A Large-scale Study of Spatiotemporal Representation Learning with a New Benchmark on Action Recognition

Andong Deng* Taojiannan Yang* Chen Chen
Center for Research in Computer Vision
University of Central Florida, USA
andong.deng@ucf.edu, taoyang1122@knights.ucf.edu, chen.chen@crcv.ucf.edu

Abstract

*The goal of building a benchmark (suite of datasets) is to provide a unified protocol for fair evaluation and thus facilitate the evolution of a specific area. Nonetheless, we point out that existing protocols of action recognition could yield partial evaluations due to several limitations. To comprehensively probe the effectiveness of spatiotemporal representation learning, we introduce BEAR, a new **BE**nchmark on video **A**ction **R**ecognition. BEAR is a collection of 18 video datasets grouped into 5 categories (anomaly, gesture, daily, sports, and instructional), which covers a diverse set of real-world applications. With BEAR, we thoroughly evaluate 6 common spatiotemporal models pre-trained by both supervised and self-supervised learning. We also report transfer performance via standard finetuning, few-shot finetuning, and unsupervised domain adaptation. Our observation suggests that the current state-of-the-art cannot solidly guarantee high performance on datasets close to real-world applications, and we hope BEAR can serve as a fair and challenging evaluation benchmark to gain insights on building next-generation spatiotemporal learners. Our dataset, code, and models are released at: https://github.com/AndongDeng/BEAR*

1. Introduction

Learning good spatiotemporal representations [45, 70, 24, 60, 16, 76] is fundamental for video understanding tasks. In action recognition, a common evaluation protocol is to first evaluate the model performance on large-scale video datasets such as Kinetics-400 [30], then show its effectiveness of transfer learning to different downstream tasks [5, 17, 38, 2, 41, 72]. Many video datasets [56, 32, 30, 20, 9] have been introduced over the past few years to advance the field. However, there are several major limitations: (1) These datasets are similar in terms of domains and actions. Most of them only contain daily or sports actions because these categories are easy to collect from the web. Yet many important real-world applications, such as anomaly detection and industrial inspection, are rarely included. (2) Each of these datasets has its own characteristics (*e.g.* appearance-focused [30], motion-focused [20], fine-grained [52], egocentric [9]). Previous works usually conduct evaluations on a few datasets. However, without evaluating a suite of datasets, we cannot fully diagnose a model and make further improvements. (3) The held-out test set for these datasets either does not exist or is not commonly adopted. This will affect the transfer performance because models tuned on a test set using hyperparameter optimization or neural architecture search might achieve good performance but cannot transfer well due to overfitting.

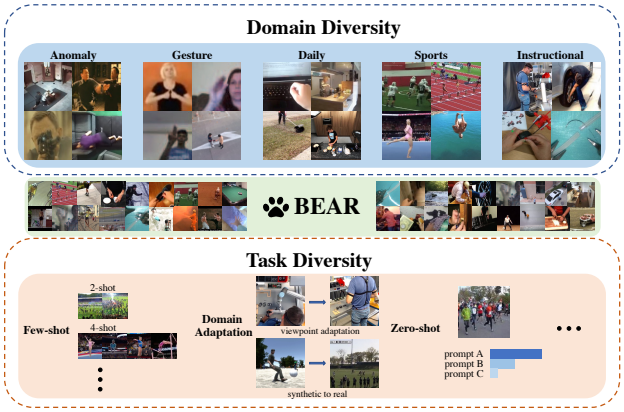

Figure 1: BEAR is a collection of 18 video action recognition datasets grouped into 5 categories (Anomaly, Gesture, Daily, Sports, and Instructional). It enables various evaluation settings, e.g., standard finetuning, few-shot finetuning, unsupervised domain adaptation, and zero-shot learning.

In light of this, we propose a unified and challenging BEnchmark on video Action Recognition, named BEAR, to better evaluate spatiotemporal representation learning. We define good representations as those that can achieve strong

*Equal Contribution.

transfer learning performance on diverse, unseen domains even with limited data. To this end, we build BEAR by collecting a suite of 18 video action recognition datasets grouped into 5 categories (Anomaly [57, 78], Gesture [42], Daily [53, 10], Sports [28], and Instructional [58]), which cover a diverse set of real applications. The datasets in BEAR are also diverse in video sources (*e.g*. YouTube, CCTV cameras, self-collected) and viewpoints (*e.g*. egocentric, 3rd person, drone, and surveillance). In addition, we split each dataset into train and test sets, *strictly keeping the test set held out during training in all of our experiments*. We will also provide an online evaluation server to enable fair comparisons.

With BEAR, one can probe spatiotemporal representation learning methods from a much more diverse perspective and answer many important questions. Does the good performance on commonly-used large-scale datasets translate to real applications? Do recent transformer-based models consistently outperform simple 2D models in different domains? How sensitive is the model to domain and viewpoint change? Could the model achieve good performance when downstream data is limited? In this work, we comprehensively investigate 6 representative video models pretrained by both supervised and self-supervised learning in various settings (*e.g*. full-shot, few-shot, domain adaptation). Our study quantifies existing intuition and uncovers several new insights: (1) Simple 2D video models can outperform recent transformer-based models when equipped with strong backbones. (2) The previous evaluation protocols are constrained to downstream datasets that resemble Kinetics-400. However, the high performance of these datasets does not necessarily transfer to other application domains. (3) Viewpoint shift has a dramatic impact on downstream task performance. Even the recent domain adaptation methods cannot address the problem to satisfactory. This suggests we may need to go beyond domain adaptation and shift attention to building more comprehensive pre-training datasets. (4) Self-supervised spatiotemporal representation learning still lags remarkably behind supervised learning. Even the SoTA VideoMAE [60] fails to outperform simple supervised models in diverse domains. Our goal is to provide a unified and challenging evaluation benchmark to evaluate spatiotemporal representation learning from various perspectives, which hopefully could guide future development in video understanding.

