Improving Video Model Transfer with Dynamic Representation Learning

Yi Li Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego

Abstract

Temporal modeling is an essential element in video understanding. While deep convolution-based architectures have been successful at solving large-scale video recognition datasets, recent work has pointed out that they are biased towards modeling short-range relations, often failing to capture long-term temporal structures in the videos, leading to poor transfer and generalization to new datasets. In this work, the problem of dynamic representation learning (DRL) is studied. We propose dynamic score, a measure of video dynamic modeling that describes the additional amount of information learned by a video network that cannot be captured by pure spatial student through knowledge distillation. DRL is then formulated as an adversarial learning problem between the video and spatial models, with the objective of maximizing the dynamic score of learned spatiotemporal classifier. The quality of learned video representations are evaluated on a diverse set of transfer learning problems concerning many-shot and few-shot action classification. Experimental results show that models learned with DRL outperform baselines in dynamic modeling, demonstrating higher transferability and generalization capacity to novel domains and tasks.

1. Introduction

Following the success of deep learning for image recognition [38, 48, 71, 75], convolutional neural networks have also gained popularity for video classification problems, such as action recognition [13, 25, 55, 70, 79, 84, 88], where they outperform other classification architectures, such as recurrent networks [20, 41, 89]. However, current action recognition performance is significantly below the levels of image recognition. This is, in part, due to the difficulty of learning *video representations* that generalize well. Part of this difficulty stems from current data collection practices, namely the use of action datasets collected from the web (e.g. YouTube) [10, 35, 45, 46, 59]. While these datasets have much larger size and diversity of action classes, performers and scenes than those collected in controlled settings [6, 66, 86], they are also known to exhibit various types of biases that hinder the generalization of trained video models to unseen domains [17, 53, 54, 67]. One of the most prevalent among these biases is the *spatial bias* due to the spurious correlation between action labels and the spatial appearance of video frames, in the form of background objects, scenes or human pose [53]. For example, the presence of a horse in the video is enough to infer the "horseback riding" action if that is the only action class in the dataset that involves horses. Spatial bias creates shortcuts that allow classifiers to infer action labels without modeling the temporal video component, known as *video dynamics,* leading to overoptimistic performance in popular action recognition benchmarks [54].

One of the nefarious consequences of dataset bias is that it favours certain representations over others [53]. In this context, the spatial bias of most video datasets is likely responsible for the dominant performance of convolutional architectures, known to favor local over long-range dependencies, in the action recognition field. This type of "evolutionary adaptation" of networks to dataset bias has, in fact, been documented in the object recognition literature, leading to the prevalence of networks with a strong bias towards local textures over global object shape [3, 8, 31]. We hypothesize that, in video modeling, the same phenomenon justifies the supremacy of 3D convolutional networks that implement localized spatio-temporal video representations, based on a very small number of frames, basically ignoring long-range video dynamics. This, we claim, hampers the generalization of these networks to unseen domains. For example, a horse representation is insufficient for transfer to a new domain that requires discrimination between "riding a horse" and "chasing a horse" or the classification of video into Olympic equestrian event classes. However, this problem is hard to diagnose in the predominant action recognition setting, where training and test data come from the same domain. If "horseback riding" is the only horse related class, detecting horses is enough for high performance on test data. Diagnosing the problem requires deploying the action recognizer outside of the native test set, as is usually done for tasks like few-shot learning [27, 64, 72, 74], domain adapta-

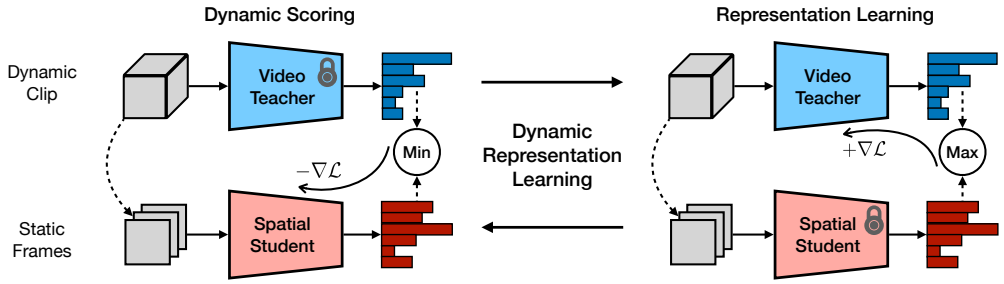


Figure 1. We propose *dynamic representation learning* (DRL), an adversarial learning strategy to enhance the modeling of video dynamics. DRL alternates between two steps: *Dynamic scoring* step quantitatively evaluates the temporal dynamics captured by a video model, defined as the expected difference been its predictions and those of a spatial student; *representation learning* step updates parameters of the video network to optimize dynamic score.

tion [29, 30, 42, 81], domain generalization [51, 52, 60, 61], etc. Since, in these settings, target domain videos might not contain the same types of bias as the training set, the action recognition system should transfer or generalize poorly. Despite the acknowledgement of dataset and model bias in action recognition, as well as efforts to overcome the locality of convolutional operators [31, 85], little effort has been devoted to the quantitative study of the spatial bias of current models, or how the reduction of this bias improves generalization performance.

