Video-FocalNets: Spatio-Temporal Focal Modulation for Video Action Recognition

Syed Talal Wasim[1*] Muhammad Uzair Khattak[1*] Muzammal Naseer[1]
Salman Khan[1,2] Mubarak Shah[4] Fahad Shahbaz Khan[1,3]

[1]Mohamed bin Zayed University of AI [2]Australian National University
[3]Linköping University [4]University of Central Florida

Abstract

Recent video recognition models utilize Transformer models for long-range spatio-temporal context modeling. Video transformer designs are based on self-attention that can model global context at a high computational cost. In comparison, convolutional designs for videos offer an efficient alternative but lack long-range dependency modeling. Towards achieving the best of both designs, this work proposes Video-FocalNet, an effective and efficient architecture for video recognition that models both local and global contexts. Video-FocalNet is based on a spatio-temporal focal modulation architecture that reverses the interaction and aggregation steps of self-attention for better efficiency. Further, the aggregation step and the interaction step are both implemented using efficient convolution and element-wise multiplication operations that are computationally less expensive than their self-attention counterparts on video representations. We extensively explore the design space of focal modulation-based spatio-temporal context modeling and demonstrate our parallel spatial and temporal encoding design to be the optimal choice. Video-FocalNets perform favorably well against the state-of-the-art transformer-based models for video recognition on five large-scale datasets (Kinetics-400, Kinetics-600, SS-v2, Diving-48, and ActivityNet-1.3) at a lower computational cost. Our code/models are released at https://github.com/TalalWasim/Video-FocalNets.

1. Introduction

State-of-the-art video recognition methods have been significantly influenced by Convolutional Neural Networks (CNNs) since the introduction of Alexnet [36]. Initially 2D [30, 48, 55] and later 3D [7, 19, 63] CNNs achieved better performance on both small-scale [37, 57] and large-scale [6, 20, 31] video recognition benchmarks. With their local connectivity and translational equivariance properties, CNNs have a better inductive bias especially useful for learning on small datasets. However, CNNs are limited in their ability to model long-range dependencies due to their limited receptive field. On the other hand, Vision Transformers (ViTs) [13] offer long-range context modeling and have been quite effective for image classification [13, 45, 46] and video recognition [2, 4, 47, 75]. ViTs are based on the self-attention [65] mechanism originally proposed in Natural Language Processing (NLP) that encodes minimal inductive biases and can model both short- and long-range dependencies. This allows ViTs to better generalize to large datasets, as shown by recent results on major video recognition benchmarks [6, 20, 31] where they have out-performed their CNN counterparts. However, ViTs come at a high computational and parameter cost [76].

Video recognition requires both short-range and long-

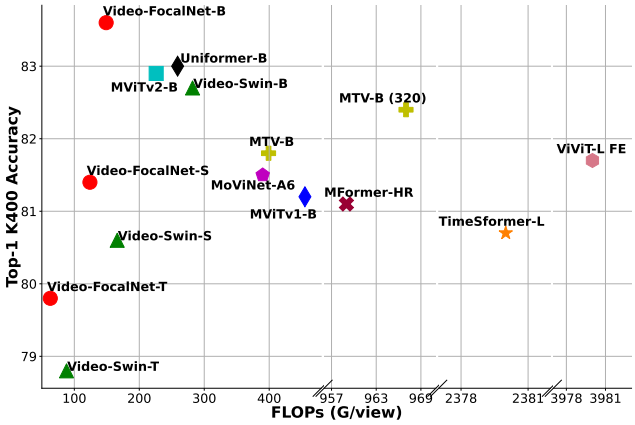


Figure 1: *Accuracy vs Computational Complexity trade-off comparison:* We show the performance of Video-FocalNets against recent methods for video action recognition. Accuracy is compared on the Kinetics-400 [31] dataset against GFLOPs/view. Our Video-FocalNets perform favorably compared to their counterparts across a range of model sizes (Tiny, Small, and Base).

*Joint first authors.
wasimtalal@gmail.com

range spatio-temporal dependencies to be accurately modeled in order to achieve high performance. However, existing methods demonstrate a trade-off between efficiency and performance. While CNNs are more efficient and suited for short-range information modeling, they are limited in their representation learning capabilities for long-range dependencies and larger datasets. ViTs resolve these issues but at an increased parametric complexity and high computational cost. The high complexity originates from the dual-step self-attention operation that first performs a query-key *interaction*, followed by an *aggregation* over the context values. The query-key interaction requires the computationally expensive step of calculating token-to-token attention scores via dot-product since the queries and keys do not contain information about the surrounding context (they are simply linear projections of the input tokens). In this context, this work seeks to optimize efficiency and performance while modeling both local and global contexts in videos.

We present an effective and efficient architecture for video recognition named Video-FocalNet (Fig. 1). Video-FocalNet proposes a spatio-temporal focal modulation architecture that reverses the steps of the self-attention operation for better efficiency. This architecture is inspired by focal modulation [76] for image recognition and extends it to videos by independently *aggregating* the surrounding spatial and temporal context for each token into spatial and temporal modulators, followed by fusing them with the queries in the *interaction* step. The aggregation is based on a hierarchical contextualization step using a stack of depthwise and pointwise convolutions for the spatial and temporal branches, respectively, followed by a gated aggregation that enables modeling both short- and long-range dependencies. The aggregation step (based on depthwise/pointwise convolutions) and the interaction step (based on element-wise multiplication) are both computationally less expensive than their self-attention counterparts i.e., query-key interactions and query-value aggregation via matrix multiplications.

We extensively explore various design configurations for optimal spatio-temporal context modeling with focal modulation. Our analysis shows that the proposed parallel spatial and temporal focal modulation design offers the best performance and is suitably efficient compared to other sequential designs. We introduce a family of Video-FocalNet architectures (tiny, small, and base) based on spatio-temporal focal modulation and demonstrate their favorable performance compared to state-of-the-art transformer-based methods on video recognition at a lower computational cost. Our major contributions are summarized as follows:

- We tackle the challenge of effective spatio-temporal modeling for video recognition. To solve this challenge, we propose a video-focal modulation block that is able to use computationally efficient depthwise and pointwise convolutions through a hierarchical context aggregation design for local-global context modeling.
- We explore various design choices for spatio-temporal focal modulation and propose a parallel design for spatial and temporal encoding that optimizes for both performance and computation cost as shown in Fig. 5.
- We achieve state-of-the-art performance on three major benchmarks: Kinetics-400 [31], Kinetics-600 [5] and Something-Something-v2 [20], surpassing comparable methods in literature by 0.6%, 1.2% and 0.6% respectively. Also, we outperform previous works on the relatively smaller Diving-48 [41] and ActivityNet-1.3 [24] datasets. We achieve an optimal trade-off between accuracy and computation cost as shown in Fig. 1.

