DirecFormer: A Directed Attention in Transformer Approach to Robust Action Recognition

Thanh-Dat Truong[1], Quoc-Huy Bui[2], Chi Nhan Duong[3], Han-Seok Seo[4]
Son Lam Phung[5], Xin Li[6], Khoa Luu[1]
[1]CVIU Lab, University of Arkansas [2]NextG, FPT Software
[3]Concordia University [4]Dep. of Food Science, University of Arkansas
[5]University of Wollongong [6]West Virginia University
tt032, khoaluu, hanseok @uark.edu, bqhuy@apcs.fitus.edu.vn
dcnhan@ieee.org, phung@uow.edu.au, Xin.Li@mail.wvu.edu

Abstract

Human action recognition has recently become one of the popular research topics in the computer vision community. Various 3D-CNN based methods have been presented to tackle both the spatial and temporal dimensions in the task of video action recognition with competitive results. However, these methods have suffered some fundamental limitations such as lack of robustness and generalization, e.g., how does the temporal ordering of video frames affect the recognition results? This work presents a novel end-to-end Transformer-based Directed Attention (DirecFormer) framework[1] for robust action recognition. The method takes a simple but novel perspective of Transformer-based approach to understand the right order of sequence actions. Therefore, the contributions of this work are threefold. Firstly, we introduce the problem of ordered temporal learning issues to the action recognition problem. Secondly, a new Directed Attention mechanism is introduced to understand and provide attentions to human actions in the right order. Thirdly, we introduce the conditional dependency in action sequence modeling that includes orders and classes. The proposed approach consistently achieves the state-of-the-art (SOTA) results compared with the recent action recognition methods [4, 18, 72, 74], on three standard large-scale benchmarks, i.e. Jester, Kinetics-400 and Something-Something-V2.

1. Introduction

Video understanding has recently become one of the popular research topics in the computer vision community. Video data has become ubiquitous and occurs in numerous daily activities and applications, e.g., movies and camera surveillance [15, 38, 47, 61]. In the field of video understanding [60], action recognition has become a fundamental problem. In action recognition, there is a need to pay more attention to the temporal structures of the video sequences. Indeed, emphasis on temporal modeling is a common strategy among most methods. It can be considered as the main difference between video and images. These works include long-short term dependencies [14, 16], temporal structure, low-level motion, and action modeling as a sequence of events or states.

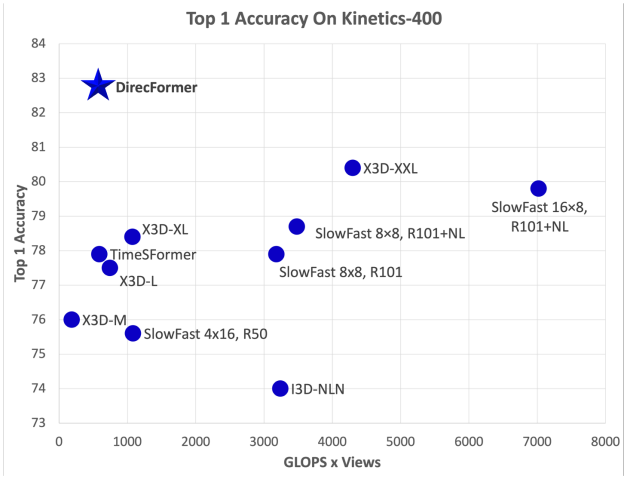

Figure 1. **Result Preview.** Top 1 accuracy against GLOPS Views of our DirecFormer compared to other methods. DirecFormer achieves SOTA performance while maintaining the low computational cost.

The current methods in video action recognition utilize 3D or pseudo 3D convolution to extract the spatio-temporal features [7, 50, 57, 70]. However, these 3D CNN-based methods suffer from intensive computation with many parameters to be learned. Others attempt to adopt two-stream structures [19, 20, 22, 53] for accurate action recognition, since information from one branch could be fused to the

[1]The implementation of DirecFormer is available at https://github.com/uark-cviu/DirecFormer

other one. Some methods in this category require computing the optical flow first, which could be time-consuming and requires a large amount of storage. Others apply 3D convolution to avoid computing the optical flow. Nonetheless, this approach also requires a large amount of computational resources to implement.

Although prior methods [1, 43, 48, 63, 66] have achieved remarkable performance, they have several limitations related to the robustness of the models. In this paper, we therefore address two fundamental questions for current action recognition models. In the first question, given a set of video frames *shuffled in a random order* and different from the original one, will it be classified as the same label as the original recognition result? If it is the case, these models have been clearly overfitted or biased to other factors (e.g. scene background), rather than learned semantic information of the actions. In the second question, we want to understand whether these action recognition models are able to *correct the incorrectly-ordered frames* to the right ones and provide an accurate prediction? Finally, we introduce a new theory to improve the robustness and generalization of the action recognition models.

1.1. Contributions of this Work

In this work, we present a new end-to-end Transformer-based Directed Attention (DirecFormer) approach to robust action recognition. Our method takes a simple but novel perspective of Transformer-based approach to learn the right order of a sequence of actions. The contributions of this work are three-fold. First, we introduce the problem of ordered temporal learning in action recognition. Second, a new Directed Attention mechanism is introduced to provide human action attentions in the right order. Third, we introduce the conditional dependency in action sequence modeling that includes orders and classes. The proposed approach consistently achieves the State-of-the-Art results compared to the recent methods [2, 44, 72] on three standard action recognition benchmarks, i.e. Jester [45], Something-in-Something V2 [23] and Kinetics-400 [31], as in Fig. 1.

2. Related Work

Video Action Classification. In recent years, video understanding has become a popular topic in Computer Vision due to its promising applications such as robotics, autonomous driving, camera surveillance or human behavior analysis. In the early days, many traditional approaches used hand-crafted features as a method to encode information of video sequences [11, 33, 36, 37, 52, 62, 67]. Among these approaches, iDT [62] achieved very good performance by utilizing dense trajectory features and became one of the most popular hand-designed methods.

