Semantic Video Transformer for Robust Action Recognition

Keval Doshi and Yasin Yilmaz
University of South Florida
4202 E Fowler Ave, Tampa, FL 33620
{kevaldoshi, yasiny}@usf.edu

Abstract

Video action recognition has attracted significant research attention over the past several years. Although adversarial effects and robustness in image classification models have been heavily investigated, robustness of action recognition models to natural or adversarial perturbations remain largely unexplored. Moreover, even though transformer based approaches have shown great promise on various vision tasks, they have yet to be evaluated in terms of their robustness. To this end, we propose a Semantic Video Transformer for Action Recognition (SeViTAR), which maps visual features obtained by a video transformer to a more robust visual-semantic representation. We extensively evaluate the proposed approach on the ROSE Challenge dataset, and outperform all baselines with a significant margin.

1. Introduction

Due to the availability of extensively annotated large datasets as well as the development of improved deep neural network (DNN) architectures, several visual recognition tasks, such as image classification and video action recognition, have made significant progress in recent years. Although they have achieved considerable success, DNNs have been discovered to be vulnerable to adversarial attacks [10, 17, 18, 21]. An extensive body of research has demonstrated that introducing perturbations into an image or video can cause a given DNN prediction to be incorrect. Specifically, it was recently shown that convolutional neural network (CNN) based architectures for video action recognition are susceptible to such adversarial attacks and added perturbations can significantly affect their overall performance. Moreover, natural perturbations to image or video can happen such as camera shaking due to wind, poor video quality due to the Internet connection problems, etc. Machine learning models should be robust to both natural and adversarial perturbations. While robust image recognition models have been extensively studied, robustness in video action recognition still remains largely unexplored.

In the existing research, video action recognition is characterized as a supervised classification problem, in which a model is trained on a collection of known classes and then used to classify a set of unknown test videos. Most existing approaches rely on extracting visual feature representations using DNN architectures and then classifying them using one-hot encoding. In this paper, we argue that using the inherent semantic information from the class labels can lead to a more robust representation. For example, classes such as *skiing* and *slalom skiing* might have similar visual feature representations, and thus not be discriminative enough to be classified correctly. Hence, we propose leveraging the semantic information of the class labels to generate a more robust feature representation.

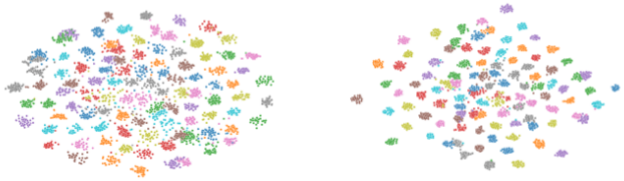

Figure 1. t-SNE visualization of the I3D (left) and Transformer (right) features extracted from the UCF-101 dataset. Each point represents a video and various classes are represented with different colors. We see that the features learned by the proposed Transformer model (right) are semantically more separable than the I3D features (left). Best viewed in color.

Several existing video understanding works primarily leverage 3D convolutional models for extracting spatiotemporal feature representations from videos, such as C3D [22] or I3D [2]. However, convolution-based approaches have a number of flaws caused due to inductive biases such as local connection, translation invariance, and a locally constricted receptive field, which in particular severely limits the learning capacity of convolutional models on massive datasets. Convolutional kernels are also incapable of capturing spatiotemporal correlations that span a large number of time instances. Finally, despite advancements in hard-

ware acceleration, training and evaluating deep CNNs on large video datasets remains a computationally demanding task. Self-attention based transformer models, on the other hand, are able to overcome these limitations due to a relatively larger receptive field [28]. Furthermore, because self-attention based models are computationally more efficient compared to convolutional approaches, they can process longer video sequences, capturing long-range dependencies more effectively. In Fig. 1, we show the t-SNE representation of the extracted visual features using the proposed transformer based approach as compared to the I3D model. We see that the transformer model learns more semantically separable features as compared to I3D.