2. Related Work

Video Recognition: Early methods in video recognition are feature-based [34, 38, 66]. However, with the startling success of 2D CNNs [23, 36, 56, 62] on Imagenet [12], they were also introduced to the task of video recognition [30, 48, 55]. Later, after the release of large-scale datasets, such as Kinetics [31], the 3D CNN-based methods were introduced [7, 19, 63]. These were much more effective in modeling spatio-temporal relations and outperformed 2D CNN-based methods. However, the computational cost for these 3D CNN-based methods was quite prohibitive. Therefore, various variants of the 3D CNNs were introduced [14, 17, 18, 40, 44, 53, 58, 60, 64, 72], which decreased computation cost and improved performance. With the success of the Vision Transformer [13] for image recognition, they were also introduced to video recognition. The first methods in this area used a combination of Vision Transformers and CNNs [35, 69, 70], including transformer blocks to model the longer range context. Later advancements then introduced fully transformer based architectures [2, 4, 16, 42, 47, 51, 75, 83], which outperformed all previous methods across multiple benchmarks. Recently, a new method [39] has been proposed, combining CNNs and ViTs, which achieves comparable performance to state-of-the-art fully transformer-based methods.

Global Context Modeling: Due to the localized nature of 2D CNNs, global context modeling was lacking in pure 2D CNN-based computer vision methods. Self-attention [65] was introduced to model long-range dependencies for visual inputs. However, self-attention comes at a high computation cost due to the required matrix multiplications. Various approaches have been introduced to address this problem. These include local window based attention [10, 45, 46, 50, 52, 77, 81], along with variants that add global tokens to model global information [1, 3, 29, 49, 79]. To reduce the computation cost, some methods used computationally efficient patterns for attention such as strided [8] and axial [27] patterns, as well as attention computed along

the channel dimension rather than the token dimension [15]. Other methods also combined convolution and self-attention for local and global modeling [21, 43, 71, 74]. Various methods for linearizing self-attention were also investigated, including the projection of token dimensions [33, 68], factorizing the softmax-attention kernel [9, 54, 73]. Using CNNs, a new method for modeling global context, termed focal modulation [76], has been proposed recently. To model local and global information, focal modulation employs hierarchical context aggregation to combine information from increasing sizes of receptive fields.

3. Methodology

Let us assume a video input is encoded to produce a feature representation $\mathbf{X}_{st} \in \mathbb{R}^{T \times H \times W \times C}$ with T frames, $H \times W$ spatial resolution, and C channels respectively. To obtain the spatio-temporal context enriched representation $\boldsymbol{y}_i \in \mathbb{R}^C$ for a given token (query) $\boldsymbol{x}_i \in \mathbb{R}^C$ in the input spatio-temporal feature map $\mathbf{X}_{st}$, it is necessary to perform an *interaction* between the query and its neighboring spatial and temporal tokens, and then *aggregate* the resulting information over the surrounding spatio-temporal contexts. To effectively model a spatio-temporal input, it is important to encode both short-range and long-range dependencies for the enriched context modeling for videos.

Self-attention [65] which is used in state-of-the-art video recognition methods [2, 4, 16, 47, 51, 75, 83], uses a First Interaction, Last Aggregation (FILA) process which involves initially calculating the attention scores through the query and key interaction $\mathcal{T}_1$, followed by aggregation $\mathcal{M}_1$ over the contexts as shown in Eq. 1.

$$\boldsymbol{y}_i = \mathcal{M}_1(\mathcal{T}_1(\boldsymbol{x}_i, \mathbf{X}_{st}), \mathbf{X}_{st}). \tag{1}$$

Since the query and keys during the interaction process are simple linear projections of the input feature map, self-attention involves computationally expensive token-to-token attention score calculation through query-key interactions because individual keys do not contain information about the surrounding context.

Recently, a new encoding method, Focal modulation [76] has been proposed, which follows an early aggregation process by First Aggregation, Last Interaction (FALI) mechanism. Essentially, both self-attention and focal modulation involve the *interaction* and *aggregation* operations but differ in the sequence of operation. In the case of focal modulation, context aggregation $\mathcal{M}_2$ is performed first, followed by the interaction $\mathcal{T}_2$ between the queries and the aggregated features, as shown in Eq. 2.

$$\boldsymbol{y}_i = \mathcal{T}_2(\mathcal{M}_2(i, \mathbf{X}_{st}), \boldsymbol{x}_i). \tag{2}$$

The output of the aggregation is known as the *modulator* which encodes the surrounding context for each query. Note that the operator $\mathcal{M}_2$ in focal modulation is based on convolutions, which are computationally more efficient compared to $\mathcal{M}_1$ in self-attention. Similarly, the interaction operator $\mathcal{T}_2$ is a simple element-wise multiplication, compared to the token-to-token attention score computation in self-attention which has quadratic complexity.

The focal modulation process given by [76] works well on images by extracting the *spatial* context around a query token. However, to model spatio-temporal information, both the *spatial* and *temporal* contexts surrounding a single query token have to be extracted. To achieve this we propose our architecture, Video-FocalNets which explicitly models both intra-frame (spatial) and inter-frame (temporal) information. Our approach aims to independently model the spatial and temporal information by proposing a two-stream spatio-temporal focal modulation block, in which one branch learns spatial information and the other models the temporal information. By decoupling the spatial and temporal branches, we are able to separately extract and aggregate spatial and temporal context for each query token, generating spatial and temporal *modulators*. These modulators are then fused with the query tokens to build the final feature map.

Our design transfers the desirable qualities of late aggregation in focal modulation to the video tasks. Particularly, Focal Modulation is performed for each target token with the context centered around it, hence it is translationally invariant. It also decouples the queries from the context around them, allowing the queries to preserve fine-grained information, while the coarser context surrounding it is extracted. Focal modulation uses a hierarchical-gated aggregation method, to aggregate information across multiple levels of granularity. This allows for modeling both short- and long-range dependencies within a video while improving computational and parameter efficiency.

We now present our approach on spatio-temporal focal modulation in Sec. 3.1, specifying the *Hierarchical Contextualization* and *Gated Aggregation* processes for videos in Sec. 3.1.1 and Sec. 3.1.2, respectively. For consistency, we maintain the same terminology as proposed in [76]. Finally, we outline our network architecture variants in Sec. 3.2.