2. Related Work

Human action recognition is to distinguish the ongoing actions (or sometimes events) in a video. Different from image classification, video action recognition requires effective temporal modeling [64], awareness of the action hierarchies [52], and the interaction between the subjects and objects [20]. In early years, video models simply inherit the 2D convolution structures [55, 25] and process temporal information either by extending 2D convolutions into 3D [61, 5, 75] or including optical flow [54]. However, optical flow-based approaches suffer from costly flow precomputation, thus 2D CNNs with more sophisticated temporal modeling are designed [64, 82, 38, 73, 71]. For 3D CNNs, factorized architectures [46, 62, 74, 83] are introduced to improve the model efficiency and reduce overfitting. Recently, Transformer [63] continues to showcase its capability from language to image and also to video [2, 1, 80, 41, 77, 68]. Top performers on most video action recognition datasets are transformer-based. In this work, we fairly evaluate 6 popular video models belonging to 2D CNN, 3D CNN, and Transformer, respectively. With comparable backbones, we surprisingly reveal that 2D CNNs can sometimes outperform transformer models.

Spatiotemporal representation learning is advancing rapidly in the last few years, especially in a self-supervised manner. Self-supervised pre-training is appealing because it could learn visual knowledge from massive unlabeled data, which alleviates the annotation burden compared with its supervised counterpart. Most approaches design a pretext task to learn the intrinsic spatiotemporal feature within the video data, such as sorting the shuffled video sequence [33], next frame prediction [23], predicting the frame rate [14], contrastive learning [12, 18, 43, 31, 45], mask modeling [60, 16], etc. Despite their promising performance, a recent work [59] points out that video self-supervised pre-training is less robust than its supervised counterpart when the downstream setting varies. In this work, we also compare supervised pre-training with self-supervised ones in terms of both standard finetuning and few-shot finetuning on our benchmark.

Vision benchmark is often designed as a testbed, which consists of multiple datasets from different domains. Each benchmark might have its own motivation, but they share the same goal of providing a unified protocol for evaluation and thus facilitating the evolution of a specific area. Many well-established benchmarks have been proposed in different research areas [66, 79, 34, 36, 22]. However, there is no such comprehensive benchmark for video action recognition. Two works that are the closest to ours are VTAB [79] and SEVERE-benchmark [59]. VTAB contains 19 datasets that cover a broad spectrum of domains and semantics. All tasks are formulated as the image classification problem for the sake of a homogeneous task interface. Inspired by VTAB, we build the first comprehensive evaluation benchmark for video action recognition. BEAR includes 18 datasets across 5 domains towards real applications. It enables fair comparison and thorough investigation of existing video models, which allows us to address interesting open questions. SEVERE-benchmark [59] investigates how sensitive video self-supervised learning is to the current

Table 1: Statistics of the selected datasets used in our video benchmark. We collect 18 datasets covering 5 common data domains for comprehensive benchmarking. In the column of video viewpoint, "sur." means surveillance videos, and "dro." means drone videos.

Dataset	Domain	Label classes	Clip *num.*	Avg Length (sec.)	Training data per class (min, max)	Split ratio	Video source	Video viewpoint
XD-Violence [69]	Anomaly	5	4135	14.94	(36, 2046)	3.64:1	Movies, sports, CCTV, etc.	3rd, sur.
UCF Crime [57]	Anomaly	12	600	132.51	38	3.17:1	CCTV Camera	3rd, sur.
MUVIM [11]	Anomaly	2	1127	68.1	(296, 604)	3.96:1	Self-collected	3rd, sur.
WLASL100 [35]	Gesture	100	1375	1.23	(7, 20)	5.37:1	Sign language website	3rd
Jester [42]	Gesture	27	133349	3	(3216, 9592)	8.02:1	Self-collected	3rd
UAV Human [37]	Gesture	155	22476	5	(20, 114)	2:1	Self-collected	3rd, dro.
CharadesEgo [53]	Daily	157	42107	10.93	(26, 1120)	3.61:1	YouTube	1st
Toyota Smarthome [10]	Daily	31	14262	1.78	(23, 2312)	1.63:1	Self-collected	3rd, sur.
Mini-HACS [81]	Daily	200	10000	2	50	4:1	YouTube	1st, 3rd
MPII Cooking [50]	Daily	67	3748	153.04	(5, 217)	4.69:1	Self-collected	3rd
Mini-Sports1M [29]	Sports	487	24350	10	50	4:1	YouTube	3rd
FineGym99 [52]	Sports	99	20389	1.65	(33, 951)	2.24:1	Competition videos	3rd
MOD20 [44]	Sports	20	2324	7.4	(73, 107)	2.29:1	YouTube and self-collected	3rd, dro.
COIN [58]	Instructional	180	10426	37.01	(10, 63)	3.22:1	YouTube	1st, 3rd
MECCANO [48]	Instructional	61	7880	2.82	(2, 1157)	1.79:1	Self-collected	1st
INHARD [8]	Instructional	14	5303	1.36	(27, 955)	2.16:1	Self-collected	3rd
PETRAW [26]	Instructional	7	9727	2.16	(122, 1262)	1.5:1	Self-collected	1st
MISAW [27]	Instructional	20	1551	3.8	(1, 316)	2.38:1	Self-collected	1st

conventional benchmark in terms of domain, samples, actions, and tasks. Compared to SEVERE-benchmark [59], we study both supervised and self-supervised learning in more domains (anomaly, instructional), with more datasets (18 vs 8) and more settings (few-shot, zero-shot, and unsupervised domain adaptation).

3. BEAR

Despite new datasets being introduced every year, the most widely adopted benchmarks in the video action recognition community are Kinetics-400/600/700 [3, 4, 30], Something-something-v1/v2 [20], UCF-101 [56] and HMDB-51 [32]. However, these datasets share a high similarity in that they are mostly composed of daily and sports actions. Models that achieve good performance on these datasets may not generalize well to the challenging real-world scenarios due to dramatic domain shifts. For example, anomaly videos are often captured from surveillance cameras, which look quite different from daily videos due to viewpoint change. Ideally, a video model is expected to cope with diverse real-world applications.

To comprehensively evaluate the generalization capability of video models, we present BEAR, a new benchmark for human action recognition. As shown in Table 1, BEAR is a collection of 18 action recognition datasets, carefully designed towards *practical use*, *data diversity*, and *task diversity*. Compared to existing video action recognition datasets, BEAR has the following desirable properties.

