How Transferable are Video Representations Based on Synthetic Data?

by

Yo-whan Kim

S.B. in Computer Science and Engineering, Massachusetts Institute of Technology (2022)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author . . . .
Department of Electrical Engineering and Computer Science
May 6, 2022

Certified by . . . .
Aude Oliva
Director of Strategic Industry Engagement
Thesis Supervisor

Certified by . . . .
Rogerio Schmidt Feris
Principal Scientist and Manager
Project Supervisor

Accepted by . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

How Transferable are Video Representations Based on Synthetic Data?

by

Yo-whan Kim

Submitted to the Department of Electrical Engineering and Computer Science
on May 6, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Action recognition has improved dramatically with massive-scale video datasets. Yet, these datasets are accompanied with issues related to curation cost, privacy, ethics, bias, and copyright. Compared to that, only minor efforts have been devoted toward exploring the potential of synthetic video data. In this work, as a stepping stone towards addressing these shortcomings, we study the transferability of video representations learned solely from synthetically-generated video clips, instead of real data. We propose a novel benchmark for action recognition, in which a model is pre-trained on synthetic videos rendered by various graphics simulators, and then transferred to a set of downstream action recognition datasets, containing different categories than the synthetic data. Our extensive analysis on this benchmark reveals that the simulation to real gap is closed for datasets with low object and scene bias, where models pre-trained with synthetic data even outperform their real data counterparts. We posit that the gap between real and synthetic action representations can be attributed to contextual bias and static objects related to the action, instead of the temporal dynamics of the action itself.

Thesis Supervisor: Aude Oliva
Title: Director of Strategic Industry Engagement

Thesis Supervisor: Rogerio Schmidt Feris
Title: Principal Scientist and Manager

Acknowledgments

Foremost, I would like to thank my two supervisors, Rogerio Feris and Aude Oliva, for their endless insights and guidance throughout my program, and for giving me the opportunity to learn from the best. I would not have been able to carry out this research successfully without their invaluable advice.

I would like to thank SouYoung Jin for her overwhelming support and for providing me with all the necessary resources to carry out this research. I would also like to thank Rameswar Panda for thoughtful discussions and assistance, as well as for facilitating the model training process.

I would like to thank Richard Chen, Hilde Kuehne, and Leonid Karlinsky for their amazing suggestions, tips, and analysis throughout the research process.

I would like to thank the MIT Computational Perception & Cognition Lab, and MIT-IBM Watson AI Lab for sponsoring my two MEng semesters.

Finally, I would like to express my gratitude to my family and friends who stood by me through the ups and downs of my MEng journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Large-scale pre-training using massive datasets, containing hundreds of thousands or even millions of video clips, have brought significant progress in action recognition [24, 33, 1, 14]. High capacity models trained on such large datasets have shown remarkable generalization performance to downstream tasks where training data is limited [58, 14].

While this progress is exciting, these large-scale video datasets also have shortcomings. Collecting and annotating videos is expensive, tedious, and time-consuming. As a result, methods that learn feature representations from unlabeled videos, including self-supervised [54], weakly-supervised [14], and semi-supervised approaches [46], have received significant attention in recent years. However, these works do not address other important ethical, legal, and technical issues related to processing real-world data, as described below:

- **Privacy concerns**. Video samples may include human interactions and activities, but often, the individuals' sensitive information, such as faces, license plates, or location indicators, may be captured along.

- **Proprietary issues**. Massive-scale datasets containing millions or billions of images and videos, such as IG65M [14] and JFT3B [59], are not publicly available, preventing the large community to reproduce results, which hinders research progress.

- **Ethical issues and bias**. Ethical issues related to skin tone and gender [5], as well as unwanted contextual bias are difficult to control in existing large-scale datasets. Consequently, state-of-the-art models may fail to predict actions such as a person dancing in a mall [8], or a woman snowboarding [17].

- **Data protection and copyright issues**. Data collected without consent, which is common for existing massive-scale datasets, may violate copyright as well as data protection laws such as the General Data Protection Regulation (GDPR).

A promising way to address these issues is using computer-generated synthetic videos for pre-training. By leveraging 3D models of humans and scenes, an arbitrary number of videos can be generated by varying simulation parameters such as lighting, texture, and background, while enabling the control of sensitive attributes of humans, such as gender and race. This approach of training with synthetic data has a long history in computer vision [12, 29, 34]. Recent efforts on action recognition [19, 10, 51] have used completely synthetic datasets or synthetic/real hybrid datasets to train deep neural network models. However, these works rely on domain adaptation techniques that assume *the same label set* for both synthetic data and real data. This might not always be feasible as each action class requires motion capture or simulation capacities. To the best of our knowledge, no previous work has studied the transferability of action representations based on synthetic data to diverse downstream tasks, where the synthetic and real domains have disjoint label sets.

In this work, we introduce a benchmark that addresses this important problem. As shown in Figure 1-1, our pre-training dataset consists solely of synthetic video clips. We used various graphics simulators and synthetic videos [19, 10, 51] to create a dataset with 150 action categories, where each category has 1,000 samples. We have six downstream datasets: UCF101 [47], HMDB51 [26], Something-Something V2 [15], Diving48 [28], Ikea Furniture Assembly (IkeaFA) [48], and UAV-Human [27]. Both UCF101 and HMDB51 datasets are based on YouTube videos depicting a broad variety of actions. As a result, they exhibit a high object and scene bias [28], i.e.,

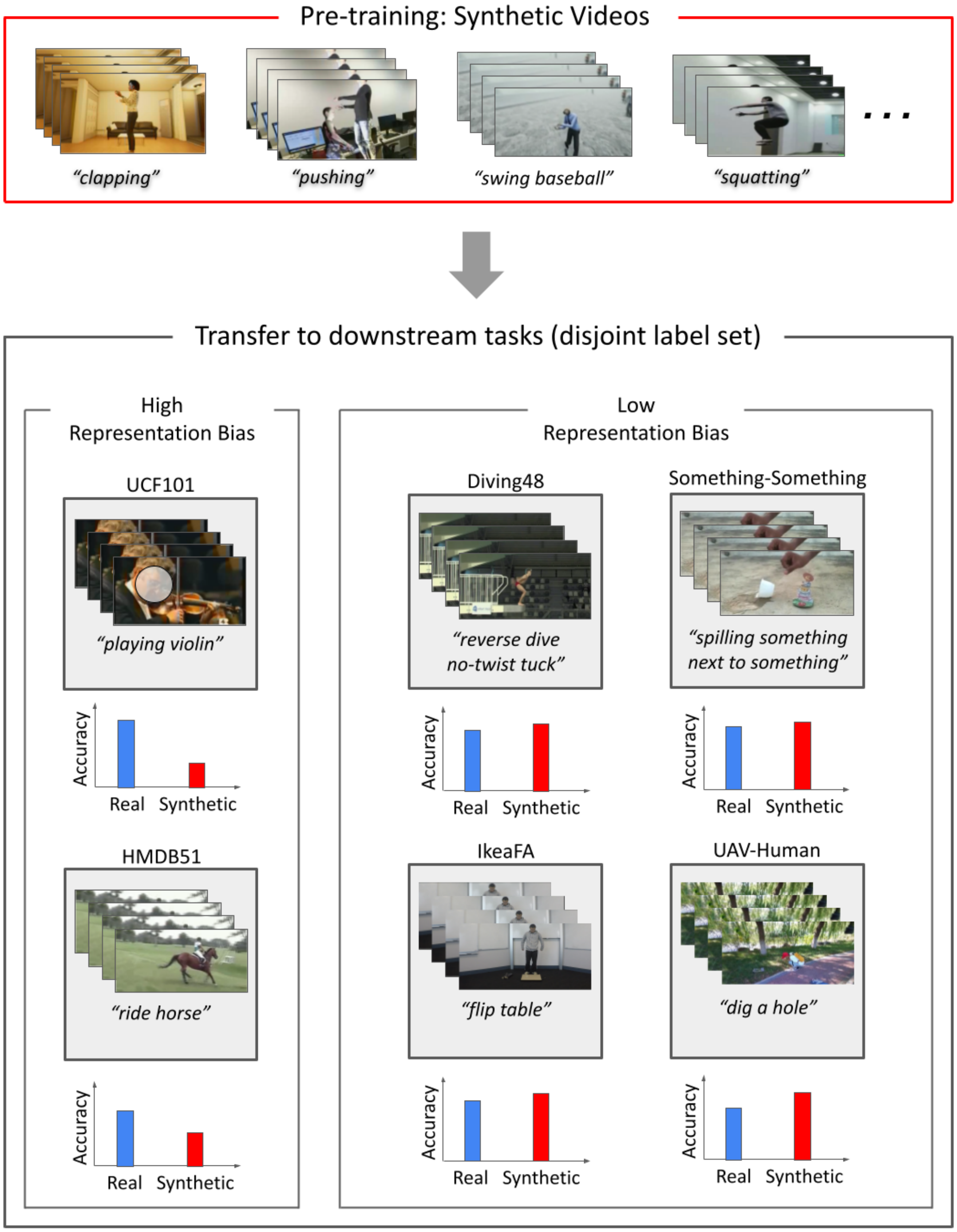


Figure 1-1: Schematic representation of benchmark pipeline. We introduce a novel action recognition benchmark to pre-train a model on synthetic videos and transfer learned knowledge to downstream tasks with disjoint label sets. We observe that the models pre-trained on synthetic videos even beat those pre-trained on real videos when the downstream datasets have low object and scene bias.

several actions can be recognized by just looking at static objects or the background, as opposed to the action itself. For example, the action “playing violin” could likely be recognized by detecting the object violin instead of understanding the temporal dynamics of the action. On the other hand, the remaining four datasets have low object and scene bias, as understanding temporal dynamics is needed to correctly recognize actions in these datasets.