3.1. Spatio-Temporal Focal Modulation

To model the spatial and temporal dimensions, we propose a two-stream spatio-temporal focal modulation block. The overall architecture is presented in Fig. 2 and the design of the spatio-temporal focal modulation is presented in Fig. 3. We validate its effectiveness via detailed ablations and comparisons with alternate design choices in Sec. 4.3. For an input spatio-temporal feature map $\mathbf{X}_{st} \in \mathbb{R}^{T \times H \times W \times C}$, the two-stream spatio-temporal encoding process involves independent aggregations along the spatial and temporal dimensions, followed by a joint interaction with the queries,

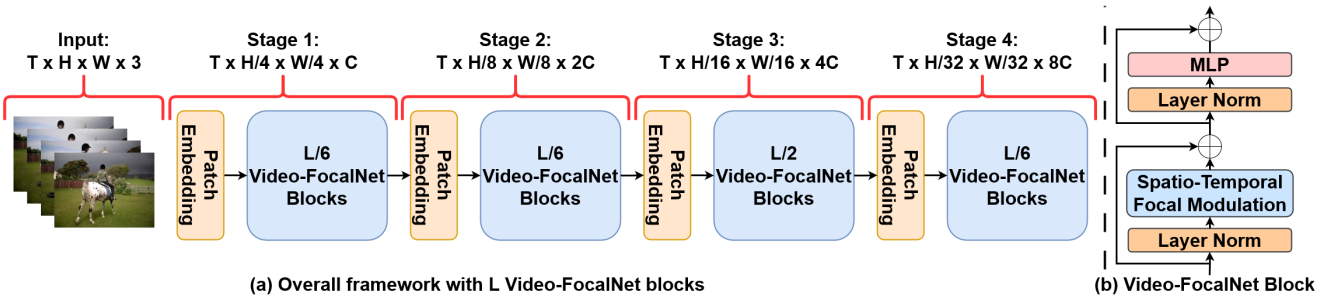


Figure 2: **(a) The overall architecture of Video-FocalNets:** Following [46, 76, 77], we define a four-stage architecture, with each stage comprising a patch embedding and a number of Video-FocalNet blocks. The total number of blocks is L, with stages one, two, three, and four having $L/6$, $L/6$, $L/2$, and $L/6$ blocks respectively. **(b) Single Video-FocalNet block:** Similar to the transformer blocks [65], we replace self-attention with Spatio-Temporal Focal Modulation.

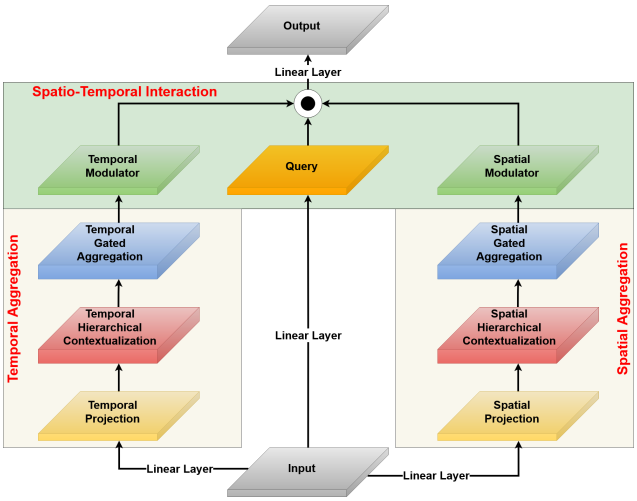


Figure 3: **The Spatio-Temporal Focal Modulation layer:** We design a spatio-temporal focal modulation block that independently models the spatial and temporal information. The input is first projected using linear layers to produce queries, spatial/temporal feature maps, and spatial/temporal gates. Then through hierarchical contextualization ($\mathcal{M}_s/\mathcal{M}_t$) and gated aggregation ($\mathbf{G}_s/\mathbf{G}_t$), the spatial and temporal modulators are produced. These then interact with the query tokens through element-wise multiplication operation ($\textcircled{\bullet}$) to form the final spatio-temporal feature map.

as shown in Eq. 3.

$$\boldsymbol{y}_i = \mathcal{T}_{st}(\mathcal{M}_s(i_t, \mathbf{X}_{st,t}), \mathcal{M}_t(i_{hw}, \mathbf{X}_{st,hw}), \boldsymbol{x}_i), \quad (3)$$

where $\mathbf{X}_{st,t} \in \mathbb{R}^{H\times W\times C}$ is a single spatial slice for the temporal dimension $t \in \{1, ..., T\}$, while i_t is the spatial location for the slice $t \in \{1, ..., T\}$. Similarly, $\mathbf{X}_{st,hw} \in \mathbb{R}^{T\times C}$ is a single temporal slice for the spatial dimensions $h \in \{1, ..., H\}$ and $w \in \{1, ..., w\}$, while i_{hw} is the temporal location. The operators $\mathcal{M}_s$ and $\mathcal{M}_t$ are based on the depth-wise convolution and point-wise convolution operators respectively and $\mathcal{T}_{st}$ is an element-wise multiplication. The spatio-temporal focal modulation process can therefore be defined as follows:

$$\boldsymbol{y}_i = q(\boldsymbol{x}_i) \odot m_s(i_t, \mathbf{X}_{st,t}) \odot m_t(i_{hw}, \mathbf{X}_{st,hw}), \quad (4)$$

where $q(\cdot)$ is a query projection function and $\odot$ is the element-wise multiplication, $m_s(\cdot)$ and $m_t(\cdot)$ are context aggregation functions, whose outputs are called *spatial modulator* and *temporal modulator* respectively. The formulation of $m_s(\cdot)$ and $m_t(\cdot)$ involves two steps: *Hierarchical Contextualization* and *Gated Aggregation*. The following Sec. 3.1.1 and Sec. 3.1.2 talk about *Spatio-Temporal Hierarchical Contextualization* and *Spatio-Temporal Gated Aggregation* respectively.

3.1.1 Spatio-Temporal Hierarchical Contextualization

We first project the input spatio-temporal feature map $\mathbf{X}_{st} \in \mathbb{R}^{T\times H\times W\times C}$ using two linear layers, producing $\mathbf{Z}_s^0$ and $\mathbf{Z}_t^0$, as defined by Eq. 5.

$$\begin{aligned} \mathbf{Z}_s^0 &= f_{z,s}(\mathbf{X}_{st}) \in \mathbb{R}^{T\times H\times W\times C}, \\ \mathbf{Z}_t^0 &= f_{z,t}(\mathbf{X}_{st}) \in \mathbb{R}^{T\times H\times W\times C}, \end{aligned} \quad (5)$$

where $f_{z,s}$ and $f_{z,t}$ are the spatial and temporal linear projection layers respectively. We then apply a series of L depth-wise convolutions (DWConv) and point-wise convolutions (PWConv) to the respective spatial and temporal projected inputs $\mathbf{Z}_s^0$ and $\mathbf{Z}_t^0$ along the spatial and temporal dimensions respectively. The outputs $\mathbf{Z}_s^\ell$ and $\mathbf{Z}_t^\ell$, at each focal level $\ell \in \{1, ..., L\}$, are therefore given as:

$$\begin{aligned} \mathbf{Z}_s^\ell &= f_{a,s}^\ell(\mathbf{Z}_s^{\ell-1}) \triangleq \mathsf{GeLU}(\mathsf{DWConv}(\mathbf{Z}_s^{\ell-1})) \in \mathbb{R}^{T\times H\times W\times C}, \\ \mathbf{Z}_t^\ell &= f_{a,t}^\ell(\mathbf{Z}_t^{\ell-1}) \triangleq \mathsf{GeLU}(\mathsf{PWConv}(\mathbf{Z}_t^{\ell-1})) \in \mathbb{R}^{T\times H\times W\times C}, \end{aligned} \quad (6)$$

where $f_{a,s}^\ell(\cdot)$ and $f_{a,t}^\ell(\cdot)$ are the spatial and temporal contextualization functions with GeLU [25] activation function.

