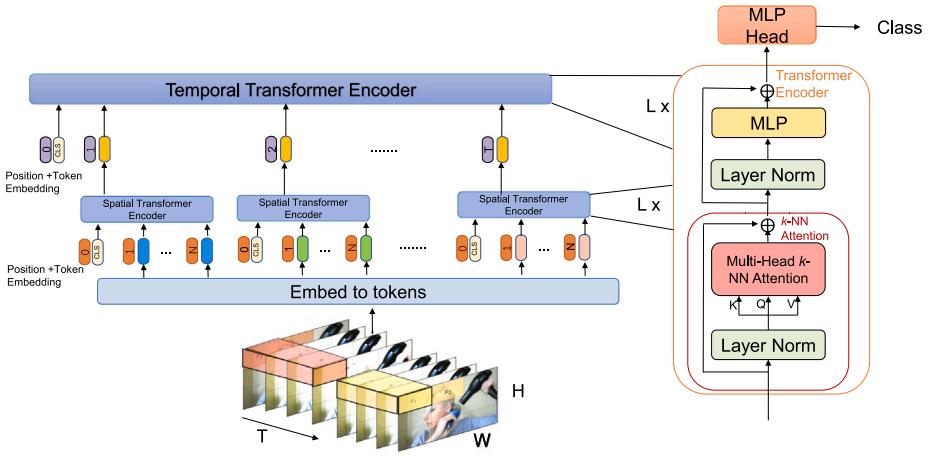

Fig. 2. The proposed decomposed k-NN encoder (Dk-ViViT), which is separated into two Transformer encoders: spatial Transformer encoder and temporal Transformer encoder.

in Uk-ViViT, $L_s = 8$ as the number of spatial Transformer layers in Dk-ViViT, and $L_t = 4$ as the number of temporal Transformer layers in Dk-ViViT. The batch size is 8 through 160 epochs. The number of input frames is 8 with the frame interval of 32. We also use the same naming scheme for these models (e.g., k-ViViT/16 × 2 denotes that ViViT is the backbone and the tubelet size is $h \times w \times t = 16 \times 16 \times 2$). Note that the height and width are equal in tubelet. Thus, the small tubelet size will contain more tokens, further increasing the computation. k is the only extremely significant parameter in k-NN attention, the general rule of k value is that assign k to around $\frac{n}{2}$ at each scale stage for simple token generation methods. The complicated token generation methods will use $\frac{2}{3}n$ or $\frac{4}{5}n$ at each scale stage, n denotes the total number of tokens. We use synchronous SGD with a momentum of 0.9, and a cosine learning rate schedule with linear warm-up during the fine-tuning is based on [17] for 160 epochs. Unfortunately, due to limitations in GPU memory and processing capacity, we were unable to conduct testing of our model on the Kinetics dataset [39]. Nevertheless, it is important to note that our model has demonstrated outstanding performance on two well-known action recognition datasets. These datasets encompass a wide range of actions, including human–human interactions and human–object interactions.

4.2.1. Datasets

We use two public action recognition datasets to evaluate our model, UCF101 [21] and HMDB51 [22].

UCF101 [21] is an action recognition dataset containing 13,320 realistic action videos from YouTube and is assigned to 101 classes, over 13k clips, and 27 h of video. It includes diverse actions boasting massive variations in camera movement [58], object details, and different environments. The dataset is briefly divided into five categories, and some clips from the same video are shown. The training dataset contains approximately 9.4k videos, and the testing dataset has around 3.7k action videos.

HMDB51 [22] consists of 6.8k action videos and 51 action classes, primarily collected from movies and YouTube. The split methods are similar to the UCF101 datasets. HMDB51 has split the datasets into training, testing, and validating datasets. We adopt the average accuracy from the three splits as the final accuracy for the action recognition in the experiments.

4.3. Ablation study

4.3.1. Input tokenization

We consider the efficiency of different input encoding methods using Uk-ViViT on HMDB51. When the frame interval is 32, we sample 8 frames and set the tubelets of length $t = 4$. We propose our model on different input encoding methods. As Table 1 shows that "central frame initialisation" gets the best performance than "uniform frame sampling" and "filter inflation" [40]. The method "uniform frame sampling" is provided by ViT, which samples the frame and embeds the 2D frame independently. However, "central frame initialisation" considers the 3D convolutional filters to extract tubelets from input videos for action recognition. Therefore, we consider using "central frame initialisation" encoding method for all the experiments.

