Spatio-Temporal Convolution-Attention Video Network

Ali Diba[1], Vivek Sharma[2,3⋆], Mohammad.M Arzani, Luc Van Gool[1,4]
[1] KU Leuven, [2] Sony AI
[3] Massachusetts Institute of Technology, [4] ETH Zurich

Abstract

*In this paper, we present a hierarchical neural network based on convolutional and attention modeling for short and long-range video reasoning, called Spatio-Temporal Convolution-Attention Video Network (**STCA**). The proposed method is capable of learning appearance and temporal cues in two stages with different temporal depths to maximize engagement of the short-range and long-range video sequences. It has the benefits of convolutional and attention networks in exploiting spatial and temporal cues for a new spatio-temporal sequence modeling. Our method is a novel mixer architecture to obtain robust properties of convolution (such as translational equivariance) while having the generalization and sequential modeling ability of transformers to deal with dynamic variations in videos. The proposed video deep neural network aims to exploit spatio-temporal information in two stages: 1.) Short Clip Stage (SCS) and 2.) Long Video Stage (LVS). SCS handles spatio-temporal cues dealing with short-range video clips and operates on video frames with 3D convolutions and multi-headed self-attention modeling. Since SCS operates on video frames, this reduces the quadratic complexity of the self-attention operation. In LVS, we mitigate the issue of modeling long-range temporal self-attention. LVS models long-range temporal reasoning using representation (i.e., tokens) obtained from SCS. LVS consists of variants of long-range temporal modeling mechanisms for learning compact and robust global temporal representations of the entire video. We conduct experiments on six challenging video recognition datasets: HVU, Kinetics (400, 600, 700), Something-Something V2, and Long Video Understanding dataset. Through extensive evaluations and ablation studies, we show outstanding performances in comparison to state-of-the-art methods on the mentioned datasets.*

1. Introduction

Over the last two decades, we have observed that human action recognition in videos has received huge attention due to potential applications in video retrieval, behavior analysis, surveillance, and video understanding tasks. Even if considerable progress in video understanding was made, in practice, the traditional video recognition models are still rather limited to tasks designed for short-range videos (e.g., 5-10 seconds in length). In order to have a holistic video understanding, we would require video recognition methods that can model temporal reasoning with different temporal resolution in addition to exploiting the short-range information.

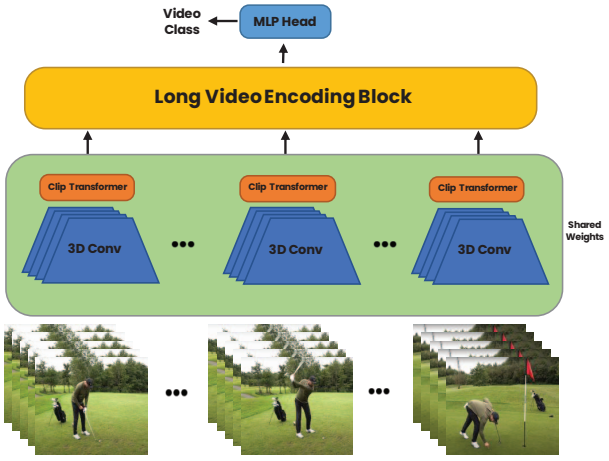

Figure 1. We propose a novel Spatio-Temporal Convolution-Attention Video Network (STCA) architecture to analyze video content in two stages, exploiting spatio-temporal clues in short-range clips and modeling long-range relations between different clips. This model is a novel mixer of 3D-ConvNet, self-attention, and gated-attention neural networks.

In action recognition, the task is to classify a video into single [34, 37, 55] or multiple [8] semantic labels. Neural networks for action recognition can be categorized into two types, namely spatial neural networks [11, 54, 65] (which operates on spatial cues using standard 2D filters and pooling kernels) and spatio-temporal neural networks [8, 26, 60] (which integrate both spatial and temporal cues at the same time using 3D filters and pooling kernels). Unfortunately, the majority of the spatial and spatio-temporal neural networks are designed for short-range videos (e.g., 5 seconds in length). In this paper, we design a model equipped for both short and long-range videos, thereby allowing us to

perform generic video understanding.

The unprecedented success of Transformers [62] in Natural Language Processing (NLP) tasks recently inspired the vision [13] and video understanding [2,70] community. The recently proposed video transformers models are partially effective at modeling temporal dependencies and allow the interactions between each pair of input sequences based on multi-headed self-attention. However, the drawback in the design of the attention mechanism: quadratic computational complexity of self-attention, has made these models either very computationally costly [62] or have left the model to operate on pre-extracted CNN features [74] or operate on image patches [13] or operate on short-range sequences; or only consider Classification (CLS) token outputs for each frame. However, doing so removes the fine-grained spatio-temporal cues, thus, degrading its application for modeling long-range video understanding tasks. This characteristic of pairwise interactions is essential for long-range video modeling. To effectively reduce the computation complexity of Transformers and capture long-range temporal reasoning, many attempts to use transformers exist in the video community. Such as STAM (Space Time Attention Model) [52] operates on a sequence of image patches extracted from video frames followed by modeling pairwise dependencies for temporal modeling. Likewise, Multi-View Transformer (MVT) [70] extracts tokens from spatio-temporal tubelets of varying dimensions and image patches from the video frames for spatio-temporal reasoning using tokens from all image patches jointly. ViVit [2] extracts spatio-temporal tokens (or tubelets) from the video clips, which are then encoded by a series of transformer layers. All these methods suffer from quadratic complexity of the self-attention operation as they operate on patches - thus limiting the temporal modeling to only short-range clips with a temporal window of 16-64 frames $\sim$ 1-2 seconds (30fps). More similar to our work are, Islam et al. [29] utilize structured state-space sequence model(S4) [24] to model long-range sequences. MetaFormer [71] uses convolutions as a token mixer in the bottom stages. Convolutional vision Transformer (CvT) [69] proposes a hierarchy of Transformers containing a convolutional token embedding and a convolutional Transformer block. See "Action Recognition" in the related work for a comprehensive review.