To obtain the global representation, a global average pooling operation is performed along the spatial and temporal dimensions on $\mathbf{Z}_s^L$ and $\mathbf{Z}_t^L$ respectively as shown in Eq. 7.

$$\begin{aligned} \mathbf{Z}_s^{L+1} &= \text{Avg-Pool}(\mathbf{Z}_s^L), \\ \mathbf{Z}_t^{L+1} &= \text{Avg-Pool}(\mathbf{Z}_t^L), \end{aligned} \tag{7}$$

where Avg-Pool is the global average pool operator.

3.1.2 Spatio-Temporal Gated Aggregation

Next, we condense the respective spatial and temporal feature maps, $\mathbf{Z}_s^\ell$ and $\mathbf{Z}_t^\ell$, into the respective spatial and temporal modulators through a gating mechanism. We obtain the respective spatial and temporal gating weights, $\mathbf{G}_s = f_{g,s}(\mathbf{X}_{st}) \in \mathbb{R}^{H\times W\times(L+1)}$ and $\mathbf{G}_t = f_{g,t}(\mathbf{X}_{st}) \in \mathbb{R}^{T\times(L+1)}$, using the linear projection layers $f_{g,s}$ and $f_{g,t}$. This is followed by a dot product between the feature maps and their respective gates, as shown in Eq. 8.

$$\begin{aligned} \mathbf{Z}_s^{out} &= \sum_{\ell=1}^{L+1} \mathbf{G}_s^\ell \odot \mathbf{Z}_s^\ell \in \mathbb{R}^{H\times W\times C}, \\ \mathbf{Z}_t^{out} &= \sum_{\ell=1}^{L+1} \mathbf{G}_t^\ell \odot \mathbf{Z}_t^\ell \in \mathbb{R}^{T\times C}, \end{aligned} \tag{8}$$

where $\mathbf{Z}_s^{out}$ and $\mathbf{Z}_t^{out}$ are the single aggregated spatial and temporal feature maps and $\mathbf{G}_s^\ell \in \mathbb{R}^{H\times W\times 1}$ and $\mathbf{G}_t^\ell \in \mathbb{R}^{T\times 1}$ are slices of $\mathbf{G}_s$ and $\mathbf{G}_t$ respectively for the level ℓ. To enable communication across different channels, another set of linear layers, $h_s(\cdot)$ and $h_t(\cdot)$, are used to obtain the *spatial modulator* ($\mathbf{M}_s = h_s(\mathbf{Z}_s^{out}) \in \mathbb{R}^{T\times H\times W\times C}$) and *temporal modulator* ($\mathbf{M}_t = h_t(\mathbf{Z}_t^{out}) \in \mathbb{R}^{T\times H\times W\times C}$) respectively.

Therefore, the spatio-temporal focal modulation process defined by Eq. 4 can be rewritten as:

$$\boldsymbol{y}_i = q(\boldsymbol{x}_i) \odot h_s(\sum_{\ell=1}^{L+1} \boldsymbol{g}_{i,s}^\ell \cdot \boldsymbol{z}_{i,s}^\ell) \odot h_t(\sum_{\ell=1}^{L+1} \boldsymbol{g}_{i,t}^\ell \cdot \boldsymbol{z}_{i,t}^\ell) \tag{9}$$

where $\boldsymbol{z}_{i,s}^\ell$/$\boldsymbol{z}_{i,t}^\ell$ and $\boldsymbol{g}_{i,s}^\ell$/$\boldsymbol{g}_{i,t}^\ell$ are the spatial/temporal visual feature and spatial/temporal gating value at location i of $\mathbf{Z}_s^\ell$/$\mathbf{Z}_t^\ell$ and $\mathbf{G}_s^\ell$/$\mathbf{G}_t^\ell$ respectively.

3.1.3 Design Variations

We further compare our proposed spatio-temporal focal modulation design against various other possible designs shown in Fig. 4. This explorative study validates the proposed design to be the optimal one. The first design, **(a)**, is a simple extension of the spatial focal modulation to videos, which passes each frame through the spatial encoder (which uses only 2D depthwise convolution) and averages along the temporal dimension. Mathematically, Eq. 3 for this case can be re-written as:

$$\boldsymbol{y}_i = \mathcal{T}_{st}(\mathcal{M}_s(i_t, \mathbf{X}_{st,t})). \tag{10}$$

A variation of this design, **(b)**, uses factorized 3D convolution (2D depthwise followed by 1D pointwise convolution).

The next **(c)** uses a factorized encoder that stacks two encoders, one spatial (using 2D depthwise convolution) and one temporal (using 1D depthwise convolution), on top of each other. This is similar to the factorized encoder design presented by [2] but replaces spatial and temporal self-attention with spatial and temporal focal modulation.

The second last design **(d)** follows the concept of divided space-time attention proposed by [4] and uses alternating spatial and temporal focal modulation.

The final design **(e)** is the proposed spatio-temporal focal modulation. The accuracy and computation requirements for each are reported on the Kinetics-400 dataset in Fig. 5. It can be seen that the proposed design is the best in terms of accuracy and computation.

3.2. Network Variants

Following [47, 76], we use the same four-stage layouts and hidden dimensions as in [76], but replace the focal modulation block with our spatio-temporal focal modulation block. In each stage, a stack of L Video-FocalNet blocks is used, divided between the four stages as $\{L/6, L/6, L/2, L/6\}$. We introduce four different versions of Video-FocalNets. The architecture hyper-parameters of these model variants are:

- Video-FocalNet-T: $C = 96$, block$_{num}$ = $\{2, 2, 6, 2\}$
- Video-FocalNet-S: $C = 96$, block$_{num}$ = $\{2, 2, 18, 2\}$
- Video-FocalNet-B: $C = 128$, block$_{num}$ = $\{2, 2, 18, 2\}$

We use non-overlapping convolution layers for patch embedding at the beginning (kernel size=4×4, stride=4) and between two stages (kernel size=2×2, stride=2), respectively. The focal levels (L) for the models are set to 2 with the kernel for the first level set to $k^1 = 3$. We gradually increase the kernel size by 2 from lower focal levels to higher ones, *i.e.*, $k^\ell = k^{\ell-1} + 2$.