Real Applications. Besides the common daily and sports categories, BEAR contains another three categories including anomaly activity, gesture, and instructional actions. These action categories have important real-world applications such as people fall detection (*e.g.* MUVIM [11]), sign language recognition (*e.g.* WLASL100 [35]), industrial inspection (*e.g.* MECCANO [48]), and surgical workflow recognition (*e.g.* PETRAW [26]).

Data Diversity. BEAR is not only diverse in application domains but also in the data source, video viewpoint, and video length. As shown in Table 1, BEAR contains videos from various sources such as movies, CCTV cameras, YouTube, and drone cameras. It also includes videos in the 1st and 3rd person views. In terms of video length, the average clip duration varies from the shortest (*e.g.* 1.23s in WLASL100 [35]) to the longest (*e.g.* 153.04s in MPII Cooking [50]). In addition, the training sample size per class varies across datasets, from the lowest (*e.g.* 1 for MISAW [27]) to the highest (*e.g.* 9592 for Jester [42]).

Few-shot Transfer. The standard finetuning protocol for transfer is to train a model on the whole training data, which is often more than thousands of videos. However, in many real applications, the annotated video data is scarce, *e.g.* anomaly recognition (rarely happens and is costly to label), medical operation (privacy concern), and industrial operation (need the expertise to label). To better evaluate a model's potential in real applications, we need to evaluate its effectiveness under few-shot learning. Hence in BEAR, besides the full datasets, we also split each dataset into 16-shot, 8-shot, 4-shot, and 2-shot versions. This allows researchers and practitioners to thoroughly evaluate a model's sensitivity to data scarcity.

Flexible Evaluation. Thanks to the data diversity in BEAR, researchers can easily evaluate video models under various settings. For example, full-shot and few-shot learning, domain adaptation from one dataset (or category) to another. Moreover, we also believe that new settings can be easily

derived based on our benchmark.

Fair Comparison. The held-out test set for most video action recognition datasets either does not exist or is not commonly adopted. This allows previous methods to conduct hyperparameter optimization or even neural architecture search directly on the test set. Test set tuning usually leads to good testing performance, but it may not translate to other datasets. *To promote fair comparison and generalization capability, we will hold the test sets and provide an evaluation server for future researchers and practitioners.*

Dataset Accessibility. We provide scripts to download and format all 18 datasets automatically. Our codebase is built upon MMAction2 [7], so researchers can easily integrate their new models by providing a model definition file without additional efforts to perform evaluations. Furthermore, the total number of video clips in BEAR is about 310K, which is comparable to Kinetics-400. *Therefore, the overall time cost is similar to training a model on Kinetics-400.*

4. Models

There has been a considerable amount of video models proposed to solve the human action recognition task. From the perspective of the basic building block, these models can be roughly classified into three categories: 2D CNNs, 3D CNNs, and transformer-based models. To investigate the efficacy of each model type, in this work, we select two representative works from each category: TSN [64] and TSM [38] for 2D CNNs, I3D [5] and 3D Non-local network [67] for 3D CNNs, TimeSformer [2] and VideoSwin [41] for transformer-based models. We would like to point out that for CNN-based models (TSN, TSM, I3D, and NL), we choose ConvNext-base [40] as the backbone because it has a similar model size and performance, as shown in Table 2, on ImageNet-1K compared to ViT-B and Swin-B, which is the backbone of TimeSformer and VideoSwin, respectively. This alleviates the impact from the backbone, thus presenting a more fair comparison among different video architectures. In this work, we finetune all the models based on both supervised and self-supervised pre-training on Kinetics-400, and the pre-training performance is shown in Table 3. The pre-training details can be found in Supplementary Sec.2. In the following sections, we will provide a comprehensive study w.r.t. transferring performance from multiple perspectives: standard finetuning, few-shot finetuning, unsupervised domain adaptation, and zero-shot evaluation.

5. Standard Finetuning

Finetuning models that are pre-trained on large-scale datasets have been a mainstream learning paradigm in deep learning, and performance on various downstream datasets can provide a more comprehensive evaluation with less bias. Thus, in BEAR, we regard standard finetuning as a basic evaluation method. Specifically, we finetune the pre-trained models on the 18 datasets to investigate: 1) the performance of different types of video models on different data domains; 2) the difference between supervised pre-training and self-supervised pre-training; 3) potential factors (*e.g.* domain shift, viewpoint shift, etc.) that have significant impacts on the performance of downstream tasks. We want to emphasize that during finetuning, we do not tune hyperparameters on the test set to avoid potential overfitting. All reported results are based on the evaluation of the last checkpoint. The Top-1 accuracy of each model is presented in Table 4. Besides the performance on each dataset, we also propose two *composite metrics* over the 18 datasets for evaluation. The first one is the macro-average accuracy which is the average of the accuracy on each dataset. The second one is micro-average accuracy, which calculates the average accuracy on the video level. Micro-average considers the size difference of the 18 datasets. We include the details of the complete finetuning results, and the previous best-reported performance, if any, for each dataset in Supplementary Sec.3.

Table 2: Comparison among ConvNeXt-base, ViT-base, and Swin-base. Params denote the parameters volume, and Top-1 acc means the top-1 accuracy in the ImageNet classification task.

Backbone	Params(M)	Top-1 acc(%)
ConvNeXt-base [40]	88.59	85.8
ViT-base [13]	86.57	81.8
Swin-base [39]	87.77	85.2

Table 3: The pre-training results of 6 models on Kinetics-400 in both supervised and self-supervised settings. The supervised results are based on the single-view test, and the self-supervised ones are based on KNN evaluation.

model	Supervised	SSL
TSN	77.6	43.1
TSM	76.4	43.2
I3D	74.2	51.3
NL	73.9	50.7
TimeSformer	75.8	50.3
VideoSwin	77.6	51.1

Model comparison. In previous studies, transformer-based video models [2, 41] have been demonstrated to be more effective than CNNs on several representative datasets. This conclusion leads the trend of model design toward more sophisticated transformers, which makes CNNs less appealing compared with the pre-transformer era. However, we argue that the current conclusion could be biased since the comparison between transformers and

Table 4: Finetuning results based on the supervised pre-trained and self-supervised pre-trained models as well as the X3D pre-trained models. Generally, from the two composite metrics (macro-average accuracy and micro-average accuracy), we can tell that TSM surprisingly outperforms other counterparts in both pre-training settings.

