TP-VIT: A TWO-PATHWAY VISION TRANSFORMER FOR VIDEO ACTION RECOGNITION

Yanhao Jing *Feng Wang**

School of Computer Science and Technology, East China Normal University, China

ABSTRACT

Recently, inspired by the success of Transformer in natural language processing tasks, a number of works have attempted to apply Transformer-based models to video action recognition. Existing works only use one RGB stream as the input for Transformer. How to use multiple pathways and multiple streams with Transformer for action recognition has not been studied. To address this issue, we present a novel structure namely Two-Pathway Vision Transformer (TP-ViT). Two parallel spatial Transformer encoders are used as two pathways with different framerates and resolutions of the input video. The high-resolution pathway contains more spatial information, while the high-framerate pathway contains more temporal information. The two outputs are fused and fed into a temporal Transformer encoder for action recognition. Furthermore, we also fuse skeleton features into our model to get better results. Our experiments demonstrate that our proposed models achieve the state-of-the-art results on both the coarse-grained dataset Kinetics and the fine-grained dataset FineGym.

***Index Terms*—** Vision transformer, video action recognition, two-pathway transformer.

1. INTRODUCTION

As an important task for video content analysis, video action recognition has been intensively studied for decades. In the past few years, deep learning based approaches have achieved encouraging results and become the state of the art [1, 2, 3, 4, 5] by effectively capturing spatial and temporal information in videos. Recently, inspired by the success of the Transformer [6] in natural language processing and image recognition tasks [7, 8], some works attempt to apply Transformers to video understanding and action recognition [9, 10, 11, 12]. How to efficiently capture and tokenize the spatio-temporal information has become the focus of these approaches.

The current Transformer-based approaches focus on studying various attention methods such as space only, joint space-time, or divided space-time [10], and do not fully utilize the abundant information in videos. For instance, skeleton feature is effective to describe human actions [13]. Video clips with different resolutions and framerates contain information of various forms [3]. How to make full use of multiple pathways and multiple streams for action recognition with Transformer has not been studied.

In this paper, we propose a novel Two-Pathway Vision Transformer framework (TP-ViT) to combine different features for better performance. Figure 1(b) illustrates our framework. Inspired by Two-Stream [14] and SlowFast [3], we take a slow and a fast samples of the videos as the inputs of two pathways respectively. Two spatial Transformers are used to process the clip patches in two pathways. A temporal Transformer is then used to fuse the two spatial tokens. Besides the RGB stream, our framework can make use of multi-stream information. We extract skeleton features and incorporate them into the Two-Pathway Transformer to improve the performance of action recognition.

The main contributions of this paper are as follows:

- We propose a two-pathway spatial-temporal Transformer which makes full use of multiple pathways and multiple streams for video action recognition.
- We propose a multi-sampling embedding method to extract tokens for the slow and the fast pathways from a video. The slow pathway with high resolutions captures more spatial information, while the fast pathway with high framerates captures more temporal information.
- We fuse skeleton features into our network and show the effectiveness of multi-stream information for action recognition with the Transformer.

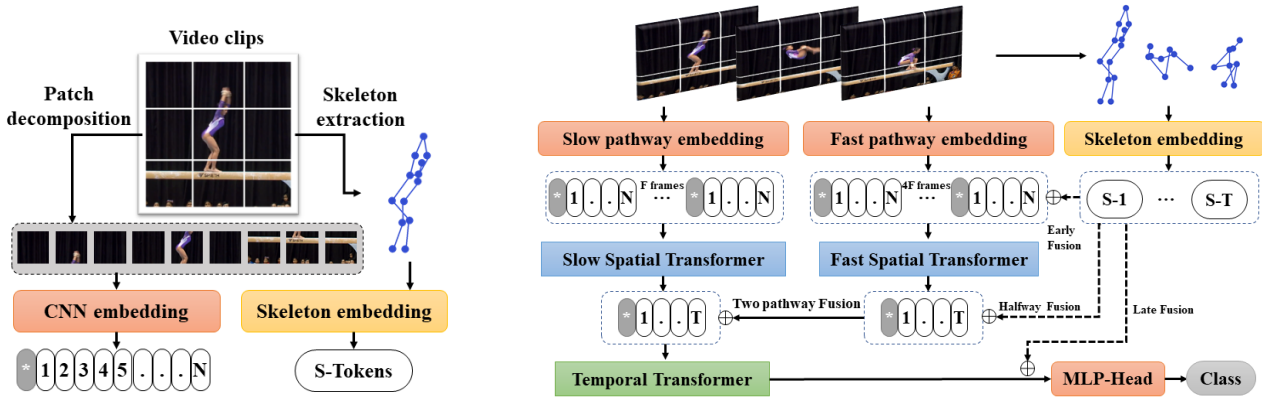
The remaining of this paper is organized as follows. Section 2 reviews the related works. Section 3 presents our TP-ViT. The experimental results are presented in Section 4. Finally, Section 5 concludes this paper.

