SOAR: Scene-debiasing Open-set Action Recognition

Yuanhao Zhai[1*], Ziyi Liu[2†], Zhenyu Wu[2], Yi Wu[2], Chunluan Zhou[2], David Doermann[1],
Junsong Yuan[1], Gang Hua[2]
[1]University at Buffalo [2]Wormpex AI Research
{yzhai6, doermann, jsyuan}@buffalo.edu, {wuzhenyusjtu, ywu.china, czhou002, ganghua}@gmail.com

Abstract

Deep learning models have a risk of utilizing spurious clues to make predictions, such as recognizing actions based on the background scene. This issue can severely degrade the open-set action recognition performance when the testing samples have different scene distributions from the training samples. To mitigate this problem, we propose a novel method, called Scene-debiasing Open-set Action Recognition (SOAR), which features an adversarial scene reconstruction module and an adaptive adversarial scene classification module. The former prevents the decoder from reconstructing the video background given video features, and thus helps reduce the background information in feature learning. The latter aims to confuse scene type classification given video features, with a specific emphasis on the action foreground, and helps to learn scene-invariant information. In addition, we design an experiment to quantify the scene bias. The results indicate that the current open-set action recognizers are biased toward the scene, and our proposed SOAR method better mitigates such bias. Furthermore, our extensive experiments demonstrate that our method outperforms state-of-the-art methods, and the ablation studies confirm the effectiveness of our proposed modules.

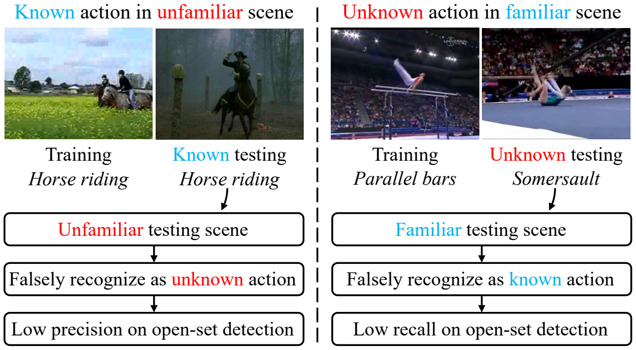

Known action in unfamiliar scene
Unknown action in familiar scene
Training
Horse riding
Known testing
Horse riding
Training
Parallel bars
Unknown testing
Somersault
Unfamiliar testing scene
Falsely recognize as unknown action
Low precision on open-set detection
Familiar testing scene
Falsely recognize as known action
Low recall on open-set detection

Figure 1. Scene-biased open-set action recognizers fail in two typical scenarios: known actions in unfamiliar scenes, and unknown actions in familiar scenes. The former leads to low precision on open-set detection, while the latter leads to low recall. Our method focuses on mitigating the scene bias to improve OSAR.

1. Introduction

Recent years have witnessed significant progress in action recognition [10, 69, 70, 41, 20, 77, 79, 26, 72]. Yet, most works follow a closed-set paradigm, where both training and testing videos belong to a set of pre-defined action categories. This limits their application as the real world is naturally open with unknown actions. Open-set recognition is proposed to identify unknown samples from known ones while maintaining classification performance on known samples [57, 31, 4]. It is challenging due to missing knowledge of the unknown world. Moreover, deep models are found to rely on spurious information to make predictions, *e.g*., classify images using local textures [24, 44] and recognize actions using background scene [40, 13]. This not only hurts the performance under the closed-set setting when training and testing sets are not independent and identically distributed, but also severely degrades the open-set recognition performance, as the distribution of the open-set testing set is unknown.

Open-set action recognition (OSAR) is especially vulnerable to the spurious information for two main reasons: (1) current benchmark datasets are found to be severely biased, and action classification using non-action information (*e.g*., scene, object, or human) achieves high accuracy [40]; (2) without a specific module design, the model tends to focus on static information learning instead of temporal action modeling [83, 16, 59, 68, 52].

This paper focuses on mitigating the scene bias in OSAR: we speculate that current OSAR methods are biased toward the scene, and the performance degrades when the testing set exhibits different scene distributions from the training set. Specifically, existing methods may fail in two typical scenarios: known action in unfamiliar scene and unknown action in familiar scene, as illustrated in Fig. 1. For the

[*]Work done during an internship at Wormpex AI Research.
[†]Corresponding author.

former scenario, a scene-biased recognizer would falsely recognize the action as unknown given the scene is unfamiliar to the training set, and lowers the OSAR precision. For the latter scenario, a scene-biased recognizer may falsely recognize the unknown action as known if a familiar scene has appeared during training, which further lowers the OSAR recall. Consequently, the two above situations degrade the overall OSAR performance. To verify our speculations, a quantitative scene bias analysis experiment is carried out in Sec. 3, and the results reveal a strong correlation between the testing scene distribution shift and OSAR performance.

To mitigate scene bias, we propose a *Scene-debiasing Open-set Action Recognition method* (SOAR), which features an *adversarial scene reconstruction module* (AdRecon) and an *adaptive adversarial scene classification module* (AdaScls). As shown in Fig. 3, we formulate the OSAR task as an uncertainty estimation problem, where the recent evidential deep learning is leveraged to quantify the second-order prediction uncertainty [58, 1, 3, 39]. To mitigate scene bias, AdRecon promotes the backbone to reduce scene information by applying adversarial learning between a decoder and the backbone. Meanwhile, AdaScls encourages the backbone to learn scene-invariant feature by preventing a scene classifier from predicting the scene type of input videos.

Specifically, for AdRecon, our intuition stems from the observation that reconstruction autoencoders prioritize reconstructing the low-frequency part of the input [29], which typically corresponds to the static scene in the video domain. Therefore, we regard the decoder that takes as input video feature and reconstructs the video as a scene information extractor. By applying adversarial learning between the decoder and the encoder, AdRecon promotes the encoder (*i.e.*, the feature backbone) to reduce scene information within the output feature. Furthermore, to reduce the noise from reconstructing the foreground motion, we propose background estimation and uncertainty-guided reconstruction to make the decoder focus on background scene reconstruction, thus preserving motion information during adversarial learning.