Table 1
Comparison of an input encoding method using Uk-ViViT model on HMDB51.

Encoding methods	Top-1 accuracy(%)
Uniform frame sampling	78.5
Filter inflation	79.3
Central frame initialization	80.2

Table 2
The top-1 accuracy tendency to vary the number of tokens on HMDB51.

Tubelets size	Uk-ViViT (Top-1 Acc%)	Dk-ViViT (Top-1 Acc%)
16 × 8	66.7	69.4
16 × 4	73.1	75.2
16 × 2	80.2	82.5

4.3.2. Varying the k value of k-NN attention

We first process the result of the accuracy in the different numbers of tokens for the temporal dimension in Table 2. We notice that using smaller tubelet sizes has higher accuracy through all our models.

According to the state-of-the-art results in ViViT, they consistently use a spatial resolution of 224. Thus, we input the resolution of 224 x 224 in our model. Then we adopt the appropriate number of tokens and testify the value of k. Intuitively, we compare the accuracy of Uk-ViViT by using different k values, $\frac{n}{2}$, $\frac{2}{3}n$ *or* $\frac{4}{5}n$ denote the computational k value on HMDB51 and UCF101. We consider using Uk-ViViT/16x2 and Dk-ViViT/16x2 as the input of Transformer encoder, the comparison is shown in Table 3. In addition, the value of k is the same between the spatial encoder and temporal encoder for the Dk-ViViT model. We compare our proposed model in the same scale stages of k on HMDB51 and UCF101.

According to the results from Section 3, we notice that k value between $\frac{2}{3}n$ and $\frac{4}{5}n$ do not show significant different. However, the accuracy gap between $\frac{n}{2}$ and $\frac{2}{3}n$ is large.

4.3.3. Model variants

We consider our proposed model variants across UCF101 and HMDB51, both in efficiency and accuracy, in Table 4. We use ViViT as the backbone and tubelet size of 16 × 2 in all the models shown

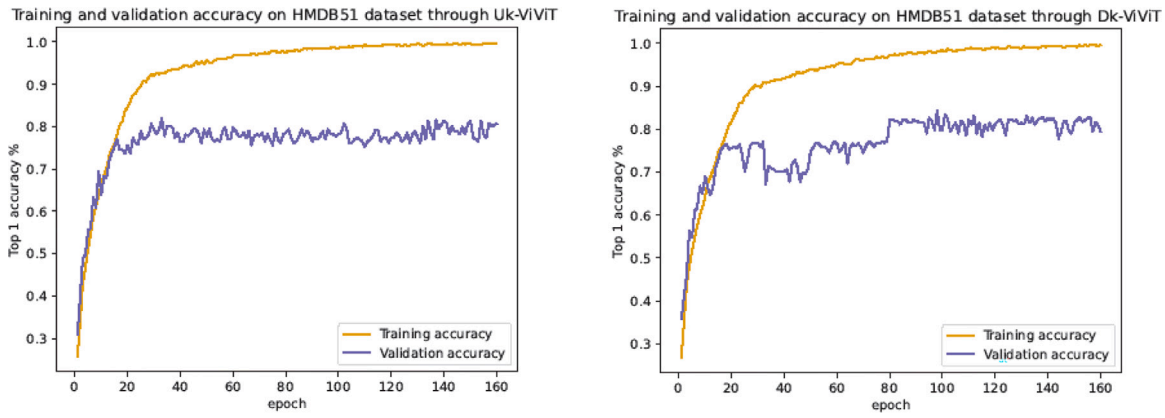

Fig. 3. The training and validation accuracy on HMDB51 through Uk-ViViT (left) and Dk-ViViT(right).

Table 3
The effect of different k values through the Uk-ViViT model and Dk-ViViT model on HMDB51 and UCF101.

k value	HMDB51		UCF101	
	Uk-ViViT (Top-1 Acc%)	Dk-ViViT (Top-1 Acc%)	Uk-ViViT (Top-1 Acc%)	Dk-ViViT (Top-1 Acc%)
$\frac{n}{2}$	79.3	80.4	87.2	90.3
$\frac{2}{3}n$	80.2	82.5	90.6	94.2
$\frac{4}{5}n$	80.0	81.0	88.4	91.5

Table 4
Comparison of model architectures based on ViViT, the tubelet size of 16 × 2. We report Top-1 accuracy and the number of parameters for different models and compare them between the backbone ViViT model and the variant model.

