Mitigating Representation Bias in Action Recognition: Algorithms and Benchmarks

Haodong Duan[1,3], Yue Zhao[2], Kai Chen[3], Yuanjun Xiong[4], and Dahua Lin[1]

[1] The Chinese University of Hong Kong
[2] The University of Texas at Austin
[3] Shanghai AI Lab
[4] Amazon AI

Abstract. Deep learning models have achieved excellent recognition results on large-scale video benchmarks. However, they perform poorly when applied to videos with rare scenes or objects, primarily due to the bias of existing video datasets. We tackle this problem from two different angles: algorithm and dataset. From the perspective of algorithms, we propose Spatial-aware Multi-Aspect Debiasing (**SMAD**), which incorporates both *explicit* debiasing with multi-aspect adversarial training and *implicit* debiasing with the spatial actionness reweighting module, to learn a more generic representation invariant to non-action aspects. To neutralize the intrinsic dataset bias, we propose **OmniDebias** to leverage web data for joint training selectively, which can achieve higher performance with far fewer web data. To verify the effectiveness, we establish evaluation protocols and perform extensive experiments on both re-distributed splits of existing datasets and a new evaluation dataset focusing on the action with rare scenes. We also show that the debiased representation can generalize better when transferred to other datasets and tasks.

1 Introduction

Human beings have cognitive bias, and so do the machine learning systems [36]. Human cognitive bias comes from the uniqueness of individual experiences (learning materials) and the tendency of brains to simplify information processing [25]. Machine learning systems are biased for similar reasons. First, the datasets used for training can be intrinsically biased: *e.g.*, sampled from a shifted distribution [44] or collected with a pre-defined ontology [38]. Even if the dataset faithfully represents the real world, there is human bias in the real world which we do not want the machine learning system to exploit, *e.g.*, gender bias [5,6]. Mitigating the bias in machine learning systems has long been a challenging yet valuable research area [3, 17, 19].

In computer vision, the efforts for building datasets that faithfully represent the real visual world never end. Better data collection and labeling strategies [14, 38, 43, 57] are designed for building less biased datasets from scratch. Besides, various tools can be applied to a built dataset (visual [47] or tabular [4] data) to detect and mitigate unwanted bias. In action recognition, [34] introduces the concept of representation bias and attempts to reduce it throughout dataset construction. However, the dataset they propose is on a small scale and in a narrow domain. We investigate existing large-scale

datasets instead and quantify the representation bias by designing different train-test splits and analyzing the performance gaps.

Besides, we propose **OmniDebias**, which uses external web media as auxiliary data to mitigate the dataset bias. On the one hand, the diversity of web data provide us with rich examples that are uncommon in existing datasets, which makes it a suitable data source for debiasing. On the other hand, web data are also severely biased to some factors, *e.g.*, *scene*. OmniDebias adopts a simple yet effective data selection strategy to sample a less biased subset from the entire dataset. Co-training with the selected subset outperforms the vanilla co-training both in performance and data efficiency.

Though effective in debiasing, constructing 'unbiased' datasets can be difficult and may cost lots of human labor, while designing debiasing algorithms is a much cheaper alternative. A series of works [35, 53] aim at devising algorithms to mitigate the bias in the learned representation, preventing the algorithms from amplifying the bias in training data. In particular, SDN [11] proposes to mitigate *scene* bias in action recognition with adversarial training and human mask confusion loss. Previous works usually restrict the debiasing algorithm to a specific factor. In the real world, the bias in the dataset can be complex and non-trivial to understand. To deal with more complicated dataset bias, we extend the single-factor adversarial training to a multi-aspect fashion, which shows better generalization capability.

To mitigate the *generic* representation bias in action recognition, we propose a spatial-aware multi-aspect debiasing framework (**SMAD**). A video can have multiple facets besides the action label, such as the background *scene* or the *object* that people interact with. Video datasets collected for different purposes may emphasize different facets. Considering this characteristic, we propose multi-aspect adversarial training (**MAAT**) to enforce the model invariant to these *non-action* facets. We also introduce Spatial-Aware Actionness Reweighting (**SAAR**) to ensure that the model learns where to focus to recognize action without being affected by features related to other facets. The framework **SMAD** proves to be generic for videos with various kinds of bias and does not depend on extra knowledge of specific datasets.

To fairly exhibit the effectiveness of the proposed debiasing algorithm, we devise a series of evaluation protocols. First, for the existing large-scale dataset Kinetics-400 [7], we re-distribute the original train splits by either *scene* or *object* such that the hidden facet does not overlap between the *re-distributed* train and test sets (**facet-based re-distribution**). Second, we collect an additional Action with RAre Scene (ARAS) dataset[5] for evaluation to simulate the **out-of-distribution** setting. Third, we follow the routine of measuring the debiasing effect by transferring the learned model to downstream tasks (**downstream-task transferring**), such as feature classification, few-shot learning, and finetuning on other datasets (as is proposed by [11]).

Our contributions are three-fold: 1. We propose SMAD, which considers multiple aspects in adversarial training and achieves better performance when complex bias exists in the training set. 2. We propose OmniDebias, which exploits the richness and diversity of web data effectively and efficiently. 3. We evaluate our method on both conventional evaluation protocols (downstream-task transferring) as well as the new

[5] Dataset released at https://github.com/kennymckormick/ARAS-Dataset.

ones (facet-based re-distribution, out-of-distribution testing). The improvements of our methods on all three benchmarks are consistent and remarkable.

2 Related Work

Action Recognition. Action recognition aims at recognizing human activities in videos. Following the success of deep learning in the image domain, two series of deep ConvNets become the mainstream architectures for action recognition, named 2D-CNN and 3D-CNN methods. 2D-CNN methods like Two-Stream [41] and TSN [49] are lightweight while lacking temporal modeling capability to some extent. 3D-CNN methods [7, 18, 45, 46] use 3D convolutions for temporal modeling and achieve the state-of-the-art on large-scale benchmarks like Kinetics-400 [7]. In this paper, we show that both architectures are vulnerable to biases. Our proposed framework can help to mitigate this problem.

