A Large-scale Robustness Analysis of Video Action Recognition Models

Madeline Chantry Schiappa [1*†], Naman Biyani [2*], Prudvi Kamtam[1]
Shruti Vyas [1], Hamid Palangi [3], Vibhav Vineet [3‡], Yogesh Rawat [1‡]
CRCV, University of Central Florida [1], IIT Kanpur [2], and Microsoft Research [3]

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.01412

Abstract

We have seen a great progress in video action recognition in recent years. There are several models based on convolutional neural network (CNN) and some recent transformer based approaches which provide top performance on existing benchmarks. In this work, we perform a ***large-scale robustness analysis*** *of these existing models for video action recognition. We focus on robustness against* ***real-world distribution shift*** *perturbations instead of adversarial perturbations. We propose* ***four*** *different benchmark datasets,* ***HMDB51-P***, ***UCF101-P***, ***Kinetics400-P***, *and* ***SSv2-P*** *to perform this analysis. We study robustness of* ***six*** *state-of-the-art action recognition models against* ***90*** *different perturbations. The study reveals some interesting findings, 1)* ***transformer*** *based models are consistently* ***more robust*** *compared to CNN based models, 2)* ***Pretraining improves robustness*** *for Transformer based models more than CNN based models, and 3) All of the studied models are* ***robust to temporal perturbations*** *for all datasets but SSv2; suggesting the importance of temporal information for action recognition varies based on the dataset and activities. Next, we study the role of augmentations in model robustness and present a real-world dataset,* ***UCF101-DS***, *which contains realistic distribution shifts, to further validate some of these findings. We believe this study will serve as a benchmark for future research in robust video action recognition* [1].

1. Introduction

Robustness of deep learning models against real-world distribution shifts is crucial for various applications in vision, such as medicine [4], autonomous driving [41], environment monitoring [60], conversational systems [36], robotics [68] and assistive technologies [5]. Distribution shifts with respect to training data can occur due to the variations in environment such as changes in geographical locations, background, lighting, camera models, object scale, orientations,

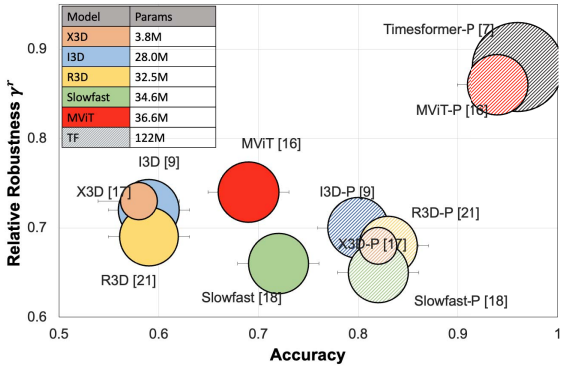

Figure 1. Performance against robustness of action recognition models on UCF101-P. y-axis: relative robustness γ^r, x-axis: accuracy on clean videos, model names appended with P indicate pre-training, and the size of circle indicates FLOPs.

motion patterns, etc. Such distribution shifts can cause the models to fail when deployed in a real world settings [26]. For example, an AI ball tracker that replaced human camera operators was recently deployed in a soccer game and repeatedly confused a soccer ball with the bald head of a lineman, leading to a bad experience for viewers [13].

Robustness has been an active research topic due to its importance for real-world applications [4, 41, 60]. However, most of the effort is directed towards images [7, 25, 26]. Video is a natural form of input to the vision systems that function in the real world. Therefore studying robustness in videos is an important step towards developing reliable systems for real world deployment. In this work we perform a large-scale analysis on robustness of existing deep models for video action recognition against common real world spatial and temporal distribution shifts.

Video action recognition provides an important test scenario to study robustness in videos given there are sufficient, large benchmark datasets and well developed deep learning models. Although the existing approaches have made impressive progress in action recognition, there are several fundamental questions that still remain unanswered in the field. Do these approaches enable effective temporal modeling, the crux of the matter for action recognition approaches? Are these approaches robust to real-world corruptions like noise

*The authors contributed equally as first authors to this paper.
†Corresponding author:madelineschiappa@knights.ucf.edu
‡The authors contributed equally as supervisors to this paper.
[1]More details available at bit.ly/3TJLMUF.

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.01412

(temporally consistent and inconsistent), blurring effects, etc? Do we really need heavy architectures for robustness, or are light-weight models good enough? Are the recently introduced transformer-based models, which give state of the art accuracy on video datasets, more robust? Does pre-training play a role in model robustness? This study aims at finding answers to some of these critical questions.

Towards this goal, we present multiple benchmark datasets to conduct robustness analysis in video action recognition. We utilize four different widely used action recognition datasets including HMDB51 [34], UCF101 [54], Kinetics-400 [8], and SSv2 [43] and propose four corresponding benchmarks; *HMDB51-P*, *UCF101-P*, *Kinetics400-P*, and *SSv2-P*. In order to create this benchmark, we introduce 90 different common perturbations which include, 20 different noise corruptions, 15 blur perturbations, 15 digital perturbations, 25 temporal perturbations, and 15 camera motion perturbations as a benchmark. The study covers 6 different deep architectures considering different aspects, such as network size (small vs large), network architecture (CNN vs Transformers), and network depth (shallow vs deep).

This study reveals several interesting findings about action recognition models. We observe that recent *transformer based models are not only better in performance*, but they are also *more robust than CNN models* against most distribution shifts (Figure 1). We also observe that *pretraining is more beneficial to transformers compared to CNN* based models in robustness. We find that all the models are very robust against the temporal perturbations with minor drop in performance on the Kinetics, UCF and HMDB. However, on the SSv2 dataset, behavior of the models is different whereas the performance drops on different temporal perturbations. These observations show interesting phenomena about the video action recognition datasets, i.e., *the importance of temporal information varies based on the dataset and activities*.