2. RELATED WORKS

2.1. Transformer-based Action Recognition

The Transformer [6] is originally proposed for machine translation and has since become the state-of-the-art method in many natural language processing tasks including machine translation [15], question answering [16, 17], and word generation [18, 19]. It has also been applied to computer vision in recent years [20, 21, 22]. For instance, ViT [7] divides an

*Corresponding author. Email: fwang@cs.ecnu.edu.cn.



(a) The input embeddings.

(b) The structure of TP-ViT.

Fig. 1. The input embeddings and the overview of TP-ViT.

image into 16×16 patches, which are embedded into one-dimensional sequence tokens as the input to the Transformer for image recognition.

The Transformer can process token sequences and capture the relationships between the tokens, which makes it suitable for video tasks since a video can be regarded as a sequence of images. Recently, there has been a number of works on using Transformer's architecture for video action recognition [9, 10]. In [10], a frame is embedded into a token sequence with patches by 16×16 following ViT. The sequences are then fed into different kinds of self-attention architectures including the space, the joint space-time, and the divided space-time. In [9], an alternate method named Tubelet embedding is proposed, which extracts non-overlapping spatio-temporal tubelets from the input video clips. The factorised encoder is used which consists of two Transformer encoders in series. These works only use a single-pathway RGB stream for action recognition.

2.2. Multi-Pathway and Multi-Stream Convolutional Neural Networks for Action Recognition

In the past few years, the convolutional neural networks (CNNs) have achieved the state-of-the-art performances in video action recognition. To make full of the abundant information in videos for action recognition, there have been a number of works on multi-pathway convolutional neural networks. In [3], the SlowFast architecture has two pathways which operate at low and high temporal resolutions respectively. The fast pathway is designed to capture the fast-changing motion but with fewer spatial details, while the slow pathway has better spatial modeling capabilities to capture more spatial information.

Besides the RGB stream, other streams are also explored to improve the performance. In [14], optical flow [23, 24] is fused with the RGB stream in the two-stream ConvNet towers and different fusion strategies are studied. In [25], a Spatial-Temporal Graph ConvNet is proposed to process the skeletons extracted by [13] for action recognition. However, skeletons are usually used alone due to their particularity, and few works attempt to combine skeletons with other streams. In this paper, we study how to effectively utilize multi-pathway and multi-stream information with Transformer for action recognition.

3. TP-VIT NETWORK

As illustrated in Figure 1(b), our TP-ViT can be described as an RGB stream architecture that operates at two different spatial resolutions and framerates. The slow pathway and the fast pathway capture more spatial and more temporal information respectively. Two spatial Transformers are used to process two-pathway inputs and then feed the classification tokens into a temporal Transformer. Furthermore, we tokenize skeleton data and fuse it into the final token sequence to improve the performance.

3.1. Multi-sampling Embeddings

There are a variety of options for the embedding of video clips. In [7, 8], a frame is partitioned into 16×16 patches and then embedded with full connections. In [10], CNNs are used to extract the features of patches. In this paper, we follow the work in [10] for video clip embedding. We sample the slow and the fast pathways at different resolutions and framerates. Figure 1(a) illustrates our implementation.

