MM-ViT: Multi-Modal Video Transformer for Compressed Video Action Recognition

Jiawei Chen
OPPO US Research Center
jiawei.chen@oppo.com

Chiu Man Ho
OPPO US Research Center
chiuman@oppo.com

Abstract

This paper presents a pure transformer-based approach, dubbed the Multi-Modal Video Transformer (MM-ViT), for video action recognition. Different from other schemes which solely utilize the decoded RGB frames, MM-ViT operates exclusively in the compressed video domain and exploits all readily available modalities, i.e., I-frames, motion vectors, residuals and audio waveform. In order to handle the large number of spatiotemporal tokens extracted from multiple modalities, we develop several scalable model variants which factorize self-attention across the space, time and modality dimensions. In addition, to further explore the rich inter-modal interactions and their effects, we develop and compare three distinct cross-modal attention mechanisms that can be seamlessly integrated into the transformer building block. Extensive experiments on three public action recognition benchmarks (UCF-101, Something-Something-v2, Kinetics-600) demonstrate that MM-ViT outperforms the state-of-the-art video transformers in both efficiency and accuracy, and performs better or equally well to the state-of-the-art CNN counterparts with computationally-heavy optical flow.

1. Introduction

Video is one of the most popular media forms due to its rich visual and auditory content. In particular, videos have accounted for 75% of the global IP traffic everyday [48], which in consequence, poses an urgent need for automated video understanding methods. Among the video analysis tasks, action recognition is a fundamental one and becomes increasingly demanding in video applications, e.g., intelligent surveillance, self-driving, personal recommendation and entertainment [49]. However, most existing action recognition methods do not fully exploit the rich multi-modal information within videos, as they either depend on a single modality [2, 6, 19] or treat modalities separately [10, 54, 55, 73]. This is partially due to
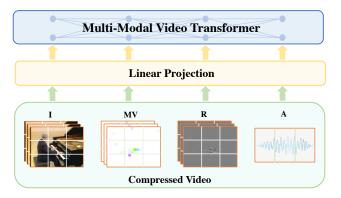

Figure 1: Our proposed MM-ViT operates exclusively in the compressed video domain, which allows collaborative fusion of appearance (I-frame), motion (Motion-vector & Residual) and audio features for action recognition. To the best of our knowledge, MM-ViT serves as the first work combining the strength of Transformer and multi-modal learning for compressed video action recognition.

the difficulty of reasoning expressive cross-modal interactions [28]. Whereas, recent advancements in cross-modal vision-language transformers [22, 52, 59] and multi-modal multi-task transformers [30] amply demonstrate the superiority of transformers in reasoning across multiple modalities. Furthermore, pure transformer models have also achieved competitive performance for vision tasks, e.g., image classification [14, 43], object detection [43, 83] and video classification [2, 6]. Such successes motivate our attempt to design a new transformer-based approach which explicitly interprets both visual and audio concepts for action recognition.

Another persistent challenge in action recognition is to effectively and efficiently model the temporal structure with large variations and complexities. Early works directly apply 2D CNNs to individual frames and then aggregate frame-level features via mid-level fusion or late-fusion [33, 68, 82]. Later, 3D/Pseudo-3D CNNs [8, 9, 31, 50, 60, 63] propose to operate convolution jointly over space and time to better integrate temporal information with CNNs. Re-

cently, video transformers equipped with spatial-temporal self-attention are proposed for video classification and show promising results [2, 6]. Nevertheless, most state-of-the-art results [17, 41, 57, 63, 75] on public benchmarks are still achieved by incorporating hand-crafted motion features, i.e., dense optical flow. It indicates that the optical flow extracted by external off-the-shelf methods contains complementary motion cues to the spatiotemporal features produced by deep learning frameworks. However, obtaining optical flow frame-by-frame is excessively time consuming, which usually causes a major computational bottleneck in recognition models. A recent line of works [54, 73, 80, 81] avoid optical flow computation by exploiting the motion information from compressed videos. Such a video contains only a few key frames (i.e., *I-frames*) and their offsets (i.e., *motion vectors* and *residuals*) for storage reduction. Specifically, these methods utilize the *motion vectors* and *residuals* present in the compressed video to model motion, and manage to achieve competitive accuracy, while run at least an order of magnitude faster than competing methods with optical flow. Despite being promising, these methods fail to take advantage of the strong interaction between modalities (i.e., treating modalities separately), and the crucial information in the audio stream.

To address the aforementioned issues in the existing methods, we propose a pure-transformer approach, named Multi-Modal Video Transformer (MM-ViT), for video action recognition. The design of MM-ViT follows the spirit of the image model "Vision Transformer" (ViT [14]), but extends its self-attention mechanism from the image space to the space-time-modalitiy 4D volume. In particular, MM-ViT operates in the compressed video domain, and thus it could exploit the readily available visual modalities, i.e. *I-frames*, *motion vectors* and *residuals*, to model appearance and temporal structure, and further leverage *audio* content for additional supervision (see Figure. 1). In order to maintain a reasonable computational cost, we design several efficient model variants which factorize the self-attention computation along the space, time and modality dimensions. Further, three distinct cross-modal attention mechanisms are proposed to facilitate learning the inter-modal interactions. We extensively evaluate our model on three action recognition datasets, and demonstrate that MM-ViT outperforms alternative video transformer architectures, and performs comparably, and in some cases superior, to the established CNN counterparts with the assistance of computationally expensive optical flow.