Next, we study the role of training with data augmentations in model robustness and analyze the generalization of these techniques to novel perturbations. To further study the capability of such techniques, we propose a real-world dataset, *UCF101-DS*, which contains realistic distribution shifts without simulation. This dataset also helps us to better understand the behavior of CNN and Transformer-based models under realistic scenarios. We believe such findings will open up many interesting research directions in video action recognition and will facilitate future research on video robustness which will lead to more robust architectures for real-world deployment.

We make the following contributions in this study,

- A large-scale robustness analysis of video action recognition models to different real-world distribution shifts.
- Provide insights including comparison of transformer vs CNN based models, effect of pre-training, and effect of temporal perturbations on video robustness.
- Four large-scale benchmark datasets to study robustness for video action recognition along with a real-world dataset with realistic distribution shifts.

2. Related work

2.1. Action recognition

Video understanding has made rapid progress with the introduction of a number of large-scale video datasets such as Kinetics [8], Sports1M [30], Moments-In-Time [44] , SSv2 [23] and YouTube-8M [2]. A number of recent models have emphasized the need to efficiently model spatio-temporal information for video action recognition. Some early approaches, inspired by image classification models [33], utilize 2D-CNN models [30] for video classification. Some recent works [37, 62, 72] have proposed effective ways to integrate image level features for video understanding. The success of 2D convolution has inspired many 3D convolution based approaches for recognizing actions in videos [12, 29]. For example, C3D [58] learns 3D ConvNets, outperforming 2D CNNs through the use of large-scale video datasets. Many variants of 3D-CNNs are introduced for learning spatio-temporal features such as I3D [10] and ResNet3D [24]. 3D CNN features were also demonstrated to generalize well to other vision tasks [1, 11, 17, 56, 65, 66]. Because 3D CNN based approaches lead to higher computational load, recent works aim to reduce the complexity by decomposing the 3D convolution into 2D and 1D convolutions [48, 59, 64], or incorporating group convolution [40]; or using a combination of 2D and 3D-CNN [12]. Furthermore, SlowFast [20] network employs two pathways to capture short-term and long-term temporal information by processing a video at both slow and fast frame rates.

Recently, transformer based models have shown remarkable success in various vision tasks, such as image classification, after the introduction of Vision Transformer (ViT) [16]. The impressive performance led to using transformer-based architectures for video domains. Video transformers have led to state-of-the-art performance on Kinetics-400 [8], SSv2 [23] and Charades [53]. Specific to video, a temporal attention encoder was added on top of ViT, further improving performance on action recognition [46]. More recently, MViT [18] was proposed; a multi-scale vision transformer for video recognition that achieved top results on SSv2. A factorized spacetime attention based approach was proposed in Timesformer [6] after analysis of various variants of spacetime attention based on compute-accuracy tradeoff. Video Swin Transformer [38] investigated spatiotemporal locality and showed that an inductive bias of locality is a better speed-accuracy trade-off compared to using global self-attention. We use both CNN-based and recent transformer-based archi-

tectures to study their robustness for action recognition.

2.2. Robustness

Many recent works on robustness in the vision community are focused on adversarial attacks, where a computed perturbation is deliberately added to the input sample [3, 71]. Different from adversarial attacks, the real-world distribution shifts in data naturally emerge from different scenarios. Some of the recent works are focused towards understanding the robustness of existing methods in the image domain against these distribution shifts [7, 25, 26, 51]. In [26], the authors analyzed different image classification models for different corruptions in ImageNet. Similarly, in [49] the authors presented a new benchmark of naturally occurring distribution shifts using ImageNet and studied the robustness of different image models. In a recent study [45], the image based transformer models were found to be more robust towards different kinds of perturbations. The benchmark in [57] analyzed natural robustness and demonstrated that data augmentation is not sufficient to improve model robustness.

Some recent works have further explored the use of data augmentation to improve the robustness of image models [21, 27, 69]. Data augmentations such as various noise types [39, 42, 50], transformations [21, 70], and compositions of these simple transformations [14, 27] are shown to be helpful in improving the robustness of deep networks. These robustness studies are mainly focused on images. There are a few works addressing the issue of adversarial robustness in videos [63] and analyzing importance of temporal aspect in videos [15, 52]. Different from these existing works, this work provides a large-scale benchmark on robustness of video action recognition models against real-world perturbations.

In a recent effort [67], an initial analysis on robustness against natural distribution shift was presented for videos extending visual augmentations [26]. This work was focused on compression specific perturbations including: bit rate, compression, frame rate, and packet loss. Different from this work, we emphasize on temporal perturbations that are not limited to compression. Moreover, this study a small scale benchmark focusing on subsets of Kinetics [9] and SSv2 [22]. In comparison, our analysis uses the full Kinetics and SSv2 dataset while additionally analyzing models on UCF101 [55] and HMDB51 [35], the most common action recognition evaluation datasets. As a result, our findings differ from their initial findings, e.g. model capacity and robustness or generalization when trained on perturbations.