Video decomposition. A raw video clip $\mathbf{X} \in \mathbb{R}^{F \times C \times H \times W}$ consists of F RGB frames of size $H \times W$ with C channels. Following the method in ViT [7], we decompose each frame into N non-overlapping patches. The size of each patch is $P \times P$, and thus $N = H \times W / P^2$. We get the patches $x_{(p,t)} \in \mathbb{R}^{3 \times P \times P}$ with $p = 1, \ldots, N$ denoting the spatial locations in the frame and $t = 1, \ldots, F$ being the index if the frame.

CNN embedding. We handle the patch $x_{(p,t)}$ with a single 2D Conv layer with the kernel of size P, and then flat-

ten the output into a vector $\mathbf{z}_{(p,t)}^{(0)} \in \mathbb{R}^D$. Following the implementation in [16], we add a special learnable vector $\mathbf{z}_{(0,t)}^{(0)} \in \mathbb{R}^D$ to embed the classification token (CLS token) in the first position of the token sequence in one frame.

Slow pathway and Fast pathway. We implement the slow pathway with the standard sampling above. For the fast pathway, we set the frame rate to 4 times that of the slow pathway and set the resolution to $1/4$ of the slow pathway. In our implementation, we set the slow pathway's input size to $8 \times 3 \times 224 \times 224$, and set the fast pathway's to $32 \times 3 \times 112 \times 112$. In this way, we use only one patch size $P \times P$, and the total patch numbers of both pathways are the same ($F \times H^2/P^2$). As the result, the computation costs of two pathways are almost identical.

3.2. Two-Pathway ViT

In our framework, two spatial Transformers are used to process the two samples generated in Section 3.1, which are then fused and fed into a temporal Transformer for action recognition.

Transformer encoder. Similar to the structure in TimeSformer [10], we use the L-layer self-attention blocks with $\mathcal{A}$ heads as the Transformer encoder. At block l, it computes the query/key/value vectors from the patch embedding as follows:

$$\mathbf{q}_{(p)}^{(l,a)} = W_Q^{(l,a)} \text{LayerNorm}\left(z_{(p)}^{(l-1)}\right)$$

$$\mathbf{k}_{(p)}^{(l,a)} = W_K^{(l,a)} \text{LayerNorm}\left(z_{(p)}^{(l-1)}\right)$$

$$\mathbf{v}_{(p)}^{(l,a)} = W_V^{(l,a)} \text{LayerNorm}\left(z_{(p)}^{(l-1)}\right)$$

where $a = 1, \ldots, \mathcal{A}$ is the index of the attention heads, p is the index of the patch in one frame, and $z_{(p)}^{(l-1)}$ is the output token of the last layer. The query/key/value latent dimension of each head is set to $\mathbb{R}^{D_h}$ where $D_h = D/\mathcal{A}$.

In each block, we only compute the spatial attention within one frame, and the self-attention weights are as follows:

$$\mathbf{a}_{(p)}^{(l,a)} = \text{softmax}\left(\frac{\mathbf{q}_{(p)}^{(l,a)\top}}{\sqrt{D_h}} \cdot \left[\mathbf{k}_{(0,0)}^{(l,a)} \{\mathbf{k}_{(p')}^{(l,a)}\}_{p'=1,\ldots,N}\right]\right).$$

Then we get the output of an attention block through a Multilayer Perceptron (MLP) by using residual connections:

$$\mathbf{z'}_{(p,t)}^{(l)} = \mathbf{a}_{(p,t)}^{(l)} \mathbf{v}_{(p,t)}^{(l)} + \mathbf{z}_{(p,t)}^{(l-1)}$$

$$\mathbf{z}_{(p,t)}^{(l)} = \text{MLP}(\text{LayerNorm}(\mathbf{z'}_{(p,t)}^{(l)})) + \mathbf{z'}_{(p,t)}^{(l)}.$$

At the end of the blocks, we get the CLS token $z_{(0)}^{(L)}$ with a LayerNorm for each frame:

$$\mathbf{y_t} = \text{LayerNorm}(\mathbf{z}_{(0,t)}^{(L)}) \in \mathbb{R}^D.$$

The output is denoted as $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_F\} \in \mathbb{R}^{F \times D}$.