4. Results and Analysis

4.1. Experimental Setup and Protocols

Datasets: We report results for video action recognition on three large-scale datasets, Kinetics-400 (K400) [31], Kinetics-600 (K600) [5] and Something-Something-v2 (SS-v2) [20]. For each dataset, we train on the training set and evaluate on the validation set. K400 consists of ∼240k training and ∼20k testing videos across 400 classes. K600 consists of ∼370k training and 28.3k testing videos across 600 classes. SS-v2 consists of 169k training and 24.7k validation

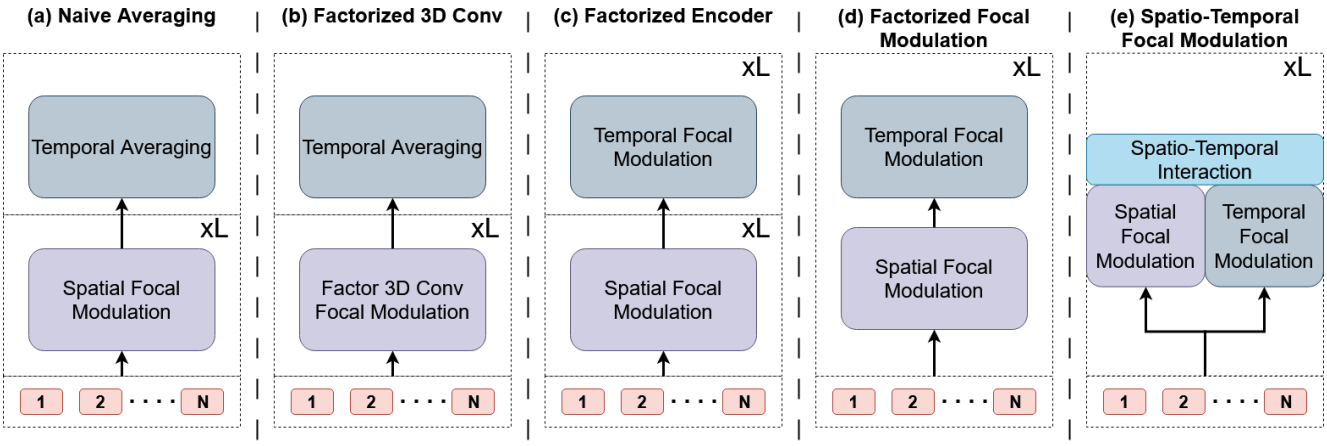


Figure 4: We show various design choices for spatio-temporal context modeling via Focal Modulation and evaluate them in Fig. 5. **(a)** A naive solution where the frames are passed through spatial focal modulation layers and averaged. **(b)** A variation of the naive solution replacing 2D depthwise convolution with factorized 3D convolution (2D depthwise followed by 1D pointwise convolution). **(c)** A factorized encoder design that stacks two encoders, one modeling spatial and the other the temporal dimension. **(d)** A design based on [4] which uses factorized spatial and temporal focal modulation. **(e)** Our proposed spatio-temporal focal modulation with parallel spatial and temporal branches followed by spatio-temporal interaction.

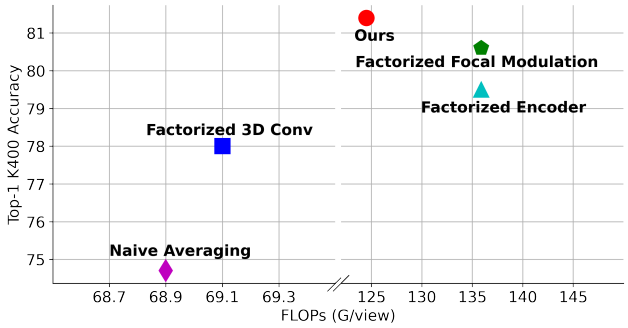


Figure 5: Comparison of various design choices for Video-FocalNet-S on Kinetics-400 [31] validation set.

videos across 174 classes. For all three datasets, we report Top-1 accuracy and compare it against the state-of-the-art.

We additionally test our Video-FocalNet on the Diving-48 (D-48) [41] and ActivityNet-1.3 (ANet-1.3) [24] datasets. D-48 is a challenging dataset of diving actions consisting of ∼15000 training and ∼2000 testing samples. Actions are only differentiated by the subtle movement of the diver across the frames with the background being majorly constant. This means that the dataset requires robust temporal modeling for good performance. In fact, it has been shown by [67], that disrupting (through random shuffling) or removing (by single frame evaluation) temporal information in this dataset can result in accuracy drops of up to $\sim 33.6\%$ and $\sim 70.2\%$ respectively. Alternatively, the ANet-1.3 dataset consists of untrimmed videos for action recognition tasks.

Implementation Details: For K400 and K600, we follow a similar training scheme to [39, 42] and train for 120 epochs with a linear warmup of 20 epochs using the SGD optimizer. We linearly scale the learning rate by $LR \times \frac{batchsize}{512}$ where $LR = 0.1$ is the base learning rate. The spatial modules are initialized from the pretrained Imagenet-1K FocalNet [76] weights, while the rest are randomly initialized. For augmentations, we follow a recipe similar to [16] with some variations. To each clip, we apply a horizontal flip, Mixup [80] (α=0.8), and CutMix [78], each with a probability of 0.5.

Table 1: Training hyperparameters for experiments in the main paper. "–" indicates that the regularisation method was not used at all. Values that are constant across all columns are listed once. Datasets are denoted as follows: K400: Kinetics-400. K600: Kinetics-600. SS-v2: Something-Something-v2. D-48: Diving 48. ANet-1.3: ActivityNet-1.3.

	K400	K600	SS-v2	D-48	ANet-1.3
Optimization					
Optimizer			SGD		
Batch size			512		
Learning rate schedule			cosine with linear warmup		
Linear warmup epochs			20		
Base learning rate			0.1		
Epochs			120		
Data augmentation					
Random crop probability			1.0		
Random flip probability	0.5	0.5	–	–	–
Scale jitter probability			1.0		
Maximum scale			1.33		
Minimum scale			0.75		
Colour jitter probability			0.8		
Other regularisation					
Stochastic droplayer rate [28]			0.1		
Label smoothing [59]			0.1		
Mixup ($\alpha = 0.8$) probability [80]			0.5		

Table 2: Comparison with state-of-the-art methods on Kinetics-400 [31].

Method	Pre-training	Top-1	Views	FLOPs (G/view)
TEA *(ICCV'21)* [40]	ImageNet-21K	76.1	10×3	70
TSM-ResNeXt-101 *(ICCV'21)* [44]	ImageNet-21K	76.3	-	-
I3D NL *(ICCV'21)* [69]	ImageNet-21K	77.7	10×3	359
VidTR-L *(ICCV'21)* [83]	ImageNet-21K	79.1	10×3	351
LGD-3D R101 *(CVPR'19)* [53]	ImageNet-21K	79.4	-	-
SlowFast R101-NL *(ICCV'19)* [18]	ImageNet-21K	79.8	10×3	234
X3D-XXL *(CVPR'20)* [17]	ImageNet-21K	80.4	10×3	194
OmniSource *(ECCV'20)* [14]	ImageNet-21K	80.5	-	-
TimeSformer-L *(ICML'21)* [4]	ImageNet-21K	80.7	1×3	2380
MFormer-HR *(NeurIPS'21)* [51]	ImageNet-21K	81.1	10×3	959
MViTv1-B *(ICCV'21)* [16]	-	81.2	3×3	455
MoViNet-A6 *(CVPR'21)* [35]	ImageNet-21K	81.5	1×1	390
ViViT-L FE *(CVPR'21)* [2]	ImageNet-21K	81.7	1×3	3980
MTV-B *(CVPR'22)* [75]	ImageNet-21K	81.8	4×3	399
MTV-B (320p) *(CVPR'22)* [75]	ImageNet-21K	82.4	4×3	967
Video-Swin-T *(CVPR'22)* [47]	ImageNet-1K	78.8	4×3	88
Video-Swin-S *(CVPR'22)* [47]	ImageNet-1K	80.6	4×3	166
Video-Swin-B *(CVPR'22)* [47]	ImageNet-1K	80.6	4×3	282
Video-Swin-B *(CVPR'22)* [47]	ImageNet-21K	82.7	4×3	282
MViTv2-B *(CVPR'22)* [42]	-	82.9	5×1	226
Uniformer-B *(ICLR'22)* [39]	ImageNet-1K	83.0	4×3	259
Video-FocalNet-T	ImageNet-1K	79.8	4×3	63
Video-FocalNet-S	ImageNet-1K	81.4	4×3	124
Video-FocalNet-B	ImageNet-1K	**83.6**	4×3	149

Table 3: Comparison with state-of-the-art methods on Kinetics-600 [5] dataset.