Motivated by these observations, we propose leveraging self-attention architectures, in particular a spatiotemporal transformer model, for extracting visual feature representation from videos. The self-attention approach, in contrast to convolutional kernels, can capture long-range dependencies and is permutation invariant while being computationally efficient during training and inference. With the help of a semantic embedding mechanism, our semantic transformer based approach is able to learn spatiotemporal features that are more robust to input perturbations compared to the convolutional models. Our contributions can be summarized as follows:

- *A novel semantic video transformer architecture.* We propose a novel architecture that maps the spatiotemporal video features learned by a transformer to a semantic embedding with the help of class labels.
- *A robust transformer based approach for video action recognition.* We show that the semantic embedding together with the spatiotemporal transformer output provide a more robust feature representation.
- *An extensive evaluation of the proposed approach on several benchmark datasets.* The proposed method outperforms existing benchmark approaches by a wide margin and is ranked 1st in the ROSE Challenge.

2. Related Works

Video action recognition has been extensively studied over the past several years [2, 7–9, 20, 24]. Early approaches leveraged temporal layers such as the LSTM [19] to learn long-term dependencies. On the other hand, 3D-Convolutional models such as C3D [22] and I3D [2] extended 2D-CNN to 3D-CNNs kernels. Recently, transformer based approaches such as the TimeSformer architecture [1] and VidTr [28] have shown promising results on video recognition tasks. While earlier approaches use handcrafted semantic features [12], recent works have primarily use Word2Vec [15] for generating the semantic embeddings from the class labels. However, such approaches are prone to suffer from the domain shift problem, which occurs when a model trained on the seen semantic labels is unable to generalize well to the unseen class labels [11].

Despite the fact that adversarial attacks on images have been studied for several years since [10], the vulnerabilities of video recognition models have only recently come to light [13, 14, 18, 25–27]. However, several recent approaches address the vulnerabilities of existing deep neural network based models to perturbations or adversarial effects. Specifically, the most common perturbation pattern is sending queries to the target model and estimating gradients to generate adversarial data. PatchAttack (V-BAD) [13] is the first proposed black-box video attack. They proposed a method to generate perturbations for each frame of a video, and then updated their perturbations with queries. Heuristic Attack [25] is another black-box adversarial attack which uses a query-based attack strategy to heuristically select the key frames to be attacked. More recently, Geo-Trap [14] leverage geometric transformations for generating perturbations using less number of queries.

3. Proposed Approach

In this section, we present the proposed approach for robust video action recognition. We introduce a novel transformer model, *SeViTAR (Semantic Video Transformer for Action Recognition)*, which leverages spatiotemporal self-attention and semantic embedding to learn feature representations that are more robust to video perturbations compared to the benchmark convolutional architectures.

3.1. Problem Definition

Traditionally, video action recognition has been defined as a supervised classification problem, where given a training set of videos X^s and labels S from seen classes $\{(x_1^s, s_1), \ldots, (x_N^s, s_N)\}$, we aim to accurately classify a set of videos $X^u = \{(x_1^u), \ldots, (x_M^u)\}$ from unknown classes $U = \{u_1, \ldots, u_M\}$, where N and M are the number of training and testing videos respectively.

3.2. Semantic Video Transformer

Given the promising performance of recently proposed video transformers, such as TimeSformer [1] and VidTr [28], we extract visual features from videos using a spatiotemporal transformer model. Similar to the TimeSformer and VidTr architectures, we leverage the divided space-time attention mechanism. Specifically, applying spatial and temporal attention at the same time requires $\mathcal{O}(n^2)$ complexity with respect to the sequence length, which quickly becomes inefficient. Hence, as shown in Fig. 2, in each encoding block, we first apply temporal attention followed by spatial attention, both of which use the standard qkv attention scheme proposed in [23]. The entire transformer consists of A parallel self-attention heads with L sequential