Mitigating Dataset Bias. All datasets, more or less, have dataset bias. In computer vision, [44] studies 12 widely used image datasets and finds their data are of different domains and distant from the real visual world. In natural language processing, gender bias occurs in corpus collected from social media and news [22, 29]. There are two main approaches to mitigate dataset bias: The first is to design better data collection and labeling strategies [38] or to calibrate the existing dataset with bias detection tools [27, 47]. The second is to compensate dataset bias with domain adaptation techniques [20, 30, 37]. In this paper, following the first approach, we propose to use diversified web media to neutralize the dataset bias.

Mitigating Algorithm Bias. Even if the dataset faithfully represents the real world, bias still exists. Due to human bias, real-world data may bias towards specific factors, while discriminative models even amplify such unwanted bias [54]. In machine learning, it is intuitive to add constraints or regularizations for the pursued fairness metric to the existing optimization objective [1, 50, 52]. However, most of these approaches are intractable in deep learning. Meanwhile, adversarial training has broader applications both in machine learning and deep learning. [53] use adversarial debiasing for bias mitigation, but the bias factor is required to be known beforehand. [33] propose to use adversarial example reweighting and achieves good performance on debiasing action recognition.

Domain adaptation. Domain adaptation (DA) aims at learning well-performing models on the target domain with training data from the source domain. To that end, many works try to find a common feature space for the source and target domains via adversarial training, both for image tasks [10, 21, 39] and action recognition [9, 12, 13]. The setting of debiasing is similar to, but not the same as DA. The main difference is that we have no access to testing videos during training. Besides, the debiasing setting does not assume the amount of testing data, while DA algorithms require a certain number of testing videos to determine the data distribution.

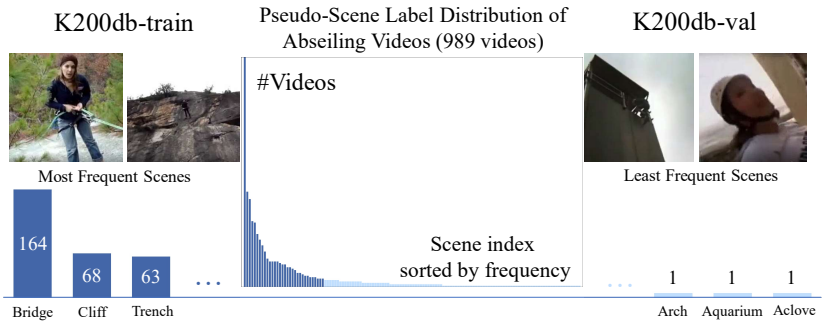

Fig. 1: **The long-tailed *scene* distribution of abseiling videos.** Most videos belong to several scene categories. In the distribution tail, there are many scene categories that rarely occur in training videos. We sample videos from the distribution head to form `K200db-train`, from the tail to form `K200db-val`.

3 Formulation

Following [53], the problem can be formulated as follows. For a video recognition dataset D, we can view each data sample as a tuple (x, y, z) drawn from the joint distribution (X, Y, Z), where x denotes the video, y denotes the *action* label, z denotes one or multiple *non-action* labels, such as *scene*, *object* or other attributes. We consider the supervised learning task, which builds a predictor $\hat{Y} = f(X)$ for Y given X.

Due to the dependence of Y and Z in the training set, the predictors learned via standard supervised learning also yield predictions $\hat{Y}$ dependent on Z given the action label Y. Such behavior will lead to poor generalization capability, severely undermine the testing performance if $P(Z|Y)$ differs a lot between the train and test split.

Our goal is to learn non-discriminatory action recognition models *w.r.t.* Z, which generalize well to testing videos with factors (*scene*, *object*, *e.g.*) that rarely appear in the training set. Non-dicrimination criterias have been of three types in fairness literature [2], *independence* ($Y' \perp Z$), *seperation* ($Y' \perp Z|Y$) and *sufficiency* ($Y \perp Z|Y'$). In the context of video recognition, we pursue EQUALIZED ODDS, similar to *separation*, which is to minimizing the variance of $P(\hat{Y} = y|Y = y, Z = z)$ for different z given y. Besides improving the z-unbiasedness, we also need to maximize $P(\hat{Y} = y|Y = y)$ to secure a good recognition model.

4 Evaluation Benchmark

4.1 Crafting Evaluation Datasets

Most existing datasets assume the joint distribution $P(Y, Z)$ identical between train and validation splits. To find out if an action recognition model is biased towards the *non-action* labels, we design two evaluation protocols based on Kinetics-400 [7]: re-distributing the existing train-val split and constructing a new validation set.

Re-distributing Train-Val Split. We start with the original Kinetics-400 train split (with $\sim$240k videos). We apply a ResNet-50 trained on Places-365 [57] to obtain the pseudo *scene* labels. As shown in Figure 1, the pseudo *scene* labels have a long-tailed

Fig. 2: **Samples of ARAS dataset.**[6]

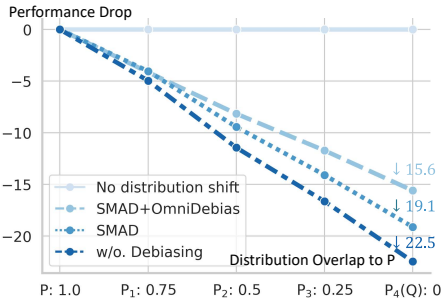

Fig. 3: **The Top1-Acc severely drops along with the scene distribution shift.**

TSN	Top1 Acc
K200-val	76.2
K200-unbias	53.7 ↓ 22.5
ARAS-64	55.8 ↓ 20.4
SlowOnly	**Top1 Acc**
K200-val	75.1
K200-unbias	51.9 ↓ 23.2
ARAS-64	51.0 ↓ 24.1

Table 1: **Top1-Acc of TSN and SlowOnly on 3 test sets.**

distribution. We take the tail as the validation set and sample a subset from the head to be the training set. To maintain the inter-class sample balance, we select 200 classes with the most training samples and construct a subset that contains 80k videos for training and 10k for validation, (denoted as `K200db-train` and `K200db-val`, `db` for debiasing). We examine the *action-scene* correlation of the two splits by calculating the normalized mutual information (NMI) of *action* and *scene*: for `K200db-train`, the NMI is 0.466 (0.397 if sampled randomly); for `K200db-val`, the NMI is 0.374 (0.488 if sampled randomly). Based on the splitting method, we can also tune the overlap of common *scene* labels in `K200db-train` and `K200db-val` for varying distribution shift.