Method	Pre-training	Top-1
SlowFast R101-NL *(ICCV'19)* [18]	ImageNet-21K	81.8
X3D-XXL *(CVPR'20)* [17]	ImageNet-21K	81.9
TimeSformer-L *(ICML'21)* [4]	ImageNet-21K	82.2
MFormer-HR *(NeurIPS'21)* [51]	ImageNet-21K	82.7
ViViT-L FE *(CVPR'21)* [2]	ImageNet-21K	82.9
MTV-B *(CVPR'22)* [75]	ImageNet-21K	83.6
MTV-B (320p) *(CVPR'22)* [75]	ImageNet-21K	84.0
Video-Swin-B *(CVPR'22)* [47]	ImageNet-21K	84.0
Uniformer-B *(ICLR'22)* [39]	ImageNet-1K	84.5
MoViNet-A6 *(CVPR'21)* [35]	ImageNet-21K	84.8
MViTv1-B *(ICCV'21)* [16]	None	83.8
MViTv2-B *(CVPR'22)* [42]	None	85.5
Video-FocalNet-B	ImageNet-1K	**86.7**

Table 4: Comparison with state-of-the-art methods on Something-Something-v2 [20] dataset.

Method	Pre-training	Top-1
SlowFast R50 *(ICCV'19)* [18]	ImageNet-21K	61.7
TimeSformer-HR *(ICML'21)* [4]	ImageNet-21K	62.5
VidTR *(ICCV'21)* [83]	ImageNet-21K	63.0
ViViT-L FE *(CVPR'21)* [2]	ImageNet-21K	65.9
MFormer-L *(NeurIPS'21)* [51]	ImageNet-21K	68.1
MTV-B *(CVPR'22)* [75]	ImageNet-21K	67.6
MTV-B (320p) *(CVPR'22)* [75]	ImageNet-21K	68.5
Video-Swin-B *(CVPR'22)* [47]	Kinetics400	69.6
Uniformer-B *(ICLR'22)* [39]	Kinetics400	70.4
MViTv1-B *(ICCV'21)* [16]	ImageNet-21K	67.6
MViTv2-B *(CVPR'22)* [42]	Kinetics400	70.5
Video-FocalNet-B	Kinetics400	**71.1**

See detailed hyperparameters in Tab. 1.

During training, we sample T frames with a stride of τ, denoted as $T \times \tau$ [18]. For the spatial domain, we follow Inception [59] and take a crop of $H \times W = 224 \times 224$, with input area selected within a scale of $[\min, \max] = [0.08, 1.00]$ and aspect ratio jitter between $3/4$ and $4/3$. During inference, we report results as an average across $N_{clip} \times N_{crops}$ where a total of N_{clip} clips are uniformly sampled from the video, and for each video, N_{crops} spatial crops are taken during inference. For K400 and K600 we use 4×3 for inference. For SS-v2, D-48 and ANet-1.3 we follow the same training recipe as K400 and K600, with slight changes as followed by [16, 39, 42, 47]. We initialize our model with the K400 pretrained weights. For augmentations, we don't use the random horizontal flip and infer on 1×3 views.

Additionally, owing to the large scale of the Kinetics-400 [31] and Kinetics-600 [5] datasets, we preprocess the videos before starting to train. Following the guidelines of [11], each video is first resized, with the shorter side resized to 256 pixels.

4.2. Comparison with State-of-the-art

Kinetics-400: On the K400 [31] dataset, we report results for the Video-FocalNet-T, Video-FocalNet-S and Video-FocalNet-B variants, comparing against recent methods in Tab. 2. Considering first the T and S variants, it can be

Table 5: Comparison with state-of-the-art methods on Diving 48 V2 [41] dataset.

Method	Pre-training	Top-1
SlowFast R50 *(ICCV'19)* [18]	ImageNet-21K	77.6
TimeSformer-L *(ICML'21)* [4]	ImageNet-21K	81.0
RSANet R50 *(NeurIPS'21)* [32]	ImageNet-1K	84.2
VIMPAC *(arXiv'21)* [61]	HowTo100M	85.5
BEVT *(CVPR'22)* [67]	Kinetics400	86.7
GC-TDN *(CVPR'22)* [22]	ImageNet-1K	87.6
ORVIT Transformer *(CVPR'22)* [26]	ImageNet-21K	88.0
TFCNET *(arXiv'22)* [82]	ImageNet-1K	88.3
Video-FocalNet-B	Kinetics400	**90.8**

Table 6: Comparison on ActivityNet 1.3 [24] dataset.

Method	Pre-training	Top1 (%, ↑)(↓)
Video-Swin-B *(CVPR'22)* [47]	Kinetics400	88.5
Video-FocalNet-B	Kinetics400	**89.8**

seen that our method surpasses the equivalent Video-Swin Transformer [47] variants by 1.0% and 0.8% respectively while reducing the TFLOPs by 25%. Our larger base model, Video-FocalNet-B, surpasses the previous state-of-the-art Uniformer-B [39] and MViTv2-B [42] by 0.6% and 0.7% respectively, while maintaining comparable TFLOPs with MViTv2-B [42] and reducing TFLOPs by about 45% compared to Uniformer-B [39].

Kinetics-600: On the K600 [5] dataset, we report results for Video-FocalNet-B against recent methods in literature in Tab. 3. Compared to the previous state-of-the-art MViTv2-B [42], our Video-FocalNet-B achieves 1.2% higher performance. Our method using the ImageNet-1K initialization also surpasses previous methods pretrained on the larger ImageNet-21K dataset while maintaining much lower TFLOPs.

Something-Something-v2: On the SS-v2 [20] benchmark we report results for Video-FocalNet-B and compare against state-of-the-art methods in Tab. 4. On this temporally challenging benchmark, our method surpasses the previous state-of-the-art MViTv2-B [42] and Uniformer-B [39] by 0.6% and 0.7% respectively. This strong performance shows that our method can effectively model the subtle temporal changes and dependencies in this challenging dataset.