For AdaScls, instead of only conducting video-level adversarial scene classification as in [13], we propose to adaptively apply weights on the background and foreground locations: higher weights on the action foreground and lower weights on the background scene, where the background and foreground locations are determined by the learned spatio-temporal uncertainty map. As a result, AdaScls prioritizes debiasing on the foreground, and promotes scene-invariant action feature learning.

Extensive experiments performed on UCF101 [63], HMDB51 [38] and MiTv2 [45] demonstrate the effectiveness of our proposed modules, and show our SOAR achieves state-of-the-art OSAR performance. Besides, quantitative scene bias analysis experiments reveal that our SOAR achieves the lowest scene bias compared to previous arts.

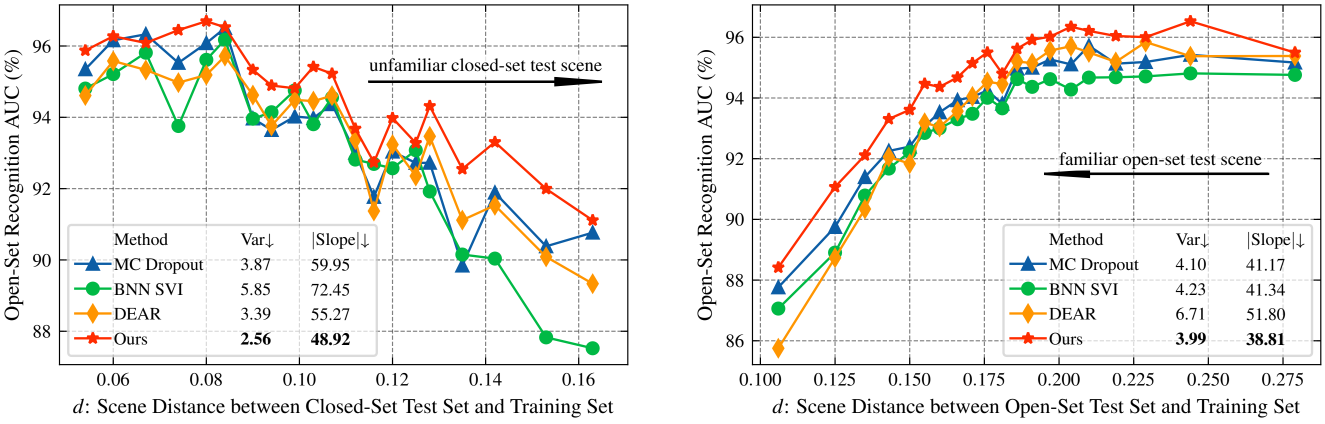
To summarize, our contributions are threefold:

- We design a quantitative experiment to analyze the scene bias of current OSAR methods. The results reveal a strong correlation between testing scene distribution shift and OSAR performances. Our SOAR achieves the lowest scene bias while outperforming state-of-the-art OSAR methods, demonstrating the effectiveness of our debias method.
- We propose an adversarial scene reconstruction module. By preventing a decoder from reconstructing the video background from the extracted feature, AdRecon forces the backbone to reduce scene information from the feature while preserving motion information.
- We propose an adaptive adversarial scene classification module, which prevents a scene classification head from predicting the scene type of the video. Benefiting from additional guidance from the learned uncertainty map, AdaScls promotes effective scene-invariant feature learning.

2. Related work

Action recognition in the closed-set setting has been widely exploited in recent years. Two-stream convolutional networks [62] use two separate networks to learn appearance and motion from RGB frames and optical flow, respectively. I3D [10] expands the 2D CNNs in the two-stream network to 3D CNNs, and significantly improves recognition performance. Due to the expensive cost of optical flow estimation, several recent works [83, 16, 59, 68, 52] try to learn motion information from raw videos directly. In this paper, we also aim to learn motion information with only RGB frames input to reduce the data process cost.

Open-set recognition aims to recognize testing samples that do not belong to the training classes [56]. There are mainly two groups of work for open-set recognition, *i.e.*, discriminative methods and generative methods [25]. For discriminative models, several traditional methods leverage support vector machines to reject the unknown [57, 31, 4]. OpenMax [5] first adopts deep learning models in the open set recognition problem, where it redistributes the softmax output to estimate the uncertainty. DOC [60] proposes a 1-vs-rest layer to replace the softmax layer and tighten the decision boundary. Recently, several methods explicitly model the potential open-set samples in the latent space, and promote a more discriminative decision boundary [12, 86, 11]. Generative methods explicitly generate samples of unknown/known classes, thus helping learn a better decision boundary [23, 18, 50, 11, 34, 80, 86, 78]. Specifically, several methods [78, 47, 65] leverage the autoencoder to reconstruct the input, and use the reconstruction error to determine open-set samples.



(a) Analysis on the known action in unfamiliar scene scenario. The performances of OSAR methods degrade when closed-set testing samples exhibit unfamiliar scenes to the training set.

(b) Analysis on the unknown action in familiar scene scenario. The performances of OSAR methods degrade when open-set testing samples exhibit familiar scenes to the training set.

Figure 2. Quantitative scene bias analysis using UCF101 [63] as known and MiTv2 [45] as unknown. Our SOAR is least affected by scene.

Most of the above methods focus on the image domain. For the OSAR problem, ODN [61] detects new categories by applying a multi-class triplet thresholding method. Busto *et al*. [9] propose an approach for open-set domain adaptation on action recognition. Several methods [36, 64, 37, 3] focus on learning the uncertainty of unknown classes. Specifically, Bayesian neural networks are widely adopted in the action domain [36, 64, 37]. Recently, evidential deep learning [58, 1, 3, 39] shows great potential in uncertainty estimation and achieves superior performances [3] in OSAR.