Dataset	Supervised pre-training						X3D	Self-supervised pre-training					
	TSN	TSM	I3D	NL	TimeSformer	VideoSwin		TSN	TSM	I3D	NL	TimeSformer	VideoSwin
XD-Violence	**85.54**	82.96	79.93	79.91	82.51	82.40	75.11	80.49	**81.73**	80.38	80.94	77.47	77.91
UCF-Crime	35.42	**42.36**	31.94	34.03	36.11	34.72	25.69	**37.50**	35.42	34.03	34.72	36.11	34.03
MUVIM	79.30	**100**	97.80	98.68	94.71	**100**	99.56	99.12	**100**	66.96	66.96	99.12	**100**
WLASL	29.63	43.98	49.07	**52.31**	37.96	45.37	44.91	27.01	27.78	29.17	**30.56**	25.56	28.24
Jester	86.31	**95.21**	92.99	93.49	93.42	94.27	92.24	83.22	**95.32**	87.23	93.89	90.33	90.18
UAV-Human	27.89	**38.84**	33.49	33.03	28.93	38.66	36.07	15.70	30.75	31.95	26.28	21.02	**35.12**
CharadesEGO	8.26	8.11	6.13	6.42	**8.58**	8.55	5.69	6.29	6.59	6.24	6.31	7.59	**7.65**
Toyota Smarthome	74.73	**82.22**	79.51	76.86	69.21	79.88	79.09	68.71	**81.34**	77.82	76.16	61.64	80.18
Mini-HACS	84.69	80.87	77.74	79.51	79.81	**84.94**	60.57	64.60	63.24	70.24	60.57	73.92	**75.58**
MPII Cooking	38.39	46.74	**48.71**	42.19	40.97	46.59	42.19	34.45	**50.08**	42.79	40.36	35.81	47.19
Mini-Sports1M	54.11	50.06	46.90	46.16	51.79	**55.34**	41.91	43.02	43.59	46.28	45.56	44.60	**47.60**
FineGym	63.73	**80.95**	72.00	71.21	63.92	65.02	68.49	54.62	**75.87**	69.62	68.79	47.60	58.94
MOD20	**98.30**	96.75	96.61	96.18	94.06	92.64	92.08	91.23	92.08	91.94	92.08	90.81	**92.36**
COIN	81.15	78.49	73.79	74.30	**82.99**	76.27	61.29	61.48	64.53	71.57	**72.78**	67.64	68.78
MECCANO	**41.06**	39.28	36.88	36.13	40.95	38.89	30.78	32.34	35.10	34.86	33.62	33.30	**37.80**
InHARD	84.39	**88.08**	82.06	86.31	85.16	87.60	84.86	75.63	**87.66**	82.54	80.81	71.28	80.10
PETRAW	94.30	95.72	94.84	94.54	94.30	**96.43**	95.46	93.18	**95.51**	95.02	94.38	85.56	91.46
MISAW	61.44	**75.16**	68.19	64.27	71.46	69.06	69.06	59.04	**73.64**	70.37	64.27	60.78	68.85
Macro Avg.	62.70	**68.10**	64.92	64.75	64.27	66.48	61.39	57.09	**63.35**	60.50	59.39	57.23	62.33
Micro Avg.	64.92	**70.82**	67.81	67.83	67.66	69.73	65.87	59.13	**68.11**	64.35	66.21	62.19	65.71

CNNs is obviously unfair. Basically, it is a widely accepted notion that the selection of different backbones can inherently yield significant differences, let alone the overall model design. To this end, as aforementioned, we carefully select ConvNeXt [40] as the CNN backbone, which is comparable with ViT [13] and Swin Transformer [39] w.r.t. both model size and ImageNet classification performance. We believe such a fair comparison could lead to more convincing and compelling conclusions. As shown in Table 4, we notice that there is no absolute winner among all the models, but surprisingly, 2D CNNs perform better on most datasets, especially TSM, which outperforms other models in 8 out of 18 datasets. This indicates that 2D video models are still competitive with transformers when equipped with strong backbones. Likewise, the two composite metrics also provide evidence that TSM outperforms other models, and transformer-based models do not exhibit clear advantages over CNN-based models.

Inspecting further, we can see that VideoSwin excels in mini-HACS and mini-Sports1M. However, as aforementioned, these datasets, along with other popular datasets such as UCF-101 and HMDB-51, share high similarities with Kinetics-400 in terms of actions and viewpoints. Thus the performance on these datasets may not fully reflect the effectiveness of the evaluated model. Indeed, as shown in Table 4, VideoSwin is only comparable or inferior to TSM in the other three categories (*i.e.*, anomaly, gesture, and instructional). This demonstrates that the impressive performance on Kinetics-400 and other similar datasets may not be consistent with downstream tasks with vastly different actions. To fully probe the effectiveness of a video model, we need to evaluate it on datasets with different distributions. Besides, we also consider the NAS-based X3D [15], which achieves good performance on Kinetics-400, to reveal the overfitting problem of tuning on the test set.

- *Despite the emergence of recent transformers, 2D video models can still be promising alternatives for action recognition if equipped with powerful backbones.*
- *Previous evaluation protocols have been limited to target datasets similar to Kinetics-400, which could potentially result in biased evaluations. However, BEAR could address this issue by including target data from five distinct domains, ensuring a more comprehensive and unbiased assessment of model performance.*

Impact of viewpoint change We also observe something interesting in terms of the data distribution. Several datasets such as UCF-Crime, UAV-Human, CharadesEGO, MPII-Cooking, and MECCANO exhibit notably low performance. Upon closer inspection of Table 1, it is evident that these datasets involve significant viewpoint changes from Kinetics-400. For instance, UCF-Crime is collected from CCTV footage, UAV-Human contains drone-view videos, CharadesEGO only contains 1st person-view videos, and MECCANO is also egocentric. This indicates that the viewpoint change in downstream tasks could dramatically damage the model performance. Therefore, leveraging pre-training datasets with rich egocentric visual knowledge, such as EGO4D [21], may offer a suitable alternative to Kinetics-400 for finetuning on egocentric data. Besides, in Sec. 6 and Sec. 7, we will further discuss the challenge caused by the viewpoint change in the target domain.