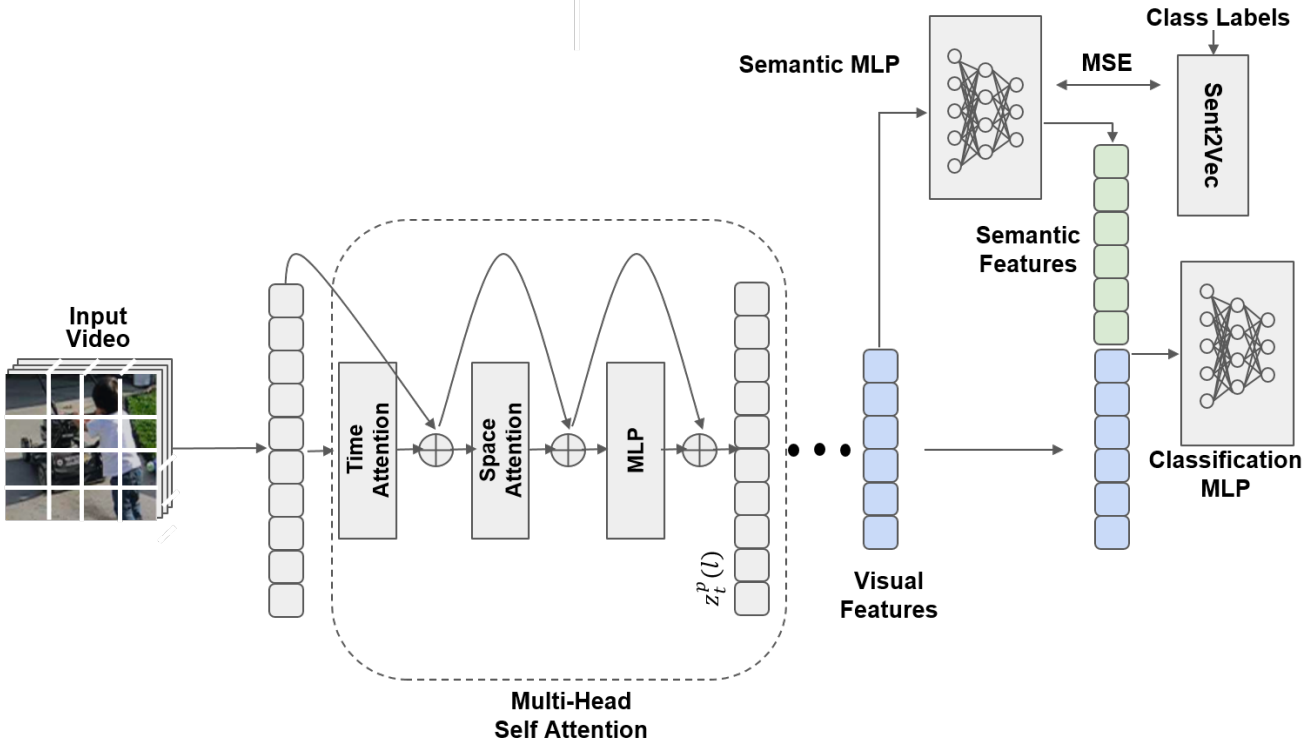

Figure 2. The proposed semantic video transformer architecture, SeViTAR, for robust action recognition.

encoding blocks in each head. The visual feature extraction procedure is as follows.

First, a clip y of F frames and size $H \times W$ is sampled from the input video x. We proceed by converting the entire video clip y to a sequence of 2D patches of size $P \times P$, given by $e_t^p \in \mathbb{R}^{3\times P^2}$. Here $p = 1, \ldots, N$ represents the spatial position of each patch (i.e., patch index), $t = 1, \ldots, F$ denotes the temporal location of each frame and 3 is the number of color channels. Finally, we flatten each patch e_t^p into $v_t^p \in \mathbb{R}^{3P^2}$ and linearly map it into a score vector using a trainable linear projection $E \in \mathbb{R}^{q\times 3P^2}$:

$$z_t^p(0) = Ev_t^p + \mu_t^p, \tag{1}$$

where $\mu_t^p \in \mathbb{R}^q$ is a latent vector learned to encode the spatiotemporal position of each (p, t) pair. Following the NLP transformer BERT [4], we add a latent vector $z_0^0(0) \in \mathbb{R}^q$ for an additional fictional patch that will be trained to represent the video's score vector through time and spatial self-attention. The input to the model is $\{z_0^0(0), z_t^p(0)\}_{p,t}$. Each block l receives $\{z_0^0(l-1), z_t^p(l-1)\}$ and produces $\{z_0^0(l), z_t^p(l)\}$. We then use the output of the last step $z_0^0(L)$ as the video's visual feature representation.