Two-pathway spatial Transformer. Our Two-pathway spatial Transformer is composed of two Transformer encoders described above with two stacks of self-attention blocks. Since the spatial Transformers are used to process the patches of the same size, the two pathway Transformer encoders can share the same parameters. After the two-pathway spatial Transformer, we get the spatial CLS tokens of two pathways. Each CLS token has the same size as the patch embedding, i.e., the output size of the slow pathway is $F \times D$, and the fast pathway is $4F \times D$.

Fusion of two pathways. We average the fast-pathway CLS tokens $\mathbf{Y}_{fast} \in \mathbb{R}^{4F \times D}$ as a group of 4 frames and get $\mathbf{Y'}_{fast} \in \mathbb{R}^{F \times D}$. These average tokens $\mathbf{Y'}_{fast}$ are then concatenated after the slow-pathway CLS tokens $\mathbf{Y}_{slow} \in \mathbb{R}^{F \times D}$ according to the frame's position.

Temporal Transformer. The temporal Transformer has the same structure as the spatial Transformers but with half the layers to receive the spatial CLS tokens and generate the temporal CLS tokens. We then input the temporal CLS tokens into the MLP head. The MLP consists of two linear projections separated by a GELU [26] and the dimension D remains fixed throughout all layers. Finally, a linear classifier is used for classification.

3.3. Fusion of Skeleton Feature

Besides RGB stream, the skeleton feature is used to capture the higher-level semantic information. We employ OpenPose [13] to extract the skeleton feature, which consists of the skeleton data of 18 points on human body. The feature of a skeleton point is described as (x, y, r), where x, y are the coordinates of the point, and r is the confidence of the point. By concatenating the features of 18 points, the token of the $i-$th frame is denoted as $\mathbf{s}_i \in \mathbb{R}^{18 \times 3}$. The tokes of the frames in a video clip are combined as $\{\mathbf{s}_1, \ldots, \mathbf{s}_F\}$ and processed with the MLP to form the final skeleton tokens $\mathbf{S} \in \mathbb{R}^{F \times D_s}$.

To fuse the skeleton and the RGB streams, we implement three fusion strategies, i.e. early fusion, halfway fusion, and late fusion. Concatenation is used for these three strategies. The early fusion directly concatenates the spatial tokens and skeleton tokens. In the halfway fusion, we concatenate every spatial CLS token generated by the spatial Transformer with the skeleton tokens of the same frame. The late fusion concatenates the output before the MLP head with all the skeleton tokens.

4. EXPERIMENTS

We experiment three architectures for action recognition, i.e. single Slow-Pathway Transformer (RGB-Slow), Two-Pathway Transformer (RGB-SF), and Two-Pathway Transformer with the skeleton features (SF+SK), and compare them with the state-of-the-art approaches.

Method Info			Kinetics-400	
Method	Modality	Clip size	Top-1	Top-5
I3D[1]	2Stream	$64 \times 224 \times 224$	71.60%	90.00%
TSM[2]	2Stream	$16 \times 224 \times 224$	74.10%	91.20%
I3D NL[4]	RGB	$32 \times 224 \times 224$	77.70%	93.30%
SlowFast[3]	RGB-SF	$64 \times 224 \times 224$	79.80%	93.90%
X3D-XXL[5]	RGB	$16 \times 448 \times 448$	80.40%	94.60%
TimeSformer[10]	RGB	$8 \times 224 \times 224$	78.00%	93.70%
TimeSformer-HR[10]	RGB	$16 \times 448 \times 448$	79.70%	94.40%
TimeSformer-L[10]	RGB	$96 \times 224 \times 224$	80.70%	94.70%
TP-ViT(ours)	RGB-Slow	$8 \times 224 \times 224$	73.34%	91.09%
TP-ViT(ours)	RGB-SF	$32 \times 224 \times 224$	80.32%	94.37%
TP-ViT(ours)	SF+SK	$32 \times 224 \times 224$	**80.81%**	**95.39%**

Table 1. Performances on the Kinetics-400 [1] dataset.