Diving-48: On D-48 [41] we report our results and compare them against recent methods in literature in Tab. 5. Video-FocalNet-B surpasses the previous state-of-the-art method TFCNET [82] by 2.5%. This shows that our method can effectively model the temporal information even when using a small number of training samples.

ActivityNet-1.3: For ANet-1.3 [24], results are presented in Tab. 6. Our proposed Video-FocalNet outperforms the baseline Video-Swin (CVPR'22) [47] model by a significant margin on the untrimmed video dataset. This demonstrates the efficacy of our method in localizing highlights and addressing the challenges posed by untrimmed videos. We appreciate your insightful suggestion and believe that evaluating our method on untrimmed video datasets further supports its potential in this challenging problem setting.

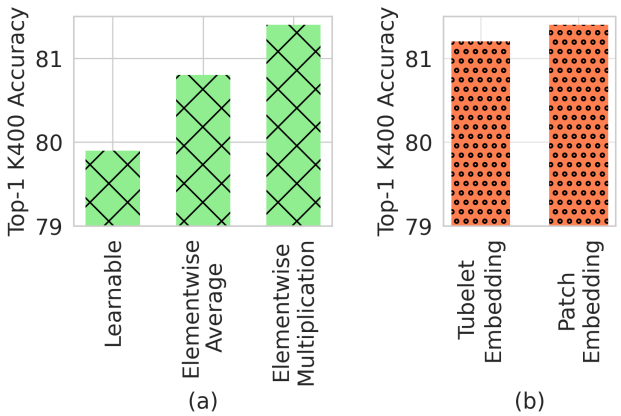


Figure 6: **(a)** Ablation of various modulator-query spatio-temporal interaction methods for Video-FocalNet-S on Kinetics-400 [31] validation set. **(b)** Ablation for using patch vs tubelet embedding for Video-FocalNet-S on Kinetics-400 [31] validation set. Note that the number of frames is adjusted to ensure that the number of tokens is the same.

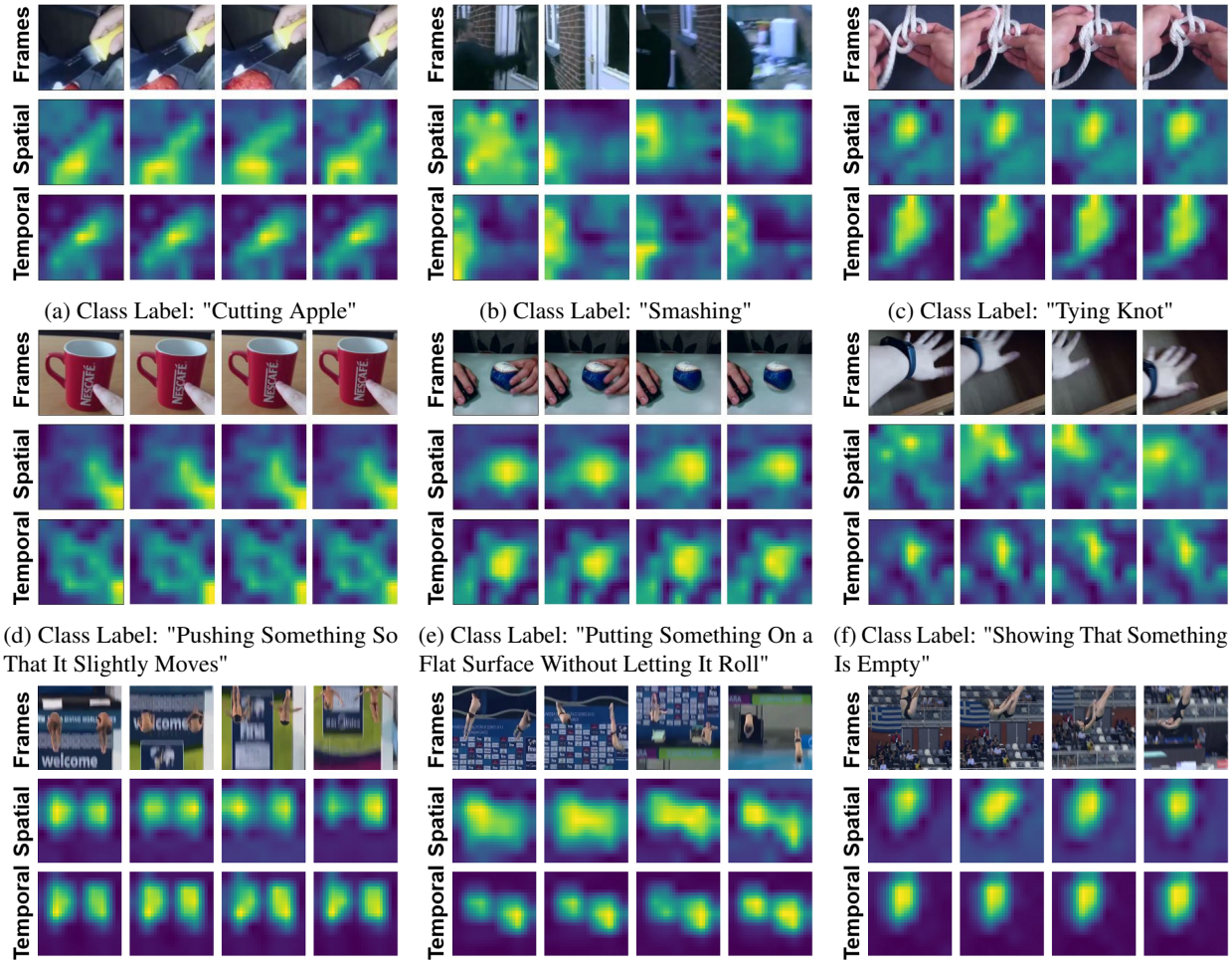
4.3. Ablations

In this section, we present an ablative analysis of various choices in our final design. Note that all ablations are performed using the Video-FocalNet-S variant on K400 using the same training settings as mentioned in Sec. 4.1.

Modulator Fusion Method: Since we propose a two-stream spatio-temporal focal modulation design, we end up with two modulators, one each for the spatial and temporal branches respectively, that need to be fused with the query tokens. We evaluate various fusion methods to see which works best. Fig. 6 **(a)** shows the comparison of three fusion techniques which include simple averaging, elementwise multiplication, and a learnable projection layer. We find that elementwise multiplication gives the best performance.

Patch Embedding vs Tubelet Embedding: Many recent works [2, 47, 75] propose encoding a tubelet of $T \times H \times W \times 3$, with $T = 2$, into a single token rather than patch embedding with $T = 1$. We evaluate this design choice for our model and find that a simple patch embedding works better for us, as shown in Fig. 6 **(b)**.