3. Distribution shifts

Existing research in action recognition is mostly focused on training and testing the proposed methods on a benchmark dataset with little to no distribution shift from training to testing samples. In most of the real-world applications, we observe different types of distribution shifts in testing environments before deployment, affecting the performance of the models. To help circumvent this issue, it is important to study robustness of existing deep learning based video action recognition models against real-world perturbations, i.e., they are not artificially created using adversarial attacks and happen naturally for example due to change in environment, different camera settings, etc. Towards this goal, we designed a set of perturbations which are frequently encountered in real-world environments. Existing datasets on action recognition do not focus on such distribution shifts and therefore it is important to construct a benchmark that covers a wide range of distribution shifts which will be beneficial for the community. We study five different categories of real-world perturbations which include, *noise*, *blur*, *digital*, *temporal*, and *camera motion*.

Noise: We define 4 categories for noise; *Gaussian*, *Shot*, *Impulse*, and *Speckle* noise. *Gaussian noise* can appear due to low-lighting conditions. *Shot noise* tries to capture the electronic noise caused by the discrete nature of light. We use Poisson distribution to approximate it. *Impulse noise* tries to simulate corruptions caused by bit errors and is analogous to salt-and-pepper noise. *Speckle noise* is additive noise where noise added is proportional to the pixel intensity.

Blur: We define three kinds of perturbations for blur effect; *Zoom*, *Motion*, and *Defocus*. *Zoom blur* occurs when the camera moves toward an object rapidly. *Motion blur* appears due to the destabilizing motion of camera. Finally, *Defocus blur* may happen when the camera is out of focus.

Digital: Recent years have seen a sharp increase in video traffic. In fact, video content consumption increased so much during the initial months of the pandemic that content providers like Netflix and Youtube were forced to throttle video-streaming quality to cope with the surge. Hence efficient video compression to reduce bandwidth consumption without compromising on quality is more critical than ever. We evaluate the models on JPEG and two other video encoding codecs and analyse the drop in accuracy due to these compression methods. *JPEG* is a lossy image compression format which introduces compression artifacts. *MPEG1* is designed to compress raw digital video without excessive quality loss and is used in a large number of products and technologies. *MPEG2* is an enhanced version of MPEG1 and is also a lossy compression for videos which is used in transmission and various other applications.

Temporal: Although CNN-based approaches have made impressive progress in action recognition, one of the major questions that still remain unanswered is whether these approaches enable more effective temporal modeling, the crux of the matter for action recognition? How different are the recent Transformers from 3D-CNN based approaches as

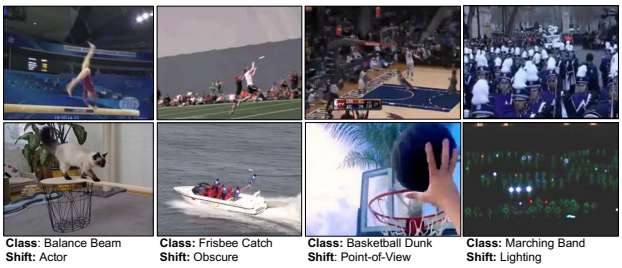

Figure 2. Sample video frames from proposed UCF101-DS dataset (bottom) compared to UCF101 (top) for 5 classes and 5 variations.

Table 1. Details of action recognition models used in this study.

Model	R3D [24]	I3D [8]	SF [20]	X3D [19]	MViT [18]	TF [6]
Params	32.5M	28.0M	34.6M	3.8M	36.6M	122M
FLOPs	55.1G	75.1G	66.6G	5.15G	70.7G	196G
# of frames	8	8	32	16	16	8
Frame rate	8	8	2	5	4	32

far as temporal modelling of video data is concerned? To compare the approaches on effective temporal modelling, we define five different temporal perturbations; *Sampling rate*, *Reversal*, *Jumbling*, *Box jumbling*, and *Freezing*. *Sampling rate* evaluates the models against varying skip-frame rates. *Reversal* perturbation reverse the video frames with varying skip-frame rates. *Jumbling* shuffles the frames in a segment-wise fashion. We utilize frame index permutation for 5 different segment sizes (4,8,16,32,64). In *Box jumbling*, we shuffle the segments instead of frames inside those segments. *Freezing* perturbation freeze video frames randomly and tries to capture the issues with video buffering.

Camera motion: To compare the approaches for robustness in the presence of irregularities due to camera motion, we define three perturbations; *Static rotation*, *Dynamic rotation*, and *Translation*. *Static rotation* uses a constant rotation angle for all the video frames. It captures effects due to tilted camera orientation. *Random rotation* rotates each frame by a varying random angle. It captures effects due to changing camera angle. *Translation* randomly crops a video frame with varying crop location across time. This is introduced to capture the random shaking motion of camera.

Severity level The natural perturbations may occur in videos at different severity levels depending on the environment/situation. Therefore, it is important to study the effect of these perturbations at different severity levels. We generate 5 levels from 1-5 where 1 refers to minimal distribution shift and 5 refers to a large distribution shift. We apply the proposed perturbations at every severity level on all the testing videos of the benchmark and save it for a consistent evaluation. More details about the implementation of these perturbations are provided in the supplementary.

4. Model variants

We perform our experiments on six different action recognition models which are based on CNN and Transformer architectures. The goal is to benchmark multiple backbones and simultaneously study the behavior of CNN and Transformer based models for robustness in video action recognition. We evaluate three most popular CNN-based action recognition models which are known to perform well, not only in action recognition, but also serve as fundamental building blocks for many other problems in the video domain. These include I3D [8], ResNet3D (R3D) [24], and SlowFast (SF) [20]. Among these, I3D and R3D are based on 3D convolutions but differ in the backbones, where I3D uses Inception-V1 and R3D uses a ResNet backbone. Slowfast is one of the best action recognition models and is based on a 3D-CNN, which can use any backbone in its two stream approach. We use a R3D backbone for both slow as well as the fast branch. We also evaluate X3D, an efficient CNN model [19] that attempts to optimize the network size and its complexity. Recently, Transformer based models have shown a great success in various vision-based tasks [32]. Several models have been proposed for video representation learning [6, 18, 38, 46]. We use the top two Transformer based models in this study, including Timesformer (TF) [6] and MViT [18]. Timesformer utilizes a factorized space-time attention whereas MViT uses pooling attention for efficient computation. More details are shown in Table 1.