Motivated by the above observations, we propose a spatio-temporal neural network called Spatio-Temporal Convolution-Attention Video Network (**STCA**). Figure 1, sketches STCA. STCA extracts appearance and temporal information in a novel way to maximize engagement of the short-range and long-range video sequences for robust reasoning in short or long videos. Specifically, in this work, we take a different path toward efficient video recognition. Our proposed method aims to exploit spatio-temporal information in two stages: 1.) Short Clip Stage (SCS) and 2.) Long Video Stage (LVS). In SCS, we make use of 3D convolution layers as a way to abstract the spatio-temporal information from video clips, together with transformers. Unlike previous works, which operated on 3D tubelet patches [2], we operate on video frames allowing us to learn a more robust representation. Further, this reduces a large number of patch-based spatio-temporal tokens, thus reducing the quadratic complexity of the self-attention operation. In LVS, we mitigate the issue of modeling long-range temporal self-attention. LVS temporally aggregates and encodes spatio-temporal dynamic cues from video clips (tokens from SCS) into a compact and robust global temporal representation of the entire video. Our method is evaluated on six challenging benchmark video recognition datasets, namely HVU, Kinetics (400, 600, 700), Something-Something V2, and Long Video Understanding dataset. We experimentally show that our video modeling (see Sec. 4) achieves superior or comparable performances compared to state-of-the-art works on Holistic Video Understanding (HVU) [8], Kinetics [34], Something-Something V2 [22], and Long-form Video Understanding datasets [68].

2. Related Works

Classical Action Recognition: Over the last two decades, a multitude of action recognition techniques in videos have been proposed by the vision community. Among the hand-engineered ones that could effectively model the appearance and motion/dynamic representations across frames in videos are HOG3D [35], SIFT3D [50], HOF [38], ESURF [67] and MBH [5], iDTs [63] and more. Despite their good performance, they have several shortcomings, they are computationally expensive and lack scalability to capture semantic concepts for the large-scale dataset. To overcome such issues, several other techniques were proposed to model the temporal structure in an efficient way, such as the actom sequence model [20]; temporal action decomposition [46]; dynamic poselets [64]; ranking machines [19].

Neural Networks based Action Recognition: In the last decade also, we have seen that Convolutional Neural Networks (ConvNets) based action recognition [18, 33, 54, 60, 65] has taken a huge leap to exploit the appearance and the temporal information in an end-to-end learning fashion. ConvNets-based methods operate on 2D (individual image level) or 3D (video clips or snippets of K frames). In the 2D setting, spatial and/or temporal information are modeled via LSTMs/RNNs to capture long-term motion cues [12, 72], or via encoding methods such as Bilinear models [11], Fisher encoding (FVs) [59], and Vector of Locally Aggregated Descriptors (VLAD) [21]. While in the 3D settings, the network directly extracts spatio-temporal features from raw videos using 3D ConvNets. The filters

and pooling kernels operate on (x, y, time) i.e. 3D convolutions ($s \times s \times d$) [72] where d is the kernel temporal depth, and s is the kernel spatial size. Simonyan et al. [54] proposed a two-stream network, cohorts of RGB and flow ConvNets. Tran et al. [60] explored 3D ConvNets on video streams for spatio-temporal feature learning for clips and filter kernel of size $3 \times 3 \times 3$. In this way, they avoid calculating the optical flow explicitly and still achieve good performance. Tran et al. in [61] further extended the ResNet architecture with 3D convolutions. Sun et al. [58] proposed a factorized spatio-temporal ConvNet and decomposed the 3D convolutions into 2D spatial and 1D temporal convolutions. Similar to [54] and [60] is Feichtenhofer et al.'s [18] work, where they propose 3D pooling. Wang et al. [65] use multiple clips sparsely sampled from the whole video as input and then combine the scores in a late fusion approach. Carreira *et al.* proposed inception [27] based 3D CNNs, which they referred to as I3D [4], where they convert a pre-trained 2D ConvNet [28] to 3D ConvNet by inflating the filters and pooling kernels with an additional temporal dimension d. All these architectures have fixed homogeneous temporal 3D kernel depths throughout the whole architecture. T3D [7] models variable temporal convolution kernel depths over shorter and longer temporal ranges. Furthermore, in [6], Diba et al. propose spatio-temporal channel correlation networks. DynamoNet [9], learn dynamic motion filters for modeling an effective internal motion representation using dynamic filter networks [30, 31, 53]. HATNet [8] exploits both 2D ConvNets and 3D ConvNets to learn an effective spatio-temporal feature representation, similar to slow-and-fast networks [16].