Based on this setting, we conducted an extensive analysis on the transferability of pre-trained video models based on synthetic data, including the effect of linear probing and fine-tuning, number of classes, number of samples per class, and their relative performance with respect to pre-trained ImageNet models, which are used to measure the object and scene representation bias of each dataset. We solidify our findings by replicating the experiments with models of various capacities and complexity, and further perform rigorous hyperparameter sweeping on the downstream tasks.

We show that the transferability gap between synthetic and real action recognition models is directly related to the object and scene bias of the datasets. Models pre-trained on Kinetics clearly outperform Synthetic pre-trained models on datasets with high bias (UCF101, HMDB51). The gap is closed for datasets with low bias (Something-Something, Diving48, IkeaFA, UAV-Human), where Synthetic pre-trained models achieve similar or better accuracy than their real counterparts.

In summary, the main contributions of this work are as follows:

1. We propose a novel benchmark for studying the transferability of synthetic video representations for action recognition. To the best of our knowledge, no previous work has investigated this problem before. This is a promising direction to mitigate ethical and legal issues with existing large-scale datasets of real images.

2. We show that the simulation to real gap is simply closed for datasets with low object and scene bias, but still exists for datasets with high bias. This result suggests that the gap between real and synthetic action representations exists largely due to contextual bias and static objects related to the action, instead of the temporal dynamics of the action itself.

Chapter 2

Related Work

2.1 Action Recognition Benchmarks

Video datasets have rapidly evolved from small-scale benchmarks such as KTH [43] and Weizmann [4], with a few thousand video clips, to medium-scale datasets such as UCF101 [47] and HMDB51 [26], and recently to large-scale datasets containing hundreds of thousands or millions of annotated videos, such as Kinetics [24], YouTube 8M [1], and the Moments in Time dataset [33]. It is well-established that pre-training on such large datasets followed by fine-tuning on downstream tasks boosts performance, especially when the target datasets are small [47, 26, 21, 57, 16, 15, 45]. With the challenges of curating and defining label taxonomies for massive-scale datasets, the focus has shifted to pretraining on unlabeled videos [54, 46], or video datasets accompanied by weak supervision such as social media hashtags [14] or narrated instructions [30], which can be obtained without expensive data curation. Compared to existing action recognition benchmarks that use real-world datasets, we propose a novel benchmark that aims at studying *pre-training and transfer from synthetic videos*, as a stepping stone to mitigate issues related to privacy, bias, ethics, and copyright.

2.2 Learning from Synthetic Data

Synthetic data has been widely used to solve various computer vision problems by replacing real-world training data [12, 31, 35, 36, 39, 53, 11, 51, 22, 25, 42, 55]. While many of these works have tried to generate synthetic data as similar as real data, Baradad et al. [2] has shown that synthetic images with structured noise can be used for representation learning as the diversity of training images is as important as naturalism. Further, approaches to optimizing simulator parameters have been explored to learn better synthetic data for specific tasks [3, 41, 23], or even tasks not seen during training [32].

Only a few works attempted to learn action recognition from synthetic data. ElderSim [19] generates realistic videos of elders' daily activities in households to augment limited publicly available elder activity data. SURREACT [51] introduces a novel data generation methodology that reconstructs 3D human body models from videos to render synthetic videos for unseen viewpoints at various angles. ThreeDWorld [12] is a synthetic video simulation platform for interactive multi-modal physical simulation, and also supports human-agent interactions. In our work, *we compile a large-scale synthetic video dataset* to explore a mixture of these simulators, as well as pre-made synthetic video datasets, such as Procedural Human Action Videos (PHAV) [10].

Existing approaches to use simulators for action recognition [12, 19, 51] have shown performance improvement by adding the simulated videos to the original training datasets. However, in contrast to our proposed benchmark, no prior work has studied the transferability of synthetic video representations to other domains that may have different action categories than the synthetic datasets.

2.3 Domain Knowledge Transfer from Synthetic Data

Many approaches have been proposed to transfer knowledge from synthetic to real domains, generally relying on standard domain adaptation methods [9] to bridge the gap between the two domains. Examples include generative models to improve the realism of synthetic images and videos [38, 18], as well as methods that operate in the feature space, such as adversarial methods which encourage domain confusion to learn domain-invariant features [37, 13, 50], and discrepancy-based approaches that align feature distributions of the two domains [40, 60]. More recently, Syn2Real [56], a large-scale synthetic-to-real benchmark has been introduced for unsupervised domain adaptation.

These domain adaptation methods assume the same label set between the synthetic and real domains. By contrast, in our work, we remove this assumption and instead consider *multiple downstream tasks with a disjoint label set.* In addition, while prior work has been focused on adapting video representations from the synthetic to real domains, we show that the gap in action recognition performance between these domains is directly related to the object and scene bias of the downstream datasets.

Chapter 3

Proposed Benchmark

To mitigate the issues that come with pre-training models on real videos, synthetic data can be leveraged. To this end, we systematically explore the impact of synthetic data pre-training. We propose a novel benchmark for action recognition to use synthetic data only for pre-training models. We construct a benchmark consisting of three synthetic video datasets generated by multiple simulators (Section 3.1) and pre-train models on the Synthetic dataset. We then transfer the models to downstream tasks depicting various properties (Section 3.3).

3.1 Synthetic Dataset Sources

We create our Synthetic dataset by merging three publicly available assets: 1) ElderSim [19], 2) SURREACT [51], and 3) PHAV [10]. Figure 3-1 shows some synthetic videos and action categories used in our work.

3.1.1 ElderSim

ElderSim [19] generates realistic videos of elders' daily activities in households, along with 2D and 3D skeleton trajectories, with a goal of augmenting limited publicly available elder activity data. Authors of ElderSim also claim that by combining the generated videos with real videos, they were able to achieve state-of-the-art elder action recognition performance.

ElderSim has four realistic 3D rendered furnished residential house models for background with flexible lighting and camera viewpoint options. There are 15 different human agents with varying skin color, outfits, gender, etc.

3.1.2 SURREACT

SURREACT [51] introduces a novel data generation methodology that reconstructs 3D human body models from videos to render synthetic videos for unseen viewpoints at various angles. By augmenting a real dataset with such rendered synthetic videos, the authors were able to improve the state-of-the-art performance on human action multi-view benchmarks. For this work, we use SURREACT generated videos on two datasets: UESTC [20] and NTU [44].

SURREACT only supports static images as background. We generate 8 different videos with varying viewpoints, human agent body shape, clothes, and gender per source video.

3.1.3 PHAV

Procedural Human Action Videos (PHAV) [10] is a large scale synthetic pre-made dataset generated using modern game engines, thus providing physically plausible motions and actions. PHAV contains actions performed by 20 artist-designed human models at seven different large-scale environment backgrounds. Four lighting settings based on period of day, as well as four weather options are available in the pre-made dataset. Around 40,000 videos are provided, with at least 1,000 examples per class.