Later, with the success of deep learning architecture [25, 34, 54–56] using computing hardware, i.e., GPU, TPU, and the introduction of various large-scale datasets, e.g. Sport1M [30], Kinetics [6, 31] and AVA [24], video understanding, especially video action recognition, has become easier to approach in the research community, resulting in an introduction of a series of deep learning frameworks. These methods mainly focus on learning spatio-temporal representations in an end-to-end classification manner. [12] proposed to model the temporal relationship using LSTM [26] to corporate 2D CNN features. [30] presented an approach that fuses the information from the temporal dimension while suggesting applying a single 2D CNN to each frame of the video sequence. However, this method cannot handle well the motion change and perform weaker than the hand-crafted features methods.

The remaining approaches for video action recognition could be divided into two categories. The first group contains models that adopt the conventional two-stream structure [53] to improve the temporal modeling capability. A spatial 2D CNN is used to learn semantic features and a temporal 2D CNN is applied in the other branch to analyze the motion content of the video sequences using the optical flow as input. Both streams are trained in parallel and the scores are averaged to make the final predictions. [20–22] studied different combinations to fuse spatio-temporal information between both streams. TSN [64] proposed sampling sparse frames from evenly divided segments of the video clip to capture long-range dependencies. These dual-path methods require additional computation of the optical flow, which is time-consuming and demand a considerable amount of storage. However, our proposed method can operate without the need of optical flow modality, thus, reducing the complexity of the network.

The second category for action recognition is 3D CNN based methods and the (2+1)D CNN variants. C3D [57] was the first work to apply 3D convolutions to model the spatial and temporal features together. I3D [7] was proposed to inflate 2D convolutional kernels into 3D to capture spatio-temporal features. However, the major drawback of 3D CNNs is the large number of parameters involved. To cope with the intensive computation of 3D CNN, various methods adopted the 2D + 1D paradigm. P3D [50] decomposes 3D convolution into a pseudo-3D convolutional block. R(2+1)D [59] and S3D-G [70] factorize the 3D convolution to enhance accuracy and reduce the complexity. TRN [76] introduced an interpretable relational module to replace the average pooling operation. TSM [42] shifts part of the features forward and backward along the temporal dimension, allowing the network to achieve the performance of 3D CNN but maintains the complexity of 2D CNN. Non-local neural network [65] proposed a special non-local operation for better capturing the long-range temporal dependencies between video frames. SlowFast [19] adopted a dual-path network to model the spatio-temporal information

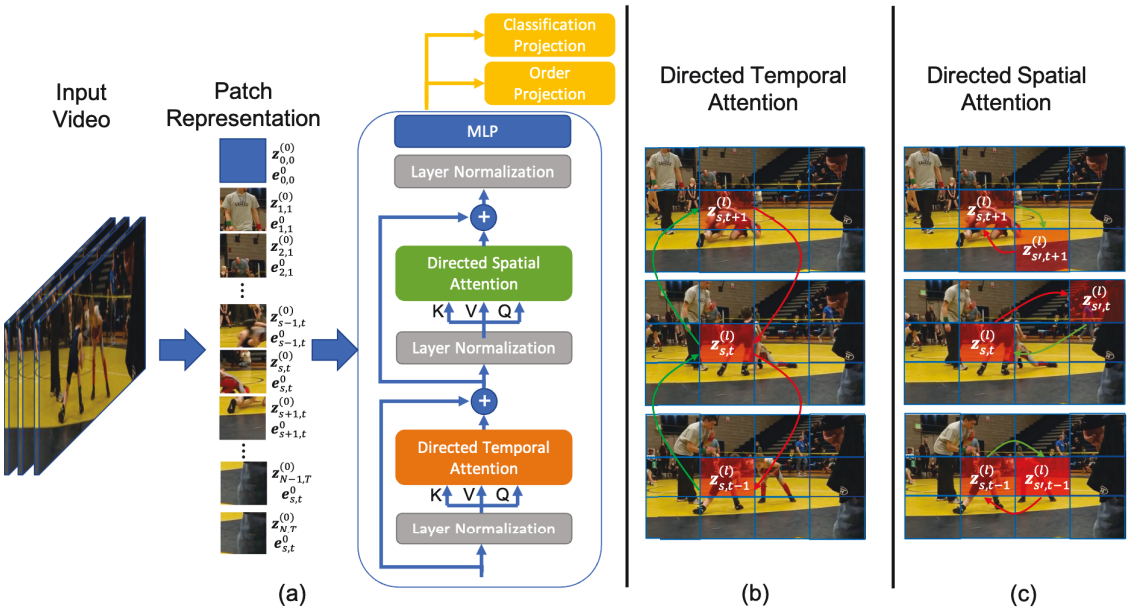

Figure 2. **The Proposed Framework.** (a) The Proposed DirecFormer. (b) Directed Temporal Attention. (c) Directed Spatial Attention. The green arrows in (b) and (c) denotes for positive correlation and the red arrows denotes for negative correlation.

at two different temporal rates, with mid-level features being interactively fused. In general, our method also learns to approximate the spatio-temporal representation at the feature level with the help of the knowledge distillation process.

More recently, significant improvement in terms of efficiency has been reported for action recognition in [8]; it was found that 2D-CNN and 3DCNN models behave similarly in terms of spatio-temporal representation ability and transferability. More efficient action recognition can be achieved by focusing more on making the most of selected frames by dynamic knowledge propagation [32] or exploiting spatio-temporal self-similarity [35]. Most recent works such as elastic semantic network (Else-Net) [40] and memory attention network (MAN) [39] also reported promising improvement in terms of recognition accuracy.

Video Ordering Several prior works [27, 46, 71] have considered the frame orders into account. Although these prior works have partially addressed some aspects of order prediction, their losses only provide a weak supervision, i.e. binary label for in- or out-of-order [27,46] or sub-clip based order [71]. Moreover, there is no explicit mechanism to enforce the model focus on the motion information rather than particular background information of the scene.