Next, for a more robust representation, we augment the visual features with semantic features. Specifically, we obtain a semantic embedding of the input video x by mapping visual features $z_0^0(L)$ to a semantic space through a multilayer perceptron (MLP). This semantic MLP is trained by minimizing the mean squared error (MSE) between the output semantic feature vector $\hat{s}$ and the Sent2Vec embedding s_{vec} of the video class label, s. Finally, the semantic feature vector $\hat{s}$ is concatenated with the visual feature vector $z_0^0(L)$ to form the visual-semantic feature representation $[\hat{s}, z_0^0(L)]$. In the final stage, the classification MLP trains on these visual-semantic representations to classify the input videos using the cross-entropy loss.

3.3. Implementation Details

The proposed SeViTAR model is built upon the SlowFast [3] and TimeSformer packages [1]. We preprocess the input videos by resizing the shorter side to 256 pixels and then a random crop is applied to create a 224×224 ($H\times W$) video snippet. Following the Vision Transformer (ViT) [5] model, we use a patch size of 16×16, which results in a total of $N = 196$ patches in a frame. During each encoding block for each patch, the size of the score vectors ($z_t^p(l)$) at each encoding block is $q = 768$, and the number of self-attention heads is set as $A = 12$. The number of input frames is set as 8. We extracted the features using 4 NVIDIA A6000 GPUs with a batch size of 8. The loss function is minimized via synchronized SGD with a learn-

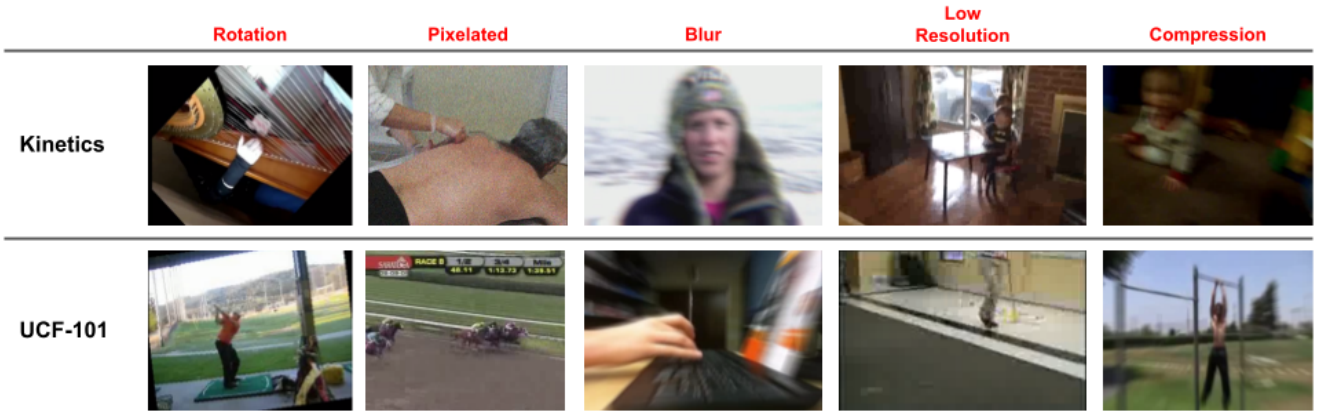

Figure 3. Different types of perturbations from the ROSE Challenge dataset.

ing rate of 0.002. To extract semantic embeddings, we use the Sent2Vec algorithm proposed in [16].

4. Experiments

4.1. Datasets

The UCF-101, HMDB-51, and the Kinetics datasets are three publicly available benchmark datasets that have been commonly used in the majority of recent studies to evaluate the action recognition performance. The UCF-101 dataset contains 13,320 videos from 101 classes, with the majority of the videos focusing on five different types of actions. In the HMDB-51 dataset, there are 6767 videos divided into 51 classes based on everyday human actions. The Kinetics-400 dataset is one of the most comprehensive action recognition datasets available with a total of over 250K videos sourced from YouTube belonging to 400 different classes. To evaluate the performance of the proposed approach in terms of robustness, we use the ROSE Challenge dataset, which consists of perturbations such as spatial, temporal, camera-related, and compression corruptions added to the benchmark datasets. A sample set of the perturbations are shown in Fig. 3. The introduced perturbations make it significantly more difficult to recognize activities. Moreover, the intensity of each perturbation is different for each video.