Constructing New Validation Set. Beyond being restricted to the original dataset, we can further construct a new dataset for evaluation. This resembles the real-world scenario: the trained model is fixed while the environment changes at deployment. We begin with *action* labels in Kinetics and consider some *rare scenes*. The combinations of *actions* and *rare scenes* are used as queries to obtain web videos from YouTube. We manually examine the web videos and obtain around ten videos for each class in 104 Kinetics classes, denoted as Action with RAre Scenes (`ARAS-104`). For `K200db`, there are 64 overlapped classes (`ARAS-64`). Figure 2 shows several examples. We use `ARAS` to simulate the out-of-distribution testing for *scene*-debiasing evaluation.

[6] ARAS video samples in: https://youtu.be/j1LA3y-UuEA.

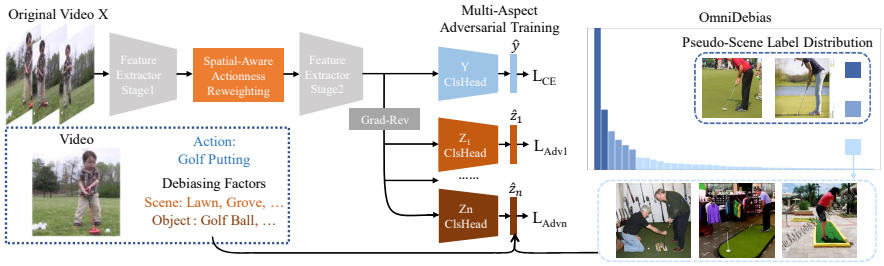

Fig. 4: **SMAD & OmniDebias. Left: SMAD framework**. Multiple adversarial heads are used in SMAD for Multi-Aspect debiasing. The SAAR module (Figure 5) is inserted in the backbone to improve the spatial modeling capability. **Right: OmniDebias**. OmniDebias only uses the `unbias` part of web media for joint training, achieving better performance and efficiency.

4.2 Evaluation of Existing Methods

We first evaluate existing methods on the new benchmarks, including a 2D-CNN method (TSN-3seg-R50) [49] and a 3D-CNN one (SlowOnly-8x8-R18) [18]. From Table 1, we observe that the Top-1 accuracies on both `K200db-val` and `ARAS-64` are significantly lower than the original validation split `K200-val`. This reflects models learned with vanilla training cannot handle the large discrepancy of the *action-scene* joint distribution between train/val splits. `K200db-[train/val]` is an extreme case that has disjoint scene labels. We can also vary the overlap of scene labels between `K200db-val` and `K200db-train`. Figure 3 demonstrates that the drop of accuracy is positively correlated to the distribution shift. That performance drop can be largely mitigated by **SMAD** and **OmniDebias**, which will be detailed in the following section.

5 Method

We devise Spatial-aware Multi-Aspect Debiasing (**SMAD**) which seeks to learn a representation invariant to multiple aspects of videos, *e.g.*, *scene*, *object*, and other attributes, with adversarial training. Besides, we propose OmniDebias to harness the richness and diversity of web data efficiently, to improve the expressive power of the learned representation. We integrate the two complementary aspects into a unified framework, as illustrated in Figure 4.

5.1 Spatial-aware Multi-Aspect Debiasing

SMAD incorporates both **explicit** debiasing using Multiple Aspects as Adversarial Training objectives (MAAT) and **implicit** debiasing with Spatial-Aware Actionness Reweighting (SAAR).

Multi-Aspect Adversarial Training. We denote each input as a tuple $(x, y, z_1, \cdots, z_M) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_M$, where we pre-define M aspects in addition to the set of *action* labels $\mathcal{Y}$. We use a ConvNet f_Θ parameterized by Θ for feature extraction. On top of f_Θ are $(M+1)$ classification heads: one head $h_\mathcal{Y}$ (parameterized by $\theta_\mathcal{Y}$) to predict the *action* y and M adversarial heads $h_{\mathcal{Z}_i}$ (parameterized by $\theta_{\mathcal{Z}_i}$) to recognize

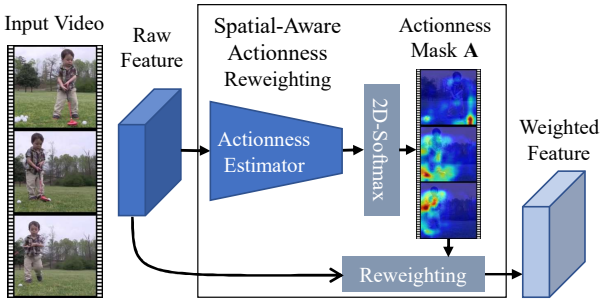

Fig. 5: **SAAR module.** The Spatial-Aware Actionness Reweighting module learns an actionness mask $\mathbf{A}$ to reweight features at different locations. The learned mask has low values on irrelated scenes or objects to suppress these features.

tags belong to the aspect $\mathcal{Z}_i$. We use the standard cross-entropy loss $L_{\text{ce},\mathcal{Y}}$ to train $h_{\mathcal{Y}}$, use adversarial losses $L_{\text{adv},\mathcal{Z}_i}$ for the rest *non-action* heads $h_{\mathcal{Z}_i}$.