Visualizations: We visualize the spatial and temporal modulators for sample videos across two datasets, K600 and SS-V2 in Fig. 7. We note that our modulators focus on the salient parts and essential dynamics of the video which are relevant to the end task. The spatial modulator tends to



(a) Class Label: "Cutting Apple"

(b) Class Label: "Smashing"

(c) Class Label: "Tying Knot"

(d) Class Label: "Pushing Something So That It Slightly Moves"

(e) Class Label: "Putting Something On a Flat Surface Without Letting It Roll"

(f) Class Label: "Showing That Something Is Empty"

(g) Class Label: ["Back", "25som", "15Twis", "PIKE"]

(h) Class Label: ["Forward", "25som", "2Twis", "PIKE"]

(i) Class Label: ["Reverse", "Dive", "NoTwis", "PIKE"]

Figure 7: We visualize the spatial and temporal modulators for sample videos from Kinetics-600 [5] (top row), Something-Something-V2 [20] (middle row) and Diving-48 [41] (bottom row). Note how the temporal modulator fixates on the *global* motion across frames while the spatial modulator captures *local* variations. For example in Fig. 7a, the temporal modulator specifically focuses on the point where the knife meets the apple, while the spatial modulator shifts focus from frame to frame based on the knife's position. For Diving-48 (bottom row), we can see that the model can specifically fixate on the area where the action happens in each frame, regardless of the camera movement and small region of interest. More interestingly, the temporal modulator can separate the two regions of action in Fig. 7g and Fig. 7h.

shift to the *local* spatial changes in individual frames, while the temporal modulator fixates to the *global* region across frames where the majority of the motion happens.

5. Conclusion

To learn spatio-temporal representations that can effectively model both local and global contexts, this paper introduces Video-FocalNets for video action recognition tasks. This architecture is derived from focal modulation for images and can effectively model both short- and long-term dependencies to learn strong spatio-temporal representations. We extensively evaluate several design choices to develop our proposed Video-FocalNet block. Specifically, our Video-FocalNet uses a parallel design to model hierarchical contextualization by combining spatial and temporal convolution and multiplication operations in a computationally efficient manner. Video-FocalNets are more efficient than transformer-based architectures which require expensive self-attention operations. We demonstrate the effectiveness of Video-FocalNets via evaluations on five representative large-scale video datasets, where our approach outperforms previous transformer- and CNN-based methods.

References

[1] Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, , and Li Yang. Etc: Encoding long and structured inputs in transformers. In *EMNLP*, 2020. 2

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 1, 2, 3, 5, 7, 8

[3] Iz Beltagy, Matthew E Peters, , and Arman Cohan. Longformer: The long-document transformer. In *arXiv preprint arXiv:2004.05150*, 2020. 2

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 1, 2, 3, 5, 6, 7, 8

[5] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2, 5, 7, 8, 9

[6] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. In *arXiv preprint arXiv:1907.06987*, 2019. 1

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 2

[8] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. In *arXiv preprint arXiv:1904.10509*, 2019. 2

[9] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *arXiv preprint arXiv:2009.14794*, 2020. 3

[10] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS*, 2021. 2

[11] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 7

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 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. In *ICLR*, 2021. 1, 2

[14] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *ECCV*, 2020. 2, 7

[15] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Herve Jegou. XCit: Cross-covariance image transformers. In *NeurIPS*, 2021. 3

[16] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 2, 3, 6, 7

[17] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 2, 7

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 2, 7, 8

[19] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. Spatiotemporal residual networks for video action recognition. In *NeurIPS*, 2016. 1, 2

[20] 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 *ICCV*, 2017. 1, 2, 5, 7, 8, 9

[21] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *CVPR*, 2022. 3

[22] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. Group contextualization for video recognition. In *CVPR*, 2022. 8

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[24] Fabian Caba Heilbron et al. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2, 6, 8

[25] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). In *arXiv preprint arXiv:1606.08415*, 2016. 4

[26] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *CVPR*, 2022. 8

[27] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. In *arXiv preprint arXiv:1912.12180*, 2019. 2

[28] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 6

[29] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, , and Joao Carreira. Perceiver: General perception with iterative attention. In *arXiv preprint arXiv:2103.03206*, 2021. 2

[30] 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. 1, 2

[31] 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. In *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 5, 6, 7, 8

[32] Manjin Kim, Heeseung Kwon, Chunyu Wang, Suha Kwak, and Minsu Cho. Relational self-attention: What's missing in attention for video understanding. In *NeurIPS*, 2021. 8

[33] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020. 3

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

[35] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *CVPR*, 2021. 2, 7

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 1, 2

[37] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 1

[38] Ivan Laptev. On space-time interest points. In *IJCV*, 2005. 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. In *ICLR*, 2022. 2, 6, 7, 8

[40] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *CVPR*, 2020. 2, 7

[41] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, 2018. 2, 6, 8, 9

[42] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 2, 6, 7, 8

[43] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. In *arXiv preprint arXiv:2104.05707*, 2021. 3

[44] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 2, 7

[45] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022. 1, 2

[46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1, 2, 4

[47] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 1, 2, 3, 5, 7, 8

[48] 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. 1, 2

[49] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022. 2

[50] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018. 2

[51] Mandela Patrick, Dylan Campbell, Yuki M Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, Jo Henriques, et al. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*, 2021. 2, 3, 7

[52] Jiezhong Qiu, Hao Ma, Omer Levy, Scott Wen tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding. In *arXiv preprint arXiv:1911.02972*, 2019. 2

[53] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion. In *CVPR*, 2019. 2, 7

[54] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng L. Efficient attention: Attention with linear complexities. In *WACV*, 2021. 3

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

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 2

[57] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Hmdb: A large video database for human motion recognition. In *CRCV-TR-12-01*, 2012. 1

[58] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *ICCV*, 2015. 2

[59] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 6, 7

[60] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 2

[61] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. VIMPAC: Video pre-training via masked token prediction and contrastive learning. In *arxiv preprint, arXiv:2106.11250*, 2021. 8

[62] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 2

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

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

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 2, 3, 4

[66] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. In *IJCV*, 2013. 2

[67] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *CVPR*, 2022. 6, 8

[68] Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. In *arXiv preprint arXiv:2006.04768*, 2020. 3

[69] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2, 7

[70] Xiaofang Wang, Xuehan Xiong, Maxim Neumann, AJ Piergiovanni, Michael S Ryoo, Anelia Angelova, Kris M Kitani, and Wei Hua. Attentionnas: Spatiotemporal attention cell

search for video classification. In *ECCV*, 2020. 2

[71] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *ICCV*, 2021. 3

[72] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 2

[73] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *arXiv preprint arXiv:2102.03902*, 2021. 3

[74] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *ICCV*, 2021. 3

[75] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. 1, 2, 3, 7, 8

[76] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. In *NeurIPS*, 2022. 1, 2, 3, 4, 5, 6

[77] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. In *NeurIPS*, 2021. 2, 4

[78] S. Yun, D. Han, S. Chun, S. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 6

[79] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *NeurIPS*, 2021. 2

[80] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 6

[81] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *ICCV*, 2021. 2

[82] Shiwen Zhang. Tfcnet: Temporal fully connected networks for static unbiased temporal reasoning. In *arxiv preprint, arXiv:2203.05928*, 2022. 8

[83] 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 *ICCV*, 2021. 2, 3, 7