Video Transformer Transformer approaches have filled an important role by acquiring competitive accuracy while maintaining computational resources compared with the traditional convolution method. A pure-transformer based model (ViViT) was demonstrated in [2] handling spatio-temporal tokens from a long sequence of frames by factorizing space-time dimension inputs efficiently on both large and small datasets. The divided spatial and temporal attention within each block, TimeSformer [4] reduces training time while achieving comparable test efficiency. A Spatial-Temporal Transformer network (ST-TR) was developed in [49, 75] for skeleton-based action recognition. A transformer-based RGB-D egocentric action recognition framework called Trear was proposed in [41] showing dramatic improvement over the existing state-of-the-art results. A multiscale pyramid network called MViT was proposed in [17] to extract information from low-level to high levels of attention. When compared with many other successful applications of transformers, their potential in action recognition has still largely remained unexplored.

3. The Proposed Method

Let $\mathbf{x} \in \mathbb{R}^{T\times H\times W\times 3}$ be the input video and $\mathbf{y}$ be the corresponding label of the video $\mathbf{x}$. H, W and T are the height, the width and the number of frames of a video, respectively. Let $\mathbf{o} \in \mathbb{N}^T$ be the permutation representing the reordering of video frames and $\mathbf{i}$ be the indexing associated with the permutation. Our goal is to learn a deep network to classify the actions and infer the permutation simultaneously as in Eqn. (1).

$$\arg\max_{\theta} \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{o},\mathbf{i}} \left(\log(p(\mathbf{y}|\mathbf{x};\theta)) + \log(p(\mathbf{i}|\mathcal{T}(\mathbf{x},\mathbf{o});\theta))\right) \tag{1}$$

where θ is the parameters of the deep neural network, and $\mathcal{T}$ is the permutation function. Given a video $\mathbf{x}$ and the permutation $\mathbf{o}$, the goal is to learn the class label $\mathbf{y}$ of the ordered video and learn the ordering $\mathbf{i}$ of the video after permutation $\mathcal{T}(\mathbf{x},\mathbf{o})$.

To effectively predict the class label $\mathbf{y}$ and the indexing of the permutation $\mathbf{i}$, a Transformer with Directed Attention is introduced to learn the directed attention in both spatial and temporal dimensions. The proposed DirecFormer is therefore formulated as in Eqn. (2).

$$\begin{aligned} \hat{\mathbf{y}} &= \phi_{cls} \odot \mathcal{G}(\mathbf{x}) \\ \hat{\mathbf{i}} &= \phi_{ord} \odot \mathcal{G}(\mathcal{T}(\mathbf{x}, \mathbf{o})) \end{aligned} \tag{2}$$

where $\mathcal{G}$ is the proposed DirecFormer; ϕ_{cls} and ϕ_{ord} are the projections that map the token outputted from DirecFormer to the predicted class label $\hat{\mathbf{y}}$ and the predicted ordering index $\hat{\mathbf{i}}$, respectively; and $\odot$ is the functional composition. Fig. 2 illustrates our proposed framework. The proposed DirecFormer method will be described in detail in the following section.

3.1. Patch Representation

Given a video frame, it is represented by N non-overlapped patches of $P \times P$ ($N = \frac{HW}{P^2}$) as in [13]. Let us denote $\mathbf{x}_{s,t} \in \mathbb{R}^{3P^2}$ as a vector representing the patch s of the video frame t, where s ($1 \leq s \leq N$) denotes the spatial position and t represents the temporal dimension ($1 \leq t \leq T$). To embed the temporal information into the representation, the raw patch representation is projected to the latent space with additive temporal representation as in Eqn. (3).

$$\mathbf{z}_{s,t}^{(0)} = \alpha(\mathbf{x}_{s,t}) + \mathbf{e}_{s,t} \tag{3}$$

where α is the embedding network and $\mathbf{e}_{s,t}$ is the spatial-temporal embedding added into the patch representation. The output sequences $\{\mathbf{z}_{s,t}^{(0)}\}_{s=1,t=1}^{N,T}$ represent the input tokens fed to our DirecFormer network. We also add one more learnable token $\mathbf{z}_{0,0}$ in the first position, as in BERT [10] to represent the classification token.

3.2. Directed Attention Approach

The proposed DirecFormer consists of L encoding blocks. In particular, the current block l takes the output tokens of the previous block $l-1$ as the input and decomposes the token into the key $\mathbf{k}_{s,t}^{(l)}$, value $\mathbf{v}_{s,t}^{(l)}$, and query $\mathbf{q}_{s,t}^{(l)}$ vectors as in Eqn. (4).

$$\begin{aligned} \mathbf{k}_{s,t}^{(l)} &= \beta_k^{(l)}\left(\tau_k^{(l)}\left(\mathbf{z}_{s,t}^{(l-1)}\right)\right) \\ \mathbf{v}_{s,t}^{(l)} &= \beta_v^{(l)}\left(\tau_v^{(l)}\left(\mathbf{z}_{s,t}^{(l-1)}\right)\right) \\ \mathbf{q}_{s,t}^{(l)} &= \beta_q^{(l)}\left(\tau_q^{(l)}\left(\mathbf{z}_{s,t}^{(l-1)}\right)\right) \end{aligned} \tag{4}$$

where $\beta_k^{(l)}$, $\beta_v^{(l)}$ and $\beta_q^{(l)}$ represent the key, value, and query embedding, respectively; $\tau_k^{(l)}$, $\tau_v^{(l)}$ and $\tau_q^{(l)}$ are the layer normalization [3].

In the traditional self-attention approach, the attention matrix is computed by the scaled dot multiplication between key and query vectors. Although scaled dot attention has shown its potential performance in video classification, this attention is non-directed because it is unable to illustrate the direction of attention. In particular, the scaled dot attention simply indicates the correlations among tokens and ignores the temporal or spatial ordering among tokens. It is noticed that the ordering of frames in a video sequence does matter. The recognition of actions in a video is highly dependent on the ordering of video frames. For example, the same group of video frames, if ordered differently in time, may result in different actions, e.g. walking might become running. However, traditional Softmax attention can not fully exploit the ordering of video frames because it does not contain the directional information of the correlation.