Transformers based Action Recognition: Inspired by the unprecedented success of Transformers [62] in sequence modeling in the field of Natural language Processing (NLP) community. Recently transformer-based models have been successfully used for vision tasks such as image classification [13], video captioning [51], multimodal representation learning [42, 43] and video classification [2, 56, 70]. Attempts to use transformers exist in the vision domain, such as 1.) The majority of the works [3, 13, 56] use only attention-based layers (similar to those employed in NLP) instead of the commonly used convolutional layers and produce SOTA results on image classification [13]; 2.) Combine pre-trained CNNs features with transformers for object detection [74]; video classification [32]. This methods [13] applies the Transformer model on the image pixel level. More similar to our work, [2, 29, 70] applies a Transformer model in the domain of action recognition and operates on modeling video frame sequences. Specifically, [29] uses a standard Transformer encoder to process 2D images in video frames and then uses a multi-scale temporal structured state-space sequence (S4) layer for subsequent video classification. STAM (Space Time Attention Model) [52] operates on a sequence of image patches extracted from video frames followed by modeling spatio-temporal dependencies between distinct frames. ViVit [2] extracts spatio-temporal tokens (or tubelets) from the video clips, which are then encoded by a series of transformer layers. Multi-View Transformer (MVT) [70] extracts tokens from spatio-temporal tubelets of varying dimensions. All these methods are partially effective at modeling temporal dependencies because they all operate on clips with short-range sequences. They model short-range sequences in order to minimize the quadratic computational complexity of self-attention. In contrast to these prior works, our work differs substantially in scope and technical approach. In our work, we propose an architecture to learn clip-level representations (i.e. tokens) using 3D ConvNets, followed by modeling clip-level temporal dependencies using MEGA [44]. To the best of our knowledge, our architecture is the first end-to-end deep network that captures local-global contextual information for long-range temporal reasoning.

Finally, it is also worth noting the work MetaFormer [71] and convolutional vision Transformer (CvT) [69]. MetaFormer uses convolutions as a token mixer in the bottom stages and vanilla self-attention in the top stages for image classification. CvT proposes a hierarchy of Transformers containing a convolutional token embedding and a convolutional Transformer block. Adding convolution allows dropping the position embedding from the network without hurting performance. These works brought inspiration for attempts at mixer neural network architectures for video and action recognition.

3. Method

Our proposed method is a spatio-temporal neural network, which extracts appearance and temporal information in a novel way to maximize engagement of the short-range and long-range video sequences. The motivation for proposing this method is deeply rooted in the need for long-range video reasoning. Our proposed method aims to exploit spatio-temporal information in two stages: 1.) Short Clip Stage (SCS) and 2.) Long Video Stage (LVS). SCS focuses on short-range video clips, primarily exploiting short-range dynamics and spatial context, and LVS models long-range temporal reasoning using cues from all short-range clips jointly. SCS consists of 3D convolution blocks and multi-headed attention modeling, while LVS consists of variants of long-range temporal modeling mechanisms for a global video understanding. We name our method *Spatio-Temporal Convolution-Attention Video Network* ***(STCA)***. Figure. 2 sketches our method.

3.1. Short Clip Stage (SCS)

In this section, we describe the SCS. SCS has two modules they are 1.) 3D convolution blocks; and 2.)