Debias has been a challenging task in machine learning. Previous works in the image domain include mitigating the gender bias [7, 81, 8, 28], and texture bias [24, 30]. Several methods address this problem with adversarial learning [76, 82, 19, 75, 33, 13, 74, 73], where the label of the debias target can be extracted with off-the-shelf pretrained models. ContraCAM [44] alleviates the scene bias in image object, where contrastive learning is used to automatically determine the discriminative regions. In action recognition, Resound [40] analyzes the scene/object/people bias existing in current datasets. DEAR [3] introduces ReBias [2] to open-set action recognition, and mitigates static bias by forcing the features learned from the original video and shuffled/static videos to be independent. Notably, ReBias [2] requires simultaneously training several backbones, while we only train one backbone with a light-weight decoder and classification heads, greatly reducing the computational burden. Choi *et al*. [13] also leverages an adversarial scene classification module; however, they conduct adversarial learning on the whole frame instead of considering specific scene locations. We propose a guide loss to direct the adversarial classification on the foreground, thus to promote effective scene-invariant feature learning. We show in the supplement that our SOAR outperforms the aforementioned debias methods [2, 13, 44].

3. The effect of the scene in OSAR

As illustrated in Fig. 1, we speculate that existing OSAR methods are vulnerable to scene bias under two typical settings: known action in unfamiliar scene and unknown action in familiar scene. To measure the severity of existing methods affected by the two problems, we conduct the following quantitative experiments.

Settings. The following describes the experimental setup to analyze the first known action in unfamiliar scene problem, and the setup for the second problem can be conducted similarly. Our essential goal is to build different combinations of testing sets, such that each testing set contains different closed-set testing samples that exhibit different scene similarities to the training set while keeping the open-set testing samples the same, and observe how these different combinations of testing sets affect performances of existing OSAR methods. Specifically, we first use an off-the-shelf scene classifier to extract the scene features $\boldsymbol{f}_{\text{scene}}$ on each training and testing video. After that, for each closed-set testing video, we compute its scene feature cosine distance to all training videos, and use the minimal distance to indicate the scene distance between this testing video and the training set. Subsequently, we sort all closed-set testing videos with their scene distance to the training set, and divide them into several non-overlapping equal-sized subsets. For each closed-set testing subset with size L, we define its scene distance to the training set d as the average of minimal scene feature cosine distance between each testing video and all training videos:

$$d = \frac{1}{L}\sum_{i=1}^{L}\min_{j}\left(1 - \boldsymbol{u}_i\boldsymbol{v}_j\right), \tag{1}$$

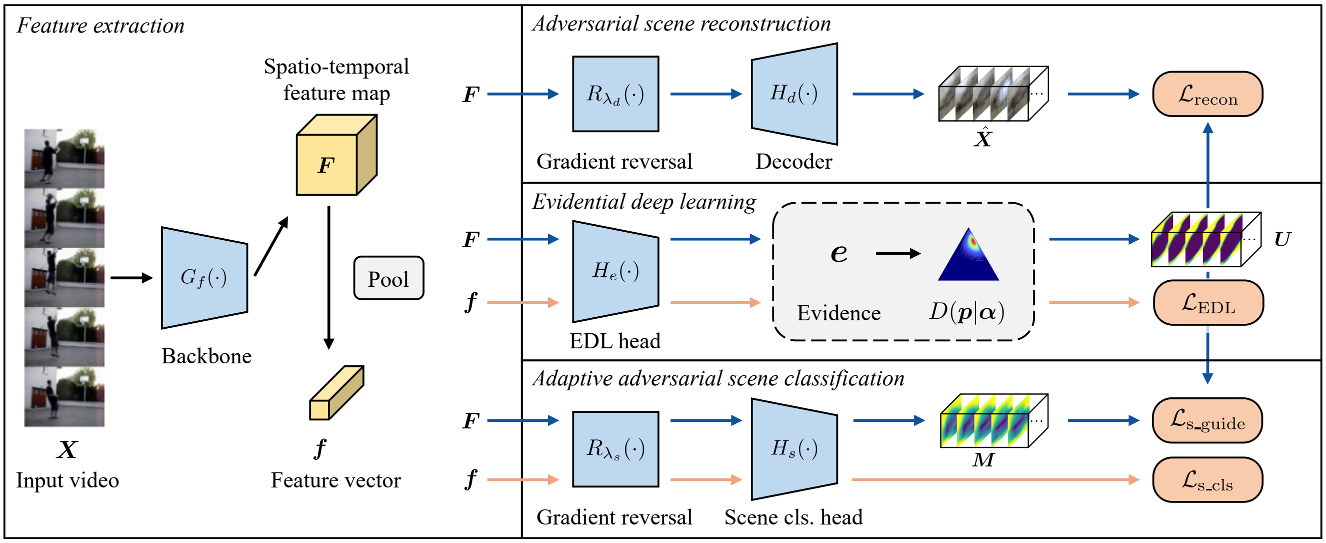


Figure 3. Framework overview. Our SOAR consists of four major modules: the feature extraction module extracts spatio-temporal features from the input video; the evidential deep learning module estimates the prediction uncertainty and outputs the spatio-temporal uncertainty map; the adversarial scene reconstruction module (AdRecon) reconstructs the video background; the adaptive adversarial scene classification module (AdaScls) predicts the scene in the video. The latter two modules are trained in an adversarial way to learn scene-invariant features.

where the unit vector $\boldsymbol{u}_i$, and $\boldsymbol{v}_j$ are normalized scene feature $\boldsymbol{f}_{\text{scene}}$ of the i-th testing video and the j-th training video, respectively. Finally, we fix the open-set testing set and combine it with different closed-set testing subsets, and observe how the performances change. Note that we additionally ensure that each closed-set subset is class-balanced, such that all testing set combinations achieve the same openness [56], which measures how open the testing environment is.