- *Prior evaluation protocols, limited in the scope of tar-*

get data, fail to capture the impact of domain gap, particularly in regard to the viewpoint, on transfer performance. However, we have identified that such a distribution shift can significantly degrade the quality of spatiotemporal representation, which further undermines the transfer performance. Hence, we recommend that future studies should include pre-training datasets beyond Kinetics-400 to provide more robust representations to improve transferability.

Self-supervised vs. supervised pre-training As can be seen from Table 4, it is notable that the overall finetuning performance of the self-supervised pre-training is less competitive than its supervised counterpart even for TSM. The most pronounced accuracy drop can be found in WLASL and FineGym. The performance of 3D Nonlocal network on WLASL drops from 52.31% to 30.56% and the performance of TimeSformer also decreases more than 15%. To reveal the potential reason behind this, we further scrutinize the data distribution gap between the selected 18 target datasets and Kinetics-400. We observe different types of domain shifts, such as UAV-Human containing only drone-view data and the egocentric MECCANO which differs significantly from Kinetics-400. We conclude that self-supervised pre-training is more susceptible to domain shifts between Kinetics-400 and the target datasets than supervised pre-training. In Sec. 6, we take a step forward on this topic by investigating few-shot settings, which are more likely to occur in real-world scenarios.

- *Self-supervised finetuning generally cannot outperform its supervised counterpart and TSM consistently performs well under the self-supervised setting.*

6. Few-shot Learning

Compared with standard finetuning where abundant annotations can be utilized, few-shot learning is of more practical significance since annotating massive amounts of videos is notoriously expensive. To extend the investigation mentioned in Section 5, we thoroughly investigate the capability of the selected 6 models on BEAR under a few-shot setting given both supervised and self-supervised pre-trained weights. Specifically, we consider (2,4,8,16)-shot settings, and for each setting, we randomly generate 3 splits and report the mean and standard deviation. Due to space constraints, we only select TSM, 3D NonLocal, and Video Swin to represent each model type for illustration as they perform generally better. Complete few-shot results and the training details are in Supplementary Sec.4.

Model comparison. The rankings of the six models in few-shot finetuning exhibit distinct variations compared to the standard finetuning. In contrast to the dominance of TSM in standard finetuning across both pre-training settings, the most effective models differ significantly across datasets in few-shot finetuning. Figure 2 demonstrates that TSM no longer clearly outperforms other models in most datasets, and the two composite metrics (which are presented in the Supplementary due to space limitations) support this conclusion. Specifically, TSM and TimeSformer exhibit similar performance in supervised pre-training, whereas I3D and VideoSwin perform better in self-supervised learning. These findings further reveal the limitations of previous simple evaluation protocols, which may not provide a fair assessment of video models. These results also confirm the necessity of BEAR, which emphasizes the importance of diverse downstream datasets and various settings for unbiased evaluation.

- *The ranking relations between models could exhibit differently between standard and few-shot finetuning even within the same datasets. This finding further emphasizes the importance of our proposed BEAR benchmark, which advocates for a comprehensive evaluation approach that considers both dataset diversity and finetuning settings.*

Impact of viewpoint change As in standard finetuning, viewpoint change also has a severe impact when it comes to few-shot learning. Comparing the results in Figure 2 with those in Table 4, we can see that the few-shot learning performance decreases drastically in general, especially in datasets that have less in common with Kinetics-400, such as UAV-Human, which is constructed by videos captured from unmanned aerial vehicles, FineGym, which contains fine-grained gym-related videos, and PETRAW and MISAW, which are simulated medical operations in the 1st person view. Conversely, in datasets that are more similar to Kinetics-400, these performance gaps are notably reduced. For example, even the 2-shot performance on Mini-HACS and MOD20 can reach approximately 60% and 85%, and the models achieve satisfying performance on the 16-shot setting on COIN. In previous works, the homogeneity of the pre-training and downstream data hindered the timely identification of such phenomena in few-shot learning. Our investigation highlights the challenge of few-shot learning and underscores the importance of bridging the gap (as aforementioned, introducing extra data, such as Ego4D) between pre-training and the target data.

Moreover, in the few-shot setting, self-supervised pre-training is more susceptible to viewpoint change. In challenging datasets such as UAV-Human and WLASL, few-shot learning can hardly obtain satisfying results based on self-supervised pre-trained weights, while in the 16-shot setting, supervised pre-training could provide comparable performance compared with standard finetuning. Similarly, in MOD20, the performance experiences a sharp decline in few-shot settings with self-supervised pre-training, while supervised pre-trained TSN and TSM can achieve accuracy