The optimization can be divided into two parts: classification heads and the backbone. For classification heads, the objective is to minimize $L_{\text{ce},\mathcal{Y}}$ and $L_{\text{adv},\mathcal{Z}_i}$:

$$\theta_{\mathcal{Y}}, \theta_{\mathcal{Z}_1}, \cdots \theta_{\mathcal{Z}_M} = \underset{\theta_{\mathcal{Y}}, \theta_{\mathcal{Z}_1} \cdots \theta_{\mathcal{Z}_M}}{\operatorname{argmin}} (L_{\text{ce},\mathcal{Y}} + \sum_{i=1}^{i=M} \lambda_i L_{\text{adv},\mathcal{Z}_i}). \tag{1}$$

λ_i is the weight of the adversarial loss. For the backbone, since we aim for feature that is both discriminative for $\mathcal{Y}$ and invariant for Z_i, the objective is to minimize $L_{\text{ce},\mathcal{Y}}$ and *maximize* $L_{\text{adv},\mathcal{Z}_i}$:

$$\Theta = \underset{\Theta}{\operatorname{argmin}} (L_{\text{ce},\mathcal{Y}} - \sum_{i=1}^{i=M} \lambda_i L_{\text{adv},\mathcal{Z}_i}). \tag{2}$$

By inserting a gradient reversal layer [21] before $h_{\mathcal{Z}_1}, \cdots, h_{\mathcal{Z}_M}$, we can simultaneously optimize the backbone f_Θ along with all heads efficiently using the standard stochastic gradient descent.

Choice of Adversarial Losses. The type of the adversarial loss depends on the label format of $\mathcal{Z}_i$. For soft-label $\mathcal{Z}_i$, we can use soft cross-entropy loss (**SoftCE**, Eq. 3) or KL-divergence loss (**KLDiv**, Eq. 4). For multi-label $\mathcal{Z}_i$, we use binary cross-entropy loss (**BCE**).

$$L_{\text{adv},\mathcal{Z}_i} = -\sum_{k=1}^{k=|\mathcal{Z}_i|} z_{i_k} \log[h_{\mathcal{Z}_i}(f(x;\Theta))]_k. \tag{3}$$

$$L_{\text{adv},\mathcal{Z}_i} = \sum_{k=1}^{k=|\mathcal{Z}_i|} z_{i_k} \log \frac{z_{i_k}}{[h_{\mathcal{Z}_i}(f(x;\Theta))]_k}. \tag{4}$$

Source of Non-Action Labels. The labels for $\mathcal{Z}_1, \cdots, \mathcal{Z}_M$ are needed in adversarial training. However, these annotations are usually unavailable for most action recognition datasets. To handle this, we use off-the-shelf ConvNets trained on the specific domain to obtain the pseudo labels. For example, we use ResNet trained on ImageNet [14] and Places365 [57] to assign the pseudo labels for *object* and *scene*, respectively.

Spatial-Aware Actionness Reweighting. The idea of adversarial training is simple but turns out to be fragile, especially when considering multiple aspects. It would be hard for the vanilla algorithm to converge if the adversarial loss weight λ_i is set as a relatively large number. One possible conjecture is that the underlying network pools feature *uniformly* across all positions. Since we can only mitigate the model bias instead of eliminating the dataset bias, the inherent bias from the data would contradict the adversarial objective unless the model selectively attends to the action-related region. To this end, inspired by the idea of actionness estimation [48,55], we propose a Spatial-Aware Actionness Reweighting module (SAAR), illustrated in Figure 5.

For a feature map $\mathbf{F} \in \mathbf{R}^{C\times T\times H\times W}$, we first estimate an actionness mask $\mathbf{A} \in \mathbf{R}^{T\times H\times W}$, where the scalar for each location represents how much the feature is related to the human action. In experiments, we use a 2D ResNet-Layer with a small bottleneck width for actionness feature extraction, and use another 2D 3×3 convolution as the actionness head, which outputs a 1-channel actionness score map $\mathbf{A}$. On top of the score map, we apply 2D-softmax across the spatial dimensions for normalization: $\mathbf{A}'(t,h,w) = \frac{e^{\mathbf{A}(t,h,w)}}{\sum_{h',w'} e^{\mathbf{A}(t,h',w')}}$. The final modulated feature map is the element-wise multiplication between $\mathbf{F}$ and $\mathbf{A}'$:

$$\mathbf{F}'(c,t,h,w) = (H\times W)\cdot \mathbf{F}(c,t,h,w) \odot \mathbf{A}'(t,h,w) \tag{5}$$

where the coefficient $H \times W$ is used to preserve the magnitude of feature maps after reweighting. We insert SAAR before the last ResNet-Layer in the backbone. Operating on a small feature map (14×14), the SAAR module adds up to 2% additional computation.

Experiments show that spatial-aware actionness reweighting can not only benefit convergence of training but also lead to better performance. It is worth noting that the benefit of SAAR is much larger when combined with MAAT than used alone, indicating that the adversarial training objective incurs weak supervision implicitly.

5.2 Exploiting Web Media with OmniDebias

Instead of restricting to labeled datasets, we also propose to leverage webly-supervised datasets for bias mitigation via co-training, considering their richness and diversity.

We use `GoogleImg` (`GG`) and `InsVideo` (`IG`) from the OmniSource dataset [16] as the web data source. Following the same pipeline as the original work, to construct the auxiliary dataset for joint training, we train a teacher network to filter web data and keep high-confidence examples. Joint training with the built auxiliary dataset can lead to much larger improvements on our evaluation benchmarks (`K200db-val`, `ARAS`), compared to the improvement on standard validation sets (`K200/400-val`), mostly because web media contain novel $z \in \mathcal{Z}$ that does not exist in the training set.

However, there is a drawback to the naïve approach. For web data, the distribution over z can be even more imbalanced than the distribution for Kinetics videos. For example, the average entropy of pseudo *scene* distributions of 400 actions is 3.02 for Kinetics, and 2.79 for `GoogleImg` (larger $\rightarrow$ more diversified). Figure 6 demonstrates pseudo *scene* distributions of 3 action classes. Co-training with such an unbalanced dataset is sub-optimal.