In this work, we address this problem by introducing a new approach to *dynamic representation learning* (DRL) based on the explicit measurement and minimization of spatial bias. As shown in Figure 1, DRL is based on an adversarial optimization between the video network and a *spatial student,* i.e. a 2D network that processes video frames independently. Video network and spatial student are optimized alternately. In a *dynamic scoring* step, shown in the left of the figure, the student is optimized to mimic the predictions of the video network. The expected difference between the predictions of the two networks reflects how dynamic the video representation is. This expected difference is denoted as the *dynamic score* of the video network, measuring how much it relies on dynamic, over spatial, cues for classification. The lower this score, the more similar the video model is to a spatial classifier, and the greater its spatial bias. While the dynamic score is naturally measured by training the student by knowledge distillation [40], we also propose an optimization-free approach based on the preprocessing of video clips to remove temporal information, which is more computationally efficient. In the *representation learning* step, shown on the right of Figure 1, the video model trained to maximize its dynamic score with respect to a learned spatial student. Two approaches are proposed for this purpose. The first is based on adversarial augmentation, using the spatial student to derive perturbations that, when added to the video, obscure spatial cues. The second poses DRL as a min-max game between the video network and the spatial student to directly optimize the dynamic score of the former. The two methods share the same key idea—to penalize the video model for leveraging spatial shortcuts to action recognition.

We hypothesize that, when pre-trained on the same dataset with a given architecture, models of larger dynamic score should transfer and generalize better to unseen video domains. To evaluate this hypothesis, we conduct a set of experimental evaluations on the robustness and transferability of the learned video representation. This consists of a) adapting the video network to a set of datasets with different action vocabulary, using linear classification over learned features or fine-tuning; b) few-shot action classification using learned representations directly; and c) applying the classifier on a set of video actions in absence of their spatial context. Experimental results show that DRL significantly improves all three tasks. For example, 5-shot gesture recognition on the Jester dataset [58] is improved by 5% using 3D ResNet architecture [37], and up to 8% with TSM network [55].

The contributions of this paper is three-fold: First, we propose *dynamic score*, a quantitative measure of temporal modeling capacity of video neural networks. Second, we propose *dynamic representation learning* (DRL), a pre-training strategy that aims to optimize dynamic scores for video models. Finally, we present a comprehensive set of experiments to measure the transferability and generalization of learned video representations, which empirically validates the importance of dynamic modeling on video transfer learning and demonstrates the advantage of DRL over baseline pre-training methods.

2. Related Work

Deep video architectures. Following the success of deep neural networks for image recognition, convolutional architectures have dominated video action recognition. While some video neural networks use 2D spatial convolutions similar to those of image CNNs [26, 55, 70, 84], others rely on 3D convolutions operating in the spatiotemporal dimensions [13, 24, 37, 44, 79], or factorize the 3D convolution into a 2D spatial and a 1D temporal convolution [63, 80, 88]. An alternative to convolutions is to use

recurrent modules, such as the Long Short-term Memory (LSTM) [41], to model video dynamics [20, 89]. Attention mechanisms have also been studied to overcome the tendency of convolutional neural networks to favor short-range correlations over long-term dependencies. These include pooling convolutional features with self-attention [16, 85], or in more recent works [1, 5, 23], replacing all convolutional layers with Transformer blocks [21, 83].

Dataset bias. Computer vision datasets are known to exhibit biases, in that their image composition does not accurately resemble the real-world data distributions [47, 76, 77, 82]. In the context of video action recognition, Sigurdsson *et al*. [69] identified the domain gap between videos of human activities retrieved from the internet and our everyday actions. Li *et al*. [53] showed that many internet retrieved datasets have *representation* bias, favoring representations that capture spurious correlations between action labels and contextual cues, such as objects or scenes [46, 49, 73]. New datasets have been collected to overcome this limitation: Charades [69] and VLOG [28] used videos of daily activities, while Diving48 [53] and FineGym [68] considered sports domains rich in fine-grained action categories.

Model bias. Various forms of algorithmic bias have been discovered in a wide range of machine learning tasks. For example, studies have found that *gender* and *racial* bias in datasets can be exploited and sometimes amplified by machine learning models [7, 9, 39, 91]. In image recognition and object detection, *contextual* bias from background objects or scenes have been shown to encourage the learning of models with poor generalization performance on novel test environments [4, 18, 65]. The local connections of convolutional neural networks can also lead to bias towards *short-range* features (e.g. colors, textures) v.s. long-range dependencies (e.g. object shape) [31, 85]. Strategies previously explored to mitigate model bias include resampling of training samples [14, 54], adversarial training [22, 56, 90], construction of adversarial input data [17, 31], or use of regularization losses during learning [2, 11].

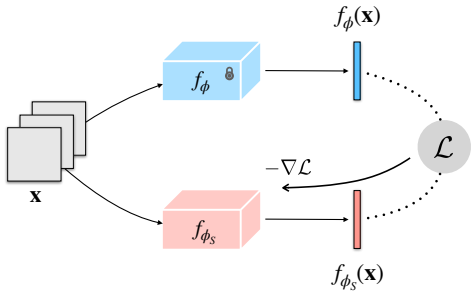
Knowledge distillation. First introduced by Hinton *et al*. [40], knowledge distillation is a method to transfer information from a teacher model to a (usually weaker) student model. Initially introduced as a solution to model compression, the technique has since been adopted to other problems, including adversarial defenses [62] and cross-modal transfer [36].

3. Dynamic Scoring

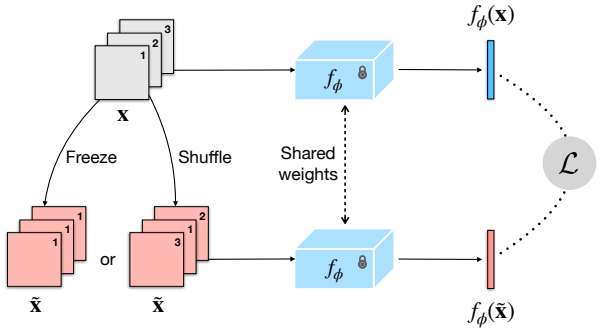
In this section, we introduce the *dynamic score*, a measure of how a representation captures video dynamics.