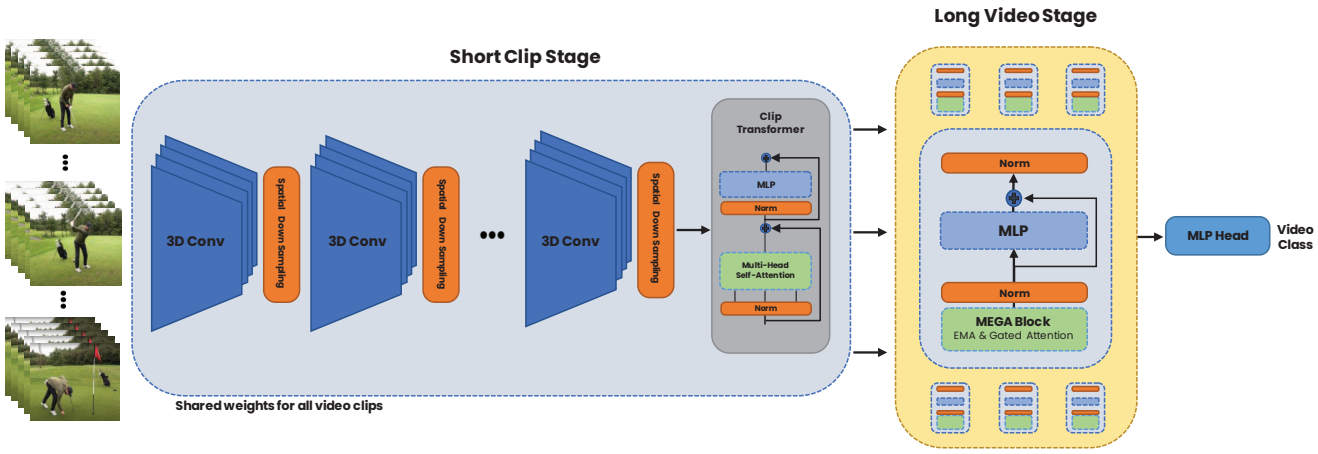

Figure 2. **Spatio-Temporal Convolution-Attention Video Network (STCA)**. Our STCA architecture is composed of two stages: 1.) Short Clip Stage (SCS) and 2.) Long Video Stage (LVS). SCS focuses on short-range video clips, primarily exploiting short-range dynamics and spatial context, and LVS models long-range temporal reasoning using cues from all short-range clips jointly. SCS consists of 3D-Conv blocks and Transformer layers, while LVS consists of long-range temporal modeling method for a global video understanding.

multi-headed self-attention layers. The 3D Convolutions (3DConv) module handles spatio-temporal cues dealing with short-range video clips (temporal window of 8 frames each). 3D-Conv aims to capture the relative temporal information between frames. We use R3D and 3D-ResNet architectures 3D modules. In Table 1, we show the architectural details of the 3D-Conv module. Our 3DConv operates on video frames rather than 3D tubelet patches [2]; this allows us to learn a spatio-temporal clip-level representation. Further, this reduces a large number of patch-based spatio-temporal tokens, thus reducing the quadratic complexity of the self-attention operation to a great extent. The 3D-Conv representation is fed into the standard transformer block with multi-headed self-attention, where the representation of each frame (token) interacts with every other frame (tokens), thus leading to a powerful spatio-temporal short-range clip representation.

We divide the video into non-overlapping $\mathcal{C}$ clips with F frames each. The $\mathcal{C}$ short-range clips are fed as input to the parallel 3D ConvNets. The 3DConvNet weights are shared. The output feature maps are then fed to the transformer block uniquely. In SCS, each clip ($V \in \mathbb{R}^{F\times H\times W\times 3}$) is processed separately using shared blocks of 3D ConvNets and Transformer. The output feature maps of the 3D convolutions fed to the transformers are of size $Z \in \mathbb{R}^{F\times D}$ where D is the feature dimension. In other words, for a given video clip v, there are F tokens $(\mathbf{z}_1^v, \ldots, \mathbf{z}_i^v, \ldots, \mathbf{z}_F^v)$, $\mathbf{z}_i^v \in \mathbb{R}^D$ with D feature dimension. The transformer block follows the architecture from ViT [13] transformer encoder. The transformer encoder is a stack of several transformer blocks. Each layer in the transformer block comprises of Multi-Headed Self-Attention (MSA), layer normalization (LN), and MLP blocks. These operations are as follows:

$$\begin{aligned} \mathbf{y} &= \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{z}\right)\right) + \mathbf{z} \\ \mathbf{o} &= \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{y}\right)\right) + \mathbf{y} \end{aligned} \tag{1}$$

Block	Layer	output size $T \times S^2$
raw clip	-	$8 \times 244 \times 224$
conv_1	$5 \times 7 \times 7$, 8, stride 1, 2, 2	$8 \times 112 \times 112$
pool_1	$1 \times 3 \times 3$, max, stride 1, 2, 2	$8 \times 56 \times 56$
res_2	$\begin{bmatrix} 3\times1\times1, 8 \\ 1\times3\times3, 8 \\ 1\times1\times1, 32 \end{bmatrix} \times 2$	$8 \times 56 \times 56$
res_3	$\begin{bmatrix} 1\times1\times1, 64 \\ 1\times3\times3, 64 \\ 1\times1\times1, 256 \end{bmatrix} \times 2$	$8 \times 28 \times 28$
res_4	$\begin{bmatrix} 1\times1\times1, 128 \\ 1\times3\times3, 128 \\ 1\times1\times1, 512 \end{bmatrix} \times 2$	$8 \times 14 \times 14$
res_5	$\begin{bmatrix} 3\times1\times1, 256 \\ 1\times3\times3, 256 \\ 1\times1\times1, 1024 \end{bmatrix} \times 2$	$8 \times 7 \times 7$
	spatial global average pool	$8 \times 1 \times 1$

Table 1. **3D ConvNet architecture of SCS** operating on a video clip with 8 frames. The proposed architectures incorporate 3D filters and pooling kernels. Each convolution layer shown in the table corresponds the composite sequence BN-ReLU-Conv operations.

3.2. Long Video Stage (LVS)

In this section, we describe the LVS. LVS models long-range temporal reasoning using representation (i.e. tokens)

obtained for $\mathcal{C}$ short-range video clips from SCS. LVS temporally aggregates and encodes spatio-temporal dynamic cues coming from $\mathcal{C}$ video clips into a compact and robust global video feature representation. The LVS block is a high-level local-global video encoder and, in practice, can handle variable video lengths for reasoning.