Therefore, we propose a new Directed Attention using the cosine similarity. Formally, the attention weights $\mathbf{a}_{(s,t)}^{(l)}$ for a query $\mathbf{q}_{s,t}^{(l)}$ can be formulated as in Eqn. (5).

$$\mathbf{a}_{(s,t)}^{(l)} = \left[\cos\left(\frac{\mathbf{q}_{s,t}^{(l)}}{\sqrt{D}}, \mathbf{k}_{0,0}^{(l)}\right) \left\{\cos\left(\frac{\mathbf{q}_{s,t}^{(l)}}{\sqrt{D}}, \mathbf{k}_{s',t'}^{(l)}\right)\right\}_{s'=1,t'=1}^{N,T}\right] \tag{5}$$

where D is the dimensional length of the query vector $\mathbf{q}_{s,t}^{(l)}$, $\mathbf{a}_{p,t}^{(l)} \in \mathbb{R}^{NT+1}$ denotes the directed attention weights. This attention is computed over the spatial and temporal dimensions. As a result, this operator suffers a heavy computational cost. We therefore divide and conquer the Directed Attention in the spatial dimension and temporal dimension sequentially as in [4].

More specifically, we first implement the attention mechanism over the time dimension ($\mathbf{a}_{(s,t)}^{(l)-time}$) as in Eqn. (6).

$$\mathbf{a}_{(s,t)}^{(l)-time} = \left[\cos\left(\frac{\mathbf{q}_{s,t}^{(l)}}{\sqrt{D}}, \mathbf{k}_{0,0}^{(l)}\right) \left\{\cos\left(\frac{\mathbf{q}_{s,t}^{(l)}}{\sqrt{D}}, \mathbf{k}_{s,t'}^{(l)}\right)\right\}_{t'=1}^{T}\right] \tag{6}$$

Then, the directed temporal attention information is accumulated to the current token representations as in Eqn. (7).

$$\begin{aligned} \mathbf{s}_{s,t}^{(l)-time} &= \mathbf{a}_{(s,t),(0,0)}^{(l)-time}\mathbf{v}_{0,0}^{(l)} + \sum_{t'=1}^{T}\mathbf{a}_{(s,t),(s,t')}^{(l)-time}\mathbf{v}_{s,t'}^{(l)} \\ \mathbf{z}'^{(l)-time}_{s,t} &= \mathbf{z}_{s,t}^{(l-1)} + \gamma^{(l)-time}\left(\mathbf{s}_{s,t}^{(l)-time}\right) \end{aligned} \tag{7}$$

where $\gamma^{(l)-time}$ denotes the temporal projection. Secondly, the temporally attentive vector $\mathbf{z}'^{(l)-time}_{s,t}$ is projected to the new key, value, and query to drive Spatial Directed Atten-

tion as in Eqn. (8).

$$\begin{aligned}\mathbf{k}'^{(l)}_{s,t} &= \beta'^{(l)}_{k}\left(\tau'^{(l)}_{k}\left(\mathbf{z}'^{(l)-time}_{s,t}\right)\right)\\ \mathbf{v}'^{(l)}_{s,t} &= \beta'^{(l)}_{v}\left(\tau'^{(l)}_{v}\left(\mathbf{z}'^{(l)-time}_{s,t}\right)\right)\\ \mathbf{q}'^{(l)}_{s,t} &= \beta'^{(l)}_{q}\left(\tau'^{(l)}_{q}\left(\mathbf{z}'^{(l)-time}_{s,t}\right)\right)\end{aligned} \tag{8}$$

Next, the Directed Attention over the spatial dimension ($\mathbf{a}^{(l)-space}_{(s,t)}$) can be computed as in Eqn. (9).

$$\mathbf{a}^{(l)-space}_{(s,t)} = \left[\cos\left(\frac{\mathbf{q}'^{(l)}_{s,t}}{\sqrt{D}}, \mathbf{k}'^{(l)}_{0,0}\right)\left\{\cos\left(\frac{\mathbf{q}'^{(l)}_{s,t}}{\sqrt{D}}, \mathbf{k}'^{(l)}_{s,t'}\right)\right\}_{t'=1}^{T}\right] \tag{9}$$

The Spatial Directed Attention is then embedded to the temporal attentive features $\mathbf{z}'^{(l)-time}_{s,t}$ to obtain a new spatial attentive feature $\mathbf{z}'^{(l)-space}_{s,t}$ as in Eqn. (10).

$$\begin{aligned}\mathbf{s}^{(l)-space}_{s,t} &= \mathbf{a}^{(l)-space}_{(s,t),(0,0)}\mathbf{v}^{(l)}_{0,0} + \sum_{s'=1}^{N}\mathbf{a}^{(l)-space}_{(s,t),(s',t)}\mathbf{v}^{(l)}_{s',t}\\ \mathbf{z}'^{(l)-space}_{s,t} &= \mathbf{z}'^{(l)-times}_{s,t} + \gamma^{(l)-space}\left(\mathbf{s}^{(l)-space}_{s,t}\right)\end{aligned} \tag{10}$$

Finally, the Spatial-Temporal Attentive features $\mathbf{z}'^{(l)-space}_{s,t}$ are projected to the output token, getting ready for the next transformer block.

Formally, the output of the current transformer block ($\mathbf{z}^{(l)}_{s,t}$) can be formed as in Eqn. (11).

$$\mathbf{z}^{(l)}_{s,t} = \varphi^{(l)}\left(\tau^{(l)}\left(\mathbf{z}'^{(l)-space}_{s,t}\right)\right) + \mathbf{z}'^{(l)-space}_{s,t} \tag{11}$$

where $\varphi^{(l)}$ is a projection mapping implemented using a multi-layer perception network, and $\tau^{(l)}$ denotes the layer normalization [3].

3.3. Classification Embedding

The final representation is obtained in the final block of DirecFormer. Then, the class index and the order index of the video are predicted using linear projections as follows:

$$\begin{aligned}\hat{\mathbf{y}} &= \phi_{cls}\left(\tau_{cls}\left(\mathbf{z}^{(L)}_{0,0}\right)\right)\\ \hat{\mathbf{i}} &= \phi_{odr}\left(\tau_{odr}\left(\mathbf{z}^{(L)}_{0,0}\right)\right)\end{aligned} \tag{12}$$

where ϕ_{cls} and ϕ_{ord} are the classification projection and order projection, respectively; τ_{cls} and τ_{ord} are the layer normalization [3].