Datasets and evaluation. We perform the experiments with the UCF101 training set for training [63], UCF101 validation set as closed-set testing set and MiTv2 validation set as open-set testing set [45]. Two metrics are used to quantify the scene bias: the variance of the OSAR AUCs under different testing combinations, and the absolute value of the linear fitting slope of the performance change curves. We note that all methods exhibit Pearson correlation coefficients between OSAR AUC and d larger than 0.7, which indicates a strong linear correlation and justifies the use of linear slope for evaluation. We divide the closed-set/open-set testing set into 20 subsets to conduct the two evaluations, respectively.

Analysis. Our SOAR is compared to previous methods [21, 36, 3] in Fig. 2. The analysis of known action in unfamiliar scene is presented in Fig. 2a. A clear performance decrease trend is observed as the scene distance between the closed-set testing set and the training set increases. The results suggest that current OSAR methods rely on the scenes to make predictions: known actions with familiar scenes are easier to recognize, while those with unfamiliar scenes are harder to recognize. Fig. 2b analyzes the unknown action in familiar scene counterpart, where we also observe a clear increasing trend as the scene distance between the open-set testing set and the training set increases, indicating unknown actions with unfamiliar scenes are easier to recognize. Both figures show a strong correlation between the scene distance and the OSAR performance, suggesting that the scene is an essential cue for open-set recognition. Moreover, we find that our SOAR achieves the lowest variance and absolute slope, showing its scene-debiasing capability.

4. Method

The overview of our proposed SOAR is illustrated in Fig. 3. Given an input video, SOAR predicts an uncertainty score that measures how likely this video contains known actions that are used for training. To mitigate scene bias, we aim to suppress the performance of scene-related tasks (reconstruction and classification) while maintaining the action recognition performance. This reduces the scene information in the learned debiased representation, such that the following uncertainty estimation process will be less dependent on the scene.

4.1. Evidential deep learning

To distinguish the known and unknown samples, a scoring function is needed to measure the likelihood that the samples are unknown. To this end, we leverage the recent evidential deep learning (EDL) methods [17, 32, 58, 1, 3, 39] for uncertainty estimation, which mitigates the over-confident [46, 67] and computationally costly [6, 21, 15] problems of existing uncertainty estimators. Essentially, for the C-way classification, EDL first collects the evidence that supports the given sample to be classified into a particular class and then builds a Dirichlet class probability distribution parameterized over the evidence. The resulting distribution models the second-order class probabilities and uncertainty. We refer readers unfamiliar with EDL to [58].

Specifically, denote the input video as $\boldsymbol{X} \in \mathbb{R}^{H \times W \times T \times D}$, where H, W, T and D represent height, width, number of frames and channels, respectively. The backbone $G_f(\cdot)$ maps it into a spatio-temporal feature map $\boldsymbol{F} \in \mathbb{R}^{H' \times W' \times T' \times D'}$, which is further average pooled as a feature vector $\boldsymbol{f} \in \mathbb{R}^{D'}$. The EDL head $H_e(\cdot)$ takes as input the feature vector $\boldsymbol{f}$, and predicts a non-negative evidence vector $\boldsymbol{e} = H_e(\boldsymbol{f}) \in \mathbb{R}^C_{\geq 0}$, which parameterizes the following Dirichlet class probability distribution:

$$D(\boldsymbol{p}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \Pi_{j=1}^{C} p_j^{\alpha_j - 1} & \text{for } \boldsymbol{p} \in \mathcal{S}_C, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathcal{S}_C$ is the C-dim unit simplex, $\alpha_j = e_j + 1$, and $B(\boldsymbol{\alpha})$ is the C-dim multinomial beta function [58]. Thus, the cross entropy action classification loss reduces to the following:

$$\mathcal{L}_{\text{EDL}} = \sum_{i=1}^{C} y_i \big(\log S - \log \alpha_i \big), \tag{3}$$

where $\boldsymbol{y}$ is the one-hot label vector, $S = \sum_{i=1}^{C} \alpha_i$ is the the total strength of the Dirichlet distribution. During inference, the class probability is given as the mean of the Dirichlet distribution $\boldsymbol{p} = \boldsymbol{\alpha}/S$, and the prediction uncertainty is deterministically given as $u = C/S$. As the EDL head estimates the uncertainty relying on the feature vector $\boldsymbol{f}$, we aim to reduce scene information in $\boldsymbol{f}$, so that the uncertainty estimation process is less dependent on the scene. Scene debias is accomplished by the following two modules via adversarial training.

4.2. Adversarial scene reconstruction

We take inspiration from reconstruction-based video anomaly detection methods, where locations with abnormal motions typically incur high reconstruction errors [54]. In action recognition, we empirically find that such reconstruction prioritizes reconstructing static background scene, while achieving low reconstruction quality on the action-related foreground. Thus, our AdRecon adds a decoder to reconstruct the video background. By regarding the decoder as a scene information extractor, we force the learned feature $\boldsymbol{F}$ not to contain scene information to hinder scene reconstruction in an adversarial learning manner.

Specifically, given spatio-temporal feature tensor $\boldsymbol{F}$, we feed it into a decoder $H_d(\cdot)$ to reconstruct the raw video frames: $\hat{\boldsymbol{X}} = H_d(R_{\lambda_d}(\boldsymbol{F}))$. $R_{\lambda_d}(\cdot)$ is a gradient reversal layer [22, 13] that acts as an identity function during forward propagation, and reverses the gradient by a factor of λ_d during backward propagation: $\frac{\mathrm{d}R_{\lambda_d}(\boldsymbol{X})}{\mathrm{d}\boldsymbol{X}} = -\lambda_d \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix. In this way, the reconstruction loss is adversarial in that it forces the backbone $G_f(\cdot)$ to reduce the scene information contained in the output feature $\boldsymbol{F}$ to hinder reconstruction (*i.e.*, maximize the loss), while encouraging the decoder to extract the scene information from the feature for reconstruction (*i.e.*, minimize the loss). Despite its simplicity, such adversarial reconstruction inevitably loses motion information, as the action-related foreground is also involved in this process. We address this problem by enforcing the decoder to focus on reconstructing the background with two additional designs: background estimation and uncertainty-weighted reconstruction.