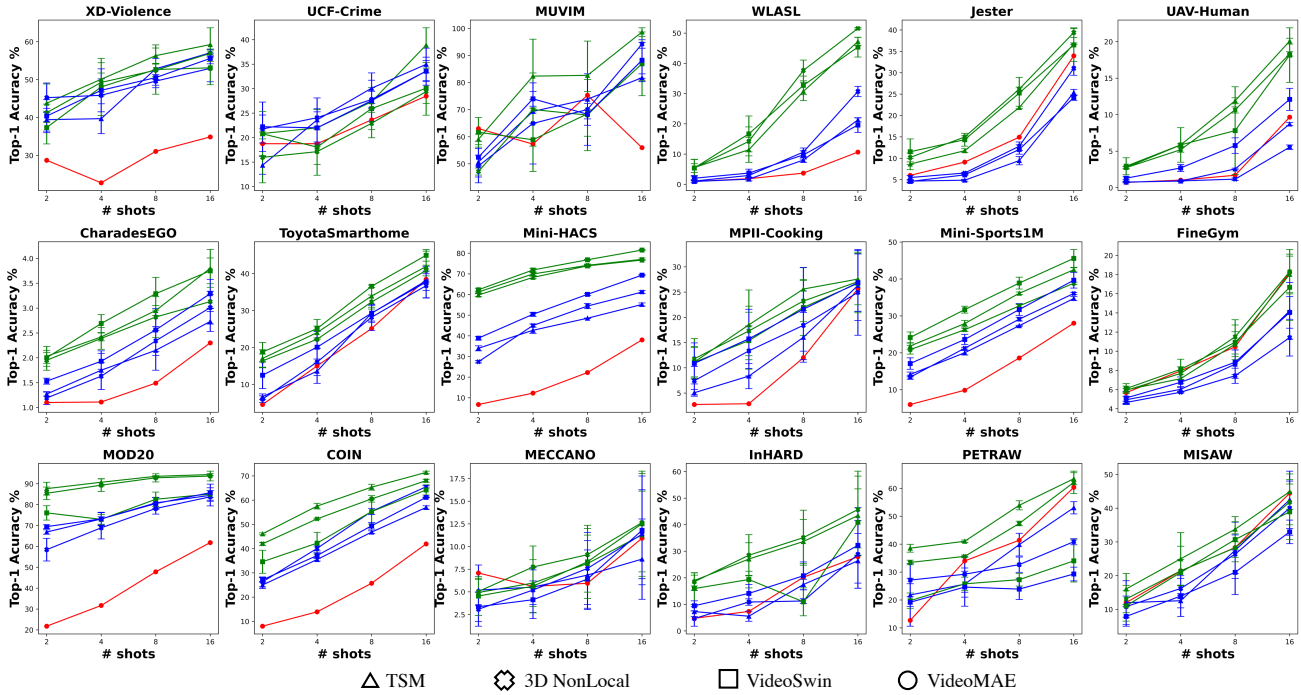

Figure 2: Results of few-shot learning based on supervised and self-supervised pre-training. The green curves represent supervised pre-training and the blue curves represent ρMoCo self-supervised pre-training. We illustrate the results of TSM, 3D NonLocal, and VideoSwin for both pre-training methods. Additionally, we add the SOTA self-supervised pre-training method VideoMAE, represented by the red curves, for comparison. It could be obvious that even the VideoMAE could lag a lot behind in the few-shot setting.

exceeding 90% in the 16-shot.

- *Few-shot finetuning remains a significant challenge in real-world scenarios. The performance drops dramatically compared to standard finetuning especially when there is a large domain gap between pre-training and target data. However, when downstream datasets are similar to source data, the performance drop could be mitigated.*
- *In few-shot learning, self-supervised pre-training is more vulnerable to viewpoint shift, while supervised pre-trained models can achieve favorable performance compared with standard finetuning on the 16-shot setting.*

Self-supervised vs. supervised pre-training. Comparing the blue curves to the green curves in Figure 2, we can see that self-supervised pre-training is generally less effective than supervised pre-training, which is consistent with the conclusion in Sec. 5. The performance gaps are pronounced in gesture datasets and are less significant in Mini-Sports1M, ToyotaSmarthome, etc. The performance gap is also different across different models. The largest gap appears in TSN and TimeSformer (the complete results are provided in Supplementary Table 6-11). One reason for the poor performance of self-supervised learning may be the limitation of ρMoCo. Therefore, to consolidate our conclusion, we further consider VideoMAE [60], which is the SoTA self-supervised method and has demonstrated even better performance than supervised models on multiple datasets. Here, we use the officially released VideoMAE ViT-B model, which achieves 81.5% Top-1 accuracy on Kinetics-400. However, comparing the results with our 6 supervised pre-trained models in Figure 2 (red vs. green curves), we show that VideoMAE could only be comparable with the best supervised pre-trained models in less than half of the datasets.

- *Supervised pre-training shows consistent advantages over self-supervised ones in few-shot finetuning. Even the SoTA VideoMAE can hardly outperform simple supervised pre-trained models in diverse domains.*

7. Unsupervised Domain Adaptation

In real-world scenarios, it is possible to transfer knowledge from similar datasets which are well-annotated to others with only limited labels. For instance, there are a lot of existing datasets that include samples of the same categories in the corresponding real-world tasks and thus can be used to facilitate model training. Nonetheless, due to the domain

Table 5: The unsupervised domain adaptation accuracy on our UDA datasets: Toyota Smarthome-MPII-Cooking (T: Toyota Smarthome, M: MPII-Cooking), Mini-Sports1M-MOD20 (MS: Mini-Sports1M, MOD: MOD20), UCF-Crime-XD-Violence (U: UCF-Crime, X: XD-Violence), PHAV-Mini-Sports1M (P: PHAV, MS: Mini-Sports1M), Jester, InHARD (I: InHARD, T: Top, L: Left, R: Right).

Settings	Inter-dataset							Intra-dataset						
Dataset	T→M	M→T	MS→MOD	MOD→MS	U→X	X→U	P→MS	Jester	IT→IL	IT→IR	IL→IR	IL→IT	IR→IT	IR→IL
Source only	5.32	7.36	18.25	12.76	54.20	33.33	61.45	68.73	4.18	30.39	19.01	22.65	24.14	12.42
TA^{3}N [6]	11.17	15.38	23.77	19.15	59.91	44.44	65.79	71.44	5.78	41.83	27.91	28.08	35.66	14.68
CoMix [51]	12.63	15.32	24.48	21.56	60.17	47.22	64.83	75.86	6.32	39.79	30.45	31.74	32.94	14.83
Supervised target	70.21	65.13	34.08	35.52	75.06	63.89	94.40	97.61	26.00	83.55	83.55	85.52	85.52	26.00

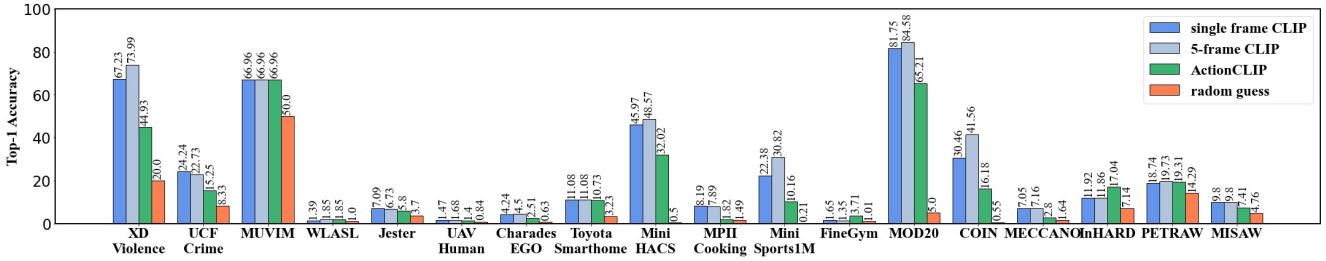

Figure 3: Results of zero-shot evaluation. For most datasets in our benchmark, CLIP-based models still cannot provide reasonable results, especially for those challenging datasets with severe viewpoint shifts and fine-grained datasets.