5. Robustness benchmarks and evaluation

Datasets We use four action recognition benchmark datasets for our experiments including UCF101 [54], HMDB51 [34], Kinetics-400 [31], and SSv2 [43]. **UCF101** is an action recognition dataset with 101 action classes. There are a total of 13K videos, with around 100 videos per class. The length of videos in this dataset ranges from 4-10 seconds. **HMDB51** has 7K videos with 51 classes. For each action, at least 70 videos are for training and 30 videos are for testing. **SSv2** is a large collection of videos with focus on humans performing basic actions with everyday objects. There are 174 classes and it contains 220,847 videos, with 168,913 in the training set, 24,777 in the validation set and 27,157 in the test set. **Kinetics-400** is another large-scale action recognition benchmark dataset with 400 classes. Each action category has at least 400 videos and each video clip last around 10 seconds. It covers a broad range of action classes including human-object interactions and human-human interactions.

We apply the proposed 90 perturbations to the test set of these datasets to create robustness benchmarks which we refer to as **HMDB51-P**, **UCF101-P**, **Kinetics400-P**, and **SSv2-P**. HMDB51-P consists of 137,610 videos, UCF101-P consists of 340,380 videos, Kinetics400-P consists of 1,616,670

videos, and SSv2-P consists of 2,229,930 videos. These benchmarks are not used for training.

We additionally propose a new dataset that focuses on real-world distribution shifts, UCF101 Distribution Shift (UCF101-DS). For classes in the UCF101 dataset, we collected videos of uncommon or isolated variations for a number of distribution shifts that are categorized into higher-level groups such as: "style", "lighting", "scenery", "actor", "occlusion". More details about this dataset can be found in the supplementary. Some examples of these variations are in Fig. 2. We have a total of 63 distribution shifts organized into 15 categories for 47 classes for a total of 4,708 clips.

Implementation details We train R3D, I3D, SlowFast, X3D, and MViT models for HMDB51 and UCF101 with and without pre-trained weights. The pre-trained weights from Kinetics-400 are used to initialize for the first variation. Furthermore, we consider I3D, Slowfast, X3D and Timesformer models for evaluation on SSv2 dataset since pretrained weights for these four models are publicly available. We use the official implementations available with pre-trained weights with the same experimental setup as described in these works. More details in Table 1.

Evaluation protocol To ensure fair comparison and facilitate reproduciblity, we evaluate all the models under similar protocol. We use clips with a resolution of 224×224 for all the datasets. For evaluation, in Kinetics dataset, we follow the protocol of taking 10 uniform temporal crops for each video and applying center crop for each of these 10 crops. The videos in UCF101 and HMDB51 are shorter in comparison to Kinetics-400, so we take 5 uniform temporal crops for each video and apply center crop for each clip. For UCF101 and HMDB51, we also evaluated models when they are pre-trained on a large-scale dataset, such as Kinetics-400, before finetuning on these smaller datasets. For SSv2 we used a single spatial crop and uniformly sampled the number of frames as used in the original model implementation.

Evaluation metrics To measure robustness, we use two metrics; one for absolute accuracy drop and the other for relative accuracy drop. If we have a trained model f, we first compute the accuracy A_c^f on the clean test set. Next, we test this classifier on a perturbation p at each of the severity levels s, and obtain accuracy $A_{p,s}^f$ for perturbation p and severity s. The absolute robustness γ^a is computed for each severity level s and perturbation p as $\gamma_{p,s}^a = 1 - (A_c^f - A_{p,s}^f)/100$. The aggregated performance of a model can be obtained by averaging all severity levels to get γ_p^a and over all perturbations to get γ^a. Different models provide varying performance on the same test videos and therefore absolute drop in performance will also depend on the models performance on clean videos. To take this into account, we compute relative performance drop to measure models robustness. The relative robustness γ^r is computed for each severity level s and perturbation p as $\gamma_{p,s}^r = 1 - (A_c^f - A_{p,s}^f)/A_c^f$ which is the difference normalized to the accuracy of the model on the test set without perturbation.

Table 2. γ^a and γ^r robustness scores of the models on Kinetic-400P benchmark dataset. For both, higher is better.

	Noise		Blur		Temporal		Digital		Camera		Mean	
Network	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r
R3D	.71	.61	.78	.70	**.98**	**.97**	.91	.88	.89	.85	.85	.80
I3D	.72	.61	.80	.72	.97	.96	.91	.87	.89	.85	.86	.80
SF	.64	.53	.80	.73	.95	.93	.91	.89	.86	.81	.83	.78
X3D	.71	.62	.81	.75	.96	.94	.90	.86	.88	.84	.85	.80
TF	.87	.84	.84	.79	.97	.94	**.94**	.92	**.95**	**.93**	.91	.88
MViT	**.93**	**.91**	**.86**	**.82**	.96	.95	**.94**	**.93**	.94	.92	**.93**	**.91**

Table 3. γ^a and γ^r robustness scores of the models on SSv2P.