Models	k-NN	UCF101	HMDB51	Params ($\times 10^6$)
ViViT - model 1 [17]	×	76.2	73.2	89.9
Uk-ViViT	✓	90.6	80.2	110.6
ViViT - model 2 [17]	×	78.7	75.0	99.8
Dk-ViViT	✓	94.2	82.5	123.3

in Table 4. The number of Transformer layers to be set in the original ViViT (spatio-temporal) model is the same as Uk-ViViT, which is set to 12. Meanwhile, we set the number of spatial Transformer layers to 8 in factorized encoder model. In addition, the number of temporal Transformers is set to 4 in Factorized encoder and Decomposed k-NN encoder. One notable disadvantage is that the input frame patches inherently include non-informational tokens and extraneous information, potentially causing misinterpretation in action recognition. Therefore, we proposed k-NN-based attention spatial and temporal factorized model to select the most relevant tokens from the entire frame, which can speed up the training process and improve the accuracy of action inference. According to the result in Table 4, the Decomposed k-NN encoder performs the best on UCF101 and HMDB51. We compare the result of Top-1 accuracy between ViViT (model 1, model 2) [17], Uk-ViViT, and Dk-ViVIT on UCF101 and HMDB51. The experimental results demonstrate the effectiveness of k-NN attention on Dk-ViViT for action recognition. As shown in Table 4, our proposed model improves 7.5% and 15.4% on HMDB51 and UCF101 for decomposing the spatial encoder and the temporal encoder, respectively. This demonstrates that the k-NN attention mechanism can further improve the accuracy based on ViViT model even with dropping the irrelevant tokens.

According to the shape of Fig. 3, we find that our proposed models have a fast convergence rate on both training and validation over the HMDB51 dataset, Fig. 3 (left) also illustrates the training loss of Uk-ViViT model increase following the number of epochs. Dk-ViViT model has a similar loss value when the model processes half epochs in Fig. 3 (right). To sum up, Dk-ViViT model outperforms the initial ViViT model and Uk-ViViT model on HMDB51 and UCF101.

Table 5
Comparison of k-NN attention in spatial and temporal domains using Top-1 accuracy.

Models variants	UCF101	HMDB51
Dk-ViViT-S	93.1	81.9
Dk-ViViT-T	92.3	80.7
Dk-ViViT	94.2	82.5

4.3.4. *Effectiveness in spatial and temporal domain*

We have conducted additional ablation studies to evaluate the effectiveness of the k-NN attention mechanism in both spatial and temporal domains by applying it separately to the spatial and temporal encoders of Dk-ViViT. We have detailed the configuration and results in Table 5. Specifically, Dk-ViViT-S refers to the utilization of k-NN attention exclusively in spatial encoders, with temporal encoders retaining self-attention. On the other hand, Dk-ViViT-T denotes the use of k-NN attention exclusively in temporal encoders, while spatial encoders continue to employ self-attention. Dk-ViViT represents the implementation of k-NN attention in both spatial and temporal domains. As shown in Table 5, Dk-ViViT has achieved the highest classification score among all four configurations, indicating the effectiveness of k-NN in both spatial and temporal domains. We believe that in the spatial domain, k-NN can effectively filter out noisy and non-informative tokenized patches, while in the temporal domain, it excels at selecting the most action-relevant frames. Therefore, in our final configuration of Dk-ViViT, we have applied k-NN attention to both the spatial and temporal encoders.

4.4. *Comparison to state-of-the-art*

We compare the performance of several state-of-the-art models for action recognition on UCF101 and HMDB51, as shown in Table 6. The comparison is listed based on three aspects: the type of backbone, the year of publishing, and the accuracy of Top-1. The comparison methods include models built on CNNs and models constructed on pure-Transformer.

We initially employ the pre-trained models and subsequently fine-tune the weight parameters on the HMDB51 and UCF101 datasets. While it is worth noting that our models outperform several pure-Transformer models on HMDB51 and UCF101 datasets. According to Table 6, we compared our model with the state-of-the-art model which was released in 2023. The TTSN model [66] only employed a temporal transformer encoder and used ResNet-50 as a hybrid backbone of the model, which achieved relatively higher accuracy on UCF101. However, the performance on the HMDB51 dataset is not as remarkable, and we attribute this to the insufficiency of the spatial feature learning of the model.