3.2 Synthetic Dataset Curation

Using the generators/dataset described in section 3.2, we create our Synthetic dataset with 150 classes in total. 55 actions from ElderSim, 100 actions from SURREACT, and 35 actions from PHAV are collected. Overlapping classes are manually screened

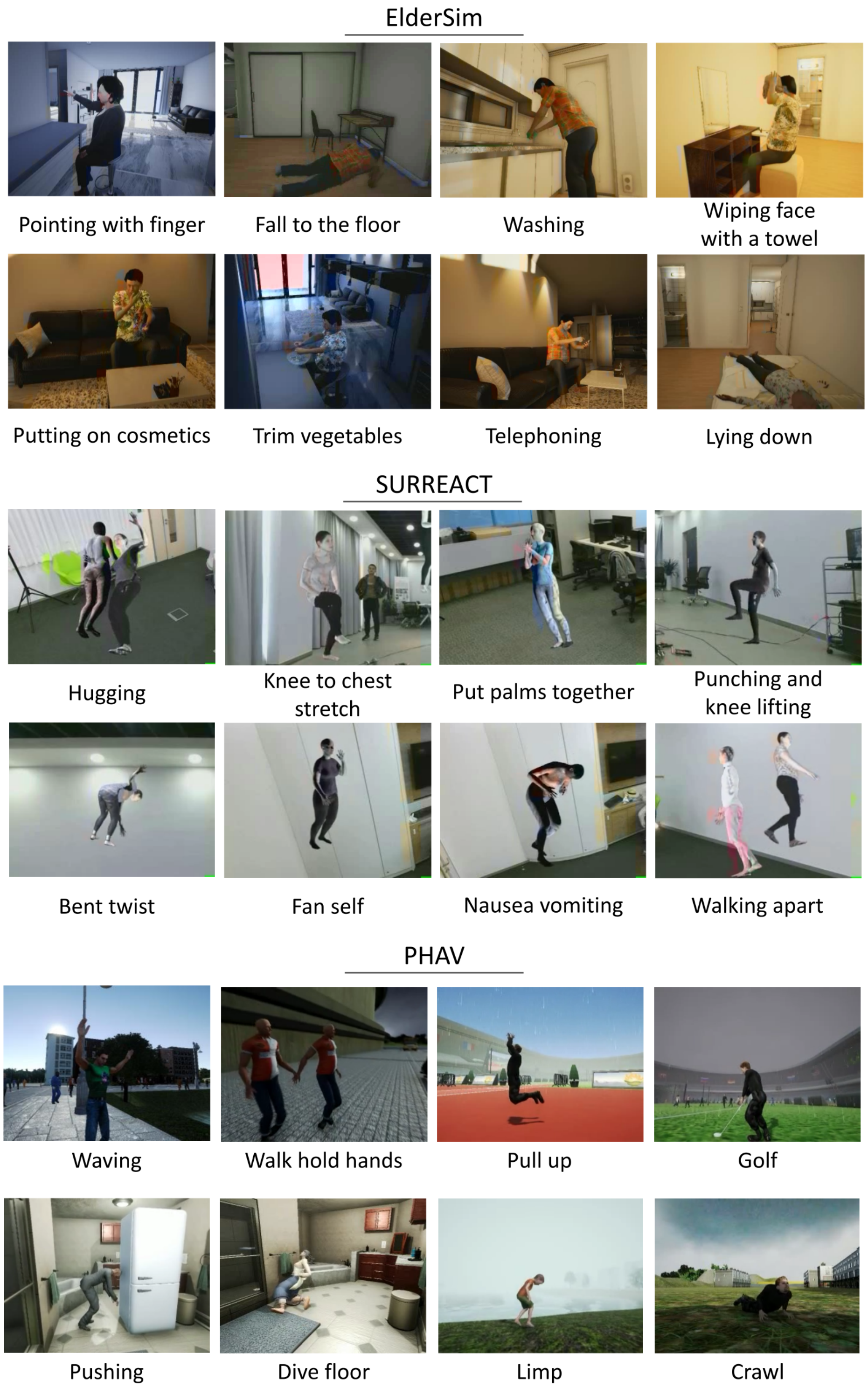


Figure 3-1: Examples of synthetic videos rendered by various simulators. We emphasize that synthetic datasets also cover action categories, such as "falling to the floor", which are not easy to obtain from the real datasets.

Table 3.1: Dataset statistics for downstream tasks.

Datasets	**# of Videos**	**# of Actions**	**Video Source**	**Domain**
UCF101 [47]	13,320	101	YouTube	General
HMDB51 [26]	6,849	51	Movies/YouTube	General
Mini-SSV2 [7]	93,000	87	User-Provided	General
Diving48 [28]	18,404	48	Web	Diving
IkeaFA [48]	111	12	Self-collected	Assembly
UAV-Human [27]	22,476	155	Flying UAV	General

and combined, and 1000 samples are randomly selected for each class. For classes consisting of samples from multiple assets, an equal number of videos is sampled from each asset to maintain an adequate ratio. While samples may have varying resolution and aspect ratio, we extract frames from all samples with a constant frames per second (fps).

3.3 Downstream Tasks

To assess the transferablity of video representations based on synthetic data, we fine-tune and linear probe the pre-trained models on six different downstream tasks. In this subsection, we describe the details of datasets used for the downstream tasks. We also show the statistics in Table 3.1.

3.3.1 UCF101

UCF101 [47] is a human-action dataset collected from YouTube, consisted of 101 action classes with 13,320 videos in total. Thanks to its variety in realistic action classes, as well as subdivided organization methodology (i.e. action categories are further divided into five types and 25 groups, in which videos in a same group have common qualities such as background or viewpoint), it has been appreciated by the computer vision community since its publication in 2012.

3.3.2 HMDB51

HMDB51 [26] presents 51 human activities with refined quality, light conditions, and accurate surrounding features, and is thus smaller than UCF101 with only 6,849 clips. HMDB51 is further divided into five types, including rather detailed action classes such as “smiling” or “laughing”.

3.3.3 Mini Something-Something V2

Something-Something V2 [15] was introduced to test the ability of a model to understand temporal dynamics rather than relying on objects or background in scenes. The dataset consists of 174 classes with around 220,000 videos of humans performing basic actions with common objects, in which action labels are independent of the objects themselves (e.g. “putting something behind something”). For our experiments, we use a reduced version of this dataset named Mini-SSV2 [7], which consists of only half of the action labels. 87 labels are chosen at random, resulting in around 93,000 videos.

3.3.4 Diving48

Diving48 [28] is a collection of diving competition videos, made up of around 18,000 videos which are divided into 48 dive sequences. Since all videos share a similar background and object features, Diving48 is considered a fine-grained dataset and is often used to test the robustness of video models.

3.3.5 Ikea Furniture Assembly

Ikea Furniture Assembly [48], or IkeaFA, provides 111 videos, each two to four minutes long. Summing up to around 480,000 frames worth of data, IkeaFA is a collection of GoPro furniture assembly videos, all of which are collected under a constant background by 14 individuals, either on a table or on the floor. There are 12 action classes in IkeaFA, including “pick leg”, “attach leg”, and “flip table”.

3.3.6 UAV-Human

As suggested by the name, UAV-Human [27] dataset is collected using an Unmanned Aerial Vehicle, thus providing a collection of videos from unique viewpoints. While the full UAV-Human dataset comes with multi-modality options (i.e. fisheye videos and night-vision videos), we only utilize 22,476 RGB videos for this work. This large-scale dataset contains 115 action classes, collected from 119 subjects. Note that for all reported numbers in the following sections, we use cross-subject-v1 evaluation method as described in [27].

Chapter 4

Representation Bias Analysis

The following sections describe the experimental setup and analyze the transferability of the Kinetics and Synthetic pre-trained models with respect to representation bias of the six downstream tasks listed in Section 3.3.

4.1 TSN ResNet-18 Experimental Setup

All experiments reported in this chapter are performed with the Temporal Segment Network (TSN) with ResNet-18 backbone [52]. TSN is an efficient 2-Dimensional Convolution Neural Network architecture designed for action recognition tasks especially with limited training samples. TSN aims to model long-range temporal structure by dividing a video input into K segments, and randomly sampling short snippets from each segment; these sparsely sampled snippets are then passed through two-stream (spatial and temporal) Convolutional Neural Networks, and fused to derive a video-level prediction.

We chose TSN as our baseline model since it is lightweight yet models both spatial and temporal information efficiently, and is suitable for downstream tasks involving less training data. Note that we train our models from scratch without ImageNet pre-trained weights.

Table 4.1: Transfer experiment top-1 accuracy results via fine-tuning (FT) and linear probing (LP) on downstream tasks. Pre-trained datasets each consist of 150 classes, up to 1,000 samples per class.

Pre-trained Dataset	**Transferred Dataset**											
	UCF101		HMDB51		Mini-SSV2		Diving48		IkeaFA		UAV-Human	
	FT	LP	FT	LP	FT	LP	FT	LP	FT	LP	FT	LP
Kinetics	**79.99**	**50.83**	**45.36**	**24.64**	38.22	6.01	31.78	8.21	38.41	28.66	12.87	1.53
Synthetic	76.32	19.35	37.65	12.35	**40.37**	**7.28**	**33.40**	**8.63**	**40.24**	**31.71**	**27.60**	**3.41**

4.1.1 Hyperparameter Settings

Both the Kinetics and Synthetic baseline models, as well as downstream tasks transfer experiments are trained until convergence using SGD optimizer with momentum of 0.9 and weight decay of $5e^{-4}$. The dropout ratio for the final layer is set to 0.5. Initial learning rate of 0.01 and $1e^{-4}$ are used for baseline models training and transferring experiments, respectively, with cosine decay learning rate scheduler. In terms of data loading settings, 8 frames are sampled per clip, with a batch size of 256 clips.