2. Related Work

Video Action Recognition: The tremendous success of deep learning methods on image-based recognition tasks has inspired significant advance in video action recognition. Most state-of-the-art solutions are based on CNNs and can be broadly classified into 2D- and 3D-CNN methods. The 2D-CNN approaches [20, 32, 33, 40, 55, 41, 65, 68, 82] process individual frames to extract frame-level features and then aggregate them with some kind of temporal modeling including late-fusion/mid-level fusion [33, 68, 82], LSTM [39, 65], channel-shift [41], feature temporal differencing [32, 40]. Alternative multi-stream formulations add additional CNN streams to incorporate auxiliary data modalities, e.g., optical flow [17, 20, 37, 55], audio [23, 36], human skeleton estimates [15, 76]. In contrast, 3D-CNN methods [8, 18, 19, 27, 31, 60] jointly learn spatial-temporal features by conducting convolution over a sequence of frames. Although better accuracy can be achieved, 3D convolution are computationally heavy, making the deployment difficult. To alleviate this problem, recent research [9, 50, 62, 63] proposes a family of pseudo-3D convolutions to reduce computational cost while preserve accuracy.

Compressed Video Action Recognition: Hand crafted motion representation, i.e. dense optical flow, contains complementary temporal information to the deep learning features, and thus employing optical flow almost guarantees to improve the recognition performance [17, 37, 41, 51, 63]. However, optical flow is unfortunately expensive to compute. To bypass this problem, pioneering works [80, 81] replace optical flow with *motion vectors* which present in the compressed video for encoding the movements of pixel blocks. Following up works achieve better performance by further exploiting *residuals* [73] or refining motion vectors with the supervision of optical flow during training [54]. Despite obtaining promising results, unlike our model, these methods ignore the rich inter-modal relations and fail to leverage *audio* signal.

Deep Multi-Modal Representation Learning: Riding on the success of deep learning, many deep multi-modal representation learning approaches have been proposed. Some suggest to fuse features from different modalities in a joint latent space by bilinear pooling [21], outer-product [78], and statistical regularization [3]. Inspired by the strong progress made in learning contextualized language representations using transformers [13, 42, 53, 77], several recent works have extended uni-modal transformers to the multi-modal setting by adding additional transformer stream(s) to learn joint vision-language representations [44, 52], vision-audio representations [38] and vision-language-audio representations [64, 79]. Our model is also built on the transformer, while it differs to the previous works by using a single transformer to model both intra- and inter-modal features with the help of effective cross-modal attention.

Vision Transformer: Self-attention has been extensively employed in computer vision tasks, e.g., image classification [5, 14, 43], object detection [7, 29, 83], and video classification [11, 25, 70]. Typically, it is used in conjunc-

tion with CNNs to augment convolutional features [5, 7, 29, 46, 83]. Until recently, Vision Transformer (ViT) [14] proposes a pure transformer architecture replacing all convolutions with self-attention, and it outperforms CNN counterparts in several downstream image tasks. Built on the ViT design, Timesformer [6] and ViViT [2] extend 2D spatial self-attention to the 3D spatial-temporal volume for video classification. In order to maintain a manageable computational cost, both works propose model variants which factorize different components of the transformer over the spatial- and temporal-dimension. In this paper, we present a new pure-transformer approach for video action recognition. Similar to [2, 6], our architecture stems mainly from the ViT design. However, instead of using solely decoded RGB frames, our model directly applies to the compressed videos and exploits both visual and audio modalities in the compressed domain. We empirically demonstrate that MM-ViT outperforms Timesformer [6] and ViViT [2] on three public action recognition benchmarks, while being more computationally efficient (in terms of FLOPs).

3. Methodology

3.1. Representations in Compressed Videos

Common video compression algorithms, such as MPEG-4, H.264 and HEVC, enable highly efficient video compression by splitting a video into *I-frames* (intra-coded frames), *P-frames* (predictive frames) and/or *B-frames* (bi-directional frames). An *I-frame* is a regular image and compressed as such. A *P-frame* holds only changes in the image from the previous frame, thus saving space. In practice, a *P-frame* comprises *motion vectors* and a *residual*. The *motion vectors* represent the movements of block of pixels (typically a frame is divided into 16x16 macroblocks during video compression) from the source frame to the target frame, thus roughly resemble coarse-grained optical flows. A *residual* retains the RGB pixel difference between a *P-frame* and its reference *I-frame* after motion compensation based on motion vectors, therefore, it usually shows changes of appearance and motion boundaries. A *B-frame* can be regarded as a special *P-frame*, where *motion vectors* are computed bi-directionally. In addition to the visual modalities, a compressed video also includes an audio stream which, in many cases, could play a critical role in recognition [23, 45].

3.2. Embedding Compressed Video Clips

MM-ViT operates on a compressed video clip $\mathcal{V}$. The vision modalities consist of T sampled *I-frames*, *motion vectors* and *residuals* of height H and width W. The audio modality contains a 1D waveform of length T'. Formally, $\mathcal{V} = \{\mathcal{I} \in \mathbb{R}^{T \times 3 \times H \times W}, \mathcal{M} \in \mathbb{R}^{T \times 2 \times H \times W}, \mathcal{R} \in \mathbb{R}^{T \times 3 \times H \times W}, \mathcal{A} \in \mathbb{R}^{T'}\}$, where $\mathcal{I}$, $\mathcal{M}$, $\mathcal{R}$ and $\mathcal{A}$ repre-

sents *I-frame*, *motion vector*, *residual* and *audio* modality, respectively. In order to (roughly) align the visual and audio signals, we partition the 1D audio waveform to T segments and project each segment to a 128-dimensional vector using a pretrained VGGish model [24].