Table 6
Comparisons to the state-of-the-art through multiple methods on UCF101 and HMDB51 for action recognition.

Method	Backbone	Year	HMDB51 (Top-1 Acc%)	UCF101 (Top-1 Acc%)
Two stream [59]	CNN	2014	59.4	88.0
C3D [60]	CNN	2015	–	85.2
L^2STM [61]	CNN	2017	66.2	93.6
T3D+TSN [62]	CNN	2017	63.5	93.2
VideoLSTM [63]	CNN	2018	56.4	89.2
I3D [40]	CNN	2018	66.4	93.4
STS [64]	CNN	2020	62.4	90.1
SVT [16]	Transformer	2022	67.2	93.7
SCT-L [65]	Transformer	2022	**84.6**	**98.7**
TTSN [66]	CNN+Transformer	2023	80.2	96.8
U*k*-ViViT (ours)	Transformer	–	80.2	90.6
D*k*-ViViT (ours)	Transformer	–	82.5	94.2

SCT-L [65] achieved the highest accuracy on HMDB51 and UCF101 attributed to its pre-training on Kinetic-400 before fine-tuning on HMDB51 and UCF101. This highlights the significance of initializing the model with a large-scale training set, a practice that significantly enhances the capabilities of the Transformer model. In summary, the existing results prove the effectiveness of our proposed *k*-ViViT model for action recognition. Our proposed D*k*-ViViT model obtains 82.5% accuracy on HMDB51, which surpasses the I3D model by 16.1%, surpasses the STS model by 20.1%, surpasses the SVT model by 15.3%, and surpasses the TTSN model by 2.3%. In addition, our proposed D*k*-ViViT model obtains 94.2% on UCF101, which surpasses the I3D model by 0.8%, surpasses the STS model by 4.1%, and surpasses the SVT model by 0.5%. Our proposed D*k*-ViViT model outperforms several concurrent models based on CNNs and pure-Transformer.

5. Conclusion and future work

In this paper, we propose a *k*-NN attention-based video vision Transformer (*k*-ViViT) model for action recognition, which speeds up the training process and neglects the irrelevant or noisy tokens in the input video. The proposed models outperform various state-of-the-art results achieved by a pure-Transformer model on two action recognition benchmarks while maintaining computational efficiency. Furthermore, we testify to an efficient data augmentation strategy to improve performance. Our future work will focus on the development of self-supervised pre-trained Transformer models for action recognition, as well as for more complex tasks such as video captioning and sign language translation.