	Noise		Blur		Temporal		Digital		Camera		Mean	
Network	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r	γ^a	γ^r
I3D	.63	.40	.85	.76	.69	.50	.69	.51	.78	.65	.78	.64
SF	.51	.22	.85	.76	.68	.48	.67	.48	.74	.59	.75	.58
X3d	**.80**	**.67**	**.90**	**.67**	.77	.61	**.86**	**.78**	.80	.67	.85	.76
TF	.78	.59	.85	.74	**.89**	**.78**	.85	.73	**.88**	**.78**	**.87**	**.77**

6. Experiments

We analyze robustness of models against 5 different kinds of perturbations and what that means for model behavior on the UCF101-P, Kinetics-P, HMDB51-P and SSv2-P. A summary of model robustness across severities and perturbation categories is shown in Figures 4 and 3 and Tables 2 and 3.

6.1. Robustness analysis

Spatial Here we focus on *Noise*, *Camera* and *Blur* perturbations. In Figure 4 we observe for Kinetics-P that spatial perturbations have the largest drop in performance as severity increases. For all three categories, we see that the transformer-based Timesformer and MViT models are typically more robust than CNN-based models. For example, performance of Timesformer and ResNet based R3D drops by ∼5% and ∼30% respectively. In Figure 3, surprisingly models are more robust to variable rotation compared to a static rotation. This may be because randomly rotating may provide some frames closer to the expected but if the static rotation is far from the expected, performance drops. Behavior on SSv2 data is similar, however, in Figure 5 we observe that MViT and Timesformer models are typically less robust than X3D. This may indicate that with a more temporal-specific dataset, the CNN-based models are more robust. In summary, *all models struggle with spatial-based perturbations and the Transformer-based architectures are typically more robust than CNN-based architectures.*

Temporal To study the effect of temporal perturbation on videos, we perform experiments after applying the different types of perturbations: *jumbling*, *box jumbling*, *jumbling*,

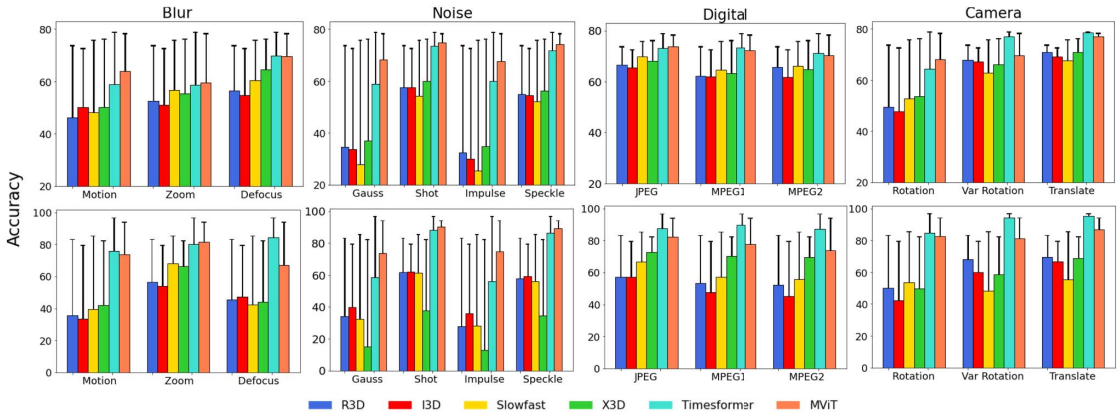

Figure 3. Robustness analysis of different action recognition models on the Kinetics-P (top row) and UCF101-P (bottom row) benchmark for various perturbations. Each bar plot corresponds to one category of perturbations showing performance drop of each model. The bar shows accuracy on perturbed dataset and the extension indicates performance drop from accuracy on clean data.

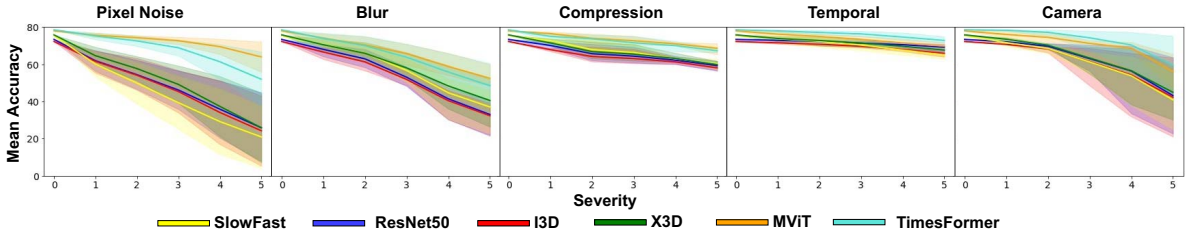

Figure 4. The mean performance on Kinetics-400P across perturbation types and severity for all models.

sampling, *reverse sampling* and *freezing* of frames. The results are presented in Fig. 4, 5, and 6. To our surprise, we observed different behaviors on different datasets. Models are typically robust on the UCF101-P, Kinetics-P, and HMDB51-P datasets while not robust on the SSv2 dataset. In order to gain further insights in their behaviors, we visualize features of CNN and transformer models using t-SNE [61] features. In Fig. 8 we visualize t-SNE features of Timesformer, X3D and Slowfast models under reverse temporal perturbation for 5 action classes and their respective opposite classes in the arrow of time from SSv2. We observe that CNN-based, X3D and Slowfast, models confuse between classes, but Timesformer model clusters different classes properly even at high severity levels. To understand class confusion further, Fig. 8 also visualizes a confusion matrix of SSv2 classes for freeze and reverse sampling between Timesformer and Slowfast. We see a noticeable different between transformer-based model and CNN-based model on over-predictions, which are visible by the dark vertical lines. This again indicates transformers may be more robust to temporal perturbations.