First, we aggregate the representations (i.e. tokens) obtained for $\mathcal{C}$ short-range video clips from SCS by simply averaging tokens of each video clip, given as:

$$\mathbf{s}^v = \frac{1}{F}\sum_{i=1}^{F}\mathbf{o}_i^v \tag{2}$$

where $\mathbf{s} \in \mathbb{R}^D$ and $S \in \mathbb{R}^{C\times D}$ We found that using only classification tokens from each video clip drops the performance due to the loss of contextual low-level spatio-temporal information within each clip.

In specific, the aggregated matrix S is fed as input to the LVS mechanism $E : S \rightarrow P$, resulting in an encoded representation P, $P \in \mathbb{R}^{C\times D}$, where $\mathbf{p}^0$ is the classification token, and D denotes the encoded feature dimension. The advantage of the encoding is that every short clip interacts with other clips, thus leading to a powerful global temporal representation of the entire video. In this work, we investigate three long-range temporal modeling methods E in the long video stage; they are:

- **MEGA**: Moving average equipped gated attention mechanism, namely MEGA [44], which is a gated single-head attention block equipped with an exponential moving average. MEGA is as expressive as the most commonly used multi-head attention and is used as a drop-in replacement for regular multi-head self-attention in Transformers. MEGA effectively models long-range sequential context by simply applying gated attention to each local chunk-wise attention and can go beyond the chunk boundary. Further, MEGA offers linear time and space complexity with minimum loss of contextual information. The operation of MEGA is given as:

$$\begin{aligned}\mathbf{y} &= \mathrm{LN}(\mathrm{MEGA}\,(\mathbf{s}))\\ \mathbf{p} &= \mathrm{LN}\,(\mathrm{MLP}\,(\mathbf{y}) + \mathbf{y})\end{aligned} \tag{3}$$

- **S4**: MEGA is also closely related to structured state-space sequence (S4) [24]. S4 leverages the HiPPO framework [23] to initialize its low-rank structured state matrices. S4 captures complex long-range dependencies in the sequential data. It has linear time and space complexity wrt. to the input sequence length, and it significantly reduces the computational cost of processing long video sequences. The architecture of S4 is given as follows:

$$\begin{aligned}\mathbf{x}_{s4} &= \mathrm{S4}(\mathrm{LN}\,(\mathbf{s}))\\ \mathbf{x}_{mlp} &= \mathrm{MLP}(\mathrm{Pooling}\,(\mathbf{x}_{s4}))\\ \mathbf{p} &= \mathbf{x}_{mlp} + \mathbf{s}\end{aligned} \tag{4}$$

- **Standard Transformer block**: Same as Section 3.1, we apply multi-headed self-attention (MSA) for long-range temporal reasoning using tokens from all short-range clips jointly. MSA has quadratic computational complexity.

Compared to the standard transformer block, S4 and MEGA offer linear time computational complexity and also robustness for long sequence modeling. Further, MEGA and S4 perform more robustly than MSA layers, apart from computational efficiency as shown in our evaluations. Thus, the scope of long-range sequence modeling on top of short-range clips is effective since real-world applications often span over larger intervals. One can readily employ other temporal modeling methods too.

The classification token ($\mathbf{p}^0$) yielded from the LVS block is then mapped to one of the action classes by the MLP head using a softmax classifier.

4. Experiments

In this section, we introduce the datasets and implementation details of our proposed method. Following, we demonstrate an extensive ablation study. Finally, we compare our method with the state-of-the-art methods on challenging video recognition datasets.

4.1. Datasets

We have evaluated the proposed video understanding architecture on six challenging benchmark video recognition datasets, namely HVU [8], Kinetics (400, 600, 700) [34], Something-Something V2 [22], and Long Video Understanding dataset [68]. We use the pre-defined training/testing splits and protocols provided originally.

Kinetics [34] is considered one of the largest video datasets focusing on human activities. Kinetics video samples are 10 seconds on average, and there are three versions of this dataset consisting of 400, 600, and 700 human action classes. We report the experimental results on all versions of Kinetics.

Something-Something [22] has more than 220,000 video clips of human interactions with commonly-used objects.

Long-form video understanding benchmark (LVU) [68] consists of the publicly available MovieClip dataset [1], which has 30K videos from 3000 movies. The main different property of this dataset is the temporal length. Each video sample is from one to three minutes long. The

benchmark includes nine tasks covering (1) content understanding, which consists of ('relationship', 'speaking style', 'scene/place') prediction, (2) metadata prediction, which includes ('director', 'genre', 'writer', and 'movie release year') classification, and (3) user engagement, which requires predicting ('YouTube like ratio' and 'YouTube popularity'). We have focused on the first two sets of tasks for our evaluations because of their classification-based nature.

Holistic Video Understanding(HVU) [8] is a large-scale video understanding dataset that focuses on multiple visual semantic categories like objects, scenes, actions, attributes and etc. This dataset has over 500K video clips ranging from 5 to 10 seconds in duration. It is considered one of the largest datasets which consider holistic visual understanding in videos from static and dynamic perspectives.