gap, models directly trained on one dataset cannot be well generalized on the target data. In such case, unsupervised domain adaptation (UDA) [19] can largely alleviate this distribution shift issue by learning the domain-invariant feature when labeled source data is available, learning representations that would promote the performance on the target domain. In BEAR, we construct several dataset pairs for UDA based on two different paradigms: inter-dataset adaptation and intra-dataset adaptation. Given that one of the features of our benchmark is that we collect several datasets with obvious viewpoint shifts, we also focus on this point when we build our UDA datasets. The details of the dataset statistics can be found in Supplementary Sec.5. We provide two common baseline results: 'Source only' and 'Supervised target'. The former directly evaluates the model trained on the source training set with the target test set, and the latter is the supervised learning performance on the target domain. Besides, we also evaluate two recent UDA algorithms on our benchmark: TA^{3}N [6] and CoMix [51].

Inter-dataset adaptation. Inter-dataset is constructed based on two different datasets that have different distributions, especially viewpoint change, but share common categories. Toyota Smarthome contains videos captured from 7 different cameras deployed in an apartment, while MPII-Cooking consists of videos from a down-view camera. Specifically, we select 6 new categories, which contains original action classes in Toyota Smarthome and MPII-Cooking, for the new Toyota Smarthome-MPII-Cooking dataset. The number of videos is 5,233 and 943 for Toyota Smarthome and MPII-Cooking, respectively. Similarly, for Mini-Sports1M and MOD20, we select 15 categories to build the new dataset. In contrast to Toyota Smarthome-MPII-Cooking, the data distribution in Mini-Sports1M-MOD20 is much more balanced. There are 1,650 videos for Mini-Sports1M and 1,767 for MOD20. We also consider the anomaly detection dataset. Basically, there are three shared action categories in UCF-Crime and XD-Violence: *abuse*, *fighting*, and *shooting*. The domain shift in this dataset is also conspicuous: all the videos in UCF-Crime are from surveillance footage, where the target objects in video frames can only be in a small region, while most videos in XD-Violence are collected from action movies, which could record an action with abundant details. To provide a dataset for synthetic-to-real transfer, which is of great significance in real-world scenarios, we also include the simulated dataset PHAV [49] to construct PHAV-Mini-Sports1M dataset. We combine 15 classes from Mini-Sports1M into 6 categories (*playing soccer*, *playing golf*, *playing baseball*, *shooting gun*, *shooting archery* and *running*) existing in PHAV to build the paired dataset.

Intra-dataset adaptation. Intra-dataset, on the contrary, is built within one dataset that records the same actions differently. We include Jester(S-T), which is initially introduced by [51], in BEAR since it has been a well-established dataset for domain adaptation. Each identical action in Jester with a contrary direction is merged into one category. We also construct a three-view dataset based on InHARD. Basically, each original frame in InHARD con-

tains three distinguished views (i.e., top, left, and right). We simply split the frames according to the view and construct three sub-datasets as InHARD-Top, InHARD-Left, and InHARD-Right. We keep the category the same as the original dataset.

Challenging viewpoint adaptation. As shown in Table 5, domain adaptation can be obviously challenging, especially in viewpoint change cases. For instance, ToyotaSmarthome and MPII-Cooking share similar attributes w.r.t. their actions, since they both record kitchen events. However, videos in ToyotaSmarthome are recorded via different cameras in the living room, while videos in MPII-Cooking are recorded by a down-view camera. The performance between these two datasets is far lower than the 'supervised target'. Similar observations can also be obtained in InHARD. Although the adaptation is conducted within the dataset, recent methods still fail to perform well when adapting from one viewpoint to another. However, the gap between supervised target and UDA methods is much smaller in other UDA datasets where the viewpoint change is smaller. These results, along with the observations in Secs. 5 and 6, reveal that viewpoint change has a critical impact on transfer performance, which is hard to mitigate even with recent UDA algorithms.

8. Zero-shot Learning

Direct finetuning on annotated datasets is a commonly adopted paradigm for action recognition, but the recent success of vision-language models, which leverage the rich correspondence between natural language and visual content, has provided a new learning paradigm for vision tasks in a zero-shot setting, which is severely required in applications without labeled data. Therefore, we also provide the zero-shot evaluation on BEAR using the recent CLIP-based [47, 65] models.

Basically, we provide two different settings for frame-level CLIP evaluations, *i.e.* single-frame, which follows the settings in [47] and 5-frame, where we sample 5 frames from the input video and fuse the model output of each frame. Similarly, we also construct multiple templates for each dataset to obtain ensemble textual embeddings. Considering the inconsistent label domains for the selected datasets, we provide different templates given their distinct attributes of both data and labels. For instance, UCF-Crime [57] is mostly constituted of surveillance videos in a crime scene; thus, a sentence like *'a photo from a surveillance camera showing a criminal doing {} in a crime scene.'* is utilized as a part of the prompts. Additionally, we evaluate all the datasets via ActionCLIP [65], which is pre-trained on Kinetics-400 based on video and label-text correlation, to unmask the difference of zero-shot performance between image-based models and video-based models.

As illustrated in Figure 3, different from its versatility in the image domain, most of the zero-shot results based on CLIP are still far lower than those of supervised learning. For example, WLASL shows poor correlations between frames and the corresponding labels, which can be partly explained by the large visual gap between the visual information of sign languages and the label itself. Surprisingly, for most datasets, ActionCLIP, which leverages more frames, performs even worse than CLIP. Part of the reason could be that ActionCLIP finetunes CLIP on Kinetics-400, which leads to catastrophic forgetting and overfitting. However, for some datasets, zero-shot learning could outperform few-shot learning, such as XD-Violence and MOD20, which even approaches supervised learning. This may be partly because the high vision-text correlation existed in these datasets, and this also demonstrates the potential of language supervision in action recognition.