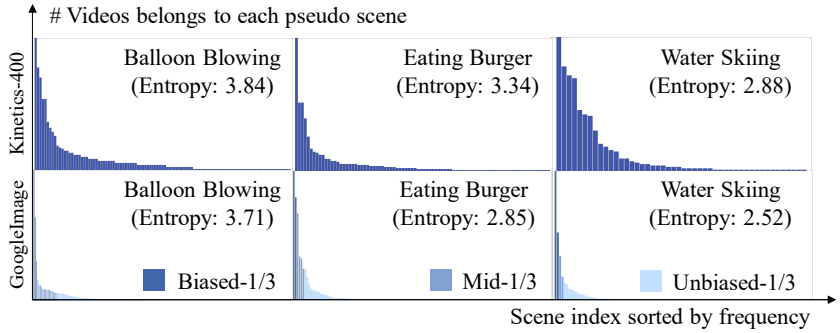

Fig. 6: **Pseudo-scene distributions.** Visualization of pseudo scene distributions of 3 action categories in Kinetics-400 and GoogleImg. GoogleImg has more imbalanced distributions.

Component		SMAD			OmniDebias			Combination		
Model	Train / Test	ARAS-64	K200db-val	K200-val	ARAS-64	K200db-val	K200-val	ARAS-64	K200db-val	K200-val
2D	K200db-train	60.3 ↑ 4.5	55.7 ↑ 2.0	75.5 ↓ 0.7	68.3 ↑ 12.5	60.0 ↑ 6.3	77.4 ↑ 1.2	70.9 ↑ 15.1	62.0 ↑ 8.3	78.1 ↑ 1.9
3D	K200db-train	58.4 ↑ 7.4	55.0 ↑ 3.1	74.6 ↓ 0.5	69.2 ↑ 18.2	60.7 ↑ 8.8	78.7 ↑ 3.6	71.6 ↑ 20.6	62.7 ↑ 10.8	78.4 ↑ 3.3
Model	Train / Test	ARAS-104	-	K400-val	ARAS-104	-	K400-val	ARAS-104	-	K400-val
2D	K400-train	56.2 ↑ 2.1	-	70.0 ↓ 0.6	60.2 ↑ 6.1	-	71.3 ↑ 0.7	61.8 ↑ 7.7	-	71.4 ↑ 0.8
3D	K400-train	55.0 ↑ 3.5	-	68.2 ↓ 0.1	60.2 ↑ 8.7	-	71.0 ↑ 2.7	61.2 ↑ 9.7	-	70.8 ↑ 2.5

Table 2: **The individual and joint effects of SMAD and OmniDebias.** We report the Top-1 accuracies on three test sets: `ARAS`, `K200db-val` and `K200/400-val`. ↑ and ↓ denote the improvement or decline to the baseline w/o. debiasing.

Thus we propose OmniDebias to utilize web media more efficiently. In OmniDebias, we use a simple data selection strategy to select a subset of the entire web dataset for joint training. Specifically, based on the same approach introduced in **Benchmark**, we sort the samples in a same action class by the descending order of z-frequency [7]. Based on the z-frequency, we divide the auxiliary dataset into 3 equal-sized parts, *i.e.* `[web]-bias`, `[web]-mid` and `[web]-unbias` (`[web]` can be `GG`, `IG`, *etc.*), and use `[web]-unbias` only for joint training. OmniDebias consistently outperforms not only using other parts but also the union, indicating its efficacy and efficiency.

6 Experiments

6.1 Experiment Setting

Acquisition of non-action labels. For debiasing, *non-action* labels can be either pseudo labels inferred by a pretrained model or ground-truth labels from a multi-label dataset. To acquire pseudo labels for debiasing, we use ResNet50 [26] pretrained on ImageNet and Places365 as the pseudo label extractor for *scene* and *object*. We also tried ResNet18 and DenseNet161 [28] as the extractor for *scene* labels but observe a subtle difference ($\leq 0.3\%$). For ground-truth labels, we use the HVU dataset [15], which annotates Kinetics videos with three additional tag categories: *context*, *attribute*, *event*.

[7] For z = `scene`, if 20 out of 100 samples have the scene label 'cliff', the z-frequency of each of the 20 samples is 0.2 (20 / 100).

Method	Access	K200db-val	ARAS-64
Baseline	✗	51.9	51.0
AdaBN [32]	✗	52.0	52.3
FrameShuffle [8]	✗	52.4	51.3
SDN [11]	✗	54.0	55.3
DANN [21]	✓	52.4	53.3
MMD [23]	✓	52.7	53.1
SAVA [13]	✓	53.4	54.0
SMAD	✗	55.0	58.4

Table 3: **SMAD v.s. other debiasing and domain-adaptation algorithms. Access** indicates if the algorithm needs to access validation data in training.

Metric	Independence $I(Y', S)$	Separation $I(Y', S\|Y)$	Sufficiency $I(Y, S\|Y')$
Baseline	0.498 ↑ 0.011	0.376	0.356
SMAD	0.490 ↑ 0.003	0.373	0.366
OmniDebias	0.496 ↑ 0.009	0.334	0.321
Combination	0.491 ↑ 0.004	0.321	0.327

Table 4: **Fairness metrics on K200-val.** We train the SlowOnly-R18-8x8 on `K200db-train`, I denotes normalized mutual information (lower $\rightarrow$ more non-discriminative). Red marks denote the *scene*-bias amplified to the oracle independence $I(Y, S) = 0.487$.

Training and Evaluation. We use TSN-3seg-R50 with ImageNet pretraining as the 2D-CNN baseline, SlowOnly-8x8-R18 as the 3D-CNN one. In the choice of adversarial losses, we use a weighted combination of SoftCE and KLDiv for soft-label (better than each individual), use BCE for multi-label. The loss weight is 0.5 for SoftCE and KLDiv losses, 5 for BCE loss. For testing, we uniformly sample 25 frames for TSN or ten clips for SlowOnly with center crop and average the final predictions. In experiments, we report the Top-1 accuracy. Since ARAS is a small evaluation set, we first examine the statistical significance: we train the TSN on `K200db-train` for 10 times with different random seeds and test it on `ARAS-64`. We find the standard deviation of accuracy is around 0.3%, which means a difference larger than 0.8% is statistically significant.