Background estimation. Instead of using raw frames $\boldsymbol{X}$ as the reconstruction target, we propose to use the video background $\bar{\boldsymbol{X}}$ for the adversarial reconstruction, such that the foreground action information will not be removed from the feature. To achieve background estimation, we leverage the temporal median filter (TMF), which has been demonstrated as an effective background estimation method [43, 66, 55]. Specifically, for a given pixel location, the most frequently repeated intensity in a sequence of frames is most likely to be the background value for that scene [51, 42]. Thus, TMF takes the pixel-wise temporal median in a sliding window on a frame sequence as the corresponding background. We denote the background clip of the video as $\bar{\boldsymbol{X}}$, which is used as the reconstruction target.

Uncertainty-weighted reconstruction. Despite the simplicity and effectiveness of TMF, it extracts inferior background in videos with static foreground (*e.g.*, apply eye makeup), which further disturbs adversarial scene reconstruction. To address this problem, we leverage the spatio-temporal uncertainty map. Similarly to the class activation map [84], we apply the EDL head onto the spatio-temporal feature $\boldsymbol{F}$ to generate a spatio-temporal evidence map $\boldsymbol{E} = H_e(\boldsymbol{F}) \in \mathbb{R}^{H' \times W' \times T' \times C}$. The evidence map $\boldsymbol{E}$ can be converted to a spatio-temporal uncertainty map $\boldsymbol{U} \in \mathbb{R}^{H' \times W' \times T'}$ according to DST [17]: $u_{i,j,t} = C / \sum_c (e_{i,j,t,c} + 1)$, where $u_{i,j,t}$ is the element of $\boldsymbol{U}$ at index i, j, t.

Intuitively, similar to the class activation map that indicates discriminative locations that respond to the class label [49, 53], the obtained uncertainty map is expected to indicate locations that are discriminative for action recognition (*i.e.*, foreground) with low uncertainty, while high uncertainty indicates the background scene. Meanwhile, considering that the reconstruction task should focus on the background scene while neglecting the foreground action, the uncertainty map can serve as a weight map to guide the reconstruction. Specifically, scene locations with high uncertainties are assigned higher weights for reconstruction, so that the backbone $G_f(\cdot)$ will focus on removing information at these locations to disturb reconstruction. The final reconstruction loss $\mathcal{L}_{\text{recon}}$ is formulated as a weighted L1 loss:

$$\mathcal{L}_{\text{recon}} = \frac{1}{HWTD} \sum_{i,j,t,d} u'_{i,j,t} \|\bar{x}_{i,j,t,d} - \hat{x}_{i,j,t,d}\|_1, \quad (4)$$

where $u'_{i,j,t}$ is the element of $\boldsymbol{U}'$ at index i, j, t defined in Eq. (5). Specifically, since $\boldsymbol{U}$ has different spatio-temporal resolution from our reconstruction target $\bar{\boldsymbol{X}}$, we upsample it to have the same size as $\bar{\boldsymbol{X}}$. Additionally, min-max normalization is applied on $\boldsymbol{U}$, such that $\boldsymbol{U}'$ ranges from 0 to 1, meaning that the most confident locations will have no reconstruction loss, while the most uncertain locations have the largest reconstruction weight as 1. These steps are formulated as follows:

$$\boldsymbol{U}' = \text{up}\left(\text{norm}(\boldsymbol{U})\right), \quad (5)$$

where $\text{up}(\cdot)$ is the trilinear interpolation upsampling function, and $\text{norm}(\cdot)$ is the min-max normalization.

4.3. Adaptive adversarial scene classification

To further facilitate scene-invariant action feature learning, we propose AdaScls for adaptive adversarial video scene classification. Denote the scene label as $\boldsymbol{y}_{\text{s}} \in \mathbb{R}^N$, where N is the number of pre-defined scene classes. The scene classification head $H_s(\cdot)$ predicts the video-level scene type given feature vector $\boldsymbol{f}$ as $\hat{\boldsymbol{y}}_{\text{s}} = H_s(R_{\lambda_s}(\boldsymbol{f}))$, where λ_s is the gradient reversal weight. The adversarial scene classification is achieved via a cross-entropy classification loss $\mathcal{L}_{\text{s\_cls}}$:

$$\mathcal{L}_{\text{s\_cls}} = -\frac{1}{N} \sum_{i=1}^{N} y_{\text{s},i} \log \frac{\exp(\hat{y}_{\text{s},i})}{\sum_{j=1}^{N} \exp(\hat{y}_{\text{s},j})}. \quad (6)$$

Despite previous exploration [13], we note that blindly performing adversarial scene classification on the whole video may yield suboptimal OSAR results. As video scene classification tends to focus on static cues [40], this can cause the action foreground to be disregarded during the adversarial classification, hindering the learning of scene-invariant action feature. This issue becomes more pronounced when there is a strong correlation between the action foreground and the scene. To mitigate this problem, we propose to direct the adversarial scene classification towards the foreground locations.

In our AdaScls, we use the uncertainty map $\boldsymbol{U}$ to adaptively guide the learning of scene classification, so that the adversarial classification focuses on the foreground locations. Specifically, the scene class activation map $\boldsymbol{M} \in \mathbb{R}^{H' \times W' \times T'}$ can be obtained by passing the feature map $\boldsymbol{F}$ to the scene classification head $m_{i,j,t} = H_s(R_{\lambda_s}(\boldsymbol{F}))_{i,j,t,n}$, where $n = \operatorname*{argmax}_i y_{\text{s},i}$. The uncertainty guidance is accomplished by maximizing the difference between the normalized uncertainty map and the scene class activation map. As both terms are within range $[0, 1]$, we minimize the L1 distance between $1 - \text{norm}(\boldsymbol{U})$ and $\text{norm}(\boldsymbol{M})$ as a proxy:

$$\mathcal{L}_{\text{s\_guide}} = \frac{1}{HWT} \sum_{i,j,t} \|1 - \text{norm}(\boldsymbol{U})_{i,j,t} - \text{norm}(\boldsymbol{M})_{i,j,t}\|_1. \quad (7)$$

4.4. Model training

The overall training loss $\mathcal{L}$ is a weighted sum of the evidential learning loss $\mathcal{L}_{\text{EDL}}$, adversarial scene reconstruction loss $\mathcal{L}_{\text{recon}}$, adversarial scene classification loss $\mathcal{L}_{\text{s\_cls}}$ and the scene guide loss $\mathcal{L}_{\text{s\_guide}}$:

$$\mathcal{L} = \mathcal{L}_{\text{EDL}} + w_{\text{recon}}\mathcal{L}_{\text{recon}} + w_{\text{s\_cls}}\mathcal{L}_{\text{s\_cls}} + w_{\text{s\_guide}}\mathcal{L}_{\text{s\_guide}} \quad (8)$$

where w_{recon}, $w_{\text{s\_cls}}$ and $w_{\text{s\_guide}}$ are weight hyperparameters.