CRediT authorship contribution statement

Weirong Sun: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft. **Yujun Ma:** Conceptualization, Investigation, Software, Writing – review & editing. **Ruili Wang:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Z. Chen, R. Wang, Z. Zhang, H. Wang, L. Xu, Background–foreground interaction for moving object detection in dynamic scenes, Inform. Sci. 483 (2019) 65–81.
[2] W. Guo, G. Chen, Human action recognition via multi-task learning base on spatial–temporal feature, Inform. Sci. 320 (2015) 418–428.
[3] Y. Tian, Z. Xu, Y. Ma, W. Ding, R. Wang, Z. Gao, G. Cheng, L. He, X. Zhao, Survey on deep learning in multimodal medical imaging for cancer detection, Neural Comput. Appl. (2023) 1–16.
[4] W. Ji, R. Wang, Y. Tian, X. Wang, An attention based dual learning approach for video captioning, Appl. Soft Comput. 117 (2022) 108332.
[5] L. Kong, G. Li, W. Rafique, S. Shen, Q. He, M.R. Khosravi, R. Wang, L. Qi, Time-aware missing healthcare data prediction based on ARIMA model, IEEE/ACM Trans. Comput. Biol. Bioinform. (2022).
[6] J. Guo, P. Yi, R. Wang, Q. Ye, C. Zhao, Feature selection for least squares projection twin support vector machine, Neurocomputing 144 (2014) 174–183.
[7] H. Liu, C. Zhang, Y. Deng, B. Xie, T. Liu, Z. Zhang, Y.-F. Li, TransIFC: invariant cues-aware feature concentration learning for efficient fine-grained bird image classification, IEEE Trans. Multimed. (2023).
[8] H. Liu, T. Liu, Y. Chen, Z. Zhang, Y.-F. Li, EHPE: Skeleton cues-based gaussian coordinate encoding for efficient human pose estimation, IEEE Trans. Multimed. (2022).
[9] T. Liu, H. Liu, B. Yang, Z. Zhang, LDCNet: Limb direction cues-aware network for flexible human pose estimation in industrial behavioral biometrics systems, IEEE Trans. Ind. Inform. (2023).
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[11] F. Hou, R. Wang, J. He, Y. Zhou, Improving entity linking through semantic reinforced entity embeddings, 2021, arXiv preprint arXiv:2106.08495.
[12] Y. Liu, X. Yuan, X. Jiang, P. Wang, J. Kou, H. Wang, M. Liu, Dilated adversarial U-net network for automatic gross tumor volume segmentation of nasopharyngeal carcinoma, Appl. Soft Comput. 111 (2021) 107722.
[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[14] Y. Ma, R. Wang, Relative-position embedding based spatially and temporally decoupled transformer for action recognition, Pattern Recognit. 145 (2024) 109905.
[15] Y. Ma, R. Wang, M. Zong, W. Ji, Y. Wang, B. Ye, Convolutional transformer network for fine-grained action recognition, Neurocomputing (2023) 127027.
[16] K. Ranasinghe, M. Naseer, S. Khan, F.S. Khan, M.S. Ryoo, Self-supervised video transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2874–2884.
[17] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836–6846.
[18] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.
[19] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun, C. Schmid, Multiview transformers for video recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3333–3343.
[20] P. Wang, X. Wang, F. Wang, M. Lin, S. Chang, H. Li, R. Jin, Kvt: k-nn attention for boosting vision transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 285–302.
[21] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, 2012, arXiv preprint arXiv:1212.0402.
[22] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2556–2563.
[23] L. Wang, X. Yuan, M. Zong, Y. Ma, W. Ji, M. Liu, R. Wang, Multi-cue based four-stream 3D ResNets for video-based action recognition, Inform. Sci. 575 (2021) 654–665.
[24] T. Jiang, M. Zong, Y. Ma, F. Hou, R. Wang, MobileACNet: Acnet-based lightweight model for image classification, in: International Conference on Image and Vision Computing New Zealand, Springer, 2022, pp. 361–372.
[25] Y. Ma, B. Zhou, R. Wang, P. Wang, Multi-stage factorized spatio-temporal representation for RGB-D action and gesture recognition, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 3149–3160.
[26] G. Wang, Y. Zhou, Z. He, K. Lu, Y. Feng, Z. Liu, G. Wang, Knowledge-guided pre-training and fine-tuning: Video representation learning for action recognition, Neurocomputing (2023) 127136.
[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[28] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2005) 107–123.
[29] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, Int. J. Comput. Vis. 103 (2013) 60–79.