In order to study the transferability of synthetic video models, we perform both fine-tuning and linear probing on our downstream tasks. For fine-tuning experiments, either Kinetics or Synthetic pre-trained weights are loaded and the entire network is trained, while for linear probing, we freeze the pre-trained weights after loading and only train the final output layer.

4.2 Main Results

We first present the transfer learning experiments top-1 accuracies in Table 4.1. Note that the Synthetic dataset used for pre-training consists of 150 classes with 1000 samples per class, and the Kinetics dataset used for pre-training had been downscaled to match the Synthetic dataset's statistics. All classes and samples for the downsized Kinetics dataset were randomly selected from full Kinetics.

We would like to emphasize that our goal is not to obtain state-of-the-art results on the downstream datasets, given the reduced pre-training dataset sizes and lightweight backbone as described above. Instead, we aim at providing a fair comparison between real and synthetic models, using the same pre-training dataset sizes. We show both fine-tuning and linear probing transfer results.

While the Kinetics pre-trained model is preferable for UCF101 and HMDB51 transfers, our Synthetic pre-trained model outperforms the Kinetics model when transferring on Mini-SSV2, Diving48, IkeaFA, and UAV-Human. Qualitatively, we theorize that UCF101 and HMDB51 are more prone to object and scene representation bias than the other four datasets. We further believe that the Synthetic dataset is more robust to bias than Kinetics since clips are generated on either shared background image/rendering or without surrounding objects in relation to the action class, which forces the model to focus on the actions over possible biases. For example, if we refer back to Figure 3-1, PHAV's "golf" example generates an agent performing golf swing in a stadium with other agents strolling behind. Every generated video will have different background scene and features. However, models trained on real videos may learn to classify the category "golf" from scene (e.g. golf course) or objects (e.g. golf carts) that have high correlation to golf rather than from the swing action itself. This property of generated synthetic videos helps Synthetic pre-trained models to outperform real video pre-trained models on datasets with low representation bias.

Borrowing the representation bias definition from [28], we quantify representation bias for each downstream dataset using the following equation:

$$\mathcal{B}(\mathcal{D}, \phi) = \log \frac{\mathcal{M}(\mathcal{D}, \phi)}{\mathcal{M}_{rnd}}, \tag{4.1}$$

where bias $\mathcal{B}$ for dataset $\mathcal{D}$ using representation ϕ is directly related to the ratio of the performance of subject representation $\mathcal{M}(\mathcal{D}, \phi)$ to random chance performance, $\mathcal{M}_{rnd}$. We calculate $\mathcal{M}(\mathcal{D}, \phi)$ by measuring the performance of a linear action recognition classifier trained on top of a frozen ImageNet model. The intuition is that ImageNet features encode static cues, such as objects, and therefore $\mathcal{M}(\mathcal{D}, \phi)$ is related to the amount of action categories that can be recognized solely by static cues in the videos, without any temporal dynamics.

Table 4.2: Representation bias for each downstream task dataset calculated using ImageNet representation. LP stands for linear probing.

	Transferred Dataset					
	UCF101	HMDB51	Mini-SSV2	Diving48	IkeaFA	UAV-Human
ImageNet LP Accuracy, $\mathcal{M}(\mathcal{D}, \phi)$	48.53	27.78	9.03	8.68	33.54	2.52
Representation Bias, $\mathcal{B}(\mathcal{D}, \phi)$	5.62	3.83	2.97	2.06	2.01	1.96

Table 4.3: Performance gap between Synthetic and Kinetics pre-trained model is calculated by taking the accuracy ratio. A gap value greater than 1.0 represents Synthetic pre-trained model outperforms.

Synthetic to Kinetics Performance Gap	**Transferred Dataset**					
	UCF101	HMDB51	Mini-SSV2	Diving48	IkeaFA	UAV-Human
Linear Probing	0.95	0.50	1.21	1.05	1.11	2.23
Fine-tuning	0.38	0.83	1.06	1.05	1.05	2.14

Table 4.2 summarizes the representation bias measured using an ImageNet pre-trained model for downstream tasks. As theorized, UCF101 and HMDB51 have high representation bias scores of 5.62 and 3.83, respectively, while Mini-SSV2, Diving48, IkeaFA, and UAV-Human have much lower representation bias scores of 2.97, 2.06, 2.01, and 1.96, respectively. It is expected that UCF101 and HMDB51 have high biases as they are composed of daily human actions with related objects and scene features in them. In addition, while Mini-SSV2 shows lower bias score than UCF101 and HMDB51 as its action categories focus on temporal movement/change of objects rather than objects themselves, it still has higher bias than Diving48 as models can learn unintentional object bias since some objects are more prone to specific action categories than others (e.g. round objects are more inclined to roll). Every IkeaFA video is taken under the same setting, with identical objects present in the frame throughout the entire video, and this consistency is reflected by its low representation bias. Finally, UAV-Human also exhibits low representation bias as UAV's far-distance viewpoint encompasses vast scene and object information, degrading the model's

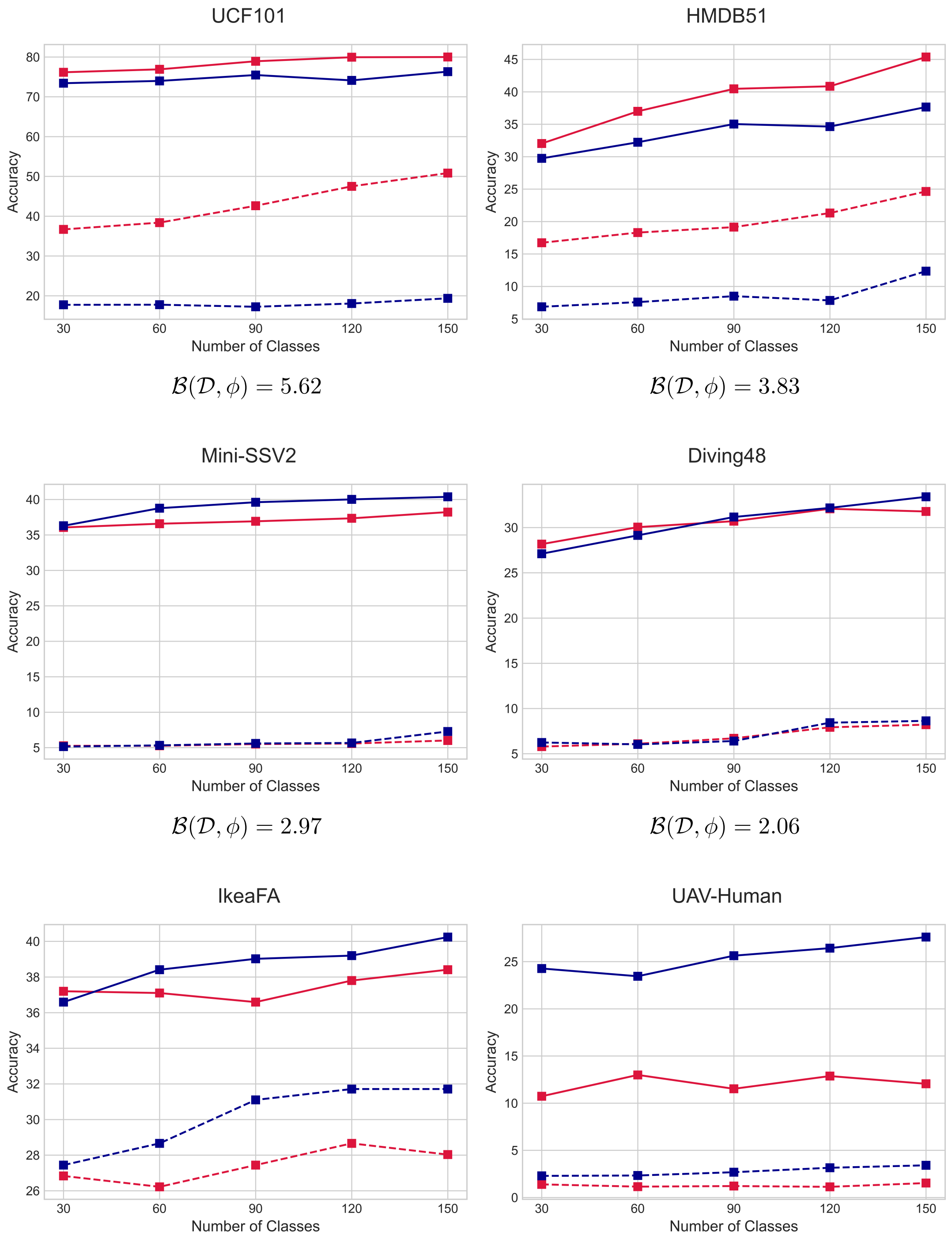
ability to predict an action based on such information.