5. Experiments

Datasets. We follow DEAR [3] to use three datasets for evaluation: UCF101 [63], HMDB51 [38] and MiTv2 [45]. We use the training split 1 from UCF101 for training, which consists of $9,537$ videos from 101 classes. For testing, validation split one from UCF101 is used as known samples, and testing split 1 of HMDB51 and the testing set of MiTv2 are respectively used as unknown samples. The HMDB51 testing set consists of $1,530$ videos from 51 classes, thus the openness of testing combination UCF101 + HMDB51 is 10.6%; the MiTv2 testing set consists of $30,500$ videos from 305 classes, achieving an openness of 36.9%. Note that openness measures how open the testing environment is, and increases as the number of testing classes increases [56]. We only use the MiTv2 dataset for the scene bias evaluation, as reported in Sec. 3. Note that HMDB51 is a relatively small dataset, which prevents it from splitting into multiple subsets to perform the scene bias evaluation.

Evaluation metrics. We use accuracy on the closed testing set for closed-set classification evaluation. For binary open-set recognition, we use the area under the receiver operating characteristic curve (AUC), false alarm rate at a true positive rate of 95% (FAR@95) and the true positive rate at a false positive rate of 10% (TPR@10) for evaluation. For the C + 1 way classification (*i.e.*, the C known classes and the unknown class), we follow DEAR [3] to report the mean and variance of open macro F1 (open maF1), which weighted sums the macro F1 for the C+1 way classification under different openness points. Note that open macro F1 is a threshold-dependent metric and the uncertainty threshold is set to the maximal training uncertainty. We use AUC as the main metric, as FAR@95 and TPR@10 are only applicable for particular points on the ROC curve, while open maF1 is sensitive to the threshold value.

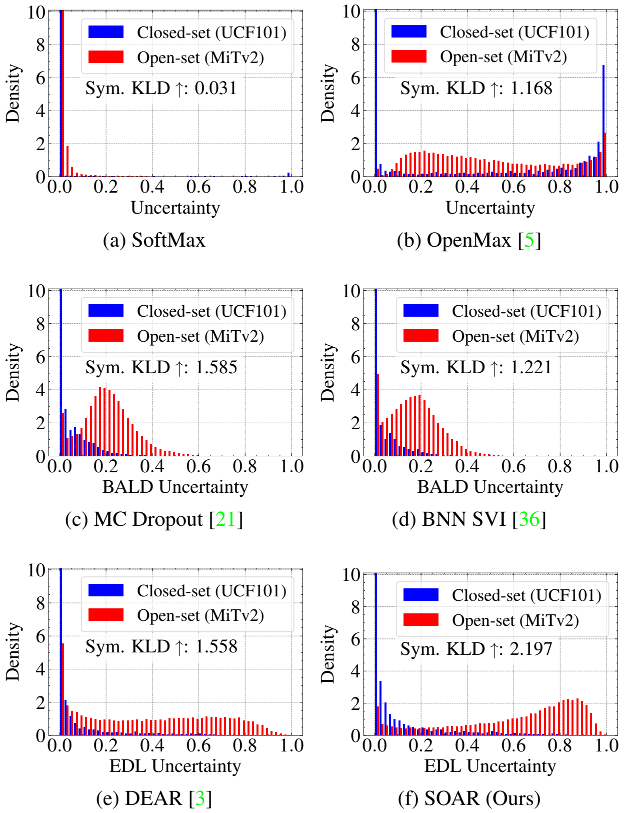
Implementation details. Our method is implemented with MMAction2 [14], a toolbox based on PyTorch [48]. Kinetics400 [10] pre-trained ResNet50-based I3D [27, 10] is

Methods	UCF101 [63]+MiTv2 [45]				UCF101 [63]+HMDB51 [38]				Closed-set
	AUC ↑	FAR@95 ↓	TPR@10 ↑	Open maF1 ↑	AUC ↑	FAR@95 ↓	TPR@10 ↑	Open maF1 ↑	Accuracy
SoftMax	44.47	96.93	8.85	55.50 ± 0.45	44.34	97.91	3.66	73.13 ± 0.12	94.10
OpenMax [5]	63.96	45.89	3.78	66.21 ± 0.16	63.67	80.53	6.54	67.81 ± 0.12	56.54
MC Dropout [21]	93.66	25.43	85.72	68.12 ± 0.20	86.11	77.50	70.13	71.13 ± 0.15	94.13
BNN SVI [36]	93.16	25.88	79.36	67.96 ± 0.19	85.63	71.52	66.14	71.57 ± 0.17	93.89
DEAR [3]	93.52	29.53	84.03	75.12 ± 0.27	87.12	71.32	72.21	88.07 ± 0.20	93.97
SOAR (Ours)	**94.60**	**25.33**	**86.47**	$\mathbf{76.22 \pm 0.32}$	**88.10**	**69.57**	**72.75**	$\mathbf{89.55 \pm 0.22}$	**95.24**

Table 1. Comparison with state-of-the-art methods. All methods are trained on UCF101 [63], and evaluated on two different open sets where unknown samples are from HMDB51 [38] and MiTv2 [45], respectively. Performances with different backbones are listed in the supp.