3.4. Self-supervised Guided Loss For Directed Temporal Attention Loss

In this stage, we are given the permutation of the current input video. To further reduce the burden of the network when learning the temporal attention, we propose a new self-supervised guided loss to enforce the temporal attention learning from the prior order knowledge. Formally, the self-supervised loss can be formulated as in Eqn. (13).

$$\mathcal{L}_{self} = \frac{1}{LNT^2}\sum_{l=1}^{L}\sum_{s=1,t=1}^{N,T}\sum_{t'=1}^{T}\left(1 - \mathbf{a}^{(l)-time}_{(s,t),(s,t')}\right)\varsigma(\mathbf{o}_t, \mathbf{o}_{t'}) \tag{13}$$

where $\varsigma(\mathbf{o}_t, \mathbf{o}_{t'}) = 1$ if the index $\mathbf{o}_t < \mathbf{o}_{t'}$, otherwise $\varsigma(\mathbf{o}_t, \mathbf{o}_{t'}) = -1$. The guided loss $\mathcal{L}_{self}$ helps to indicate the attention learning the correct direction during the training process. Finally, the total loss function of DirecFormer is defined as in Eqn. (14).

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{ord}\mathcal{L}_{ord} + \lambda_{self}\mathcal{L}_{self} \tag{14}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{ord}$ are the cross-entropy losses of the classification projection (ϕ_{cls}) and order projection (ϕ_{ord}), respectively; $\{\lambda_{cls}, \lambda_{ord}, \lambda_{self}\}$ are the parameters controlling their relative importance.

Figure 3. The Video Samples of Three Datasets: (a) Jester, (b) Something-Something V2, and (c) Kinetics-400.

4. Experiments

In this section, we present the evaluation results with DirecFormer on three popular action recognition benchmarking datasets, i.e. Jester [45], Something-Something V2 [23], and Kinetics 400 [31]. Firstly, we describe our implementation details and datasets used in our experiments. Secondly, we analyze our results with different settings shown in the ablation study on the Jester dataset. Lastly, we present our results on Something-Something V2 and Kinetics compared to prior state-of-the-art methods.

4.1. Implementation Details

The architecture of DirecFormer consists of $L = 12$ blocks. The input video consists of $T = 8$ frames sampled at a rate of $1/32$ and the resolution of each frame is

( ). The patch size is set to ; therefore, there are —— patches in total for each frame. The embedding network is implemented by a linear layer in which the output dimension is set to . All values ( ), key ( ), query ( ) embedding networks, and projections ( ) are also implemented by the linear layers. Similar to [4, 13], we adopt the multi-head attention in our implementation, where the number of heads is set to . The network is implemented as the residual-style multi-layer perceptron consisting of two fully connected layers followed by a normalization layer. Finally, the classification projection ( ) and the order projection ( ) are implemented as the linear layer. We set the control parameters of loss to , i.e. .

There will be a total of permutations of the video frames. Therefore, learning with all permutations is ineffective. Moreover, the permutation set plays an important role. If these two permutations are very far from each other, the network may easily predict the order since the two permutations have significant differences. However, if all the permutations are close to each other, learning the temporal attention is more challenging since the two permutations have minor differences in order. Therefore, we select 1,000 random permutations from permutations so that the Hamming distance between permutations is as minimum as possible. Similar to [5], we use a greedy algorithm to generate the set of permutations.

In the evaluation, following the protocol of other papers [4, 18, 19], the single clip is sampled in the middle of the video. We use three spatial crops (top-left, center, and bottom-right) from the temporal clip and obtain the final result by averaging the prediction scores for these three crops.

4.2. Datasets

Jester. [45] This dataset is a large-scale gesture recognition real-world video dataset that includes videos of actions. Each video is recorded for approximately 3 seconds. Fig 3(a) illustrates the video samples of Jester.

Something-in-Something V2. [23] The dataset is a large-scale dataset to show humans performing predefined basic actions with everyday objects, which includes classes. It contains videos, with videos in the training set, videos in the validation set, and videos in the testing set. Similar to other work [4,68,73,77], we report the accuracy on the validation set. Fig. 3(b) shows two examples of two different class of Something-Something V2. The licenses of Something-Something V2 and Jester datasets are registered by the TwentyBN team that are publicly available for academic research purposes.

Kinetics-400. [31] The dataset contains 400 human action classes, with at least 400 videos for each action. In particular, Kinetics-400 contains training videos and validation videos. The videos were downloaded from youtube and each video lasts for 10 seconds. There are different types of human actions: *Person Actions* (e.g. singing, smoking, sneezing, etc.); *Person-Person Actions* (e.g. wrestling, hugging, shaking hands, etc.); and *Person-Object Actions* (e.g. opening a bottle, walking the dog, using a computer, etc.). Fig. 3(c) illustrates the video examples of Kinetics-400. In our experiment, following the protocol of other papers [4, 18, 18, 19, 68], we report the accuracy on the validation set. The license of Kinetics is registered by Google Inc. under a Creative Commons Attribution 4.0 International License.

Table 1. **Ablation Study On Jester.** denotes for the attention types of temporal and spatial dimension, respectively. (and) could be either : Softmax or : Cosine.

Models	Attention Time-Space	$\mathcal{L}$	$\mathcal{L}$	Top 1	Top 5
I3D [7]				91.46	98.67
3D SqueezeNet [28]				90.77	
ResNet 50 [25]				93.70	
ResNet 101 [25]				94.10	
ResNeXt [69]				94.89	
PAN [72]				96.70	
STM [29]				96.70	
ViViT-L/16x2 320 [2]				81.70	93.80
TimeSFormer [4]				94.14	99.19
DirecFormer				94.52	99.26
DirecFormer				94.65	99.25
DirecFormer				95.52	99.20
DirecFormer				96.28	99.45
DirecFormer				97.55	97.54
DirecFormer				96.15	99.38
DirecFormer				97.48	99.48
DirecFormer				98.15	99.57

4.3. Ablation Study

Effectiveness Of Directed Attention To show the effectiveness of our proposed Directed Attention, we consider three different types of the temporal-spatial attention: (i) Softmax Temporal Attention followed by Cosine Spatial Attention (DirecFormer), (ii) Cosine Temporal Attention followed by Softmax Spatial Attention (DirecFormer), and (iii) Cosine Temporal Attention followed by Cosine Spatial Attention (DirecFormer). The method is also compared with TimeSformer where the softmax attention is applied for both time and space. Table 1 illustrates the results of the DirecFormer with different settings compared to TimeSFormer and other approaches. In all configurations, our proposed DirecFormer outperforms the prior methods.