4.2. Results

In Table 1, the performance of the proposed approach is compared with existing state-of-the-art approaches such as SlowFast [6], I3D [2] and the TimeSformer [1]. We used the I3D and TimeSformer models on the ROSE Challenge dataset to obtain benchmark results. The SlowFast results are provided by the ROSE Challenge organizers. It is clearly seen that the proposed SeViTAR model considerably outperforms all benchmark results. Specifically, we see an improvement of 8.44% on the HMDB-51P and 10.96% on the Kinetics-400P datasets as compared to the next best

Method	UCF-101P	HMDB-51P	Kinetics-400P
SlowFast [6]	-	-	55.48
I3D [2]	76.9	-	-
TimeSformer [1]	88.5	70.48	54.26
Team Simplexisigil	59.46	44.50	-
SeViTAR (Ours)	**89.35**	**78.92**	**65.22***

Table 1. Video action recognition performance on the perturbed datasets of the ROSE Robustness Challenge. The proposed semantic video transformer method outperforms the existing methods with the first rank in the Challenge.

Method	UCF-101PMini	HMDB-51PMini	
Team Simplexisigil	51.57	43.40	-
SeViTAR (Ours)	**90.54**	**80.98**	

Table 2. Video action recognition performance on the mini versions of the perturbed datasets of the ROSE Robustness Challenge.

results. These results demonstrate the robustness of the proposed approach as compared to existing models. In particular, the improvement over the TimeSformer results show the contribution of the semantic embedding part.

It should be noted that the performance of the SeViTAR approach on the Kinetics dataset leverages only 80K training videos due to time and computational constraints. It is highly anticipated that the accuracy of 65.22% will further improve when trained on the entire Kinetics dataset. The performance of our method on the mini sets released for the ROSE Challenge are also provided in Table 2.

5. Conclusion

In this work, we introduce a semantic video transformer for robust video action recognition, called SeViTAR. Specifically, our goal is to highlight the robustness of the

visual-semantic embeddings extracted using the proposed SeViTAR model as compared to the existing 3D-CNN and transformer based visual models. Through experiments on the ROSE Robustness Challenge datasets, we show that the proposed approach significantly outperforms the existing approaches by a wide margin.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 2, 3, 4

[2] 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. 1, 2, 4

[3] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3435–3444, 2019. 3

[4] 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. 3

[5] 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. 3

[6] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 4

[7] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4768–4777, 2017. 2

[8] 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. 2

[9] Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5378–5387, 2015. 2

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[11] Meera Hahn, Andrew Silva, and James M Rehg. Action2vec: A crossmodal embedding approach to action learning. *arXiv preprint arXiv:1901.00484*, 2019. 2

[12] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 2

[13] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. Black-box adversarial attacks on video recognition models. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 864–872, 2019. 2

[14] Shasha Li, Abhishek Aich, Shitong Zhu, Salman Asif, Chengyu Song, Amit Roy-Chowdhury, and Srikanth Krishnamurthy. Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[16] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. 4

[17] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 1

[18] Roi Pony, Itay Naeh, and Shie Mannor. Over-the-air adversarial flickering attacks against video recognition networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 515–524, 2021. 1, 2

[19] Hasim Sak, Andrew W. Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 338–342, 2014. 2

[20] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 2

[21] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[22] 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. 1, 2

[23] 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. 2

[24] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S Yu. Spatiotemporal pyramid network for video action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1529–1538, 2017. 2

[25] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang. Heuristic black-box adversarial attacks on video recognition models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12338–12345, 2020. 2

[26] Huanqian Yan, Xingxing Wei, and Bo Li. Sparse black-box video attack with reinforcement learning. *arXiv preprint arXiv:2001.03754*, 2020. 2

[27] Hu Zhang, Linchao Zhu, Yi Zhu, and Yi Yang. Motion-excited sampler: Video adversarial attack with sparked prior. In *European Conference on Computer Vision*, pages 240–256. Springer, 2020. 2

[28] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13577–13587, 2021. 2