These observations provide an interesting phenomena about the action recognition datasets. Firstly, *temporal information is more important for action recognition on the SSv2*, where activities can often be reversed and become a different activity. Secondly, *temporal learning may not be required for shorter clips* that do not have any potential of a reversal of activities. We believe such findings will open up many interesting research directions in video action recognition.

Spatio-Temporal. Here we focus on *Compression* perturbations which affect both spatial and temporal signals in a video. In Figure 3 and 4, we observe that models are typically robust to these perturbations but struggle more on UCF101-P. For UCF101-P and HMDB51-P, we do see that the transformer-based models are typically more robust. On SSv2-P, we observe that models struggle more with compression than compared to the other datasets (Figure 5). This further indicates that SSv2-P requires more temporal learning compared to the other datasets, and therefore models struggle when temporal perturbations are present.

6.2. Effect of pretraining on robustness

We conducted experiments on the UCF101-P and HMDB51-P benchmarks, where models are pretrained on the Kinetics-400 dataset. The mean relative robustness scores across perturbation categories are shown in Fig. 7 where the closer to the center, the less robust. A breakdown of the results for each perturbation type is shown in the Supplementary for both datasets. Overall, *we observe that pretraining models results in higher robustness*. We also observe that *the relative benefit of pretraining is more evident in Transformer models compared to CNN models* (Figure 1).

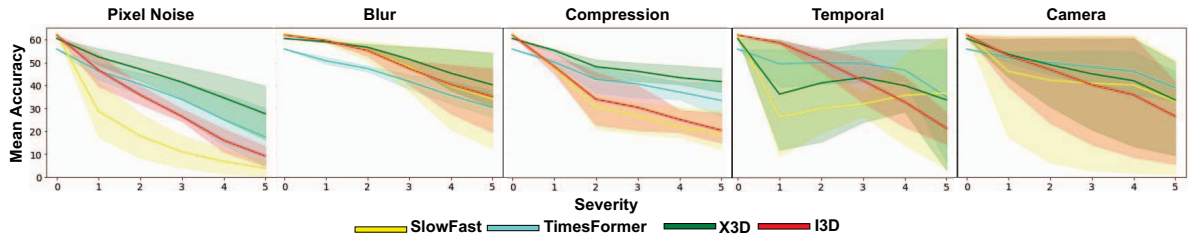

Figure 5. The mean performance on SSV2-P across perturbation types and severity for all models.

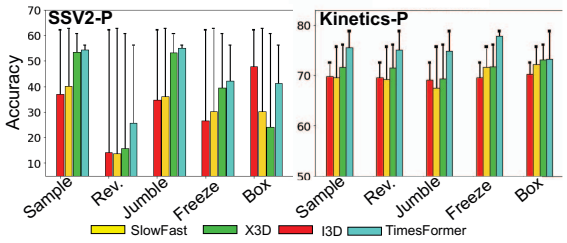

Figure 6. Robustness analysis for temporal perturbations. Each bar shows accuracy on clean data and drop on perturbed benchmarks.

6.3. Model capacity vs. robustness

To understand how model capacity might impact robustness, we compare accuracy and relative robustness γ^r in Figure 1. Models trained from scratch are solid colors while those pre-trained are dashed. The size of the dot for each model is based on the model capacity as shown in Table 1. While most of the models have around 35M parameters, the X3D [19] is significantly smaller and is still comparatively robust when compared to other models. When comparing the pre-trained MViT and Timesformer, we again see that a model with significantly less parameters is just as accurate and robust as one with significantly more parameters. In contrast to the findings in [67], our analysis *indicates that high model capacity does not necessarily mean more robustness*.

6.4. Augmentations for robustness

In this experiment we study the role of augmentations on model robustness. We explore the use of perturbations as augmentation and analyze both CNN and Transformer based models. We also experiment with PixMix [28], which is one of the recent approach for robust model learning. We use some perturbations for training and keep others for evaluation. Similarly, we use severity of 1,2, and 3 for training and 4 or 5 for testing. For PixMix [28], we apply the augmentation at severity 3 for each frame individually, in which a different fractal image is chosen for each.

The overall results for these experiments are shown in Figure 9. We observe that certain perturbations may be more beneficial for different architectures. ResNet50 becomes less robust when trained on a mix of perturbations but is more robust when trained on Spatial and PixMix. To understand changes to the networks when trained on perturbed data, we use CKA [47] to compare layer activations for the ResNet50 model on different perturbed data. Fig 9 shows a comparison between a model trained on temporal versus spatial perturbations when evaluated on UCF101-P for temporal or spatial perturbations. We find both variations of the ResNet50 are more similar for temporal perturbations compared to spatial based on the resulting scales. Both variations are also more similar at the initial layers, where most changes are in the middle or final layers. Our results indicate that *the CNN-based models may benefit more from spatial perturbations during training than transformer-based models*.

6.5. Robustness analysis on real-world videos

To better understand model behavior under natural distribution shifts, we evaluate the CNN-based model ResNet50 and the Transformer-based model MViT on UCF101-DS. The results are shown in Figure 10. When trained on UCF101, The MViT model is typically more robust to UCF101-DS compared to the ResNet50. MViT is more robust to ethnicity variations, occlusion, and changes in scene while ResNet50 is more robust to natural variations in play speed and age variations. Similar to our findings on UCF101-P in Figure 9, we find that training on perturbed data is less beneficial for MVIT compared to ResNet50. When not trained on UCF101-P, we find the MViT model is more robust to natural distribution shifts. However, when trained on UCF101-P, ResNet50 becomes more robust than MViT. *This further supports that training transformer-based models on perturbed data may not benefit robustness while it does on CNN-based approaches*. The results also indicate that these models are not typically robust to natural distribution shifts.