3.1. Definition

A video representation is a mapping $\phi : \mathcal{X} \to \mathcal{Z}$ from some space $\mathcal{X}$ of videos to a feature space $\mathcal{Z}$. A K-way



(a) Measurement by *knowledge distillation*. A spatial classification model f_{ϕ_s} is trained to predict outputs of f_ϕ. Dynamic score is defined as disagreement between standard model output $f_\phi(\mathbf{x})$ and the spatial model output $f_{\phi_s}(\mathbf{x})$.



(b) Measurement by *input modulation*. Standard model output $f_\phi(\mathbf{x})$ is compared to that of an modulated input $\tilde{\mathbf{x}}$ with temporal information removed, either by freezing the video clip $\mathbf{x}$ at one frame or shuffling its frames.

Figure 2. Measuring the dynamic score $\gamma(f_\phi, p_\mathcal{D})$ of video classifier f_ϕ. Refer to section 3.2 for details.

video classifier is the mapping $f_\phi = h \circ \phi$ composed by a feature representation ϕ and a linear classifier $h : \mathcal{Z} \to \mathbb{S}^K$, where $\mathbb{S}^K$ is the K-dimensional probability simplex. For a generic video classifier that processes video clips of T frames of dimension D, $\mathcal{X} = \mathbb{R}^{T\times D}$. The video classifier is denoted as *purely spatial* if it applies a spatial representation ϕ_s to video frames independently, i.e. if

$$f_{\phi_s} = \frac{1}{T}\sum_{i=1}^{T} g_{\phi_s}(x_i) \tag{1}$$

for some image classifier g_{ϕ_s}. Let $\mathcal{F}^S$ be the set of all such classifiers and $\mathcal{L} : \mathbb{S}^K \times \mathbb{S}^K \to [0, \infty)$ a similarity score for model predictions. The *dynamic score* of model f_ϕ with respect to a distribution $p_\mathcal{D}$ of video clips is then defined as

$$\gamma(f_\phi; p_\mathcal{D}) = \min_{f_{\phi_s}\in\mathcal{F}^S} \mathbb{E}_{\mathbf{x}\sim p_\mathcal{D}} \mathcal{L}(f_{\phi_s}(\mathbf{x}), f_\phi(\mathbf{x})). \tag{2}$$

This is zero when f_ϕ is purely spatial ($f_\phi \in \mathcal{F}^S$) and increases with the ability of the representation to capture video dynamics, i.e. temporal features of the video. While the definition above supports any similarity score between probability distributions, we use the Kullback–Leibler (KL)

divergence

$$\mathcal{L}(\tilde{\mathbf{y}}, \hat{\mathbf{y}}) = \sum_{i=1}^{K} \hat{y}_i \log \frac{\hat{y}_i}{\tilde{y}_i}. \tag{3}$$

In this case, $\gamma(f_\phi; p_\mathcal{D})$ can be intuitively interpreted as the amount of dynamic information captured by ϕ. The definition of dynamic score can also be easily generalized beyond classification problems by application of an appropriate similarity metric $\mathcal{L}$.

3.2. Measuring the dynamic score

The empirical measurement of $\gamma(f_\phi; p_\mathcal{D})$ requires finding the spatial model f_{ϕ_s} of minimum disagreement with f_ϕ, as measured by $\mathcal{L}(\cdot)$. Since it is impractical to search the entire space $\mathcal{F}^S$, we consider a few strategies to find a near-optimal f_{ϕ_s} efficiently.

Knowledge distillation. If the search space $\mathcal{F}^S$ is constrained to deep neural networks of a specific architecture, equation 2 reduces to knowledge distillation [40] from the video teacher model f_ϕ to the spatial student classifier f_{ϕ_s}. As illustrated in figure 2a, the student model is trained to predict the outputs of the video network $f_\phi(\mathbf{x})$; and $\gamma(f_\phi; p_\mathcal{D})$ is the distillation loss on the test set. We use spatial models of the form of equation 1, using a standard 2D convolutional neural network as function g.

Input modulation. Knowledge distillation requires training a student network f_{ϕ_s} different from f_ϕ from scratch. Alternatively, we consider a training free procedure, where the spatial model f_{ϕ_s} is derived from the network f_ϕ itself, by preprocessing its input video clips in a way that removes temporal information. As shown in figure 2b, this can be achieved by either "freezing" the clip, i.e. sampling a single frame and repeating it along the temporal dimension, or by reshuffling the frames in a random order. With $\tilde{\mathbf{x}}$ denoting the frozen or re-shuffled clip, dynamic score is approximated by

$$\gamma_a(f_\phi; p_\mathcal{D}) = \mathbb{E}_{\mathbf{x}\sim p_\mathcal{D}} \mathcal{L}(f_\phi(\tilde{\mathbf{x}}), f_\phi(\mathbf{x})). \tag{4}$$