Following ViT [14], we decompose each RGB *I-frame* into N non-overlapping patches of size $P \times P$. Then, we project those patches into token embeddings using a learnable linear embedding layer $\mathbf{E}^{\mathcal{I}} \in \mathbb{R}^{d \times 3P^2}$. Additionally, a spatiotemporal positional encoding $PE^{\mathcal{I}}_{(p,t)} \in \mathbb{R}^d$ is added to each patch token in order to preserve the positional information. Same operations are applied to tokenize the *motion vectors* and *residuals* as well:

$$z^{\mathcal{I}}_{(p,t)} = \mathbf{E}^{\mathcal{I}} \cdot \mathcal{I}_{(p,t)} + PE^{\mathcal{I}}_{(p,t)} \tag{1}$$

$$z^{\mathcal{M}}_{(p,t)} = \mathbf{E}^{\mathcal{M}} \cdot \mathcal{M}_{(p,t)} + PE^{\mathcal{M}}_{(p,t)} \tag{2}$$

$$z^{\mathcal{R}}_{(p,t)} = \mathbf{E}^{\mathcal{R}} \cdot \mathcal{R}_{(p,t)} + PE^{\mathcal{R}}_{(p,t)} \tag{3}$$

where $z^{\mathcal{I}}_{(p,t)}, z^{\mathcal{M}}_{(p,t)}, z^{\mathcal{R}}_{(p,t)}$ are the resulting vision tokens ($p = 1, ..., N, \quad t = 1, .., T$). For the audio feature, we first apply a linear layer $\mathbf{E}^{\mathcal{A}} \in \mathbb{R}^{d \times 128}$ to project it to the same dimensional space as vision tokens, then a temporal positional encoding $PE^{\mathcal{A}}_{(t)}$ is added:

$$z^{\mathcal{A}}_{(t)} = \mathbf{E}^{\mathcal{A}} \cdot \Phi(\mathcal{A}_{(t)}) + PE^{\mathcal{A}}_{(t)} \tag{4}$$

where the transformation function Φ is parameterized by the VGGish model [24]. To facility fully spatiotemporal self-attention across visual and audio modalities, we replicate each audio token $z^{\mathcal{A}}_{(t)}$ N times along the spatial dimension, thus $z^{\mathcal{A}}_{(p,t)} = z^{\mathcal{A}}_{(p',t)}$ for $p, p' \in \{1, ..., N\}$.

The resulting token sequences $z^{\mathcal{I}}_{(p,t)}, z^{\mathcal{M}}_{(p,t)}, z^{\mathcal{R}}_{(p,t)}$ and $z^{\mathcal{A}}_{(p,t)}$ for $p = 1, ..., N, t = 1, ..., T$, and a special "CLS" token $z^{CLS}_{(0,0)}$ constitute the input to MM-ViT. Following BERT [13], we use the output embedding of $z^{CLS}_{(0,0)}$ as the aggregated representation for the entire input sequence.

3.3. Multi-modal Video Transformers

In this section, we introduce four multi-modal video transformer architectures. We begin with an architecture which simply adopts the standard self-attention mechanism to measure all pairwise token relations. We then present three efficient model variants, which factorize the self-attention computation over the space-time-modality 4D volume with distinct strategies (see Figure. 2).

MM-ViT I - Joint Space-Time-Modality Attention: Similar in spirit to the "Joint Space-Time Attention" in Timesformer [6] and the "Spatio-Temporal Attention" in ViViT [2], each transformer layer of this model measures pairwise interactions between all input tokens. Concretely, MM-ViT I consists of L transformer layers. At each layer,
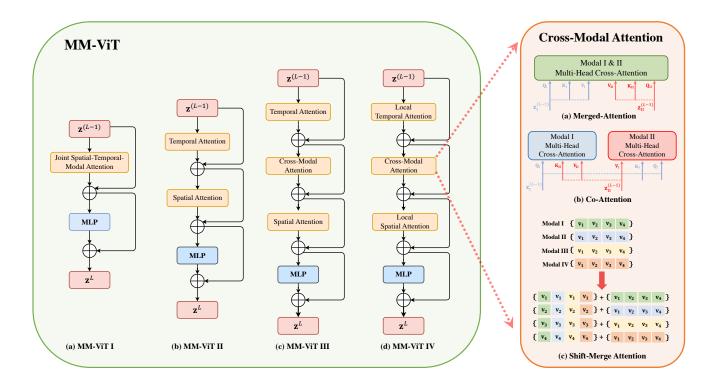
Figure 2: Overview of the investigated self-attention blocks in MM-ViT and the proposed cross-modal attention mechanisms.

a set of query(q), key(k) and value(v) vectors are first computed for every input token embedding $z_{(p,t)}^{(l-1,s)}$ from the preceding layer as below:

$$q_{(p,t)}^{(l,s)} = \mathbf{W}_Q^l \cdot \text{LN}(z_{(p,t)}^{(l-1,s)}) \tag{5}$$

$$k_{(p,t)}^{(l,s)} = \mathbf{W}_K^l \cdot \text{LN}(z_{(p,t)}^{(l-1,s)}) \tag{6}$$

$$v_{(p,t)}^{(l,s)} = \mathbf{W}_V^l \cdot \text{LN}(z_{(p,t)}^{(l-1,s)}) \tag{7}$$

where $\mathbf{W}_Q^l, \mathbf{W}_K^l, \mathbf{W}_V^l \in \mathbb{R}^{d_h \times d}$ are learnable embedding matrices, $\text{LN}(\cdot)$ denotes layer normalization [4], $s \in \mathbb{S} = \{\mathcal{I}, \mathcal{M}, \mathcal{R}, \mathcal{A}\}$. Then, the self-attention weights for query patch $q_{(p,t)}^{(l,s)}$ are given by:

$$\alpha_{(p,t)}^{(l,s)} = \text{Softmax}\left(\frac{q_{(p,t)}^{(l,s)}}{\sqrt{d_h}} \cdot \left\{k_{(p',t')}^{(l,s')}\right\}_{\substack{p'=0,\dots,N,s'=\mathbb{S} \\ t'=0,\dots,T}}\right) \tag{8}$$

The output token $z_{(p,t)}^{(l,s)}$ is further obtained by first computing the weighted sum of the value vectors based on the self-attention weights, followed by a linear projection through a MLP block. *Residual connection* is employed to promote robustness.