7. Discussion and conclusion

We have conducted a large-scale robustnesss analysis on standard CNN and Transformer based action recognition models. We created benchmark datasets based on Kinetics400, SSv2, UCF101 and HMDB51. We proposed a new dataset, UCF101-DS, that captures real-world distribution shifts in areas like scenery, point-of-view and more. Our study provides the following initial insights:

- Transformer models are generally more robust than CNN.

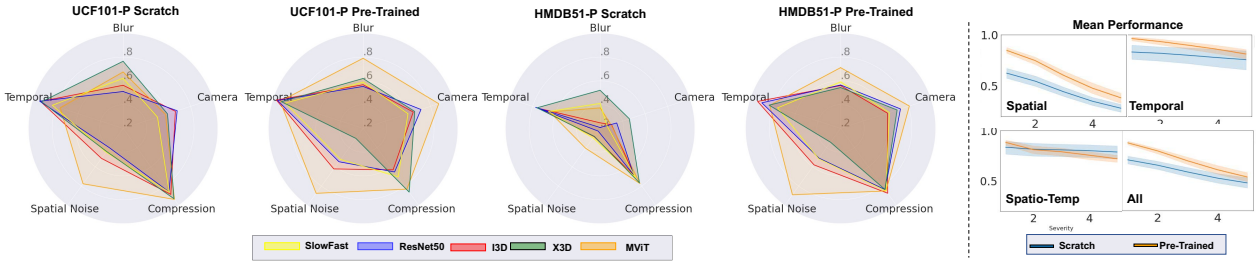

Figure 7. Left: Mean accuracy for each perturbation comparing pre-trained models to models trained from scratch on UCF101-P and HMDB51-P. Right: Mean accuracy across models for temporal, spatial and spatio-temporal perturbations on HMDB51-P and UCF101-P.

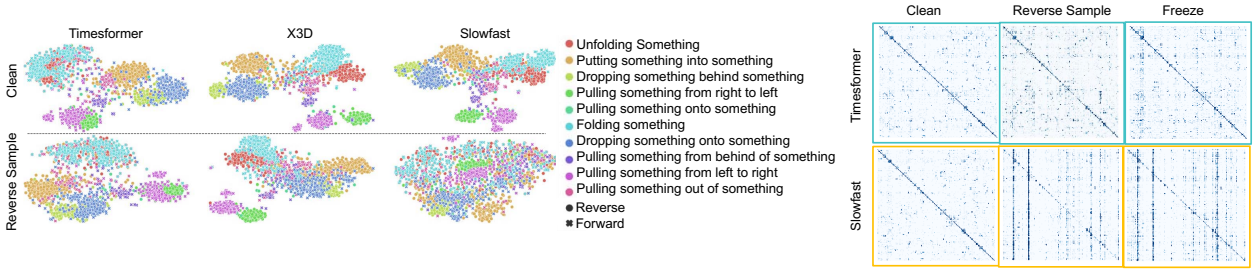

Figure 8. Comparing clean to temporally perturbed videos. Left: we visualize the feature space of a subset of classes that have a reverse activity class for SSv2. Right: The confusion matrices of SSv2 classes for different perturbations at severity 4.

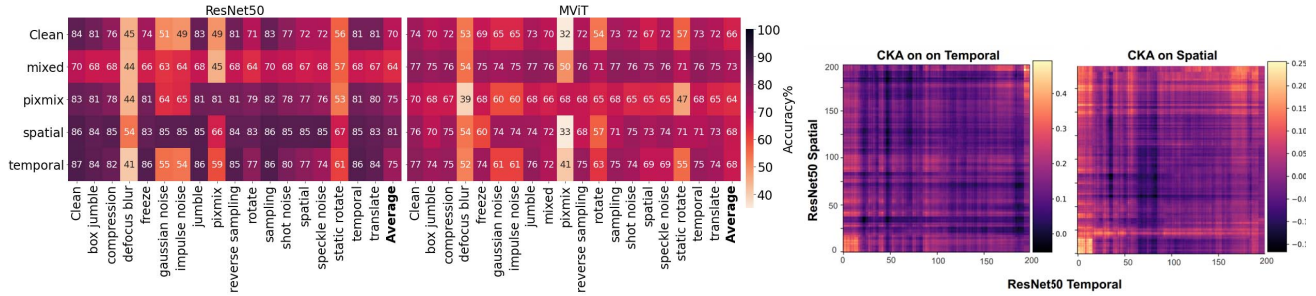

Figure 9. Left: Accuracy for each perturbation (x-axis) compared to what data models were trained on (y-axis). We compare a CNN to a transformer model when trained on clean data or different combinations of perturbations. Right: A heatmap of CKA values [47] for ResNet50 when trained and tested on either spatial or temporal perturbations.

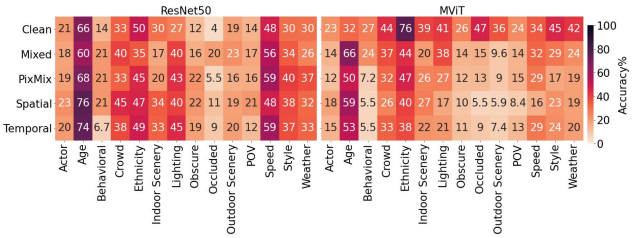

Figure 10. Overall results on our proposed UCF101-DS dataset.