3.3. Relation to dataset bias

While the dynamic score is a measure for video representations, it is closely related to prior measures of the *spatial bias* of video action datasets, which translates into unexpectedly high action recognition performance by purely spatial classifiers [43, 53]. By replacing the video model output $f_\phi(\mathbf{x})$ in equation 2 with the ground-truth label $\mathbf{y}$, the formula becomes a measure of dataset bias,

$$\gamma(p_\mathcal{D}) = \min_{f_{\phi_s} \in \mathcal{F}^S} \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_\mathcal{D}} \mathcal{L}(f_{\phi_s}(\mathbf{x}), \mathbf{y}), \tag{5}$$

quantifying the difficulty posed by dataset $\mathcal{D}$ to purely spatial recognition models. Analogous to the dynamic score, a higher score $\gamma(p_\mathcal{D})$ indicates that more temporal modeling is required to correctly classify videos of $\mathcal{D}$, as even the optimal spatial classifier performs poorly on the dataset. Importantly, this implies that the dynamic score $\gamma(f_\phi; p_\mathcal{D})$ also reflects the static bias of the data. If the videos of $\mathcal{D}$ comprise mostly spatial cues (i.e. $\mathcal{D}$ has a large static bias), model predictions $f_\phi(\mathbf{x})$ can be easily fitted by a purely spatial model f_{ϕ_s}, resulting in lower $\gamma(f_\phi; p_\mathcal{D})$. In fact, the dynamic score of the oracle classifier f^*, which predicts action class $\mathbf{y}$ with 100% accuracy, is equivalent to the dataset bias, i.e. $\gamma(f^*, p_\mathcal{D}) = \gamma(p_\mathcal{D})$.

Algorithm 1: DRL iteration by adversarial augmentation.

Input: Mini-batch $\mathcal{B} \subset \mathcal{D}$, video model f_ϕ w/ parameters θ, spatial model f_{ϕ_s} w/ parameters ψ; learning rate η, perturbation strength ϵ, distillation weight α, adversarial input weight β

for $(\mathbf{x}, \mathbf{y}) \in \mathcal{B}$ **do**

Optimize spatial model by knowledge distillation

$$\psi \leftarrow \psi - \eta \nabla_\psi \Big[\alpha \mathcal{L}_{\text{kd}}(f_{\phi_s}(\mathbf{x}; \psi), f_\phi(\mathbf{x}; \theta)) + (1-\alpha)\mathcal{L}_{\text{cls}}(f_{\phi_s}(\mathbf{x}; \psi), \mathbf{y}) \Big]; \tag{6}$$

Generate adversarial perturbation $\tilde{\mathbf{x}}$, e.g. using equation 8 for FGSM [33] attacks.

Optimize classification loss of video model on clean and augmented data

$$\theta \leftarrow \theta - \eta \nabla_\theta \Big[\beta \mathcal{L}_{\text{cls}}(f_\phi(\tilde{\mathbf{x}}; \theta), \mathbf{y}) + (1-\beta)\mathcal{L}_{\text{cls}}(f_\phi(\mathbf{x}; \theta), \mathbf{y}) \Big]. \tag{7}$$

end

Output: Updated model parameters (θ, ψ)

4. Dynamic Representation Learning

In this section we discuss two approaches to *dynamic representation learning* (DRL). DRL by *data augmentation* applies adversarial perturbations to the input data to increase the difficulty of pure spatial modeling. DRL by *direct optimization* solves a min-max problem involving the parameters of the video and spatial networks.

4.1. DRL by augmentation

Inspired by the success of adversarial training to improve model robustness [33, 50, 57, 78], this DRL approach is based on the generation of training samples adversarial to the spatial model f_{ϕ_s}. This reduces the effectiveness of spatial modeling, forcing the video network f_ϕ to model video dynamics. Adversarial perturbations are generated by any

Algorithm 2: DRL iteration by min-max optimization.

Input: Mini-batch $\mathcal{B} \subset \mathcal{D}$, video model f_ϕ w/ parameters θ, spatial model f_{ϕ_s} w/ parameters ψ; learning rate η, distillation weight α, dynamic loss weight λ

for $(\mathbf{x}, \mathbf{y}) \in \mathcal{B}$ **do**

Update spatial model by knowledge distillation

$$\psi \leftarrow \psi - \eta \nabla_\psi \Big[\alpha \mathcal{L}_{\text{kd}}(f_{\phi_s}(\mathbf{x}; \psi), f_\phi(\mathbf{x}; \theta)) + (1-\alpha)\mathcal{L}_{\text{cls}}(f_{\phi_s}(\mathbf{x}; \psi), \mathbf{y}) \Big]; \tag{10}$$

Update video model to maximize dynamic score

$$\theta \leftarrow \theta - \eta \nabla_\theta \Big[\mathcal{L}_{\text{cls}}(f_\phi(\mathbf{x}; \theta), y) - \lambda \cdot \mathcal{L}_{\text{kd}}(f_{\phi_s}(\mathbf{x}; \psi), f_\phi(\mathbf{x}; \theta)) \Big]. \tag{11}$$

end

Output: Updated model parameters (θ, ψ)

of the techniques in the adversarial attack literature, such as FGSM [33] or PGD [57]. The perturbation is adversarial to the purely spatial model f_{ϕ_s}, so as to weaken any spatial cues for action recognition. For example, for FGSM, perturbations are generated with

$$\tilde{\mathbf{x}} = \mathbf{x} + \epsilon \, \text{sgn}\Big[\nabla_{\mathbf{z}} \mathcal{L}_{\text{kd}}(f_{\phi_s}(\mathbf{z}), f_\phi(\mathbf{x}))|_{\mathbf{z}=\mathbf{x}} \Big], \tag{8}$$

where ϵ controls the size of the adversarial perturbation and the gradient is only backpropagated through f_{ϕ_s}. As usual for adversarial training, the video model is then trained with a combination of classifications losses on the original and perturbed data

$$\min_{f_\phi \in \mathcal{F}} \mathbb{E}_{\mathbf{x}, y \sim p_\mathcal{D}} \Big[\beta \mathcal{L}_{\text{cls}}(f_\phi(\tilde{\mathbf{x}}), \mathbf{y}) + (1-\beta)\mathcal{L}_{\text{cls}}(f_\phi(\mathbf{x}), \mathbf{y}) \Big], \tag{9}$$

where β is an hyperparameter. To maximize training efficiency, the video f_ϕ and spatial f_{ϕ_s} models are trained jointly, using algorithm 1. At each step, f_{ϕ_s} is first trained to mimic the predictions of f_ϕ by distillation. As usual, this includes a combination of a class label prediction and a KL loss, weighted by a factor α. The adversarial example $\tilde{\mathbf{x}}$ is then generated against f_{ϕ_s} and f_ϕ is finally updated. The whole procedure can be seen as a defense mechanism that forces f_ϕ to learn a dynamic representation.