Considering the effectiveness of the directed attention in time and space, the directions of the attention over the spatial dimension are important in some cases. For example, if performs an action to then receives an action from . Considering the mentioned example, the spatial attention should involve directions so that the model can learn

Table 2. **Order Correction By Hamilton Algorithm Performance On Jester.** $X-Y$ denotes for the attention types of temporal and spatial dimension, respectively. X (and Y) could be either S: Softmax or C: Cosine.

Models	Attention Time-Space	$\mathcal{L}_{ord}$	$\mathcal{L}_{self}$	OrderAcc
TimeSFormer [4]	$S-S$	–	–	52.84
TimeSFormer [4]	$S-S$	✓	–	72.57
DirecFormer	$C-S$			75.04
DirecFormer	$C-S$	✓		87.16
DirecFormer	$C-S$	✓	✓	90.02
DirecFormer	$C-C$			76.16
DirecFormer	$C-C$	✓		88.96
DirecFormer	$C-C$	✓	✓	**90.19**

the actor(s) performing actions in a video. However, the order of the temporal dimension plays a more important role in a video compared to the spatial dimension, since the order of the frames represents how the action is happening. As in Table 1, the results of DirecFormer $C-S$ are better than DirecFormer $S-C$ confirming our hypothesis about the importance of time and space. When the Directed Attention is deployed in both temporal and spatial dimensions, the results of DirecFormer $C-C$ were significantly improved and achieved the SOTA performance on the Jester dataset.

Effectiveness Of Losses With the order prediction loss $\mathcal{L}_{ord}$, the performance of the DirecFormer in all settings has been improved, since the prediction loss influences the way that network learns the Directed Temporal Attention. Moreover, the performance of DirecFormer is improved by employing the self-supervised guided loss $\mathcal{L}_{self}$. This self-supervised loss further enhances the directed temporal attention learning during the training. Consequently, the performance of DirecFormer is consistently improved by using our proposed losses, as in Table 1.

Order Correction To illustrate the ability of order learning of DirecFormer, we conduct an experiment in which, given a random temporal order video, we show our approaches can retrieve back the correct order of the video from the directed temporal attention. In this experiment, we use the temporal attention of the last block and average this temporal attention over the spatial dimension. Then, we perform a search algorithm to find the Hamiltonian path on the temporal attention to find the correct order. In particular, we consider the temporal attention as the adjacency matrix of the graph, in which each frame is the node of the graph. The Hamilton path is the path that goes through each node exactly once (no revisit). Since our attention represents both direction and correlation among the frames, the higher (positive) correlation is, the higher the possibility of correct order should be between frames. Therefore, the Hamilton path with maximum total weight is going to represent the order of the video should be.

Let $\hat{\mathbf{o}}$ be the order obtained by the Hamilton algorithm, the accuracy of the order retrieval can be defined as follows:

$$\text{OrderAcc} = \frac{\text{LCS}(\hat{\mathbf{o}}, \mathbf{o})}{T} \times 100 \tag{15}$$

where $\text{LCS}(\hat{\mathbf{o}}, \mathbf{o})$ is the longest common subsequence between $\hat{\mathbf{o}}$ and $\mathbf{o}$. In this evaluation, for each video, we randomly select a permutation of $\{1, ..., N\}$ as the order of the input video. To be fair between benchmarks, we set the same random seed value at the beginning of the testing script so that every time we conduct the evaluation, we obtain the same permutation for each video.

As shown in Table 2, we use the Softmax attention of TimeSFomer to retrieve the order of the video. The order accuracy of the TimeSFormer is only 52.84. In other words, the Softmax attention of TimeSFormer can only predict the correct order of approximately 4 frames over 8 frames. With the support of order prediction loss, the order accuracy of TimeSFormer is improved to 72.57%. However, without the order prediction loss, our DirecFormer $C-S$ and DirecFormer $C-C$ have already correctly predicted the order of approximately 6 frames over 8 frames (75.04% and 76.16%). When we further employ the order prediction and self-supervised guided losses, the performance of DirecFormer is significantly improved. Particularly, with the order prediction loss only, DirecFormer in all settings gains more than 87.0% (which is approximately 7 frames over 8 frames). When both losses ($\mathcal{L}_{ord}$ and $\mathcal{L}_{self}$) are employed, the order accuracy of both DirecFormer $C-S$ and DirecFormer $C-C$ is improved to 90.02% and 90.19%, respectively. It should be noted that the performance of DirecFormer $C-C$ is only minorly greater than DirecFormer $C-S$ as the directed attention over the space does not largely affect the temporal order predictions.

4.4. Comparison with State-of-the-Art Results

Something-Something V2 Table 3 illustrates the performance of our proposed approaches evaluated on Something-Something V2 compared to prior SOTA approaches. In this experiment, similar to other approaches [4], we use the DirecFormer pretrained on ImageNet-1K [9]. As in Table 3, our results in all settings outperform

Table 3. **Comparison with the SOTA methods on Something-Something V2.** $X-Y$ denotes for the attention types of temporal and spatial dimension, respectively. X (and Y) could be either S: Softmax or C: Cosine.