AdRecon	AdaScls	UCF101 [63]+MiTv2 [45]				UCF101 [63]+HMDB51 [38]				KAUS		UAFS	
		AUC ↑	FAR@95 ↓	TPR@10 ↑	Open maF1 ↑	AUC ↑	FAR@95 ↓	TPR@10 ↑	Open maF1 ↑	Var ↓	|Slope| ↓	Var ↓	|Slope| ↓
-	-	91.73	28.84	78.96	68.55 ± 0.34	85.63	78.59	68.10	87.73 ± 0.22	6.12	75.51	6.17	75.52
✓	-	94.13	27.52	85.72	73.49 ± 0.35	87.49	**69.41**	72.31	89.52 ± 0.21	3.82	59.20	4.16	49.20
-	✓	93.58	26.43	83.16	72.16 ± 0.30	87.22	71.45	69.80	87.47 ± 0.19	4.43	63.62	4.49	63.63
✓	✓	**94.60**	**25.33**	**86.47**	$\mathbf{76.22 \pm 0.32}$	**88.10**	69.57	**72.75**	$\mathbf{89.55 \pm 0.22}$	**2.56**	**48.92**	**3.99**	**38.81**

Table 2. Ablation study on the proposed AdRecon and AdaScls. The last four columns analyzes the scene bias under the known action in unfamiliar scene (KAUS) and the unknown action in familiar scene (UAFS) scenarios.



(a) SoftMax

(b) OpenMax [5]

(c) MC Dropout [21]

(d) BNN SVI [36]

(e) DEAR [3]

(f) SOAR (Ours)

Figure 4. Uncertainty distributions visualization on UCF101 [63] + MiTv2 [45]. Our method achieves the best open-set and closed-set uncertainty separation with the highest symmetric KLD. Uncertainties are normalized to $[0, 1]$ for better visualization.

adopted as the backbone. We follow [13] to use off-the-shelf Places365 [85] pretrained ResNet50 [27] to extract video scene feature and label. We implement the evidential learning head as a single-layer MLP followed by ReLU activation following DEAR [3], implement the decoder as five consecutive 3D transpose convolutional layers and implement the scene recognition head as a five-layer MLP following [13]. We use the SGD optimizer with an initial learning rate of 0.001, which decreases by a factor of 0.1 for every 20 epochs with a total epoch of 50. All hyperparameters are determined via a grid search: $\lambda_d = 1$, $\lambda_s = 10$, $w_{\text{recon}} = w_{\text{s\_cls}} = 1$, and $w_{\text{s\_guide}} = 0.1$.

5.1. Comparison with the state-of-the-art

Our SOAR is superior to previous methods in two aspects: lower scene bias and higher performance.

Scene bias evaluation is analyzed in Fig. 2, where two typical scenarios are evaluated: known action in unfamiliar scene and unknown action in familiar scene. Our SOAR achieves the lowest variance and absolute slope in both scenarios, showing that our method is least affected by the scene bias compared to previous methods. Notably, the performance improvement is more significant when the testing scene distribution is distinct from the training (*i.e.*, right part of Fig. 2a and left part of Fig. 2b). *Similar trends are also observed when using Open maF1 for the scene bias evaluation as well as using different backbones (details are in the supp.).* Furthermore, we show that our method surpasses several debias methods [13, 2, 44] in the supp. Such results demonstrate that our SOAR learns better scene-invariant action features and strong scene-debiasing capability.

Performance comparison with the state-of-the-art is listed in Tab. 1, where both OSAR and closed-set classification performances are reported. The results reveal that our SOAR outperforms all previous methods under all metrics in both OSAR and closed-set classification tasks. Furthermore, we visualize the uncertainty distributions in Fig. 4, where the separation between closed-set and open-set uncertainties is quantified with symmetric Kullback-Leibler divergence (sym. KLD). We observe that our SOAR generates a noticeable bimodal distribution and the highest sym. KLD between

Method	Biased (Kinetics [10])		Unbiased (Mimetics [71])	
	Top1 ↑	Top5 ↑	Top1 ↑	Top5 ↑
EDL Baseline	91.11	99.27	25.32	69.62
DEAR [3]	91.18	99.54	34.58	75.00
SOAR (Ours)	**92.37**	**99.69**	**37.18**	**78.92**

Table 3. Classification accuracy on biased and unbiased datasets.

AdRecon	AdaScls	CKA [35] ↓		
		UCF101 [63]	HMDB51 [38]	MiTv2 [45]
-	-	0.34	0.41	0.37
✓	-	0.27	0.37	0.34
-	✓	**0.23**	0.30	0.28
✓	✓	**0.23**	**0.28**	**0.27**

Table 4. Feature similarity between the learned action feature $\boldsymbol{f}$ and the scene feature $\boldsymbol{f}_{\text{scene}}$ on the closed-set testing set (UCF101 [63]) and the open-set testing sets (HMDB51 [38] and MiTv2 [45]). The similarity is measured with centered kernel alignment (CKA) [35].

AdRecon	Bg. Est.	Unc. Weight	AUC ↑	FAR@95 ↓	TPR@10 ↑	Open maF1 ↑
-	-	-	91.73	28.84	78.96	68.55 ± 0.34
✓	-	-	92.12	28.67	79.69	69.38 ± 0.34
✓	-	✓	93.66	27.59	82.13	72.46 ± 0.32
✓	✓	-	92.73	28.33	81.84	71.58 ± 0.33
✓	✓	✓	**94.13**	**27.52**	**85.72**	**73.49 ± 0.35**

Table 5. Ablation study on the adversarial reconstruction on UCF101 [63] + MiTv2 [45] datasets.

closed-set and open-set uncertainties. *We further show our SOAR achieves state-of-the-art OSAR performance with different backbones in the supp.* Such a clear performance advantage demonstrates the effectiveness of our method.