4.2. DRL by direct optimization

This approach *maximizes* the dynamic score of the video network directly during training. This translates to a min-max game between video model (teacher) and spatial model

Algorithm 3: Fast DRL iteration by min-max approximation.

Input: Mini-batch $\mathcal{B} \subset \mathcal{D}$, video model f_ϕ, learning rate η, dynamic loss weight λ

for $(\mathbf{x}, \mathbf{y}) \in \mathcal{B}$ **do**

Create a frozen or reshuffled copy $\tilde{\mathbf{x}}$ of $\mathbf{x}$;

Update video model to maximize dynamic score

$$\theta \leftarrow \theta - \eta \nabla_\theta \Big[\mathcal{L}_{\text{cls}}(f_\phi(\mathbf{x}; \theta), y) - \lambda \cdot \mathcal{L}_{\text{kd}}(f_\phi(\tilde{\mathbf{x}}; \psi), f_\phi(\mathbf{x}; \theta)) \Big|_{\psi=\theta} \Big]. \tag{12}$$

end

Output: Updated model parameters θ

(student)

$$\min_{f_\phi \in \mathcal{F}} \mathbb{E}_{\mathbf{x}, y \sim p_\mathcal{D}} \mathcal{L}(f_\phi(\mathbf{x}), y) - \lambda \cdot \gamma(f_\phi; p_\mathcal{D}) = \min_{f_\phi \in \mathcal{F}} \max_{f_{\phi_s} \in \mathcal{F}^S} \mathbb{E}_{\mathbf{x}, y \sim p_\mathcal{D}} \big\{ \mathcal{L}_{\text{cls}}(f_\phi(\mathbf{x}), y) - \lambda \cdot \mathcal{L}_{\text{kd}}(f_{\phi_s}(\mathbf{x}), f_\phi(\mathbf{x})) \big\}. \tag{13}$$

As described in algorithm 2, we adopt a training procedure similar to that of adversarial networks [29, 30, 32], in this case alternating between updating the parameters of f_ϕ and f_{ϕ_s}. The maximization over f_{ϕ_s} optimizes the spatial model f_{ϕ_s} to mimic the predictions of the video model f_ϕ as closely as possible. The minimization over f_ϕ then forces the latter to produce predictions that are as different as possible from those of the former, while minimizing classification error, leading f_ϕ to learn video dynamics. This adversarial objective has a similar form to the ReBias algorithm [2]. The two methods differ primarily in the similarity criteria between spatial and temporal models: While [2] optimized the Hilbert-Schmidt Independence Criterion (HSIC) between *feature* spaces learned by the de-biased and biased models, DRL uses the distillation loss, i.e. KL divergence, between their *output* probabilities. We believe that this allows the proposed dynamic score to explicitly capture the agreement between the spatial and dynamic models in classifying the action categories.

Similarly to the score computation of section 3.2, the distillation of the static model can be replaced by a training free approach, where f_{ϕ_s} is derived from f_ϕ itself by preprocessing the input video to remove temporal information. In this case, algorithm 3 is used.

5. Experiments

In this section, we perform dynamic scoring of existing video recognition networks and evaluate how DRL improves their ability to model video dynamics.

Training mode	Dataset	# classes	# examples
Pre-training	Kinetics [46]	400	300k
	miniKinetics [88]	200	85k
Fine-tuning	UCF-101 [73]	101	13k
	HMDB [49]	51	5k
	Diving-48 [53]	48	18k
Few-shot recognition†	Sth-Sth V2 [34]	174	220k
	Jester [58]	27	148k
Domain generalization	Mimetics [87]	50	700

Table 1. **Datasets for pre-training and evaluation.** †5 classes are randomly drawn during each few-shot learning episode. Few-shot training sets consist of 5 (1-shot) or 25 (5-shot) examples. Here we report the *total* number of classes and samples of the full datasets.

5.1. Experimental setup

Datasets. We adopt the practice of pre-training video recognition models on the Kinetics-400 [46] dataset. For preliminary dynamic scoring experiments (sect. 5.2), the models are scored on a 200-class subset of Kinetics, miniKinetics [88], to reduce training time. As summarized in Table 1, the impact of DRL on model transfer (sect. 5.3) is evaluated in six datasets of different action domains. UCF-101 [73], HMDB [49], Diving-48 [53] are used to evaluate transfer by fine-tuning, while Something-Something v2 [34] and Jester [58] are used to evaluate few-shot classification. We do not test these two datasets for standard classification due to their large size, which diminishes the need for model pretraining. We also evaluate domain generalization of trained classifiers without any fine-tuning on the Mimetics [87] test set, which shares action vocabulary with Kinetics but has weaker spatial biases.

Models. DRL is evaluated on video action recognition networks with different types of convolutional modules: 3D ResNet [37] uses 3D convolution kernels, while TSM [55] is based on 2D convolutions. Sampling frame rate is adaptively chosen to ensure a constant length for input clips of 1 second, so that dynamic scores are comparable across models. Detailed training procedure included in supplemental.