9. Conclusion and Discussion

In this work, we introduce a new action recognition benchmark 🐾BEAR to address several limitations in existing video benchmarks. Aiming at benefiting both academic and industrial applications, we carefully select 18 datasets covering 5 distinct data domains. Such a wide scope could provide comprehensive assessment protocols for any video model, filling the gap in the current video action recognition benchmark that only a small number of target datasets are considered. It helps prevent models from overfitting on a specific dataset which could result in biased model evaluation. Moreover, to achieve a fair comparison, we held out test data for every dataset and avoid using it for parameter selection during training, and the evaluation is based on the last checkpoint. Meanwhile, in this work, we also pay attention to the capabilities of 2D CNNs, 3D CNNs, and transformers. Importantly, we carefully select comparative backbones for them to avoid erroneous comparisons.

Based on our extensive experiments, we have several interesting and instructive observations: 1) 2D video models are competitive with SoTA transformer models when equipped with strong backbones. 2) Previous evaluation protocols on a few similar datasets can yield biased evaluation. 3) Domain shift (especially the viewpoint shift) has a large impact on transfer learning, and the performance gap could be much more remarkable in the few-shot setting. 4) Self-supervised learning still largely falls behind supervised learning, and even the SoTA VideoMAE cannot outperform supervised models on diverse downstream datasets. Moreover, we also point out that in order to learn robust spatiotemporal representations, constructing new pre-training datasets containing videos from diverse domains could benefit the target performance on a wide range of datasets. Due to space limits, we only consider evaluation datasets and leave the art of training data construction to future work.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 2

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

[3] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 3

[4] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 3

[5] 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

[6] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6321–6330, 2019. 8

[7] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 4

[8] Mejdi Dallel, Vincent Havard, David Baudry, and Xavier Savatier. Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2020. 3

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1

[10] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 833–842, 2019. 2, 3

[11] Stefan Denkovski, Shehroz S Khan, Brandon Malamis, Sae Young Moon, Bing Ye, and Alex Mihailidis. Multi visual modality fall detection dataset. *IEEE Access*, 2022. 3

[12] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, and Jue Wang. Motion-aware self-supervised video representation learning via foreground-background merging. *arXiv preprint arXiv:2109.15130*, 2021. 2

[13] 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. 4, 5

[14] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 919–929, 2020. 2

[15] 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. 5

[16] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 1, 2

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[18] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021. 2

[19] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 8

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 1, 2, 3

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 5

[22] Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. Grit: General robust image task benchmark. *arXiv preprint arXiv:2204.13653*, 2022. 2

[23] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[24] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 1

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[26] Arnaud Huaulmé, Kanako Harada, Quang-Minh Nguyen, Bogyu Park, Seungbum Hong, Min-Kook Choi, Michael Peven, Yunshuang Li, Yonghao Long, Qi Dou, et al. Peg transfer workflow recognition challenge report: Does multi-modal data improve recognition? *arXiv preprint arXiv:2202.05821*, 2022. 3

[27] Arnaud Huaulmé, Duygu Sarikaya, Kévin Le Mut, Fabien Despinoy, Yonghao Long, Qi Dou, Chin-Boon Chng, Wenjun Lin, Satoshi Kondo, Laura Bravo-Sánchez, et al. Micro-surgical anastomose workflow recognition challenge report. *Computer Methods and Programs in Biomedicine*, 212:106452, 2021. 3

[28] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[29] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 3

[30] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 3

[31] Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3195–3204, 2021. 2

[32] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563, 2011. 1, 3

[33] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 667–676, 2017. 2

[34] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *arXiv preprint arXiv:2204.08790*, 2022. 2

[35] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020. 3

[36] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 2

[37] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16266–16275, 2021. 3

[38] 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. 1, 2, 4

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4, 5

[40] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 4, 5

[41] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 1, 2, 4

[42] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 3

[43] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 2

[44] Asanka G Perera, Yee Wei Law, Titilayo T Ogunwa, and Javaan Chahl. A multiviewpoint outdoor dataset for human action recognition. *IEEE Transactions on Human-Machine Systems*, 50(5):405–413, 2020. 3

[45] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 1, 2

[46] 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

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 9

[48] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021. 3

[49] Cesar Roberto de Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel Lopez. Procedural generation of videos to train deep action recognition networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4757–4767, 2017. 8

[50] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE conference on*

computer vision and pattern recognition, pages 1194–1201. IEEE, 2012. 3

[51] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *Advances in Neural Information Processing Systems*, 34:23386–23400, 2021. 8

[52] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625, 2020. 1, 2, 3

[53] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 2, 3

[54] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 2

[55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[56] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 3

[57] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 2, 3, 9

[58] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 2, 3

[59] Fida Mohammad Thoker, Hazel Doughty, Piyush Bagad, and Cees Snoek. How severe is benchmark-sensitivity in video self-supervised learning? 2022. 2, 3

[60] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 1, 2, 7

[61] 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

[62] 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

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[64] 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, 4

[65] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 9

[66] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7289–7298, 2019. 2

[67] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 4

[68] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2

[69] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pages 322–339. Springer, 2020. 3

[70] Wenhao Wu, Dongliang He, Tianwei Lin, Fu Li, Chuang Gan, and Errui Ding. Mvfnet: Multi-view fusion network for efficient video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2943–2951, 2021. 1

[71] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *ICCV*, 2019. 2

[72] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *AAAI*, 2023. 1

[73] Boyang Xia, Zhihao Wang, Wenhao Wu, Haoran Wang, and Jungong Han. Temporal saliency query network for efficient video recognition. In *ECCV*, pages 741–759, 2022. 2

[74] 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

[75] Taojiannan Yang, Sijie Zhu, Matias Mendieta, Pu Wang, Ravikumar Balakrishnan, Minwoo Lee, Tao Han, Mubarak Shah, and Chen Chen. Mutualnet: Adaptive convnet via mutual learning from different model configurations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):811–827, 2021. 2

[76] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. AIM: Adapting image models for efficient video action recognition. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[77] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization

and question answering. *arXiv preprint arXiv:2305.06988*, 2023. 2

[78] Shoubin Yu, Zhongyin Zhao, Haoshu Fang, Andong Deng, Haisheng Su, Dongliang Wang, Weihao Gan, Cewu Lu, and Wei Wu. Regularity learning via explicit distribution modeling for skeletal video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 2

[79] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. The visual task adaptation benchmark. 2019. 2

[80] 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

[81] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019. 3

[82] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann. Hidden two-stream convolutional networks for action recognition. In *Asian conference on computer vision*, pages 363–378. Springer, 2018. 2

[83] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018. 2