4.1. Datasets

We evaluate our approach on three video action recognition datasets including the widely used coarse-grained dataset Kinetics [1], and the fine-grained datasets FineGym-99 and FineGym-288 [27]. Kinetics-400 consists of 240k training videos and 20k validation videos of 400 human action categories. FineGym-99 has 26k training videos and 8k validation videos of 99 classes, and FineGym-288 has 29k training videos and 9k validation videos of 288 classes. We report top-1 and top-5 classification accuracy (%) on these datasets.

4.2. Implementation Details

We sample $4T$ frame clips from the original video with 256×320 pixels [4]. We randomly sample T frames for the slow pathway and randomly crop 224×224 pixels. We choose all the $4T$ frames for the fast pathway, resize the video to 128×160, and randomly crop 112×112 pixels. In our experiments, we set $T = 8$.

Our models are pre-trained on ImageNet [28] and use synchronized SGD optimizer as in [29]. Following the settings in [3], we uniformly sample 32 clips from a video along the temporal axis. We scale the shorter spatial side to 256 pixels for each clip and take 3 crops of 256×256 to cover the spatial dimensions as an approximation of fully convolutional testing. We average the softmax scores for prediction by following the work in [3, 4].

4.3. Results

Tables 1 and 2 compare our approach with the state-of-the-art approaches.

Kinetics-400. In comparison with the previous SOTA, our approach achieves similar performances on Kinetics-400. Compared with TimeSformer [10], our Two-Pathway Transformer achieves approximate performance with lower resolution and framerate. Our TP-ViT(SF+SK) has a smaller number of tokens, and thus the computation cost of TP-ViT(SF+SK) is about half that of TimeSformer-L. By further combining skeleton data, the performance is slightly improved. Due to the property of the videos, the extracted skeleton data is usually imprecise or sometimes unavailable in this dataset, and thus the performance improvement is not quite significant when skeleton is combined.

FineGym-99 and FineGym-288. As can be seen in Table 2, on these two datasets, our proposed approaches outperform the existing approaches. By using only RGB stream, our Two-Pathway ViT already outperforms I3D [1] and TSM [2] with two streams (RGB + Optical flow). By incorporating the skeleton data, the performance is further improved.

Method Info		FineGym-99		FineGym-288	
Method	Modality	Top-1	Top-5	Top-1	Top-5
I3D[1]	2Stream	75.60%	N/A	66.10%	N/A
TSM[2]	2Stream	88.40%	N/A	83.10%	N/A
I3D NL[4]	RGB	75.30%	N/A	67.00%	N/A
TP-ViT(ours)	RGB-Slow	86.88%	99.37%	85.69%	98.67%
TP-ViT(ours)	RGB-SF	88.52%	99.44%	87.37%	98.65%
TP-ViT(ours)	SF+SK	**89.25%**	**99.77%**	**89.38%**	**98.96%**

Table 2. Performance on the FineGym [27] dataset.

4.4. Ablation Studies

In Tables 1, 2, and 3, we provide ablation studies for our proposed TP-ViT.

Pathways and skeleton. As can be seen in Tables 1 and 2, our Two-Pathway structure has significant advantages over the single slow pathway. The skeleton information also increases the performance of the model by 0.5% on Top-1 for Kinetics-400 and around 2% for FineGym. This demonstrates the effectiveness of our approach for combining multiple pathways and multiple streams for action recognition with Transformer.

Fusion Strategies. Table 3 compares the performances of three strategies for fusing skeleton and RGB streams. The halfway fusion shows superior to others on three datasets.

Fusion	Kinetics-400		FineGym-99		FineGym-288	
Strategy	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Early	80.49%	94.33%	88.84%	99.47%	87.49%	98.75%
Halfway	**80.81%**	**95.39%**	**89.25%**	**99.77%**	**89.38%**	**98.96%**
Late	80.18%	94.46%	88.63%	99.42%	87.80%	98.89%

Table 3. The performances of three fusion strategies.