5.2. Dynamic scoring

Table 2 summarizes the dynamic scores of networks learned on miniKinetics [88]. The Baseline is a model trained by standard cross-entropy minimization. Dynamic scores are measured by either the distillation score of equation 2 (*Distill*), or the approximate score of equation 4, based on removing temporal information from input clips (*Freeze* and *Shuffle*). DRL is implemented with adversarial augmentation (algorithm 1), direct optimization (algorithm 2), or its approximation (algorithm 3) based on clip freezing or shuffling. Several conclusions can be drawn from the table. First, all DRL methods are effective at improving the dynamic score γ of the baseline. The gains are larger for DRL by direct optimization, which frequently doubles the baseline score. Second, as usual for min-max games, the optimization of algorithm 2 is not easy. We found approximate DRL using frozen clips to be easier to train and, as shown in the table, lead to the best dynamic score. The same does not hold for approximate DRL with frame shuffling, which has the worst performance of the three direct optimization techniques. We believe that this is because frame shuffling creates artificially high temporal frequencies that the video network cannot model. Approximate DRL by direct optimization with frozen clips is used in all subsequent experiments. Third, regarding dynamic score measures, while different approaches produced different values of γ, the relative order of the different learning methods remained constant. While knowledge distillation provides the smallest (hence the most accurate) estimate of dynamic score, it requires training a new image convnet from scratch. Since measurement by freezing or shuffling input clips merely requires a single forward pass through the test set, it is much more efficient to evaluate.

Method		Dynamic score $\gamma(f_\phi, p_\mathcal{D})$ Distill	Freeze	Shuffle
Baseline		0.33	1.06	0.74
DRL (Adv. augment)	FGSM, $\epsilon = 8$	0.45	1.25	0.88
	PGD, $\epsilon = 8$	0.45	1.26	0.89
DRL (Min-max opt.)	Distill, $\lambda = 0.5$	<u>0.61</u>	<u>1.79</u>	<u>1.29</u>
	Freeze, $\lambda = 0.5$	**0.62**	**2.21**	**1.70**
	Shuffle, $\lambda = 0.5$	0.54	1.26	1.23

Table 2. **Dynamic scores** of 3D ResNet-18 models trained on miniKinetics-200, measured by distillation or freezing/shuffling clips. "Baseline" denotes standard cross-entropy training; DRL variants are discussed in sect. 4. Best results in **bold**, runners-up <u>underlined</u>.

5.3. Transfer learning

Many-shot recognition. Table 3 compares transfer learning performance from Kinetics to the UCF-101 [73], HMDB [49] and Diving-48 [53] datasets, using either *linear evaluation* over pretrained representations or full network *fine-tuning*. We compare the DRL models, trained by approximate optimization with frozen clips (algorithm 3), to baseline cross-entropy pre-training on Kinetics. Fine-tuning accuracies are also compared to networks initialized from scratch (3D ResNet) or ImageNet weights (TSM). We observe the following: First, transfer learning from Kinetics significantly outperforms training from scratch or with ImageNet initialization. This a well known consequence of the small target dataset sizes and confirms the importance of Kinetics pre-training for many action recognition datasets in the literature. Second, DRL is quite effective at increasing the dynamic score of all models, frequently doubling or even tripling the score of the baseline. Third, DRL achieves top performance in both the linear setting

Method	Architecture input	Pretrain acc. K400	$\gamma(f_\phi, p_\mathcal{D})$	Linear acc. UCF	 HMDB	 Diving	Fine-tuning acc. UCF	 HMDB	 Diving
Scratch	3D ResNet-18 112x112x16	–	–	–	–	–	59.08	24.12	47.82
Baseline		56.40	0.81	83.27	52.29	9.95	87.34	61.24	61.22
DRL		53.32	1.34	**84.30**	**55.23**	**11.32**	**87.36**	**63.59**	**63.15**
Scratch	3D ResNet-50 112x112x16	–	–	–	–	–	46.55	20.00	33.96
Baseline		62.04	0.68	87.23	59.54	16.29	89.21	64.58	67.92
DRL		59.92	1.63	**88.24**	**61.83**	**16.55**	**90.88**	**64.64**	**68.83**
ImageNet	TSM ResNet-18 224x224x8	–	–	62.75	36.14	9.29	82.55	51.90	71.07
Baseline		64.55	0.47	73.83	46.73	12.49	**92.23**	64.64	72.74
DRL		62.20	1.46	**77.13**	**50.78**	**13.40**	91.25	**65.49**	**73.96**
ImageNet	TSM ResNet-50 224x224x8	–	–	66.98	37.58	10.96	87.02	55.36	74.97
Baseline		71.19	0.45	82.13	54.64	17.01	**95.14**	69.41	77.56
DRL		68.75	0.96	**84.91**	**58.04**	**22.34**	95.03	**72.29**	**79.04**

Table 3. **Linear classification and fine-tuning performance** of Kinetics models on UCF-101, HMDB-51 and Diving-48. By improving dynamic score $\gamma(f_\phi, p_\mathcal{D})$, DRL produces transferable video representations, both for linear discrimination and fine-tuning.