4.2. Implementation Details

We used SGD as the optimizer with a momentum of 0.9 following a cosine learning rate schedule with a linear warm-up. We only use standard data augmentations of cropping, resizing, and flipping. The input frame resolution is set to be 224 × 224 in both training and inference. We use a transformer with hidden dimension 1024. The 3D-Conv blocks were pre-trained like similar methods [10, 48] with the corresponding dataset and without extra data. The number of layers in the MEGA [44] block is set to 12. For evaluation metrics, we have used Top-1 and Top-5 classification accuracy for Kinetics and Something-Something datasets and used mean average precision for HVU and LVU datasets. The temporal stride for frame sampling from videos is set to 2. We have used 16 Nvidia V100 GPUs to train STCA models.

Inference. For inference on videos, we separate each video into non-overlapping clips. The STCA is applied over the video clips by taking a 224 × 224 center crop, and for a video-level prediction, we average the prediction scores over each batch of video clips.

Encoding method	Transformer	S4	MEGA
Kinetics-400 %acc	79.1	80.4	**82.5**

Table 2. Comparison of different temporal encoding methods for LVS block.

#video-clips	4	8	16	32
Kinetics-400	80.2	81.6	82.3	**82.5**
LVU-relationship	52.5	56.1	58.43	**59.25**
LVU-genre	50.73	51.25	54.82	**56.62**

Table 3. Effect of the increasing number of video clips on STCA.

#frames	4	8	12	16
Kinetics-400 %acc	80.2	**82.5**	82	81.9

Table 4. Effect of number-of-frames in the video clip on STCA.

# of clip transformer	1	2	4
Kinetics-400 %acc	81.7	**82.5**	82.1

Table 5. The effect of varying the number of *clip transformers* in SCS module of STCA.

4.3. Ablation study

The ablation studies were evaluated on the Kinetics-400 or LVU dataset to have a finer understanding of all aspects of the proposed network.

Long Temporal Encoder. Since most similar works have used different versions of Transformers or self-attention blocks for temporal and sequential modeling, we also studied the effect of different options in our Long Video Stage for the same purpose of modeling the sequence of video clips. The main proposed block is based on MEGA, as mentioned in the previous section, due to its less computation complexity and the higher ability for long sequence modeling. The other models which were used for the study are ViT-style transformer and structured state-space sequence (S4) layer. The S4 layer [25] proposed a computationally efficient State-Space Model (SSM) to capture long-range dependencies in the sequential data. In Table 2, the superior performance of the MEGA block as the base of our LVS block is presented. The results on Kinetics-400 show that MEGA is a better encoder for long temporal modeling than S4 or transformers. Our temporal encoding method using MEGA ensures higher performance both on classification accuracy and computational complexity. Based on this outcome, the rest of the experiments on the other datasets were done using MEGA as the backbone of LVS.

Number of video clips. Another part of our ablation study is about the efficiency of the proposed model using a different number of video clips as the input. For this experiment, we have used 4, 8, 16, and 32 video clips as input for training the pipeline. The evaluation has been done both on Kinetics-400 and LVU (relationship and genre categories) to see the effect on both relatively short and long video sources. Table 3 shows the larger number of clips has a bigger impact on the LVU dataset as it has longer videos; therefore, a model with a higher capacity of long sequence modeling perform more effectively.

Clip frame number. The frame number of video clips is also another parameter that we considered during our

studies. We have set a fixed number of video frames as 64 and divided the whole video into clips with sizes of 4, 8, 12, and 16. This way, we were able to find the optimized option for the size of the videos. Based on the results in Table 4, the clip size of 8 yield the best results in our ablation study. It is a trade-off between exploiting short-range spatio-temporal information in the SCS phase and the number of input tokens to the LVS phase for longer temporal reasoning.

Number of clip transforms. We have evaluated another hyper-parameter which is a different number of transformers in the SCS block. We have called this transformer in our method as *clip transformer* since it deals with tokens from individual frames in the clip. We observe that in Table 5, the method is not sensitive to the number of transformer layers. We have tried 1, 2, and 4 layers for this evaluation. Therefore, we have set 2 layers of *clip transformer* to guarantee robust results for all the other experiments.

Method	Top-1	Top-5	TFLOPs
I3D-NL [66]	77.7	93.3	10.77
TSM-ResNeXt-101 [41]	76.3	-	-
TEA [40]	76.1	92.5	2.10
VidTR-L [73]	79.1	93.9	10.53
LGD-3D R101 [49]	79.4	94.41	-
X3D-XXL [15]	80.4	94.6	5.82
SlowFast R101-NL [17]	79.8	93.9	7.02
MFormer-HR [47]	81.1	95.2	28.76
MViT-B [14]	81.2	95.1	4.10
TimeSformer-L [3]	80.7	94.7	7.14
ViViT-L FE [2]	81.7	93.8	11.94
MTV-B [70]	81.8	95.0	4.79
Uniformer-B [39]	83.0	-	3.1
STCA	**83.4**	**95.4**	4.65

Table 6. Comparisons of STCA with the state-of-the-art on Kinetics 400.