6.2 Main Results

Re-distributed Train-Val Splits. When the training and validation subsets have different scene distributions, our method consistently bridges the performance gap between validation sets with *scene* distribution shift and the validation set without *scene* distribution shift, as shown in Figure 3. The narrower accuracy gaps reflect the improvement made under the fairness metrics EQUALIZED ODDS. When the testing and training scene distributions are completely different, *i.e.* disjoint label sets, the effect is most significant: our methods reduce the accuracy drop by nearly $\frac{1}{3}$: from 22.5% to 15.6%.

Besides EQUALIZED ODDS, we also evaluate three commonly used fairness metrics, namely *independence*, *separation*, and *sufficiency* in Table 4. SMAD largely mitigates the bias amplified by the algorithm: without SMAD, $I(Y', S)$ increases 0.011 (the *scene*-bias amplified by algorithm) compared to $I(Y, S) = 0.487$, while SMAD reduces it to 0.003. Since Y, S are not statistically independent, the sufficiency and independence cannot both hold. Thus we observe that $I(Y, S|Y')$ increases when we apply SMAD for debiasing. For OmniDebias, since additional web media are used for joint training, all three fairness metrics are improved (lower $\rightarrow$ better).

We further study the individual and combined effects of SMAD and OmniDebias. Extensive experiments are conducted with both 2D and 3D baselines: models are trained

Setting	GYM-1shot	GYM-5shot	HMDB51	UCF101	Diving48
w/o. Debiasing	42.2	52.9	49.9	84.3	17.3
+ SMAD	45.5	56.4	50.9	84.9	18.9
+ IG-all	46.6	58.4	54.3	88.4	19.8
+ IG-unbias	47.1	59.5	55.9	88.8	20.9
+ SMAD, IG-unbias	51.7	62.1	57.2	89.9	22.3

Table 5: **Few-shot Learning & Feature Classification.** The learned representation achieves good performance on downstream tasks. We report the 3-split average for HMDB51 and UCF101.

on `K200db-train` or `K400-train` and tested on 3 test sets: `ARAS-64/104`, `K200db-val` (z-unbiased) and `K200/400-val` (normal). The results are demonstrated in Table 2. SMAD can improve the performance on z-unbiased test sets by a large margin at the cost of a little accuracy drop on the normal test set. The improvement of OmniDebias is across all 3 test sets since additional web media are used for joint training. while for `K200db-val` and `ARAS` the gain is much more noticeable. Combining SMAD and OmniDebias yields the highest accuracy on all z-unbiased test sets, indicates that the two techniques are orthogonal to each other.

A new Debiasing Benchmark. In Table 3, we evaluate multiple debiasing and domain-adaptation algorithms on our new **facet-based re-distribution** and **out-of distribution** benchmarks. The models (backbone: SlowOnly-R18) are trained on `K200db-train`, tested on `K200db-val` and `ARAS-64`. SMAD is a better solution for the debiasing problem compared to the alternatives, considering its superior performance and simple deployment.

Transferring Abilities. The debiased representation is also more useful when transferred to other tasks. We study two cases: few-shot learning and video classification. SlowOnly-8x8-R18 trained on `K200db-train` is used as the feature extractor. For each video, we uniformly sample 10 clips, extract a 512-d feature for each clip, and concatenate them into a 5120-d video-level feature.

We evaluate the few-shot learning performance on FineGYM-99 [40], a fine-grained gymnastic action recognition dataset with less scene bias. We construct 10,000 5-way episodes (1-shot or 5-shot). In each episode, the cosine similarities between the query sample and support samples are used for classification. Table 5 shows that both SMAD and OmniDebias contribute to the few-shot performance on FineGYM-99.

We evalute the performance of video classification on UCF101 [42], HMDB51 [31] and Diving48 [34] with two settings: feature classification and finetuning. For feature classification, we train a linear SVM based on the 5120-d descriptors. Table 5 shows that both SMAD and OmniDebias improve the feature classification performance. Two baselines are used in the finetuning setting. We first use ResNet3D-18 [24] trained on MiniKinetics [51] with input size 112 as the baseline, for a straightforward comparison with SDN [11] (Table 6 upper). With pseudo labels for recognition only (much cheaper than pseudo labels for human detection), SMAD can outperform SDN on three downstream tasks. By introducing web data with OmniDebias, the model can obtain much better performance. We further test the finetuning performance on SlowOnly-8x8-R18

Method	Pretrain	HMDB51	UCF101	Diving48*	Diving48
ResNet3D-16x1	MiniKinetics	53.6	83.5	18.0	-
+ SDN [11]	MiniKinetics	56.7	84.5	20.5	-
+ SMAD	MiniKinetics	57.2	84.7	20.9	-
+ SMAD, IG-unbias	MiniKinetics	61.2	88.2	22.6	-
SlowOnly-8x8	K200db-train	62.6	89.6	25.5	53.9
+ SMAD	K200db-train	64.0	90.4	26.7	55.7
+ SMAD, IG-unbias	K200db-train	67.3	93.3	28.2	59.7

Table 6: **Finetuning performance.** Our work improves the finetuning performance on 3 datasets significantly under different settings. We report the 3-split average for HMDB51 and UCF101. * denotes using the old version of Diving48 annotations.

Test Set	Baseline	MAAT	SAAR	MAAT + SAAR
ARAS-64	51.0	55.8	51.6	58.4
K200db-val	51.9	54.3	52.7	55.0

Table 7: **Performance of MAAT & SAAR.** The baseline is SlowOnly-8x8-R18 trained on `K200db-train`.

trained with K200db-train, which is the setting used across this paper (Table 6 lower). The improvement of SMAD and OmniDebias is also steady and distinct upon this much stronger baseline.