In practice, we adopt the Multi-Head Self-Attention (MSA) which yields better performance. Specifically, MSA uses h sets of $\{\mathbf{W}_Q^l, \mathbf{W}_K^l, \mathbf{W}_V^l\}$. The outputs of the h heads are concatenated and forwarded to the next layer.

Note that although this model allows interactions between all token pairs, it has quadratic computational complexity with respect to the number of tokens.

MM-ViT II - Factorized Space-Time Attention: Instead of computing self-attention across all pairs of input tokens, this model factorizes the operation along the spatial- and temporal-dimensions. As shown in Figure 2 ("MM-ViT II" in the left panel), given an token $z_{(p,t)}^{(l-1,s)}$ from layer $l-1$, we first conduct self-attention temporally (or spatially) by comparing it with all tokens at the same spatial location across all modalities. Next, a spatial attention followed by a linear projection is applied to generate the output embedding $z_{(p,t)}^{(l,s)}$ from layer l. Formally, we define our factorized space-time attention as:

$$y_{(p,t)}^{(l,s)} = \text{MSA}_{\text{time}}\left(\text{LN}(z_{(p,t)}^{(l-1,s)})\right) + z_{(p,t)}^{(l-1,s)} \tag{9}$$

$$y_{(p,t)}^{'(l,s)} = \text{MSA}_{\text{space}}\left(\text{LN}(y_{(p,t)}^{(l,s)})\right) + y_{(p,t)}^{(l,s)} \tag{10}$$

$$z_{(p,t)}^{(l,s)} = \text{MLP}\left(\text{LN}(y_{(p,t)}^{'(l,s)})\right) + y_{(p,t)}^{'(l,s)} \tag{11}$$

This architecture introduces more parameters than MM-ViT I due to one additional MSA operation. However, by decoupling self-attention over the input spatial- and temporal-dimensions, MM-ViT II reduces computational complexity per patch from $\mathcal{O}(N \cdot T \cdot |\mathbb{S}|)$ to $\mathcal{O}(N \cdot |\mathbb{S}| + T \cdot |\mathbb{S}|)$.

MM-ViT III - Factorized Space-Time Cross-Modal Attention: Our third model further factorizes self-attention over the modality dimension. At each transformer layer,

it attends to space, time and modality dimensions sequentially, thus reducing the computational complexity per patch to $\mathcal{O}(N + T + |\mathbb{S}|)$. Concretely, a patch token $z_{(p,t)}^{(l,s)}$ from layer l is calculated as follows:

$$y_{(p,t)}^{(l,s)} = \text{MSA}_{\text{time}}\big(\text{LN}(z_{(p,t)}^{(l-1,s)})\big) + z_{(p,t)}^{(l-1,s)} \quad (12)$$

$$y_{(p,t)}^{'(l,s)} = \text{MCA}\big(\text{LN}(y_{(p,t)}^{(l,s)})\big) + y_{(p,t)}^{(l,s)} \quad (13)$$

$$y_{(p,t)}^{''(l,s)} = \text{MSA}_{\text{space}}\big(\text{LN}(y_{(p,t)}^{'(l,s)})\big) + y_{(p,t)}^{'(l,s)} \quad (14)$$

$$z_{(p,t)}^{(l,s)} = \text{MLP}\big(\text{LN}(y_{(p,t)}^{''(l,s)})\big) + y_{(p,t)}^{''(l,s)} \quad (15)$$

where MCA stands for Multi-Head Cross-Attention, which is specifically designed for modeling cross-modal relations.

The main idea for this architecture is to have an effective cross-modal attention to facilitate learning from multimodal data. To this end, we develop three different cross-model attention mechanisms (see the right panel in Figure. 2). The first one is called "Merged-Attention". Given a query from one modality, it considers all of the keys and values regardless of the modality type. The output of this cross-attention module for query $q_{(p,t)}^{(l,s)}$ is defined as:

$$A(q_{(p,t)}^{(l,s)}, \mathbf{K}_{(p,t)}^{(l,\mathbf{s}')}, \mathbf{V}_{(p,t)}^{(l,\mathbf{s}')}) =$$
$$\text{Softmax}\Big(\frac{q_{(p,t)}^{(l,s)} \cdot \mathbf{K}^{\top}{}_{(p,t)}^{(l,\mathbf{s}')}}{\sqrt{d_h}}\Big) \cdot \mathbf{V}_{(p,t)}^{(l,\mathbf{s}')} \quad (16)$$

where $\mathbf{s}' = \mathbb{S}$. Alternatively, one can allow queries to interact only with keys and values from other modalities, thus $\mathbf{s}' = \mathbb{S} \setminus \{s\}$. We refer this cross-modal attention to "Co-Attention". Note that it is identical to the cross-attention operation proposed for the decoder model in Transformer [66].

Lastly, we propose a computation-free shift-based method called "Shift-Merge Attention" to assist interaction across modalities. It shares a similar spirit to the shift approaches proposed in the CNN domain (e.g., TSM [41]), seeking to strike a balance between accuracy and efficiency. More specifically, we discard queries and keys, and directly work on the value embeddings by first evenly splitting each $v_{(p,t)}^{(l,s)}$ into four portions $v1_{(p,t)}^{(l,s)}, v2_{(p,t)}^{(l,s)}, v3_{(p,t)}^{(l,s)}$, and $v4_{(p,t)}^{(l,s)}$. Then, we shift and mix the value embedding portions from different modalities, but at the same spatial and temporal index as follows:

$$r_{(p,t)}^{(l,\mathcal{I})} = v1_{(p,t)}^{(l,\mathcal{I})} \| v1_{(p,t)}^{(l,\mathcal{M})} \| v1_{(p,t)}^{(l,\mathcal{R})} \| v1_{(p,t)}^{(l,\mathcal{A})} + v_{(p,t)}^{(l,\mathcal{I})} \quad (17)$$

$$r_{(p,t)}^{(l,\mathcal{M})} = v2_{(p,t)}^{(l,\mathcal{I})} \| v2_{(p,t)}^{(l,\mathcal{M})} \| v2_{(p,t)}^{(l,\mathcal{R})} \| v2_{(p,t)}^{(l,\mathcal{A})} + v_{(p,t)}^{(l,\mathcal{M})} \quad (18)$$

$$r_{(p,t)}^{(l,\mathcal{R})} = v3_{(p,t)}^{(l,\mathcal{I})} \| v3_{(p,t)}^{(l,\mathcal{M})} \| v3_{(p,t)}^{(l,\mathcal{R})} \| v3_{(p,t)}^{(l,\mathcal{A})} + v_{(p,t)}^{(l,\mathcal{R})} \quad (19)$$

$$r_{(p,t)}^{(l,\mathcal{A})} = v4_{(p,t)}^{(l,\mathcal{I})} \| v4_{(p,t)}^{(l,\mathcal{M})} \| v4_{(p,t)}^{(l,\mathcal{R})} \| v4_{(p,t)}^{(l,\mathcal{A})} + v_{(p,t)}^{(l,\mathcal{A})} \quad (20)$$

where r denotes the resulting encoding, $\|$ represents concatenation. We also add a *residual connection* to preserve the learning capability.

MM-ViT IV - Factorized Local Space-Time Cross-Modal Attention: The last proposed architecture restricts the factorized spatial and temporal attention in MM-ViT III to non-overlapping local windows, therefore, the computational cost is further reduced. Supposing a local spatial and temporal window contains M and F patches, respectively, the computational complexity per patch becomes $\mathcal{O}(M + F + |\mathbb{S}|)$. In our experiments, we set $M = \frac{N}{4}, F = \frac{T}{2}$. However, limiting the receptive field to a local window may adversely affect model's performance. To alleviate this issue, we propose to insert a convolution layer after the temporal and spatial attention to strengthen the connection between neighboring windows. The convolution kernel size is same as the window size, and the stride size = 1.

Please note we are aware of that the order of the spatial, temporal and cross-modal attention may have an impact to model performance, and leave a further discussion to Sec. 4.2.

4. Experimental Evaluation

In this section, we first describe our experimental setup, then present ablation studies about some design choices, and finally compare with state-of-the-art methods.

4.1. Setup

Datasets: We evaluate the proposed method on three popular video action recognition datasets: UCF101 [56], Something-Something-V2 [26] and Kinetics-600 [35]. UCF-101 contains 13,320 trimmed short videos from 101 action categories. It has three training-testing splits. We report the average performance across the three splits. Kinetics-600 contains around 480K 10-seconds long videos for 600 action classes. Something-Something-v2 (SSv2) consists of about 220K videos with a time span from 2 to 6 seconds for 174 action classes. Different to other datasets, it places more emphasize on a model's ability to recognize fine-grained actions since the same background scenes can be shared aross many classes. In addition, the released version of SSv2 has no audio stream, therefore, we only extract the visual modalities from SSv2 to evaluate the proposed models. In our experiments, we convert all compressed videos to MPEG-4 codec which encodes a video into *I-frames* and *P-frames*. In average, an *I-frame* is followed by 11 *P-frames*.

Training Details: Following [54, 73], we resize all training videos to 340×256. Then, random horizontal flipping (omitted for SSv2) and random cropping (224×224) are applied to *I-frames*, *motion vectors* and *residuals* for data augmentation. Patch size is set to 16×16 across the visual modalities. The audio waveform is partitioned to 1-second long segments and projected into 128-dimensional vectors by VGGish [24]. We use ViT-B/16 [14] pretrained on ImageNet-21K as our backbone, and fine-tuned using

SGD with a batch size of 8. Learning rate starts from 0.5 and is divided by 10 when the validation accuracy plateaus. **Inference Details:** During inference, unless otherwise mentioned, the input consists of 8 uniformly sampled triplets of *I-frames*, *motion vectors*, *residuals* with a crop size of 224×224, and audio features (omitted for SSv2) that are temporally aligned with the visual features. We closely follow [6] to report accuracy from three spatial crops (left, centre, right) and average the scores for final prediction.

4.2. Ablation Studies

Analysis of the proposed model variants: In this section, we first compare the performance of the proposed model variants on UCF-101 and SSv2, in terms of accuracy and efficiency. Table 1 summarizes the detailed experimental results. We first note that MM-ViT I underperforms to the factorized alternatives (MM-ViT II & III), although it consumes more computational cost. This is consistent with the observation in [6], and we suspect it is due to the lack of dedicated parameters to model spatial, temporal and cross-modal attentions separately. Moreover, our results show that factorizing self-attention over the input dimensions consistently improve both recognition accuracy and efficiency, e.g., MM-ViT II outperforms MM-ViT I by 0.83% on UCF-101 while incurs 32% less FLOPs, meanwhile, MM-ViT III (Merged-Att.) outperforms MM-ViT II by 1.75% in accuracy and requires 3.5% less FLOPs.

Among the three proposed cross-modal attention mechanisms, "Merged-Attention" achieves the best accuracy on both UCF-101 and SSv2. This suggests that sharing keys and values across all modalities is critical to obtain a comprehensive understanding of the video content. Interestingly, the "Shift-Merge Attention" performs comparably to the "Merged-Attention" while being more efficient, making it attractive in resource constrained scenarios. When restricting the self-attention to local views (MM-ViT IV), the accuracy has a significant drop (↓3.11% on UCF101, ↓4.72% on SSv2), indicating more sophisticated cross-window connection may be needed to mitigate the information loss from using local attention views.