Method	Top-1	Top-5
X3D-XL [15]	81.9	95.5
TimeSformer-L [3]	82.2	95.6
MFormer-HR [47]	82.7	96.1
SlowFast R101-NL [17]	81.8	95.1
ViViT-L FE [2]	82.9	94.6
MViT-B [14]	83.8	96.3
MoViNet-A6 [36]	83.5	-
MTV-B [70]	83.6	96.1
Uniformer-B [39]	84.5	-
STCA	**85.8**	**97.2**

Table 7. Comparisons of STCA with the state-of-the-art on Kinetics 600.

Method	Top-1	Top-5
VidTR-L [73]	70.2	–
SlowFast R101 [17]	71.0	89.6
MoViNet-A6 [36]	72.3	–
MTV-L [70]	75.2	91.7
STCA	**77.2**	**92.6**

Table 8. Comparisons of STCA with the state-of-the-art on Kinetics 700.

Method	Top-1	Top-5
SlowFast R50 [17]	61.7	–
TimeSformer-HR [3]	62.5	–
VidTR [73]	63.0	–
ViViT-L FE [2]	65.9	89.9
MViT [14]	67.7	90.9
MFormer-L [47]	68.1	91.2
MTV-B [70]	67.6	90.1
Uniformer-B [39]	70.4	-
STCA	**70.6**	**92.7**

Table 9. Comparisons of STCA with the state-of-the-art on Something-Something V2.

4.4. State-of-the-art (SOTA) comparison

Based on the ablation studies insights in the last part, we obtained the results for all the datasets and compared our proposed video recognition model with the SOTA methods. The parameters of the model, such as the number of frames and the number of clips, were set based on the best results obtained in the ablation studies. Also, the computation complexity on all the datasets is similar since we have used a fixed setup for all the SOTA comparisons on the datasets.

Kinetics. Table 6, Table 7, Table 8 present that our STCA yielded comparable and even superior results to the SOTA on Kinetics 400, 600 and 700 with a reasonable computation complexity. We take three spatial crops (left, center, and right) following common practice by other similar works [2, 15]. Our STCA method outperforms the previous 3D-ConvNet or Transformer based SOTA like X3D [15] or ViViT [2] pre-trained on ImageNet, and also outperform [3] who proposed a pure-transformer architecture. In our experiments, we compared STCA with the models that only use ImageNet or Kinetics datasets for pre-training and not web-scaled datasets like JFT [57] or [45] datasets. Our best model using only the Kinetics dataset for training shows a robust performance with the best-reported Top-1 results on Kinetics 400, 600, and 700 to 82.5% and 85.3%, 77.2%, respectively.

Method	'relationship'	'speaking style'	'scene/place'	'director'	'genre'	'writer'	'release year'
SlowFast+NL [17]	52.40	35.80	54.70	44.90	53.00	36.30	52.50
VideoBERT [56]	52.80	37.90	54.90	47.30	51.90	38.50	36.10
Obj. Transformer [68]	53.10	39.40	56.90	51.20	54.60	34.50	39.10
Long Seq. Transformer [29]	52.38	37.31	62.79	56.07	52.70	42.26	39.16
ViS4mer [29]	57.14	40.79	67.44	62.61	54.71	48.8	44.75
STCA	59.25	41.62	69.15	66.7	56.62	52.93	53.3

Table 10. Comparisons of STCA with the state-of-the-art on Long Video Understanding (LVU) dataset,

Method	Scene	Object	Action	Event	Attribute	Concept
3D-ResNet [8]	50.6	28.6	48.2	35.9	29	22.5
3D-STCNet [8]	51.9	30.1	50.3	35.8	29.9	22.7
X3D [15]	53.2	31.7	50.9	36.5	31.4	25.2
HATNet [8]	55.8	34.2	51.8	38.5	33.6	26.1
HATNet(Multi-Task) [8]	57.2	35.1	53.5	39.8	34.9	27.3
STCA	61.6	40.2	58.1	42.4	38.7	31.2

Table 11. Comparisons of STCA with the state-of-the-art on Holistic Video Understanding HVU dataset.

Something-Something v2 (SSv2). Table 9 shows that STCA achieves SOTA Top-1 accuracy with the Convolutional-Attention modules. Our model could show a better performance than recent transformer-based methods like ViViT [2], MTV [70] and TimeSformer-HR [3], as well as Convolutional-based methods like SlowFast [17] and X3D [15] by a reasonable margin. Despite the differences between the videos from this dataset and Kinetics, the performance is coherent with what we have observed in Kinetics results. This proves that the STCA network is also highly capable of modeling human-object interactions and minor and fine scene dynamics.

Long-form video understanding(LVU). In Table 10, we compare our STCA model with the related and recent methods validated on the LVU dataset. Here, we have compared our method with a few famous 3D-CNN methods like SlowFast and Transformers like VideoBERT [56] and VIS4Mer [29]. These methods have shown they can present a reasonable performance even with longer videos like instances from LVU dataset. Table 10 shows that STCA has an outstanding performance in many tasks of LVU, proving that our two-stage network setup using MEGA block as the long-temporal modeler is more robust than large models like VideoBERT or other long sequence models like S4. This is proof that our proposed method is not only robust for action recognition but is an efficient, suitable choice for understanding different contents like meta-data or even genres from long videos.