6.3 Spatial-aware Multi-Aspect Debiasing

Ablation of SMAD. We first evaluate the efficacy of two components in SMAD, namely Multi-Aspect Adversarial Training (MAAT) and Spatial-Aware Actionness Reweighting (SAAR). Table 7 demonstrates that MAAT itself can largely improve the performance on test videos with novel scenes (`ARAS-64`, `K200db-val`). Upon this decent baseline, SAAR further boost the performance by 0.7% on `K200db-val` and 2.6% on `ARAS-64`. The improvement is non-trivial since SAAR only introduces 2% additional FLOPs and requires no additional explicit supervision. It is also worth noting that without the guidance from MAAT, the gain of SAAR is much reduced. The combination of MAAT and SAAR achieves large improvement on videos with novel scenes.

Advantages of Multi-Aspect Debiasing. Multi-aspect debiasing is more generic than the *scene*-debiasing algorithm [11]. To prove that, we design a complex dataset split (`K200-both-split`, split by both *scene* and *object*) to mimic the real-world debiasing scenario. Specifically, we first create `K200-scene-split` and `K200-obj-split` using the introduced re-distributing method, with the factor *scene* and *object* respectively. Then we sample the validation videos from the union of two validation sets and sample the training videos from the remaining videos to form `K200-both-split`. On that split, we evaluate different debiasing factors. For multi-aspect debiasing with N factors, the weight of each adversarial loss is divided by N. Table 8 shows that multi-aspect debiasing consistently outperforms the single-aspect one under this setting: using

Debias Factor	K200db-val
None	51.7
scene	53.2 ↑ 1.5
object	52.9 ↑ 1.2
scene, object	53.5 ↑ 1.8
scene, object, event, attribute, context	53.8 ↑ 2.1

Table 8: **Multi-factor v.s. single-factor debiasing.**

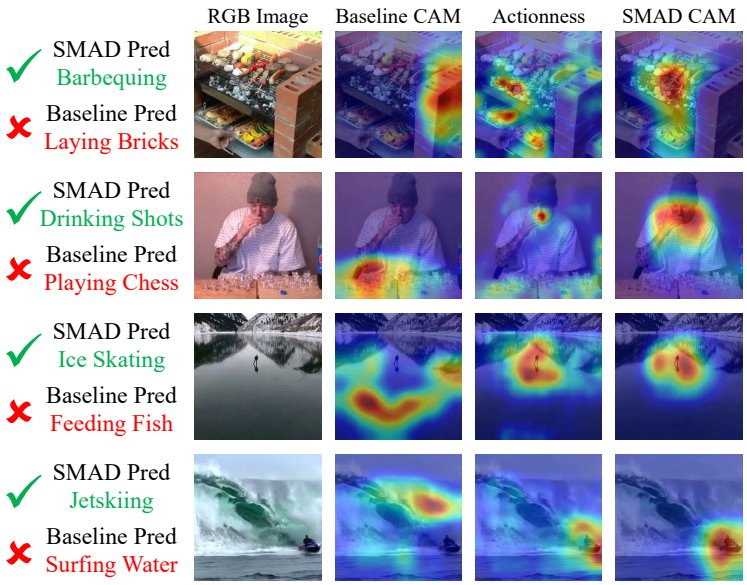

Fig. 7: **The visualization of Actionness mask and CAM.**[8]

both *scene* and *object* as debiasing factors outperforms each individual. Moreover, the best result is achieved when using all five factors for debiasing (*event*, *attribute*, *context* are not used to create `K200-both-split`).

Qualitative Results. To qualitatively show how SAAR guides feature learning, we visualize the spatial-aware actionness mask predicted by SAAR and the class activation maps (CAM) [56] of models trained with or without SMAD in Figure 7. Without debiasing, the rare scenes in action videos, *e.g.*, brick grill, many chessman-like shots, transparent ice surface, and huge waves may mislead the model to give out wrong predictions. With SMAD, models can learn to focus on human actions rather than scenes.

[8] Visualization videos in https://youtu.be/j1LA3y-UuEA.

Model	Aux-Data	ARAS-64	K200db-val	K200-val
2D	None	55.8	53.7	76.2
	GG-bias	60.0 ↑ 4.2	53.9 ↑ 0.2	76.0 ↓ 0.2
	GG-mid	60.5 ↑ 4.7	55.1 ↑ 1.4	76.3 ↑ 0.1
	GG-rand	64.1 ↑ 8.3	57.5 ↑ 3.8	77.2 ↑ 1.0
	GG-all	65.5 ↑ 9.7	58.4 ↑ 4.7	77.8 ↑ 1.6
	GG-unbias	68.3 ↑ 12.5	60.0 ↑ 6.3	77.4 ↑ 1.2
3D	None	51.0	51.9	75.1
	IG-rand	62.3 ↑ 11.3	56.8 ↑ 4.9	77.4 ↑ 2.3
	IG-all	63.4 ↑ 12.4	58.5 ↑ 6.6	78.3 ↑ 3.2
	IG-unbias	64.7 ↑ 13.7	59.8 ↑ 7.9	78.1 ↑ 3.0

Table 9: **OmniDebias.** We jointly train K200db-train with different web dataset splits. The improvement for z-unbiased test sets (K200db-val, ARAS-64) is much larger than K200-val.

6.4 Exploiting Web Media with OmniDebias

Web Data Help in Debiasing. To exploit the richness and diversity of web media, we propose joint training with both labeled datasets and unlabeled web datasets. We first try to use the entire web dataset after teacher filtering for joint training, including both web image dataset GG-all and web video dataset IG-all. Table 9 shows that the performance improved by web media is considerable for z-unbiased test sets. For both baselines, the gain on ARAS-64 is around 10%. The improvement on the normal test set K200-val, is milder ($1 \sim 4\%$) but also noticeable.