Models	Attention Time-Space	Top 1	Top 5
MSNet [77]	–	63.00	88.40
SlowFast [19]	–	63.00	88.50
SlowFast Multigrid [68]	–	63.50	88.70
TRG [73]	–	62.20	**90.30**
VidTr-L [74]	–	60.20	–
TimeSFormer [4]	$S-S$	59.10	85.60
TimeSFormer – HR [4]	$S-S$	61.80	86.90
TimeSFormer – L [4]	$S-S$	62.00	87.50
DirecFormer	$S-C$	61.70	85.20
DirecFormer	$C-S$	63.85	85.92
DirecFormer	$C-C$	**64.94**	87.90

Table 4. **Comparison with the SOTA methods on Kinetics 400.** denotes for the attention types of temporal and spatial dimension, respectively. (and) could be either : Softmax or : Cosine.

Models	Attention Time-Space	Top 1	Top 5
I3D NLN [7]	–	74.00	91.10
ip-CSN-152 [58]	–	77.80	92.80
LGD-3D-101 [51]	–	79.40	94.40
SlowFast [19]	–	77.00	92.60
SlowFast Multigrid [68]	–	76.60	92.70
X3D-M [18]	–	75.10	91.70
X3D-L [18]	–	76.90	92.50
X3D-XXL [18]	–	80.40	94.60
MViT [17]	–	78.40	93.50
TimeSFormer [4]	$S-S$	77.90	93.20
TimeSFormer – HR [4]	$S-S$	79.70	94.40
TImeSFormer – L [4]	$S-S$	80.70	94.70
DirecFormer	$S-C$	80.16	94.55
DirecFormer	$C-S$	81.69	94.62
DirecFormer	$C-C$	**82.75**	**94.86**

other candidates. With the simple design of the Transformer network with the directed attention mechanisms over time and space, our approaches achieve SOTA performance compared to traditional 3D CNN approaches [19, 77] and other Transformer approaches [4, 74] by a competitive margin.

Kinetics 400 We conduct the experiments on Kinetics 400 and compare our results with prior SOTA methods. The pretrained model on ImageNet-21K [9] for our DirecFormer is used, similar to [4]. It is noted that the prior methods [18, 19] use 10 temporal clips with 3 spatial crops of a video in the evaluation phase. However, TimeSFormer and our DirecFormer use only 3 spatial crops of a video with a single clip to achieve the solid results. In particular, our method achieves the SOTA performance compared to prior methods as shown in Table 4. The Top 1 accuracy of the best model is approximately higher than TimeSFormer-L [4] sitting at . The effectiveness of the proposed directed attention has been also proved in these experiments, as the performance of DirecFormer is consistently improved when we deploy the directed attention over time and space.

Network Size Comparison As shown in Table 5, although the number of parameters and the GFLOPS of single view

Table 5. **Network Size Comparison.** We report the computational cost of the inference phase with a single "view" (temporal clip with spatial crop) the numbers of such views used (GFLOPs views). "NA" indicates the number is not available for us.

Model	GFLOPS x Views	Params
I3D [7]	108×NA	12.0M
SlowFast 8x8 R50 [19]	36.1 × 30	34.4M
SlowFast 8x8 R101 [19]	106 × 30	53.7M
Nonlocal R50 [65]	282 × 30	35.3M
X3D-XL [18]	35.8 × 30	11.0M
X3D-XXL [18]	143.5 × 30	20.3M
ViViT-L/16x2 320 [2]	3980 × 3	310.8M
TimeSformer [4]	196 × 3	121.4M
DirecFormer	196 × 3	121.4M

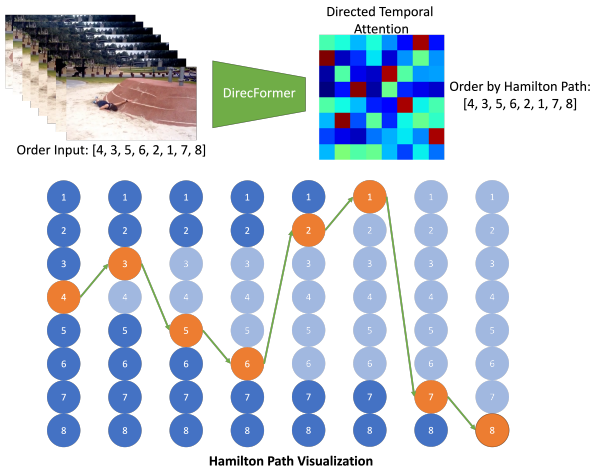

Figure 4. Visualization of Order Correction by Finding Hamilton Path On Directed Temporal Attention Map.

used in our DirecFormer are higher than the traditional 3D-CNN approaches [7, 18, 19], we only use 3 views compared to 30 views of prior approaches and maintain competitive performance. In comparison with TimeSFormer, we gain the same performance in terms of network size and inference flops; however, we achieve better accuracy on three large-scale benchmarks as shown in Tables 1, 3, and 4.

Qualitative Results Fig. 4 illustrates the Directed Attention of our proposed DirecFormer. We use a video on the validation set of Kinetics-400 and extract the attention map. We randomly permute the frames of video along the temporal dimension and correct the order of frames using the Hamilton algorithm. As in Fig. 4, we can successfully correct the frame order of the *Parkour* action video.

5. Conclusions

In this paper, we have presented a new and simple DirecFormer method with Directed Attention mechanism in Transformer over the temporal and spatial dimensions. The presented Directed Temporal-Spatial Attention not only learns the magnitude of the correlation between frames and tokens, but also exploits the direction of attention. Moreover, the self-supervised guided loss further enhances the directed learning capability of the Directed Temporal Attention. The intensive ablation study on the Jester dataset has shown the effectiveness of our proposed Directed Attention in both time and space. Furthermore, it has illustrated the impact of the proposed losses used in Directed Temporal Attention learning. The experiments on two other large-scale datasets, i.e. Something-Something V2 and Kinetics 400, have further confirmed the high accuracy performance of our proposed method.