Holistic video understanding (HVU). Table 11 shows the performance of STCA network on Holistic Video Understanding datasets with different semantic categories. STCA outperformed previous methods in all these categories. This result suggests that our proposed method can capture both appearance and dynamic clues to perceive all aspects of video content even when it is compared to an architecture like HATNet [8], which is a dual-stream 3D/2D network specifically designed to recognize appearance and motion concepts.

5. Conclusion

In this work, we handle the problem of long-range video reasoning. We propose a Spatio-Temporal Convolution-Attention Video Network (STCA), a method to efficiently capture spatio-temporal information and long-range reasoning for video recognition. Our proposed method obtains the most relevant appearance and dynamic clues from short clips and then classifies the content of videos by understanding the relations between these clips. STCA network is capable of understanding long-range videos, and it does not suffer from a quadratic computational complexity for long sequence modeling. Convolution-Attention based architecture helps STCA to keep the benefits of robustness and linear complexity while outperforming challenging video recognition datasets. Our architecture is generic and can facilitate other tasks like video captioning or video summarization that, in principle, need visual recognition models that can handle both short-range and long-range appearance and temporal relations. We believe our work opens many possibilities for further exploration.

Acknowledgements: This work was supported by GC4

Flemish AI project, and part of the work was done during Vivek Sharma's position at MIT.

References

[1] Movieclip dataset. https://www.movieclips.com. 5

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 2, 3, 4, 7, 8

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3, 7, 8

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 3

[5] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006. 2

[6] Ali Diba, Mohsen Fayyaz, Vivek Sharma, M Mahdi Arzani, Rahman Yousefzadeh, Juergen Gall, and Luc Van Gool. Spatio-temporal channel correlation networks for action classification. In *ECCV*, 2018. 3

[7] Ali Diba, Mohsen Fayyaz, Vivek Sharma, A Hossein Karami, M Mahdi Arzani, Rahman Yousefzadeh, and Luc Van Gool. Temporal 3d convnets using temporal transition layer. In *CVPR Workshops*, 2018. 3

[8] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision*, pages 593–610. Springer, 2020. 1, 2, 3, 5, 6, 8

[9] Ali Diba, Vivek Sharma, Luc Van Gool, and Rainer Stiefelhagen. Dynamonet: Dynamic action and motion network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6192–6201, 2019. 3

[10] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1502–1512, 2021. 6

[11] Ali Diba, Vivek Sharma, and Luc Van Gool. Deep temporal linear encoding networks. In *CVPR*, 2017. 1, 2

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

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 4

[14] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 7

[15] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020. 7, 8

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *ICCV*, 2019. 3

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 7, 8

[18] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 2, 3

[19] Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, 2015. 2

[20] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Temporal localization of actions with actoms. *PAMI*, 2013. 2

[21] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*, 2017. 2

[22] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The something something video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2, 5

[23] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in Neural Information Processing Systems*, 33:1474–1487, 2020. 5

[24] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2, 5

[25] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 6

[26] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *ICCV*, 2017. 1

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 3

[28] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 3

[29] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. *arXiv preprint arXiv:2204.01692*, 2022. 2, 3, 8

[30] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 3

[31] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 3

[32] M Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *European Conference on Computer Vision*, pages 731–747. Springer, 2020. 3

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

[34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 5

[35] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 2

[36] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, 2021. 7

[37] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 1

[38] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2

[39] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022. 7

[40] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2020. 7

[41] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 7

[42] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. *arXiv preprint arXiv:2204.02874*, 2022. 3

[43] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 3

[44] Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. Mega: Moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*, 2022. 3, 5, 6

[45] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 7

[46] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010. 2

[47] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021. 7

[48] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[49] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12056–12065, 2019. 7

[50] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM MM*, 2007. 2

[51] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 3

[52] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915*, 2021. 2, 3

[53] Vivek Sharma, Ali Diba, Davy Neven, Michael S Brown, Luc Van Gool, and Rainer Stiefelhagen. Classification-driven dynamic image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4033–4041, 2018. 3

[54] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1, 2, 3

[55] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[56] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 3, 8

[57] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 7

[58] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *ICCV*, 2015. 3

[59] Peng Tang, Xinggang Wang, Baoguang Shi, Xiang Bai, Wenyu Liu, and Zhuowen Tu. Deep fishernet for object classification. *arXiv:1608.00182*, 2016. 2

[60] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1, 2, 3

[61] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. Convnet architecture search for spatiotemporal feature learning. *arXiv:1708.05038*, 2017. 3

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[63] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 2

[64] Limin Wang, Yu Qiao, and Xiaoou Tang. Video action detection with relational dynamic-poselets. In *ECCV*, 2014. 2

[65] 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 *European conference on computer vision*, pages 20–36. Springer, 2016. 1, 2, 3

[66] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 7

[67] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008. 2

[68] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021. 2, 5, 8

[69] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 2, 3

[70] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022. 2, 3, 7, 8

[71] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *arXiv preprint arXiv:2210.13452*, 2022. 2, 3

[72] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015. 2, 3

[73] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13577–13587, 2021. 7

[74] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 3