Method	Architecture input	Sth-Sth v2 1-shot	 5-shot	Jester 1-shot	 5-shot
Baseline	3D ResNet-18	30.34	40.72	27.42	39.12
DRL	112x112x16	**31.24**	**43.46**	**31.79**	**44.60**
Baseline	3D ResNet-50	31.79	44.40	27.90	43.01
DRL	112x112x16	**33.85**	**46.75**	**33.24**	**47.58**
Baseline	TSM ResNet-18	30.96	41.67	28.21	39.07
DRL	224x224x8	**31.90**	**44.00**	**32.44**	**47.00**
Baseline	TSM ResNet-50	31.32	42.29	27.38	38.71
DRL	224x224x8	**32.28**	**45.02**	**31.85**	**47.04**

Table 4. **Cross-domain few-shot evaluation** of video representations on Something-Something V2 and Jester. Pre-training on Kinetics; 5-way accuracies reported in all experiments.

and with fine-tuning for almost all combinations of model and target dataset. In general, the gains with fine-tuning are smaller than that for linear classification. This is expected, as network fine-tuning can correct some of the biases of the pre-trained model. This is especially true for UCF-101, which shares very similar biases with the Kinetics training set, as both are constructed from the same data source (YouTube), and consist of similar action classes. This leads to high accuracies under the transfer setting, and smaller improvements from DRL. The gains of DRL are consistent on HMDB and Diving, where video sources and action vocabularies diverge sufficiently from Kinetics. Overall, DRL improves linear classification accuracy by 2.39% on average, and fine-tuning accuracy by 1.03%.

We also observe that an increased dynamic score does not translate into accuracy gains on the Kinetics dataset, where DRL models underperform baseline pre-training by 2–3%. This is expected since the training and test set have the same spatial biases, such that models trained to exploit all biases achieve higher accuracy. Nonetheless, when evaluated under a distribution shift, these spatial models tend to transfer worse than models trained with DRL objective.

Method	Baseline	Learned-Mixin [19]	RUBi [11]	ReBias [2]	DRL
Accuracy@1	18.9	11.4	13.4	22.4	**26.4**

Table 5. **Domain generalization** accuracy of Kinetics-pretrained 3D ResNet-18 models, evaluated on 10 classes of Mimetics.

Few-shot recognition. To evaluate the transferability of the representation *as is*, we consider the task of *cross-domain* few-shot classification. That is, a network pre-trained on Kinetics (*base*) classes, is used in another dataset (*novel* classes) to extract features that are fed to a few-shot action recognizer. The target datasets are Something-Something V2 [34] and Jester [58]. This setting presents a greater challenge than most prior work on few-shot video classification (e.g. [12, 92]), as base and novel classes are sampled from different datasets, creating a domain shift which has been found to degrade few-shot learning performance [15]. Few-shot classification is implemented using the Baseline method by Chen *et al*. [15], which optimizes a linear classifier with the support data in each episode and evaluates it on the query videos. Both 1-shot and 5-shot scenarios are tested, and 5-way accuracy is reported. Table 4 shows DRL outperforms the baseline for all combinations of model and target dataset. It also achieves large gains for most networks on the two target datasets, especially for 5-shot learning. For example, using the TSM ResNet-18 architecture we observed an improvement of 8% on 5-shot action recognition on the Jester dataset. In this setting, three DRL models achieve a rate above 47%, in comparison to a single baseline model above 40%. For Sth-Sth the gains were smaller, but between 2-3% for most models. These results support the conclusion that, despite the large size of Kinetics, models trained on this dataset are somewhat overfitted to spatial cues that do not generalize to other recognition datasets. Reducing this spatial bias, as is done by DRL, increases model robustness and transfer performance.

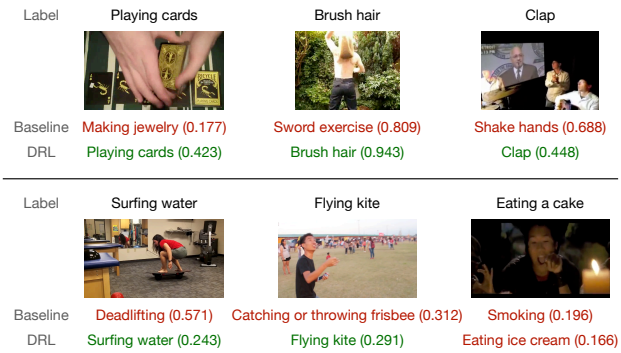


Figure 3. **Qualitative results** on Kinetics, HMDB and Mimetics videos. DRL performs better on actions out of context.

Domain generalization. We finally evaluate the video action classifiers trained on Kinetics, without fine-tuning, on the Mimetics [87] test set, which contains videos of mimed actions corresponding to 50 Kinetics classes. This leads to a test domain that shares similar action dynamics with the training data, but without co-occurring objects and scenes, reducing the advantages of spatially biased models. We follow the setup of [2] which uses 10 classes of Mimetics for evaluation. As shown in table 5, 3D ResNet-18 models trained with DRL outperforms cross-entropy pretraining baseline and prior debiasing methods by significant margin.

Qualitative examples. A closer look at model predictions in figure 3 reveal that baseline models tend to fail on actions that take place in an unfamiliar environment (e.g. "brushing hair" outdoors, or "surfing" indoors). DRL is capable of correcting these predictions by reducing the reliance on contextual cues (scene, objects etc.) and guiding the network to focus on temporal dynamics of the action itself.

5.4. Ablation studies

Dynamic loss. Figure 4 shows the influence of dynamic loss weight λ on the transfer accuracy of learned models, measured by linear probing. We compare three DRL training methods, using either direct min-max optimization (*Distill*, alg. 2) or min-max approximation by removing temporal information from input clips (*Freeze* and *Shuffle*, alg. 3). As the figure depicts, direct min-max optimization suffers from difficult optimization landscape, failing to converge for DRL weight $\lambda > 0.5$, and unable to improve representation quality for smaller values of λ. Shuffling approximation is effective at improving linear classification of learned representations for small λ, but the trend reverses as regularization weight increases, eventually leading to poorer results for $\lambda \geq 0.5$. This is likely due to the fact that frame shuffling introduces artifacts of high temporal frequency, which the convolutional networks cannot model. In contrast, min-max approximation by freezing input clips consistently benefits from greater DRL weight λ, gaining up to 5% in linear classification accuracy on HMDB, 3% on UCF and 1.5% on Diving. These results corroborate the finding from table 2 that DRL with frozen clips also lead to the greatest improvement in dynamic score of trained models.