Acknowledgement This work is supported by NSF Data Science, Data Analytics that are Robust and Trusted (DART), Arkansas Biosciences Institute (ABI) Grant Program and NSF WVAR-CRESH Grant.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021. 2

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021. 2, 3, 6, 8

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 4, 5

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 1, 3, 4, 6, 7, 8

[5] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles, 2019. 6

[6] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *CoRR*, abs/1808.01340, 2018. 2

[7] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733. IEEE, 2017. 1, 2, 6, 8

[8] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 7, 8

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 4

[11] P. Dollar, V. Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. pages 65–72, 2005. 2

[12] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634. IEEE, 2015. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4, 6

[14] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. Deep appearance models: A deep boltzmann machine approach for face modeling. *IJCV*, 2019. 1

[15] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, Nghia Nguyen, Eric Patterson, Tien D. Bui, and Ngan Le. Automatic face aging in videos via deep reinforcement learning. In *CVPR*, 2019. 1

[16] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, T Hoang Ngan Le, Marios Savvides, and Tien D Bui. Learning from longitudinal face demonstration—where tractable deep modeling meets inverse reinforcement learning. *IJCV*, 2019. 1

[17] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers, 2021. 3, 8

[18] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition, 2020. 1, 6, 8

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210. IEEE, 2019. 1, 2, 6, 7, 8

[20] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. Spatiotemporal residual networks for video action recognition. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *NeurIPS*, pages 3468–3476, 2016. 1, 2

[21] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, pages 7445–7454. IEEE, 2017. 2

[22] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941. IEEE, 2016. 1, 2

[23] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017. 2, 5, 6

[24] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056. IEEE, 2018. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE, 2016. 2, 6

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 2

[27] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7919–7929, 2021. 3

[28] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 1mb model size. *CoRR*, abs/1602.07360, 2016. 6

[29] Boyuan Jiang, Mengmeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition, 2019. 6

[30] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732. IEEE, 2014. 2

[31] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 2, 5, 6

[32] Hanul Kim, Mihir Jain, Jun-Tae Lee, Sungrack Yun, and Fatih Porikli. Efficient action recognition via dynamic knowledge propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13719–13728, 2021. 3

[33] Alexander Kläser, Marcin Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In Mark Everingham, Chris J. Needham, and Roberto Fraile, editors, *BMVC*, pages 1–10. British Machine Vision Association, 2008. 2

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *NeurIPS*, pages 1106–1114, 2012. 2

[35] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Learning self-similarity in space and time as generalized motion for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13065–13075, 2021. 3

[36] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *ICCV*, pages 432–439. IEEE, 2003. 2

[37] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pages 3361–3368. IEEE, 2011. 2

[38] T. Hoang Ngan Le, Kha Gia Quach, Khoa Luu, Chi Nhan Duong, and Marios Savvides. Reformulating level sets as deep recurrent neural network approach to semantic segmentation. *TIP*, 2018. 1

[39] Ce Li, Chunyu Xie, Baochang Zhang, Jungong Han, Xiantong Zhen, and Jie Chen. Memory attention networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 3

[40] Tianjiao Li, Qiuhong Ke, Hossein Rahmani, Rui En Ho, Henghui Ding, and Jun Liu. Else-net: Elastic semantic network for continual action recognition from skeleton data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13434–13443, 2021. 3

[41] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 2021. 3

[42] Ji Lin, Chuang Gan, and Song Han. TSM: temporal shift module for efficient video understanding. In *ICCV*, pages 7082–7092. IEEE, 2019. 2

[43] Xin Liu, Silvia L. Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C. van Gemert. No frame left behind: Full video action recognition, 2021. 2

[44] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, 2021. 2

[45] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2874–2882, 2019. 2, 5, 6

[46] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification. In *ECCV*, 2016. 3

[47] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, Ngan Le, and Marios Savvides. Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In *ICCV*, Oct 2017. 1

[48] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition, 2021. 2

[49] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *International Conference on Pattern Recognition*, pages 694–701. Springer, 2021. 3

[50] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, pages 5534–5542. IEEE, 2017. 1, 2

[51] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion, 2019. 8

[52] Sreemanananth Sadanand and Jason J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, pages 1234–1241. IEEE, 2012. 2

[53] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *NeurIPS*, pages 568–576, 2014. 1, 2

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 2

[55] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9. IEEE, 2015. 2

[56] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826. IEEE, 2016. 2

[57] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497. IEEE, 2015. 1, 2

[58] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks, 2019. 8

[59] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459. IEEE, 2018. 2

[60] Thanh-Dat Truong, Chi Nhan Duong, Ngan Le, Son Lam Phung, Chase Rainwater, and Khoa Luu. Bimal: Bijective

maximum likelihood approach to domain adaptation in semantic scene segmentation. *IICV*, 2021. 1

[61] Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, Ngan Le, and Khoa Luu. Fast flow reconstruction via robust invertible n × n convolution. *Future Internet*, 2021. 1

[62] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558. IEEE, 2013. 2

[63] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition, 2021. 2

[64] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, volume 9912, pages 20–36. Springer, 2016. 2

[65] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803. IEEE, 2018. 2, 8

[66] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition, 2021. 2

[67] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In David A. Forsyth, Philip H. S. Torr, and Andrew Zisserman, editors, *ECCV*, volume 5303, pages 650–663. Springer, 2008. 2

[68] Chao-Yuan Wu, Ross Girshick, Kaiming He, Christoph Feichtenhofer, and Philipp Krähenbühl. A Multigrid Method for Efficiently Training Video Models. In *CVPR*, 2020. 6, 7, 8

[69] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017. 6

[70] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, volume 11219, pages 318–335. Springer, 2018. 1, 2

[71] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Computer Vision and Pattern Recognition (CVPR)*. 3

[72] Can Zhang, Yuexian Zou, Guang Chen, and Lei Gan. Pan: Towards fast action recognition via learning persistence of appearance, 2020. 1, 2, 6

[73] Jingran Zhang, Fumin Shen, Xing Xu, and Heng Tao Shen. Temporal reasoning graph for activity recognition, 2019. 6, 7

[74] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions, 2021. 1, 7, 8

[75] Yuhan Zhang, Bo Wu, Wen Li, Lixin Duan, and Chuang Gan. Stst: Spatial-temporal specialized transformer for skeleton-based action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3229–3237, 2021. 3

[76] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV*, volume 11205 of *Lecture Notes in Computer Science*, pages 831–846. Springer, 2018. 2

[77] Xiaoyu Zhu, Junwei Liang, and Alexander Hauptmann. Msnet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos, 2020. 6, 7, 8