Table 1: Performance comparison of the proposed model variants on UCF101 and SSv2. Note we report FLOP numbers for UCF-101 where both visual and audio modalities are involved.

Model	Params	TFLOPs	UCF-101	SSv2
MM-ViT I	87.23M	1.26	93.41	62.55
MM-ViT II	122.69M	0.85	94.24	62.05
MM-ViT III (Merged-Att.)	158.14M	0.82	**95.99**	**64.97**
MM-VIT III (Co-Att.)	158.14M	0.82	93.87	62.10
MM-VIT III (Shift-Merge Att.)	143.98M	0.78	95.04	63.55
MM-ViT IV (Merged-Att.)	158.60M	0.72	92.88	60.25

Effect of attention order: In this part, we evaluate the ef-

fect of attention order by enumerating all possible orders of the spatial, temporal and cross-modal attention. For simplicity, we only report the results from MM-ViT III with Merged-Attention on UCF-101. As shown in Table 2, conducting temporal attention before spatial attention slightly, but consistently, performs better than the opposite, which can be explained as temporal attention provides key clues for distinguishing actions that share similar appearance features. In particular, the best performing attention order is *temporal⇒ cross-modal⇒spatial*.

Table 2: Effect of attention order on UCF101. T, S and M represent temporal, spatial and cross-modal attention, respectively.

Attention Order	Top-1	Top-5
T ⇒ S ⇒ M	95.41	99.39
S ⇒ T ⇒ M	95.24	99.40
T ⇒ M ⇒ S	**95.99**	**99.54**
S ⇒ M ⇒ T	95.31	99.50
M ⇒ T ⇒ S	95.00	99.21
M ⇒ S ⇒ T	94.85	99.25

Effect of input modality: To evaluate the importance of each data modality, we conduct an ablation study by training and evaluating the best performing model (i.e. MM-ViT III with "Merged-Attention") with different modality combinations on UCF-101. As shown in Table 3, *I-frame* is the most essential data modality as removing it alone decreases top-1 accuracy by 4.11%. The *motion vector* and *residual frame* also play important roles for video recognition. Without either modality can lead to an accuracy drop up to 2.54%. It is interesting that the audio modality has a major impact to video recognition as well, which is confirmed by the significant performance degradation (2.42% drop in top-1 accuracy) when excluding audio input. It is likely because audio contains dynamics and contextual temporal information that is beneficial for action recognition (e.g., the sound of axe hitting tree is discriminative for recognizing "Cutting Tree").

Table 3: Performance comparison of different modality combinations on UCF-101.

$\mathcal{I}$	$\mathcal{M}$	$\mathcal{R}$	$\mathcal{A}$	Top-1	Top-5
✓	✓	✓	✓	95.99%	99.54%
✗	✓	✓	✓	91.88%	98.76%
✓	✗	✓	✓	93.45%	99.20%
✓	✓	✗	✓	94.16%	99.28%
✓	✓	✓	✗	93.57%	99.21%

4.3. Comparison to State-of-the-Art

In this section, we compare our best performing architecture, i.e. MM-VIT III with "Merged Attention", to other

state-of-the-art approaches on UCF101, SSv2 and Kinetics-600 datasets. Unless otherwise specified, we follow [6] to report results from 1×3 views (1 temporal and 3 spatial crops).

UCF101: Table 4 summarizes the performance of our model and other competing methods on UCF-101. We first observe that even without using audio, MM-ViT already outperforms all other methods that operate the same compressed video modalities [80, 81, 73, 54], by up to 6.9% in top-1 accuracy. It suggests that the explicit reasoning of inter-modal relations in MM-ViT is effective. Further improvements are achieved by incorporating audio signal ($\uparrow 2.1\%$ in top-1 accuracy) and pre-training the model on Kinetics-600 ($\uparrow 3.5\%$ in top-1 accuracy). Remarkably, we surpass all CNN alternatives with or without optical flow, and thus establishes a new state-of-the-art for UCF-101.

Table 4: Performance comparison with the state-of-the-art methods on UCF-101.

	Modality	Pretrain	Top-1
C3D [60]	RGB	ImageNet1K	82.3
ActionFlowNet [47]	RGB	-	83.9
Res3D [61]	RGB	Sports-1M	85.8
PWC-Net(ResNet-18) [58]	RGB	-	90.6
ARTNet [67]	RGB	Kinetics	94.3
I3D RGB [8]	RGB	Kinetics	95.6
MF-Net [12]	RGB	Kinetics	96.0
S3D [74]	RGB	Kinetics	96.5
Two-stream [55]	RGB + Flow	-	88.0
Two-stream fusion [20]	RGB + Flow	ImageNet1K	92.5
TV-Net [16]	RGB + Flow	-	94.5
R(2+1)D [63]	RGB + Flow	Kinetics	97.3
I3D [8]	RGB + Flow	Kinetics	98.0
EMV-CNN [80]	$\mathcal{I} + \mathcal{M} + \mathcal{R}$	ImageNet1K	86.4
DTMV-CNN [81]	$\mathcal{I} + \mathcal{M} + \mathcal{R}$	-	87.5
CoViAR [73]	$\mathcal{I} + \mathcal{M} + \mathcal{R}$	ImageNet1K	90.4
DMC-Net(I3D) [54]	$\mathcal{I} + \mathcal{M} + \mathcal{R}$	ImageNet1K	92.3
MM-ViT	$\mathcal{I} + \mathcal{M} + \mathcal{R}$	ImageNet1K	93.3
MM-ViT	$\mathcal{I} + \mathcal{M} + \mathcal{R} + \mathcal{A}$	ImageNet1K	95.4
MM-ViT	$\mathcal{I} + \mathcal{M} + \mathcal{R} + \mathcal{A}$	Kinetics	**98.9**