Table 4.3 shows the performance gap between Synthetic and Kinetics pre-trained models. Specifically, we standardize the accuracy gaps by measuring the ratio of Synthetic to Kinetics pre-trained transfer accuracy. In other words, we retrieve the performance gap value by dividing accuracy of Synthetic pre-trained model over that of Kinetics pre-trained, meaning gap value greater than 1.0 representing Synthetic pre-trained model outperforming and vice versa. We can verify that UCF101 and HMDB51 have performance gap value less than 1.0 for both linear probing and fine-tuning top-1 accuracies, while Mini-SSV2, Diving48, IkeaFA, and UAV-Human have gap value greater than 1.0 for both. As a result, we observe an inverse relationship between representation bias and transfer performance gap, hence, reaching to a conclusion that the transfer gap between Kinetics pre-trained and Synthetic pre-trained models is highly influenced by the innate representation bias of the target dataset.

4.2.1 Effect of Number of Classes

We study how the number of classes in pre-training datasets influences the transferability on our downstream tasks, and analyze its relationship with representation bias. For both Kinetics and Synthetic datasets, we create four subset datasets with 30, 60, 90, and 120 classes, all with 1000 samples per class. Note that subset classes are chosen randomly, and each dataset is a superset of every other smaller dataset.

Figure 4-1 plots the fine-tuning and linear probing transfer top-1 accuracies of all pre-trained models on downstream tasks. First, we look at accuracies on UCF101, which has the highest representation bias among the downstream tasks, and note that the transferability of the Kinetics model increases as we increase the number of classes; this is specifically highlighted by the increase in linear probing accuracies. The transferability of the Synthetic pre-trained model, however, shows a less dramatic change when we increase the number of classes as merely supplying more synthetic classes does not increase the representation bias of the pre-train dataset. Again, this is further highlighted by the change in linear probing accuracies. We suspect that small increase in accuracy can be rather explained by the additional temporal information.



$\mathcal{B}(\mathcal{D}, \phi) = 5.62$

$\mathcal{B}(\mathcal{D}, \phi) = 3.83$

$\mathcal{B}(\mathcal{D}, \phi) = 2.97$

$\mathcal{B}(\mathcal{D}, \phi) = 2.06$

$\mathcal{B}(\mathcal{D}, \phi) = 2.01$

$\mathcal{B}(\mathcal{D}, \phi) = 1.96$

Figure 4-1: Fine-tuning (FT) and Linear Probing (LP) transfer results on six downstream tasks with various number of classes in pre-training datasets.
— Kinetics FT; - - - Kinetics LP; — Synthetic FT; - - - Synthetic LP.
Please refer to Appendix B for full results.

Similar observation can be made for HMDB51 accuracies, which has a relatively high representation bias, albeit being less significant.

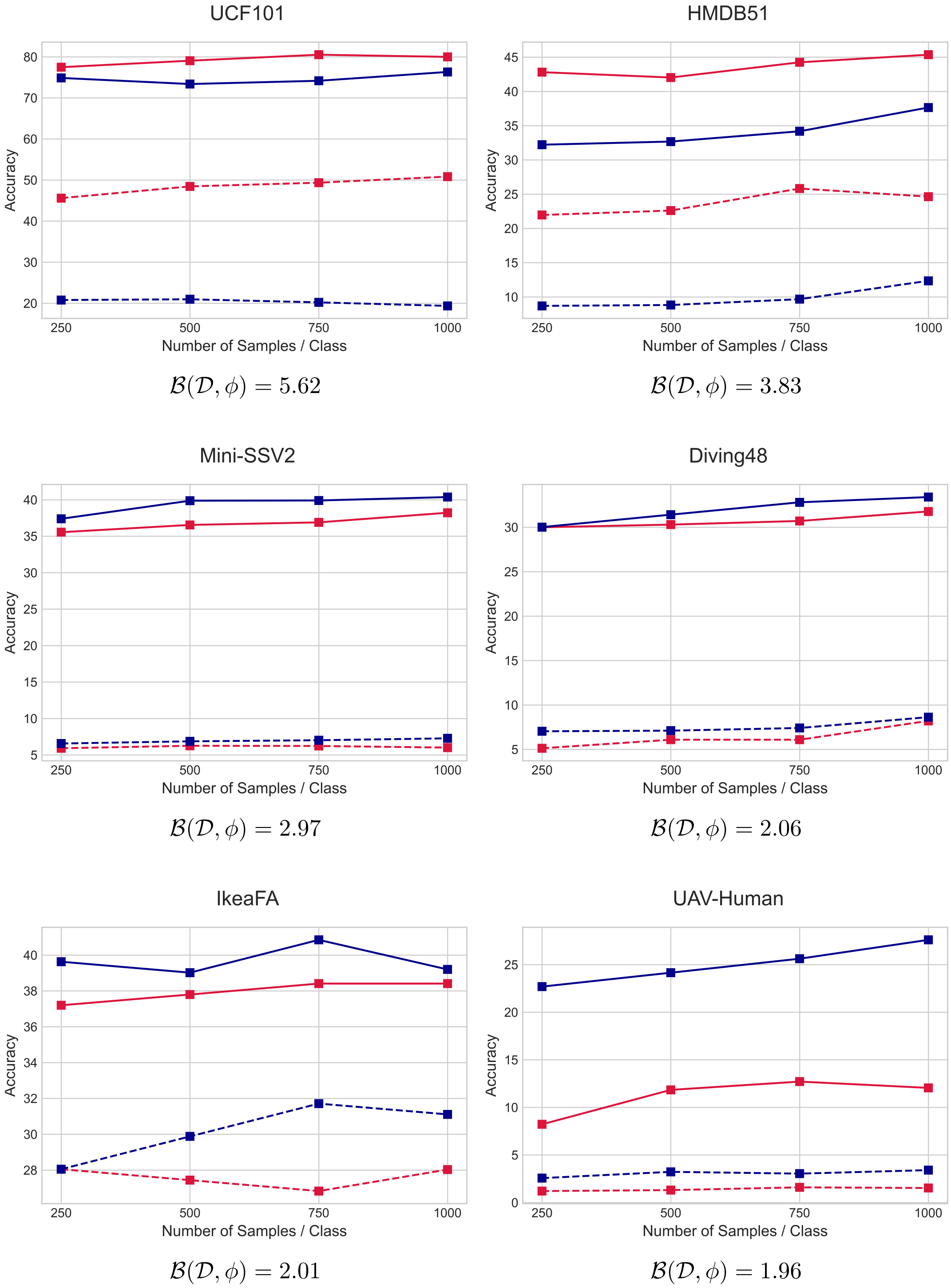
We can also see that the transferability for both Kinetics and Synthetic pre-trained models does not increase dramatically with more classes added to the pre-training datasets when transferring onto Mini-SSV2, Diving48, and UAV-Human. Additional object or scene features do not have as much of an effect when transferring onto these bias robust downstream tasks. It is also interesting to note that the Synthetic pre-trained model outperforms the Kinetic pre-trained model by a significant margin for UAV-Human, which is the downstream task with the lowest representation bias score.

For IkeaFA, we recognize improvements in linear probing accuracies for both Kinetics and Synthetic pre-trained models as we increase the number of classes from 30 to 120. We can suspect that more classes (and total video samples) are required for this downstream task for the models to learn the significant features for domain transfer.

4.2.2 Effect of Videos per Class

Next, we vary the number of samples per class for Kinetics and Synthetic pre-train datasets to examine its effects on transferability. We fix the number of classes to 150, and create three subset datasets with 250, 500, and 750 samples per class, similarly with samples being chosen at random and each dataset being a superset of every other smaller dataset.

Generally, we detect a slight increase in accuracies for all downstream tasks as we increase the number of samples per class due to more availability of pre-training samples. However, from Figure 4-2, we can observe a similar phenomenon as before; increasing Kinetics samples per class does not significantly boost the transferability compared to increasing the number of classes, as we are less likely to introduce novel representation bias with extra samples within the same class. Similarly, increasing Synthetic samples per class does not provide remarkable improvement on transferability. Although increasing the number of Synthetic samples per class implies further



$\mathcal{B}(\mathcal{D}, \phi) = 5.62$

$\mathcal{B}(\mathcal{D}, \phi) = 3.83$

$\mathcal{B}(\mathcal{D}, \phi) = 2.97$

$\mathcal{B}(\mathcal{D}, \phi) = 2.06$

$\mathcal{B}(\mathcal{D}, \phi) = 2.01$

$\mathcal{B}(\mathcal{D}, \phi) = 1.96$

Figure 4-2: Fine-tuning (FT) and Linear Probing (LP) transfer results as number of samples per class in pre-training datasets are limited to 250, 500, 750, and 1000.
— Kinetics FT; - - - Kinetics LP; — Synthetic FT; - - - Synthetic LP.
Please refer to Appendix C for full results.

variation in lighting, camera angles/position, humanoid types, and other video generation parameters, it does not deliver striking performance enhancement as it is not addressing the representation bias issue. We can further conclude that significant features required for domain transfer had saturated well before 250 samples per class for 150-class pre-trained models.