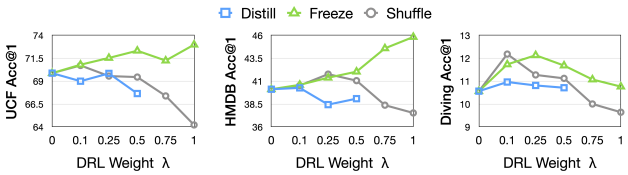


Figure 4. Linear evaluation accuracy of 3D ResNet-18 model as function of loss weight λ. All models pretrained on miniKinetics; $\lambda = 0$ corresponds to standard cross-entropy training.

	$\gamma(f_\phi, p_\mathcal{D})$	**K400** (in-dist.)	**UCF**	**HMDB**	**Diving**
Baseline	0.289	55.78	**84.35**	55.75	10.30
DRL	0.421	54.76	84.30	**57.06**	**12.34**

Table 6. Dynamic scores and linear evaluation accuracies of two-stream network with ResNet-18 backbone and 5 optical flow frames, under baseline and DRL pre-training.

Input modalities. While the main results have demonstrated that DRL reduces spatial bias and improves transferability of video CNNs that operate on raw RGB inputs, we further experiment with networks that utilize additional modalities to capture motion information explicitly. Table 6 compares the dynamic score and transfer performance of a vanilla two-stream network [70], under different pretraining strategies. It can be observed that two-stream networks suffer from the same type of bias in RGB-based models. This shows that although optical flow streams can be integrated to the networks to enhance motion capturing, they would not explicitly remove spatial bias from the appearance stream. DRL improves dynamic score of the network as well as transfer accuracy on HMDB and Diving, while achieving comparable performances to baseline on UCF.

6. Discussion and Conclusion

In this work we introduced dynamic score, a new measure of temporal dynamics modeling in video convolutional networks. Various methods to evaluate the dynamic score were discussed, either through knowledge distillation to a 2D spatial convnet or by preprocessing input clips to redact temporal information. We also proposed dynamic representation learning (DRL) to improve the dynamic score of networks, using adversarial perturbation or min-max optimization. The so trained video classifiers and their learned representations are empirically evaluated on a set of downstream tasks (linear classification, fine-tuning, few-shot recognition, domain generalization), with results demonstrating a clear benefit of enhanced temporal representation learning.

Acknowledgements. This work was funded in part by NSF awards IIS-1924937, IIS-2041009, a gift from Amazon, and a gift from Qualcomm. We acknowledge the Nautilus platform for some of the experiments discussed above.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 3

[2] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 3, 5, 7, 8

[3] Pedro Ballester and Ricardo Araujo. On the performance of googlenet and alexnet applied to sketches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. 1

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018. 3

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

[6] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1395–1402. IEEE, 2005. 1

[7] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*, 2016. 3

[8] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019. 1

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 3

[10] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[11] Remi Cadene, Corentin Dancette, Hedi Ben-Younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. *arXiv preprint arXiv:1906.10169*, 2019. 3, 7

[12] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10627, 2020. 7

[13] 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

[14] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 3

[15] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019. 7

[16] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A^2-nets: Double attention networks. *arXiv preprint arXiv:1810.11579*, 2018. 3

[17] 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. *arXiv preprint arXiv:1912.05534*, 2019. 1, 3

[18] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012. 3

[19] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019. 7

[20] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 1, 3

[21] 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. In *International Conference on Learning Representations*, 2021. 3

[22] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015. 3

[23] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021. 3

[24] 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. 2

[25] 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

[26] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 2

[27] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 1

[28] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4991–5000, 2018. 3

[29] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 5

[30] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pas-

cal 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. 2, 5

[31] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1, 2, 3

[32] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 5

[33] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 4, 5

[34] 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. 6, 7

[35] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 1

[36] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 3

[37] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 2, 6

[38] 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. 1

[39] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018. 3

[40] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3, 4

[41] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1, 3

[42] 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. 2

[43] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7366–7375, 2018. 4

[44] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 2

[45] 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. 1

[46] 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, 6

[47] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012. 3

[48] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[49] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 3, 6

[50] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 4

[51] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 2

[52] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 2

[53] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 1, 3, 4, 6

[54] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019. 1, 3

[55] 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, 6

[56] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*,

pages 3384–3393. PMLR, 2018. 3

[57] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 4, 5

[58] 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, 6, 7

[59] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 1

[60] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017. 2

[61] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013. 2

[62] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016. 3

[63] 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

[64] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017. 1

[65] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018. 3

[66] Christian Schuldt, Ivan Laptev, and Barbara 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. IEEE, 2004. 1

[67] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 535–544, 2021. 1

[68] 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. 3

[69] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 3

[70] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 1, 2, 8

[71] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[72] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017. 1

[73] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3, 6

[74] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 1

[75] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[76] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017. 3

[77] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 3

[78] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 4

[79] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1, 2

[80] 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

[81] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2

[82] Emiel Van Miltenburg. Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083*, 2016. 3

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3

[84] 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. 1, 2

[85] 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. 2, 3

[86] Daniel Weinland, Edmond Boyer, and Remi Ronfard. Action recognition from arbitrary views using 3d exemplars. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–7. IEEE, 2007. 1

[87] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision*, 129(5):1675–1690, 2021. 6, 8

[88] 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. 1, 2, 6

[89] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 1, 3

[90] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 3

[91] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017. 3

[92] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 751–766, 2018. 7