SSv2: In Table 5, we present detailed results on SSv2 including top-1&top-5 accuracy, inference resolution and computational cost (in FLOPs). Notably, MM-ViT surpasses the Timesformer [6] and ViViT [2] which also propose pure-transformer models, yet being more efficient in terms of inference FLOPs. It confirms that the additional *motion vector* and *residual* modalities used by MM-ViT provide important complementary motion features, which could benefit the classification on "temporally-heavy" dataset like SSv2. Furthermore, MM-ViT consistently performs better than the CNN counterparts that operate in the single RGB modality (by $\geq 1.5\%$ in top-1 accuracy). Although MM-ViT slightly underperforms to CNN-based MSNet-R50 [37] and blVNet [17] which use optical flow as an auxiliary modality, it eliminates the huge burden of optical flow computation and storage.

Table 5: Performance comparison with state-of-the-art methods on SSv2. The inference resolution is denoted by $M \times T \times S^2$ for the number of modalities, temporal and spatial sizes.

	Inference Resolution	Top-1	Top-5	Inference TFLOPs×views
CNN + RGB only				
TRN [82]	$1 \times 8 \times 224^2$	48.8	77.6	-
SlowFast [72]	$1 \times 64 \times 224^2$	61.7	87.8	0.07×30
TSM [41]	$1 \times 24 \times 224^2$	63.3	88.5	0.1×1
STM [32]	$1 \times 16 \times 224^2$	64.2	89.8	0.07×10
MSNet-R50 [37]	$1 \times 16 \times 224^2$	64.7	89.4	0.07×10
TEA [40]	$1 \times 16 \times 224^2$	65.1	-	0.07×30
blVNet [17]	$1 \times 32 \times 224^2$	65.2	90.3	0.13×30
CNN + RGB + Optical Flow				
TSM [41]	$2 \times 24 \times 224^2$	66.0	90.5	-
MSNet-R50 [37]	$2 \times 16 \times 224^2$	67.1	91.0	-
blVNet [17]	$2 \times 32 \times 224^2$	**68.5**	**91.4**	-
Video Transformer				
Timesformer-L [6]	$1 \times 96 \times 224^2$	62.4	-	2.38×3
ViViT-L [2]	$1 \times 32 \times 224^2$	65.4	89.8	1.45×4
MM-ViT	$3 \times 8 \times 224^2$	64.9	89.7	0.75×3
MM-ViT	$3 \times 16 \times 224^2$	66.7	90.8	1.51×3

Kinetics-600: Kinetics-600 is a larger video classification dataset that shall further demonstrate the generalization of our approach. As shown in Table 6, our MM-ViT ($T = 16$) achieves 83.5% top-1 accuracy, which makes the relative improvement over the Timesformer [6] and ViViT [2] by 1.3% and 0.5% respectively, while it remains more computationally efficient. This accuracy is also higher than CNN alternatives that either operate in the single RGB modality or use additional flow information. Again, it verifies that our model is effective in learning across multiple modalities for the complex video classification task.

However, we also notice that current video transformers incur relatively large computational overhead compared to several efficient CNN architectures [41, 32, 37, 40]. Therefore, a future line of research could be improving computational and memory efficiency upon the video transformer architectures by leveraging, e.g., low-rank approximations [69], localized attention span [43] and kernelization [34].

Table 6: Performance comparison with state-of-the-art methods on Kinetics600. The inference resolution is denoted by $M \times T \times S^2$ for the number of modalities, temporal and spatial sizes.

	Inference Resolution	Top-1	Top-5	Inference TFLOPs×views
CNN + RGB				
AttentionNAS [71]	$1 \times 16 \times 224^2$	78.5	93.9	1.03×1
LGD-3D R101 [51]	$1 \times 16 \times 224^2$	81.5	95.6	-
SlowFast R101-NL [19]	$1 \times 16 \times 224^2$	81.8	95.1	0.23×30
X3D-XL [18]	$1 \times 16 \times 224^2$	81.9	95.5	0.05×30
CNN + RGB + Optical Flow				
LGD-3D Two-Stream [51]	$2 \times 16 \times 112^2$	83.1	96.2	-
Video Transformer				
Timesformer-L [6]	$1 \times 96 \times 224^2$	82.2	95.6	2.38×3
ViViT-L [2]	$1 \times 32 \times 320^2$	83.0	96.5	3.99×12
MM-ViT	$4 \times 8 \times 224^2$	81.5	95.1	0.82×3
MM-ViT	$4 \times 16 \times 224^2$	**83.5**	**96.8**	1.66 ×3

Apply Eye Makeup

Knit

I-frame

Motion Vector

Residual Frame

" Sweeping the light purple across Lesley's **lower eyelid** "

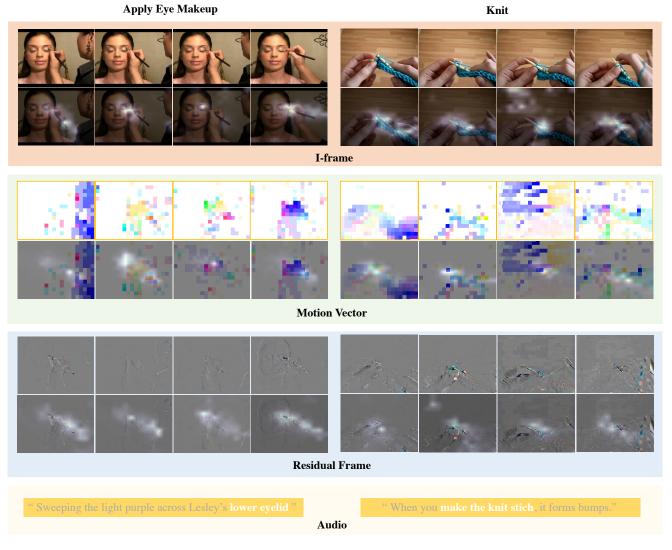" When you **make the knit stich**, it forms bumps."