Data Selection Strategy. Although web media contain novel $z \in \mathcal{Z}$ that seldomly or never occurs in the original train set, the per action category z distributions are still highly imbalanced. To study the contribution of each portion of web media, we split each web dataset into 3 equal-sized parts: bias, mid, unbias. We also randomly sample a third from the web dataset (rand) for comparison. Using bias and mid leads to worse performance than rand. Using unbias, however, not only surpasses other subsets, but also outperforms training with all web data. With OmniDebias, the performance gap between z-unbiased test sets and K200-val is largely narrowed: the gap shrinks by around 10% Top1 for ARAS-64 and around 5% Top1 for K200db-val. Besides re-distributed datasets, the improvement of OmniDebias can also be observed when trained on the full Kinetics dataset[9].

7 Conclusion

In this work, we seek to mitigate the *generic* representation bias in action recognition. We propose SMAD and OmniDebias: SMAD integrates multi-head adversarial training and spatial-aware feature reweighting for algorithm debiasing, while OmniDebias exploits the rich diversity of web data efficiently for dataset debiasing. When combined, two components lead to excellent debiasing performance and perform far better on either artificially split test sets or manually collected out-of-distribution ones.

[9] We list the detailed results in the supplementary material.

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: ICML. pp. 60–69. PMLR (2018) 3
2. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019), http://www.fairmlbook.org 4
3. Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. L. Rev. **104**, 671 (2016) 1
4. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al.: Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:1810.01943 (2018) 1
5. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. arXiv:1607.06520 (2016) 1
6. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017) 1
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017) 2, 3, 4
8. Carter, K., Shah, M.: An approach for data efficient action recognition and mitigating scene bias (2020) 10
9. Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., Zheng, J.: Temporal attentive alignment for large-scale video domain adaptation. In: ICCV. pp. 6321–6330 (2019) 3
10. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: CVPR. pp. 3339–3348 (2018) 3
11. Choi, J., Gao, C., Messou, J.C., Huang, J.B.: Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In: NeurIPS. pp. 853–865 (2019) 2, 10, 11, 12
12. Choi, J., Sharma, G., Chandraker, M., Huang, J.B.: Unsupervised and semi-supervised domain adaptation for action recognition from drones. In: WACV. pp. 1717–1726 (2020) 3
13. Choi, J., Sharma, G., Schulter, S., Huang, J.B.: Shuffle and attend: Video domain adaptation. In: ECCV. pp. 678–695. Springer (2020) 3, 10
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009) 1, 7
15. Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., Van Gool, L.: Holistic large scale video understanding. arXiv:1904.11451 **38**, 39 (2019) 9
16. Duan, H., Zhao, Y., Xiong, Y., Liu, W., Lin, D.: Omni-sourced webly-supervised learning for video recognition. arXiv:2003.13042 (2020) 8
17. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: ITCS. pp. 214–226 (2012) 1
18. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV. pp. 6202–6211 (2019) 3, 6
19. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: KDD. pp. 259–268 (2015) 1
20. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: ICCV. pp. 2960–2967 (2013) 3
21. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. pp. 1180–1189. PMLR (2015) 3, 7, 10
22. Garcia, D., Weber, I., Garimella, V.R.K.: Gender asymmetries in reality and fiction: The bechdel test of social media. arXiv:1404.0163 (2014) 3
23. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. JMLR **13**(1), 723–773 (2012) 10
24. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: CVPR. pp. 6546–6555 (2018) 11

25. Haselton, M.G., Nettle, D., Murray, D.R.: The evolution of cognitive bias. The handbook of evolutionary psychology pp. 1–20 (2015) 1
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 9
27. Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K.: The dataset nutrition label: A framework to drive higher data quality standards. arXiv:1805.03677 (2018) 3
28. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR. pp. 4700–4708 (2017) 9
29. Jia, S., Lansdall-Welfare, T., Sudhahar, S., Carter, C., Cristianini, N.: Women are seen more than heard in online newspapers. PloS one **11**(2), e0148434 (2016) 3
30. Khosla, A., Zhou, T., Malisiewicz, T., Efros, A.A., Torralba, A.: Undoing the damage of dataset bias. In: ECCV. pp. 158–171. Springer (2012) 3
31. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: ICCV. pp. 2556–2563. IEEE (2011) 11
32. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation (2016) 10
33. Li, Y., Vasconcelos, N.: Repair: Removing representation bias by dataset resampling. In: CVPR. pp. 9572–9581 (2019) 3
34. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: ECCV. pp. 513–528 (2018) 1, 11
35. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. arXiv:1802.06309 (2018) 2
36. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning (2019) 1
37. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR. pp. 1717–1724 (2014) 3
38. Ray, J., Wang, H., Tran, D., Wang, Y., Feiszli, M., Torresani, L., Paluri, M.: Scenes-objects-actions: A multi-task, multi-label video dataset. In: ECCV. pp. 635–651 (2018) 1, 3
39. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR. pp. 3723–3732 (2018) 3
40. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: CVPR. pp. 2616–2625 (2020) 11
41. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NeurIPS. pp. 568–576 (2014) 3
42. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402 (2012) 11
43. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016) 1
44. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR. pp. 1521–1528. IEEE (2011) 1, 3
45. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: ICCV. pp. 5552–5561 (2019) 3
46. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR. pp. 6450–6459 (2018) 3
47. Wang, A., Narayanan, A., Russakovsky, O.: REVISE: A tool for measuring and mitigating bias in visual datasets. In: ECCV (2020) 1, 3
48. Wang, L., Qiao, Y., Tang, X., Van Gool, L.: Actionness estimation using hybrid fully convolutional networks. In: CVPR. pp. 2708–2717 (2016) 8

49. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. pp. 20–36. Springer (2016) 3, 6
50. Woodworth, B., Gunasekar, S., Ohannessian, M.I., Srebro, N.: Learning non-discriminatory predictors. In: COLT. pp. 1920–1953. PMLR (2017) 3
51. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning for video understanding. arXiv:1712.04851 **1**(2), 5 (2017) 11
52. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: WWW. pp. 1171–1180 (2017) 3
53. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: AIES. pp. 335–340 (2018) 2, 3, 4
54. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv:1707.09457 (2017) 3
55. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: ICCV. pp. 2914–2923 (2017) 8
56. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016) 13
57. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. TPAMI **40**(6), 1452–1464 (2017) 1, 4, 7