[30] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[32] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[33] T. Liu, J. Wang, B. Yang, X. Wang, Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom, Infrared Phys. Technol. 112 (2021) 103594.
[34] Z. Zhang, C. Lai, H. Liu, Y.-F. Li, Infrared facial expression recognition via Gaussian-based label distribution learning in the dark illumination environment for human emotion detection, Neurocomputing 409 (2020) 341–350.
[35] X. Liu, T. Liu, J. Zhou, H. Liu, High-resolution facial expression image restoration via adaptive total variation regularization for classroom learning environment, Infrared Phys. Technol. 128 (2023) 104482.
[36] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[37] A. Zhou, Y. Ma, W. Ji, M. Zong, P. Yang, M. Wu, M. Liu, Multi-head attention-based two-stream EfficientNet for action recognition, Multimedia Syst. 29 (2) (2023) 487–498.
[38] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: Deep networks for video classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4694–4702.
[39] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, 2017, arXiv preprint arXiv:1705.06950.
[40] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
[41] T. Liu, Y. Ma, W. Yang, W. Ji, R. Wang, P. Jiang, Spatial-temporal interaction learning based two-stream network for action recognition, Inform. Sci. (2022).
[42] M. Zong, R. Wang, Y. Ma, W. Ji, Spatial and temporal saliency based four-stream network with multi-task learning for action recognition, Appl. Soft Comput. (2022) 109884.
[43] L. Xiong, X. Jia, Y. Ming, J. Zhou, F. Feng, N. Hu, Faster-fCoViAR: Faster frequency-domain compressed video action recognition, 2021.
[44] R. Wang, F. Hou, S. Cahan, L. Chen, X. Jia, W. Ji, Fine-grained entity typing with a type taxonomy: a systematic review, IEEE Trans. Knowl. Data Eng. (2022).
[45] E. Fish, J. Weinbren, A. Gilbert, Two-stream transformer architecture for long video understanding, 2022, arXiv preprint arXiv:2208.01753.
[46] D. Damen, H. Doughty, G.M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al., Rescaling egocentric vision, 2020, arXiv preprint arXiv:2006.13256.
[47] R. Girdhar, J. Carreira, C. Doersch, A. Zisserman, Video action transformer network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 244–253.
[48] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3202–3211.
[49] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding? in: ICML, 2021, p. 4.
[50] G. Sharir, A. Noy, L. Zelnik-Manor, An image is worth 16x16 words, what is a video worth? 2021, arXiv preprint arXiv:2103.13915.
[51] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[52] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), 2016, arXiv preprint arXiv:1606.08415.
[53] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, 2016, arXiv preprint arXiv:1607.06450.
[54] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, 2017, arXiv preprint arXiv:1710.09412.
[55] G. Huang, Y. Sun, Z. Liu, D. Sedra, K.Q. Weinberger, Deep networks with stochastic depth, in: European Conference on Computer Vision, Springer, 2016, pp. 646–661.
[56] E.D. Cubuk, B. Zoph, J. Shlens, Q.V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 702–703.
[57] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[58] C. Jing, J. Potgieter, F. Noble, R. Wang, A comparison and analysis of RGB-D cameras' depth performance for robotics application, in: 2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP), IEEE, 2017, pp. 1–6.
[59] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. 27 (2014).
[60] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, Int. Conf. Comput. Vis. 2015 (2015) 4489–4497.
[61] L. Sun, K. Jia, K. Chen, D.-Y. Yeung, B.E. Shi, S. Savarese, Lattice long short-term memory for human action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2147–2156.
[62] A. Diba, M. Fayyaz, V. Sharma, A.H. Karami, M.M. Arzani, R. Yousefzadeh, L. Van Gool, Temporal 3d convnets: New architecture and transfer learning for video classification, 2017, arXiv preprint arXiv:1711.08200.
[63] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, C.G. Snoek, Videolstm convolves, attends and flows for action recognition, Comput. Vis. Image Underst. 166 (2018) 41–50.
[64] Z. Liu, Z. Li, R. Wang, M. Zong, W. Ji, Spatiotemporal saliency-based multi-stream networks with attention-aware LSTM for action recognition, Neural Comput. Appl. 32 (18) (2020) 14593–14602.
[65] X. Zha, W. Zhu, L. Xun, S. Yang, J. Liu, Shifted chunk transformer for spatio-temporal representational learning, Adv. Neural Inf. Process. Syst. 34 (2021) 11384–11396.
[66] Y. Zhang, J. Li, N. Jiang, G. Wu, H. Zhang, Z. Shi, Temporal transformer networks with self-supervision for action recognition, IEEE Internet Things J. (2023).

Weirong Sun received a M.S. degree from The University of Auckland, Auckland, New Zealand, in 2020. She has been working toward a Ph.D. with the School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand. Her research interests include deep learning and video classification.

Yujun Ma received a M.S. degree from Hohai University, Nanjing, China, in 2019. She has been working toward a Ph.D. degree with the School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand, since 2020. Her research interests include action recognition and video analysis.

Ruili Wang received a Ph.D. degree in computer science from Dublin City University, Dublin, Ireland. He is currently the Professor of Artificial Intelligence and Chair of Research in the School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand, where he is the Director of the Centre of Language and Speech Processing. His current research interests include speech processing, language processing, video processing, data mining, and intelligent systems. Dr. Wang serves as a member and an Associate Editor of the editorial boards for international journals, such as the journals of IEEE Transactions on Emerging Topics in Computational Intelligence, Knowledge and Information Systems, Neurocomputing, and Applied Soft Computing. He was a recipient of the most prestigious research grants in New Zealand (i.e., the Marsden Fund) and the New Zealand-Singapore Data Science Research Programme Fund.