Audio

Figure 3: Visualization of attention weights to the input space. The proposed model learns to focus on relevant parts of the video for classification.

4.4. Qualitative Results

In order to qualitatively evaluate our proposed model, we visualize model attention from the output tokens to the input space via Attention Rollout method [1]. In Figure. 3, we show examples obtained from applying MM-ViT to UCF-101 videos. We can see that MM-ViT indeed attends to the relevant regions in the input space. For example, when applied to classify the video "Apply Eye Makeup", the model concentrates on the eye area and the eyeshadow brush. In some cases, MM-ViT also perceives the phrases/words which are semantically aligned with the content of the action, e.g., the model focuses on words "lower eyelid" when classifying the video "Apply Eye Makeup". The remarkable consistency of the quantitative and qualitative results again confirms the effectiveness of the proposed MM-ViT

in complex spatial-temporal-audio reasoning.

5. Conclusion

In this work, we propose **MM-ViT**, a pure-transformer approach for action recognition in the compressed video domain. By factorizing self-attention over the input space, our model enjoys the rich multi-modal information present in the compressed videos, while maintains a reasonable computational cost. Empirical results also confirm the effectiveness of the proposed cross-modal attention mechanisms in learning the inter-modal relations. Further, extensive experiments on three public benchmarks show that our model outperforms the competing uni-modal video transformers in both efficiency and accuracy, and performs better or equally well to the state-of-the-art CNN counterparts with the assistance of computationally expensive optical flow.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.

[3] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2303–2314, 2017.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[5] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295, 2019.

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[9] Jiawei Chen, Jenson Hsiao, and Chiu Man Ho. Residual frames with efficient pseudo-3d cnn for human action recognition. *arXiv preprint arXiv:2008.01057*, 2020.

[10] Jiawei Chen, Jonathan Wu, Janusz Konrad, and Prakash Ishwar. Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 139–147. IEEE, 2017.

[11] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A2-nets: Double attention networks. *arXiv preprint arXiv:1810.11579*, 2018.

[12] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Multi-fiber networks for video recognition. In *Proceedings of the european conference on computer vision (ECCV)*, pages 352–367, 2018.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] 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.

[15] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11Learning Spatiotemporal Features with 3D Convolutiona10–1118, 2015.

[16] Lijie Fan, Wenbing Huang, Chuang Gan, Stefano Ermon, Boqing Gong, and Junzhou Huang. End-to-end learning of motion representation for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6016–6025, 2018.

[17] Quanfu Fan, Chun-Fu Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. *arXiv preprint arXiv:1912.00869*, 2019.

[18] 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.

[19] 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.

[20] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[21] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[22] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer, 2020.

[23] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020.

[24] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[25] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[26] 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.

[27] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and im-

agenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[28] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *arXiv preprint arXiv:2102.00529*, 2021.

[29] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3588–3597, 2018.

[30] Ronghang Hu and Amanpreet Singh. Transformer is all you need: Multimodal multitask learning with a unified transformer. *arXiv e-prints*, pages arXiv–2102, 2021.

[31] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[32] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2000–2009, 2019.

[33] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[34] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020.

[35] 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.

[36] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.

[37] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Motionsqueeze: Neural motion feature learning for video understanding. In *European Conference on Computer Vision*, pages 345–362. Springer, 2020.

[38] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. *arXiv preprint arXiv:2012.04124*, 2020.

[39] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using lstm and cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 585–590. IEEE, 2017.

[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 Con-*

ference on Computer Vision and Pattern Recognition, pages 909–918, 2020.

[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.

[42] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[43] 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. *arXiv preprint arXiv:2103.14030*, 2021.

[44] 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. *arXiv preprint arXiv:2104.08860*, 2021.

[45] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2action: Cross-modal supervision for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10317–10326, 2020.

[46] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021.

[47] Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S Davis. Actionflownet: Learning motion representation for action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1616–1624. IEEE, 2018.

[48] Robert Pepper. Cisco visual networking index (vni) global mobile data traffic forecast update. Technical report, Cisco, Tech. Rep., Feb. 2013. Accessed: Jul. 10, 2019.[Online]. Available . . . , 2013.

[49] Oleksandra Poquet, Lisa Lim, Negin Mirriahi, and Shane Dawson. Video and learning: a systematic review (2007–2017). In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 151–160, 2018.

[50] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatiotemporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[51] 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.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[53] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[54] Zheng Shou, Xudong Lin, Yannis Kalantidis, Laura Sevilla-Lara, Marcus Rohrbach, Shih-Fu Chang, and Zhicheng Yan. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1268–1277, 2019.

[55] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

[56] 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.

[57] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 625–634, 2020.

[58] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[59] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[60] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

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

[62] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.

[63] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[64] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[65] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE access*, 6:1155–1166, 2017.

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[67] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1430–1439, 2018.

[68] 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.

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

[70] 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.

[71] 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 *European Conference on Computer Vision*, pages 449–465. Springer, 2020.

[72] Chao-Yuan Wu, Ross Girshick, Kaiming He, Christoph Feichtenhofer, and Philipp Krahenbuhl. A multigrid method for efficiently training video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 153–162, 2020.

[73] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. Compressed video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6026–6035, 2018.

[74] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv preprint arXiv:1712.04851*, 1(2):5, 2017.

[75] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[76] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[77] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[78] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[79] Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826*, 2019.

[80] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with enhanced motion

vector cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2718–2726, 2016.

[81] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with deeply transferred motion vector cnns. *IEEE Transactions on Image Processing*, 27(5):2326–2339, 2018.

[82] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.

[83] 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.