Chapter 5

Experiments with Higher Capacity Models

In this chapter, we repeat the fine-tuning experiments on downstream tasks listed in Section 3.3 using more complex video action recognition models to verify and validate our findings from the previous chapter.

5.1 Experimental Setup

In this section, we describe the two models we use to reanalyze the transferability of Kinetics and Synthetic pre-trained models.

5.1.1 TSN ResNet-50

We revisit TSN with a bigger backbone network, ResNet-50 [52]. While ResNet-18 consists of 8 residual blocks, each with two layers, ResNet-50 consists of 15 residual blocks, each with three layers. We train our models from scratch without ImageNet pre-trained weights.

5.1.2 I3D ResNet-50

We delve into a 3-Dimensional action recognition model, I3D, with ResNet-50 backbone [6]. I3D borrows designs from 2-Dimensional networks, and inflate all filters and pooling kernals with an extra dimension. Although I3Ds also benefit from 2-Dimensional counterpart's learned parameters, we train our I3D models from scratch for consistency with other architecture experiments.

5.2 Hyperparameter Sweeping

The Kinetics and Synthetic baseline models for TSN ResNet-50 and I3D are trained using SGD optimizer with momentum of 0.9, final layer dropout rate of 0.5, and number of samples of 8-frame per clip. We examine initial learning rate of [0.01, 0.02] with cosine decay, batch size of [64, 128], and weight decay rate of [0.0001, 0.0005, 0.001], resulting in total 12 combinations of hyperparameters per baseline model. We select the best-performing model for each baseline to transfer onto our downstream tasks.

For each downstream task, we use the identical optimizer, momentum, dropout rate, and number of samples. We explore initial learning rate of [0.0001, 0.0005, 0.001], batch size of [32, 64], and weight decay rate of [0.0001, 0.0005, 0.001], resulting in total 18 combinations of hyperparameters per downstream task per baseline model.

5.3 Main Results

5.3.1 Results: TSN ResNet-50

Table 5.1 shows the best performing top-1 transfer results of TSN ResNet-50 models fine-tuned on downstream tasks. Note that all downstream tasks' accuracies experience a significant improvement when using TSN with ResNet-50 backbone instead of ResNet-18.

Here, we also observe a consistent outcome as before, in which the Kinetics

Table 5.1: TSN ResNet-50 downstream tasks transfer top-1 accuracy results.

Pre-trained Dataset	**Transferred Dataset**					
	UCF101	HMDB51	Mini-SSV2	Diving48	IkeaFA	UAV-Human
Kinetics	**86.17**	**57.45**	48.50	62.84	42.07	32.45
Synthetic	83.40	54.38	**49.69**	**63.50**	**42.68**	**35.57**

Table 5.2: I3D ResNet-50 downstream tasks transfer top-1 accuracy results.

Pre-trained Dataset	**Transferred Dataset**					
	UCF101	HMDB51	Mini-SSV2	Diving48	IkeaFA	UAV-Human
Kinetics	**86.87**	**59.21**	50.08	54.82	40.85	31.13
Synthetic	82.05	55.69	**50.72**	**55.28**	**42.68**	**35.13**

pre-trained models outperform on downstream tasks with high representation bias (UCF101 and HMDB51), while the Synthetic pre-trained models outperform on those with low representation bias. However, note that the accuracy gaps between the Kinetic and Synthetic pre-trained models are more closed for all six downstream tasks when using this deeper model. For instance, the TSN ResNet-18 model pre-trained on the Kinetics outperformed the Synthetic pre-trained model on UCF101 by about 3.67%, but with ResNet-50 backbone, the gap is 2.77%. Similar observation can be made for HMDB51 results, in which the gap is closed from 7.71% to 3.07%.

Refer to Appendix D for the full hyperparameter tuning results.

5.3.2 Results: I3D ResNet-50

Best performing top-1 transfer accuracies of I3D ResNet-50 models are shown on Table 5.2. We observe a general increase in accuracies for UCF101, HMDB51, and Mini-SSV2 relative to TSN ResNet-50's performance, while there are slight drops in accuracies for Diving48, IkeaFA, and UAV-Human. However, it is still the case that the Kinetics pre-trained model performs superior on high representation bias downstream tasks, and worse on low representation bias tasks relative to the Synthetic pre-trained model.

Appendix E shows the full hyperparameter sweeping results for I3D ResNet-50.

5.3.3 Discussion

By utilizing higher capacity models and via thorough hyperparameter sweeping, we reinforce our conclusion that the Synthetic pre-trained models outperform the Kinetics pre-trained counterparts when transferred to downstream tasks with low representation bias. As expected, we notice an overall increase in transfer accuracies with TSN ResNet-50 and I3D ResNet-50 compared to TSN ResNet-18, and the accuracy gaps between the Kinetics and Synthetic pre-trained models are smaller for all six downstream tasks.

Chapter 6

Future Work

First, we would like to expand our Synthetic dataset with more classes and samples from additional video generators. It would be interesting to make our initial synthetic video dataset public to have the community add on to the dataset with their set of synthetic videos to expedite the augmentation process as well. We also seek to include more downstream tasks of various domains, such as first-person viewpoints, social media, sports, embodied perception, safety and security, and more. This work can also be extended to cross-modal learning by synthesizing videos and captions to cover a various range of domains.

Possible future directions also include pre-training models by mixing real and synthetic videos. We plan to vary the ratio of real to synthetic videos to examine whether augmenting real videos with synthetic positively or negatively impact the transferability.

We have an extensive list of experiments we have arranged to further investigate the transferability of Synthetic pre-trained models. First, we would like to replicate the transfer experiments using a 2.5-Dimensional model, such as R(2+1)D [49], and with various loss functions, such as self-supervised or contrastive loss.

Another question we pose is what makes for a good Synthetic pre-trained model. We plan to control the parameters of video generators, such as lighting, agent pose, background, lighting, camera distance, camera angle, and more to study which parameter has the greatest impact on transferability.

It is plausible that each synthetic video generator has an innate bias, and since our Synthetic dataset is composed of videos from multiple generators, we can analyze simulator bias by creating subsets of the Synthetic dataset and analyzing transferability as we increase the number of simulators included in the pre-training dataset.

Finally, another interesting application would be to incorporate the Synthetic pre-trained model in projects for real-life clients that suffer from lack of data, such as construction-related companies that are willing to monitor sites using AI action recognition deep learning models.

Chapter 7

Conclusions

In this work, we have introduced a new action recognition benchmark to mitigate the issues inherent to training models with real videos, such as privacy, bias, ethics, and copyright. Specifically, we constructed a Synthetic dataset from three publicly available assets (ElderSim, SURREACT, PHAV), trained models on the Synthetic dataset, and finally transferred the pre-trained models to various downstream tasks. Our experiments show that the models pre-trained on the Synthetic dataset outperform those pre-trained on real videos on the downstream datasets with low representation bias (Mini-SSV2, Diving48, IkeaFA, UAV-Human). This suggests that although models trained on synthetic data expose weaker object and background scene features, they do provide features with strong correlation to actions, making them more useful for downstream tasks with lower representation bias. In fact, stronger object features (inherent to models trained with real videos) may even be a nuisance factor for transfer tasks onto lower representation bias datasets.

We believe the new benchmark and the in-depth analysis on the transferability of models pre-trained on synthetic data will guide a new direction of fair and transparent study for the AI research community.

Appendix A

Statistics of Class Overlap

Table A.1: Summary of overlapping classes between pre-train Kinetics/Synthetic datasets and downstream tasks.

Pre-trained Dataset	**Transferred Dataset**											
	UCF101		HMDB51		Mini-SSV2		Diving48		IkeaFA		UAV-Human	
	# of classes	Ratio	# of classes	Ratio	# of classes	Ratio	# of classes	Ratio	# of classes	Ratio	# of classes	Ratio
Kinetics	23	0.23	11	0.22	0	0.00	0	0.00	0	0.00	27	0.17
Synthetic	13	0.13	25	0.49	0	0.00	0	0.00	0	0.00	36	0.23

Table A.1 summarizes the number of overlapping classes between Kinetics or Synthetic pre-train dataset and each of the six downstream tasks, based on their class names. Notice that Mini-SSV2, Diving48, and IkeaFA have completely disjoint action labels, and models pre-trained on Synthetic dataset outperform their respective Kinetics pre-trained models in these three datasets.

Interestingly, for HMDB51, the Synthetic pre-train dataset has more overlapping classes, yet the Kinetics pre-trained model still outperforms on this downstream task. Here, we conclude that the intersection of action labels plays a less significant role than representation bias.

Appendix B

TSN ResNet-18 Class Variation Experiment Results

Table B.1: Full fine-tuning (FT) and linear probing (LP) transfer experiment results on downstream tasks using Kinetics and Synthetic pre-trained models with varying number of classes in pre-training datasets.