5. CONCLUSION

In this paper, we have presented our Two-Pathway Vision Transformer for video action recognition, which can take multi-pathway and multi-stream information as the input. By fusing samples of different resolutions and framerates with our two-pathway structure, competitive results are achieved. Furthermore, multi-stream information can also significantly improve the performance. The experiments demonstrate the effectiveness of our TP-ViT in combining multiple pathways and multiple streams for action recognition. The concept of TP-ViT shows potentials for boosting video action recognition.

6. REFERENCES

[1] Joao Carreira and Andrew Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *CVPR 2017*. jul 2017, vol. 2017-Janua, pp. 4724–4733, IEEE.

[2] Ji Lin, Chuang Gan, and Song Han, "TSM: Temporal Shift Module for Efficient Video Understanding," in *ICCV 2019*. oct 2019, vol. 2019-Oct, pp. 7082–7092, IEEE.

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "SlowFast Networks for Video Recognition," in *ICCV 2019*. oct 2019, vol. 2019-Oct, pp. 6201–6210, IEEE.

[4] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local Neural Networks," *CVPR*, pp. 7794–7803, 2018.

[5] Christoph Feichtenhofer, "X3D: Expanding Architectures for Efficient Video Recognition," in *CVPR 2020*. jun 2020, pp. 200–210, IEEE.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *NIPS*, vol. 2017-Decem, pp. 5999–6009, 2017.

[7] 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," *arXiv e-prints*, vol. abs/2010.11929, 2020.

[8] Hangbo Bao, Li Dong, and Furu Wei, "BEiT: BERT Pre-Training of Image Transformers," *arXiv e-prints*, vol. abs/2106.08254, pp. 1–16, jun 2021.

[9] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid, "ViViT: A Video Vision Transformer," *arXiv e-prints*, vol. abs/2103.15691, 2021.

[10] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is Space-Time Attention All You Need for Video Understanding?," *arXiv e-prints*, vol. abs/2102.05095, feb 2021.

[11] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid, "Videobert: A joint model for video and language representation learning," in *ICCV*, 2019, pp. 7463–7472.

[12] Bo Jiang, Jiahong Yu, Lei Zhou, Kailin Wu, and Yang Yang, "Two-Pathway Transformer Network for Video Action Recognition," in *ICIP 2021*. sep 2021, pp. 1089–1093, IEEE.

[13] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE PAMI*, vol. 43, no. 1, pp. 172–186, jan 2021.

[14] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," in *CVPR 2016*. jun 2016, vol. 2016-Dec, pp. 1933–1941, IEEE.

[15] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes, "The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation," in *Proceedings of ACL*, Stroudsburg, PA, USA, apr 2018, vol. 1, pp. 76–86.

[16] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019*, vol. 1, pp. 4171–4186, 2019.

[17] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov, "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context," in *Proceedings of ACL*, Stroudsburg, PA, USA, 2019, pp. 2978–2988, Association for Computational Linguistics.

[18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, "Language Models are Few-Shot Learners," *NIPS*, vol. 2020-Decem, may 2020.

[19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," 2019.

[20] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit, "Scaling Autoregressive Video Models," *ICLR*, pp. 1–24, jun 2019.

[21] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans, "Axial Attention in Multidimensional Transformers," *arXiv e-prints*, vol. abs/1912.12180, pp. 1–11, dec 2019.

[22] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., Cham, 2020, pp. 108–126, Springer.

[23] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox, "FlowNet: Learning Optical Flow with Convolutional Networks," in *ICCV 2015*. dec 2015, vol. 2015 Inter, pp. 2758–2766, IEEE.

[24] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," *CVPR 2017*, vol. 2017-Jan, pp. 1647–1655, 2017.

[25] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," *AAAI 2018*, pp. 7444–7452, jan 2018.

[26] Dan Hendrycks and Kevin Gimpel, "Gaussian Error Linear Units (GELUs)," *arXiv e-prints*, vol. abs/1606.08415, pp. 1–9, jun 2016.

[27] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin, "FineGym: A Hierarchical Video Dataset for Fine-Grained Action Understanding," in *CVPR 2020*. jun 2020, pp. 2613–2622, IEEE.

[28] 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*. jun 2009, pp. 248–255, IEEE.

[29] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He, "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour," *arXiv e-prints*, vol. abs/1706.02677, jun 2017.