- Pre-training improves robustness for transformer-based models more than CNN-based models.
- Training on perturbed data benefits CNN-based models more than Transformer-based models.
- More parameters do not necessarily mean robustness.
- Unlike the other datasets studied in this benchmark, SSv2, with its reversible actions, requires temporal learning.
- Like what is seen with images, models are not robust to spatial noise but unlike with images, they are *sometimes* robust to temporal noise.

This study presented a benchmark for robustness of video models against real-world distribution shifts. The findings and the benchmark in this work can potentially open up interesting questions about robustness of video action recognition models. The benchmark introduced in this study will be released publicly at `bit.ly/3TJLMUF`.

Acknowledgements This research is based upon work supported in part by the Office of the Director of National Intelligence (IARPA) via 2022-21102100001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the US Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12487–12496, 2019. 2

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2

[3] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018. 3

[4] Victor Ardulov, Victor R Martinez, Krishna Somandepalli, Shuting Zheng, Emma Salzman, Catherine Lord, Somer Bishop, and Shrikanth Narayanan. Robust diagnostic classification via q-learning. *Scientific reports*, 11(1):1–9, 2021. 1

[5] Onur Asan, Alparslan Emrah Bayrak, Avishek Choudhury, et al. Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of medical Internet research*, 22(6):e15154, 2020. 1

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021. 2, 4

[7] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *arXiv preprint arXiv:2103.14586*, 2021. 1, 3

[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. 2, 4

[9] 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. 3

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. 2

[11] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018. 2

[12] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021. 2

[13] Cohen Coberly. Ai failure in real-world application. *Techspot*, 2020. https://www.techspot.com/news/87431-ai-powered-camera-zooms-bald-head-instead-soccer.html. [Accessed: Oct 18, 2021]. 1

[14] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. 2019. 3

[15] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, page 103406, 2022. 3

[16] 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, 2021. 2

[17] Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Videocapsulenet: A simplified network for action detection. *Advances in Neural Information Processing Systems*, 31:7610–7619, 2018. 2

[18] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers, 2021. 2, 4

[19] 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. 4, 7

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019. 2, 4

[21] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018. 3

[22] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 3

[23] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017. 2

[24] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition, 2017. 2, 4

[25] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1, 3

[26] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations.

In *International Conference on Learning Representations*, 2018. 1, 3

[27] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2019. 3

[28] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16783–16792, 2022. 7

[29] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7366–7375, 2018. 2

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

[31] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 4

[32] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 4

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 2

[34] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 2, 4

[35] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 3

[36] Han Li, Sunghyun Park, Aswarth Dara, Jinseok Nam, Sungjin Lee, Young-Bum Kim, Spyros Matsoukas, and Ruhi Sarikaya. Neural model robustness for skill routing in large-scale conversational ai systems: A design choice exploration. *arXiv preprint arXiv:2103.03373*, 2021. 1

[37] 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. 2

[38] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, 2021. 2, 4

[39] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. 2019. 3

[40] Chenxu Luo and Alan L Yuille. Grouped spatial-temporal aggregation for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5512–5521, 2019. 2

[41] Xiaobai Ma, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Improved robustness and safety for autonomous vehicle control with adversarial reinforcement learning. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1665–1671. IEEE, 2018. 1

[42] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 3

[43] Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. On the effectiveness of task granularity for transfer learning. *arXiv preprint arXiv:1804.09235*, 2018. 2, 4

[44] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 2

[45] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[46] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network, 2021. 2, 4

[47] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2020. 7, 8

[48] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 2

[49] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 3

[50] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. Increasing the robustness of dnns against image corruptions by playing the game of noise. 2020. 3

[51] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 3

[52] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 535–544, 2021. 3

[53] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in

homes: Crowdsourcing data collection for activity understanding, 2016. 2

[54] 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. 2, 4

[55] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[56] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *European Conference on Computer Vision*, pages 71–87. Springer, 2020. 2

[57] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. When robustness doesn't promote robustness: Synthetic vs. natural distribution shifts on imagenet. 2019. 3

[58] 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. 2

[59] 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. 2

[60] Silvia Liberata Ullo and GR Sinha. Advances in smart environment monitoring systems using iot and sensors. *Sensors*, 20(11):3113, 2020. 1

[61] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 6

[62] 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. 2

[63] Min Wu and Marta Kwiatkowska. Robustness guarantees for deep neural networks on videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2020. 3

[64] 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. 2

[65] Ke Yang, Peng Qiao, Dongsheng Li, Shaohe Lv, and Yong Dou. Exploring temporal preservation networks for precise temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[66] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 982–990, 2016. 2

[67] Chenyu Yi, SIYUAN YANG, Haoliang Li, Yap peng Tan, and Alex Kot. Benchmarking the robustness of spatial-temporal models against corruptions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3, 7

[68] Mark Yim, Wei-Min Shen, Behnam Salemi, Daniela Rus, Mark Moll, Hod Lipson, Eric Klavins, and Gregory S Chirikjian. Modular self-reconfigurable robot systems [grand challenges of robotics]. *IEEE Robotics & Automation Magazine*, 14(1):43–52, 2007. 1

[69] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13276–13286, 2019. 3

[70] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 3

[71] Xingwei Zhang, Xiaolong Zheng, and Wenji Mao. Adversarial perturbation defense on deep neural networks. *ACM Computing Surveys (CSUR)*, 54(8):1–36, 2021. 3

[72] 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. 2