Pre-trained Dataset			**Transferred Dataset**											
			UCF101		HMDB51		Mini-SSV2		Diving48		IkeaFA		UAV-Human	
			FT	LP	FT	LP	FT	LP	FT	LP	FT	LP	FT	LP
Kinetics	Num. Classes	30	76.16	36.67	32.03	16.73	36.05	5.26	28.17	5.79	37.20	26.83	10.73	1.39
		60	76.90	38.38	36.99	18.30	36.58	5.26	30.05	6.10	37.10	26.22	12.99	1.15
		90	78.93	42.61	40.46	19.14	36.92	5.50	30.71	6.70	36.59	27.44	11.52	1.21
		120	79.94	47.50	40.85	21.31	37.34	5.58	32.08	7.92	37.80	28.66	12.87	1.13
		150	79.99	50.83	45.36	24.64	38.22	6.01	31.78	8.21	38.41	28.03	12.05	1.53
Synthetic	Num. Classes	30	73.41	17.72	29.74	6.86	36.29	5.15	27.11	6.24	36.59	27.44	24.27	2.29
		60	73.99	17.75	32.22	7.58	38.77	5.32	29.14	6.03	38.41	28.66	23.45	2.33
		90	75.47	17.23	35.03	8.50	39.61	5.59	31.17	6.40	39.02	31.10	25.63	2.68
		120	74.12	18.05	34.64	7.84	40.01	5.66	32.18	8.43	39.20	31.71	26.43	3.15
		150	76.32	19.35	37.65	12.35	40.37	7.28	33.40	8.63	40.24	31.71	27.60	3.41

In Table B.1, we show the full experiment results with varying the number of classes for pre-training datasets. We note that for datasets with high representation bias (UCF101, HMDB51), transferability increases with more number of classes for Kinetics pre-trained models, but the effect is less pronounced for Synthetic pre-trained models. We also observe that transferability does not increase dramatically for both Kinetics and Synthetic pre-trained models on downstream tasks with low representation bias (Mini-SSV2, Diving48, IkeaFA, UAV-Human).

Appendix C

TSN ResNet-18 Sample Number Variation Experiment Results

Table C.1: Full transfer experiment results on downstream tasks when the number of samples per class in pre-training datasets is limited. Both fine-tuning (FT) and linear probing (LP) results are shown.

Pre-trained Dataset			**Transferred Dataset**											
			UCF101		HMDB51		Mini-SSV2		Diving48		IkeaFA		UAV-Human	
			FT	LP	FT	LP	FT	LP	FT	LP	FT	LP	FT	LP
Kinetics	Num. samples	250	77.48	45.60	42.81	21.96	35.55	5.92	30.00	5.127	37.20	28.05	8.22	1.21
		500	79.06	48.45	42.03	22.61	36.55	6.26	30.30	6.09	37.80	27.44	11.84	1.31
		750	80.52	49.35	44.25	25.82	36.90	6.23	30.71	6.09	38.41	26.83	12.71	1.60
		1000	79.99	50.83	45.36	24.64	38.22	6.01	31.78	8.21	38.41	28.03	12.05	1.53
Synthetic	Num. samples	250	74.86	20.78	32.22	8.69	37.39	6.57	30.02	7.04	39.63	28.05	22.69	2.57
		500	73.38	20.96	32.68	8.82	39.87	6.87	31.42	7.11	39.02	29.88	24.15	3.23
		750	74.17	20.21	34.18	9.67	39.9	7.02	32.81	7.41	40.85	31.71	25.62	3.04
		1000	74.12	18.05	34.64	7.84	40.01	5.66	32.18	8.43	39.20	31.71	26.43	3.15

In Table C.1, we show the full experiment results with varying the number of samples per class for pre-training datasets. Increasing the number of samples per class for both Kinetics and Synthetic datasets has less significant effects on transferability on all downstream tasks compared to adding classes, as we are not likely to introduce as much representation bias by having more samples within the same number of action classes.

Appendix D

TSN ResNet-50 Parameter Sweeping Results

Tables below show the full parameter sweeping results of TSN ResNet-50 on downstream tasks. 2 batch sizes, 3 learning rates (lr), and 3 weight decay rates (wd) are explored, resulting in total 18 combinations per baseline model.

D.1 UCF101

Table D.1: UCF101 TSN ResNet-50 Parameter Sweeping Results.

Pre-trained Dataset		**UCF101**					
		Batch Size 32			Batch Size 64		
		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	84.67	**86.17**	85.06	84.22	84.83	84.27
	$wd = 0.0005$	84.35	85.14	85.75	84.30	85.62	84.54
	$wd = 0.001$	85.09	85.88	85.41	83.16	84.99	84.64
Synthetic	$wd = 0.0001$	79.22	80.74	81.72	72.52	79.01	80.17
	$wd = 0.0005$	77.93	81.79	82.58	72.85	79.46	81.71
	$wd = 0.001$	78.91	82.23	**83.40**	80.39	80.02	82.02

D.2 HMDB51

Table D.2: HMDB51 TSN ResNet-50 Parameter Sweeping Results.

		HMDB51					
Pre-trained		Batch Size 32			Batch Size 64		
Dataset		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	57.12	56.08	56.08	53.20	56.41	55.95
	$wd = 0.0005$	**57.45**	56.67	55.42	52.81	56.01	56.80
	$wd = 0.001$	55.88	55.82	55.42	54.12	56.27	55.42
Synthetic	$wd = 0.0001$	46.93	53.60	**54.38**	43.66	50.64	53.59
	$wd = 0.0005$	47.84	53.40	53.60	43.34	50.13	50.72
	$wd = 0.001$	47.26	52.55	52.81	42.18	48.69	51.02

D.3 Mini-SSV2

Table D.3: Mini-SSV2 TSN ResNet-50 Parameter Sweeping Results.

		Mini-SSV2					
Pre-trained		Batch Size 32			Batch Size 64		
Dataset		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	46.39	48.17	**48.50**	43.61	45.82	47.28
	$wd = 0.0005$	46.29	46.64	47.90	44.74	45.48	47.35
	$wd = 0.001$	45.39	47.45	48.34	43.66	46.46	46.97
Synthetic	$wd = 0.0001$	47.72	48.83	48.61	43.38	46.29	47.61
	$wd = 0.0005$	47.47	49.05	49.37	45.89	47.89	48.37
	$wd = 0.001$	47.98	48.96	**49.69**	44.14	46.61	47.56

D.4 Diving48

Table D.4: Diving48 TSN ResNet-50 Parameter Sweeping Results.

		Diving48					
Pre-trained		Batch Size 32			Batch Size 64		
Dataset		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	48.48	57.66	60.56	44.37	52.54	56.45
	$wd = 0.0005$	43.50	52.08	58.22	43.65	53.30	57.41
	$wd = 0.001$	50.66	59.14	**62.84**	44.06	51.37	56.14
Synthetic	$wd = 0.0001$	51.42	57.96	59.19	42.44	53.3	57.61
	$wd = 0.0005$	50.66	57.61	57.56	44.11	55.33	57.56
	$wd = 0.001$	51.93	57.36	**63.50**	43.81	53.15	56.75

D.5 IkeaFA

Table D.5: IkeaFA TSN ResNet-50 Parameter Sweeping Results.

		IkeaFA					
Pre-trained		Batch Size 32			Batch Size 64		
Dataset		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	37.80	40.24	40.85	39.63	40.24	41.46
	$wd = 0.0005$	39.02	**42.07**	41.46	37.80	40.85	40.85
	$wd = 0.001$	37.80	40.85	**42.07**	38.41	41.46	40.24
Synthetic	$wd = 0.0001$	38.41	39.02	40.85	35.98	**42.68**	40.85
	$wd = 0.0005$	37.20	40.85	37.80	36.59	38.41	39.02
	$wd = 0.001$	39.02	40.85	40.85	37.80	39.02	39.02

D.6 UAV-Human

Table D.6: UAV-Human TSN ResNet-50 Parameter Sweeping Results.

Pre-trained Dataset		UAV-Human					
		Batch Size 32			Batch Size 64		
		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	20.41	29.80	31.61	14.73	27.78	29.57
	$wd = 0.0005$	15.72	30.25	31.59	14.88	27.72	29.77
	$wd = 0.001$	20.16	30.50	**32.45**	14.54	27.96	29.56
Synthetic	$wd = 0.0001$	20.06	33.77	**35.57**	14.15	31.74	34.65
	$wd = 0.0005$	18.80	34.26	35.13	14.04	31.47	34.31
	$wd = 0.001$	18.30	33.69	35.13	14.23	31.27	35.57

Appendix E

I3D ResNet-50 Parameter Sweeping Results

Tables below show the full parameter sweeping results of I3D ResNet-50 on downstream tasks. 2 batch sizes, 3 learning rates (lr), and 3 weight decay rates (wd) are explored, resulting in total 18 combinations per baseline model.

E.1 UCF101

Table E.1: UCF101 I3D ResNet-50 Parameter Sweeping Results.

		UCF101					
Pre-trained		Batch Size 32			Batch Size 64		
Dataset		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	70.95	86.61	86.71	53.95	85.23	86.62
	$wd = 0.0005$	70.31	86.55	86.52	54.03	85.31	86.49
	$wd = 0.001$	70.39	86.61	**86.87**	53.95	85.21	86.68
Synthetic	$wd = 0.0001$	40.31	80.21	81.69	24.12	76.02	80.57
	$wd = 0.0005$	39.17	80.78	82.02	24.37	76.21	80.35
	$wd = 0.001$	39.55	81.07	**82.05**	24.03	76.31	80.89

E.2 HMDB51

Table E.2: HMDB51 I3D ResNet-50 Parameter Sweeping Results.

		HMDB51					
Pre-trained		Batch Size 32			Batch Size 64		
Dataset		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	42.22	**59.21**	59.08	31.96	57.22	59.17
	$wd = 0.0005$	42.16	58.11	58.75	32.09	56.83	58.78
	$wd = 0.001$	42.48	58.95	58.49	32.22	57.35	58.01
Synthetic	$wd = 0.0001$	27.71	52.22	53.40	16.60	47.65	51.11
	$wd = 0.0005$	27.22	52.61	54.90	17.25	47.58	52.15
	$wd = 0.001$	27.91	52.03	**55.69**	16.93	46.73	54.18

E.3 Mini-SSV2

Table E.3: Mini-SSV2 I3D ResNet-50 Parameter Sweeping Results.

		Mini-SSV2					
Pre-trained		Batch Size 32			Batch Size 64		
Dataset		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	46.86	49.44	49.64	39.26	48.51	49.11
	$wd = 0.0005$	46.83	48.95	49.69	39.24	49.07	49.33
	$wd = 0.001$	46.91	49.65	**50.08**	39.32	48.70	49.54
Synthetic	$wd = 0.0001$	44.35	49.72	50.32	35.06	48.41	49.22
	$wd = 0.0005$	44.15	49.71	50.47	34.78	48.12	49.41
	$wd = 0.001$	44.22	49.73	**50.72**	34.51	48.67	49.68

E.4 Diving48

Table E.4: Diving48 I3D ResNet-50 Parameter Sweeping Results.

Pre-trained Dataset		**Diving48**					
		Batch Size 32			Batch Size 64		
		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	37.97	51.62	51.88	28.12	46.24	51.88
	$wd = 0.0005$	36.95	51.98	52.39	27.12	46.24	51.37
	$wd = 0.001$	37.56	52.28	**54.82**	28.07	46.04	50.15
Synthetic	$wd = 0.0001$	37.36	52.84	54.16	26.60	46.40	51.98
	$wd = 0.0005$	37.46	52.08	**55.28**	26.65	46.95	52.39
	$wd = 0.001$	37.06	51.52	54.02	27.26	46.24	51.73

E.5 IkeaFA

Table E.5: IkeaFA I3D ResNet-50 Parameter Sweeping Results.

Pre-trained Dataset		**IkeaFA**					
		Batch Size 32			Batch Size 64		
		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	38.05	38.41	39.63	37.80	39.02	38.41
	$wd = 0.0005$	38.41	39.02	**40.85**	38.41	38.41	38.41
	$wd = 0.001$	37.80	39.63	40.24	37.80	40.85	39.02
Synthetic	$wd = 0.0001$	40.24	37.80	**42.68**	31.10	37.80	41.46
	$wd = 0.0005$	35.98	40.85	38.41	30.49	36.59	40.24
	$wd = 0.001$	35.37	39.63	39.02	29.88	36.59	39.02

E.6 UAV-Human

Table E.6: UAV-Human I3D ResNet-50 Parameter Sweeping Results.

		UAV-Human					
Pre-trained		Batch Size 32			Batch Size 64		
Dataset		$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$	$lr = 0.0001$	$lr = 0.0005$	$lr = 0.001$
Kinetics	$wd = 0.0001$	9.03	25.99	29.04	15.07	29.59	**31.13**
	$wd = 0.0005$	7.35	25.00	29.56	3.84	20.09	26.80
	$wd = 0.001$	7.20	24.97	29.40	4.43	19.25	26.41
Synthetic	$wd = 0.0001$	7.40	23.82	32.66	18.40	30.66	33.82
	$wd = 0.0005$	18.46	30.17	34.74	7.59	23.57	30.88
	$wd = 0.001$	7.53	23.94	31.12	18.14	29.82	**35.13**

Bibliography

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

[2] Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. *Advances in Neural Information Processing Systems*, 34, 2021.

[3] Harkirat Singh Behl, Atilim Güneş Baydin, Ran Gal, Philip HS Torr, and Vibhav Vineet. Autosimulate:(quickly) learning synthetic data generation. In *European Conference on Computer Vision*, pages 255–271. Springer, 2020.

[4] Matteo Bregonzio, Shaogang Gong, and Tao Xiang. Recognising action as clouds of space-time interest points. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1948–1955. IEEE, 2009.

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 2018.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2017.

[7] 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.

[8] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *NeurIPS*, 2019.

[9] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[10] César Roberto de Souza12, Adrien Gaidon, Yohann Cabon, and Antonio Manuel López. Procedural generation of videos to train deep action recognition networks. 2017.

[11] César Roberto de Souza12, Adrien Gaidon, Yohann Cabon, and Antonio Manuel López. Procedural generation of videos to train deep action recognition networks. 2017.

[12] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[14] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12046–12055, 2019.

[15] Raghav Goyal, Samira Kahou, Vincent Michalski, Joanna Materzyńska, 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. *arXiv preprint arXiv:1706.04261*, 2017.

[16] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. *arXiv preprint arXiv:1705.08421*, 2017.

[17] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[19] Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, , and Ig-Jae Kim. Eldersim: A synthetic data generation platform for human action recognition in eldercare applications. In *IEEE Access*, 2021.

[20] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale varying-view RGB-D action dataset for arbitrary-view human action recognition. *CoRR*, abs/1904.10681, 2019.

[21] YG Jiang, J Liu, A Roshan Zamir, G Toderici, I Laptev, M Shah, and R Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014.

[22] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[23] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4551–4560, 2019.

[24] 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. *arXiv preprint arXiv:1705.06950*, 2017.

[25] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[26] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In Wolfgang E. Nagel, Dietmar H. Kröner, and Michael M. Resch, editors, *High Performance Computing in Science and Engineering '12*, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[27] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uavhuman: A large benchmark for human behavior understanding with unmanned aerial vehicles, 2021.

[28] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, 2018.

[29] James J Little and Alessandro Verri. Analysis of differential and matching methods for optical flow. 1988.

[30] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

[31] Hiroaki Mikami, Kenji Fukumizu, Shogo Murai, Shuji Suzuki, Yuta Kikuchi, Taiji Suzuki, Shin-ichi Maeda, and Kohei Hayashi. A scaling law for synthetic-to-real transfer: A measure of pre-training. *arXiv e-prints*, pages arXiv–2108, 2021.

[32] Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S Feris. Task2sim: Towards effective pre-training and transfer from synthetic data. *arXiv preprint arXiv:2112.00054*, 2021.

[33] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):502–508, 2019.

[34] Ramakant Nevatia and Thomas O Binford. Description and recognition of curved objects. *Artificial intelligence*, 8(1):77–98, 1977.

[35] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *Proceedings of the IEEE international conference on computer vision*, pages 1278–1286, 2015.

[36] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7249–7255. IEEE, 2019.

[37] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–771, 2018.

[38] Stephan R Richter, Hassan Abu AlHaija, and Vladlen Koltun. Enhancing photorealism enhancement. *arXiv preprint arXiv:2105.04619*, 2021.

[39] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[40] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4), 2018.

[41] Nataniel Ruiz, Samuel Schulter, and Manmohan Chandraker. Learning to simulate. *arXiv preprint arXiv:1810.02513*, 2018.

[42] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019.

[43] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36 Vol.3, 2004.

[44] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. *CoRR*, abs/1604.02808, 2016.

[45] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *CoRR*, abs/1604.01753, 2016.

[46] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *CVPR*, 2021.

[47] 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.

[48] Sam Toyer, Anoop Cherian, Tengda Han, and Stephen Gould. Human pose forecasting via deep Markov models. In *DICTA*, 2017.

[49] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition, 2017.

[50] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[51] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287, 2021.

[52] 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, 2016.

[53] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12635–12644, 2020.

[54] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021.

[55] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018.

[56] Rajeev Yasarla, Vishwanath A Sindagi, and Vishal M Patel. Syn2real transfer learning for image deraining using gaussian processes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2726–2736, 2020.

[57] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *Int. J. Comput. Vis.*, 126(2-4):375–389, 2018.

[58] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[59] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

[60] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1823–1841, 2019.