5.2. Ablation studies

Tab. 2 summarizes the ablation studies on AdRecon and AdaScls. We have the following two observations. (1) Both modules individually improve the performance over the EDL baseline, and the combination of them leads to better performance, validating their effectiveness individually and complementarily. *Notably, with only AdRecon, our method outperforms all previous OSAR methods in terms of AUC.* (2) The performance improvement from AdaScls is lower than that from AdRecon. We speculate that this is because the predicted scene label may be noisy and mislead adversarial learning. (3) The last four columns of Tab. 2 list the scene bias analysis, showing that both modules alleviate the scene bias.

Representation debiasing is analyzed in two aspects: out-of-distribution (OOD) generalization ability and feature similarity between the learned action feature and the scene feature. The OOD generalization is compared in Tab. 3, where we follow DEAR [3] to use 10 classes on Kinetics for training and biased testing, and the same categories from Mimetics [71] for unbiased testing. The results reveal that our SOAR outperforms the EDL baseline and DEAR [3] in both settings, showing our stronger debias capability. We further compare the feature similarity between the learned action feature and the scene feature in Tab. 4 with centered kernel alignment (CKA) [35], which measures the learned representation similarity between models trained on different datasets. CKA is in range $[0, 1]$, and larger value indicates higher similarity. The results reveal our modules successfully reduce the similarity between the learned action feature $\boldsymbol{f}$ and the video scene feature $\boldsymbol{f}_{\text{scene}}$ on all testing sets, showing our method reduces the scene information in the extracted feature.

Adversarial scene reconstruction. Tab. 5 analyzes the effect of different designs in AdRecon. First, we observe that simply adversarially reconstructing the raw video has minor improvement on the performance, as such training encourages the backbone to remove not only static scene information but also foreground motion information. Subsequently, our background estimation and uncertainty-weighted reconstruction individually improve the performance, and the best performance is achieved by combining both. Additional qualitative results of AdRecon are provided in Fig. 5.

Adaptive adversarial scene classification. Tab. 6 shows the effectiveness of AdaScls. $\mathcal{L}_{\text{s\_cls}}$ improves the OSAR performance as it encourages the backbone to learn scene-invariant features. Our proposed uncertainty-guidance loss

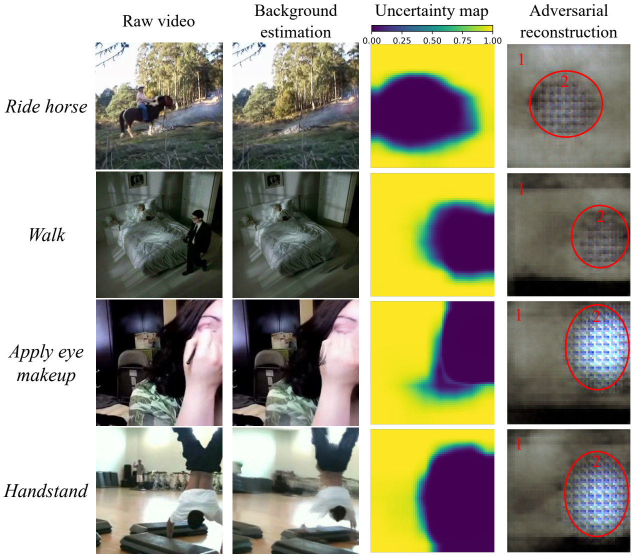


Figure 5. Qualitative results of AdRecon. The background estimation removes foreground with large motions, and the uncertainty map indicates the scene locations (*i.e.*, yellow regions with high uncertainty). AdRecon reduces the scene information within features extracted by the backbone, leading to blurry surroundings (region 1) shown in the 4-th column. As low reconstruction weight is applied on the action foreground (*i.e.*, $u'_{i,j,t} = 0$ in Eq. (4)), it is neglected during reconstruction, leading to Gaussian-noisy-like reconstruction results (region 2).

$\mathcal{L}_{\text{s\_cls}}$	$\mathcal{L}_{\text{s\_guide}}$	AUC ↑	FAR@95 ↓	TPR@10 ↑	Open maF1 ↑
-	-	91.73	28.84	78.96	68.55 ± 0.34
✓	-	92.26	29.32	82.77	71.46 ± 0.32
✓	✓	**93.58**	**26.43**	**83.16**	**72.16 ± 0.30**

Table 6. Ablation study on the adaptive adversarial scene classification on UCF101 [63] + MiTv2 [45] datasets.

$\mathcal{L}_{\text{s\_guide}}$ further improves the performance, demonstrating that the uncertainty map implicitly locates the foreground and guides adversarial scene classification learning.

6. Conclusion

In this paper, we propose SOAR to mitigate scene bias in OSAR. Specifically, we spot two typical scenarios where current OSAR methods fail, and emprically show the scene bias for existing methods. Our SOAR features an adversarial scene reconstruction module and an adaptive adversarial scene classification module. The former reduces the scene information in the extracted feature to disturb video scene reconstruction. The latter learns scene-invariant action features by preventing video scene classification with a focus on the action foreground. Our SOAR exhibits lower scene bias while achieving state-of-the-art OSAR performance.

Acknowledgements

This work is supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0124. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *NeurIPS*, 33:14927–14937, 2020. 2, 3, 4

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

[3] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *ICCV*, pages 13349–13358, 2021. 2, 3, 4, 6, 7, 8

[4] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, pages 1893–1902, 2015. 1, 2

[5] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, pages 1563–1572, 2016. 2, 7

[6] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *ICML*, pages 1613–1622, 2015. 4

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NeurIPS*, 29, 2016. 3

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

[9] Pau Panareda Busto, Ahsan Iqbal, and Juergen Gall. Open set domain adaptation for image and action recognition. *IEEE TPAMI*, pages 413–429, 2018. 3

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1, 2, 6, 8

[11] G Chen, P Peng, X Wang, and Y Tian. Adversarial reciprocal points learning for open set recognition. *IEEE TPAMI*, 2021. 2

[12] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *ECCV*, pages 507–522, 2020. 2

[13] 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. In *NeurIPS*, 2019. 1, 2, 3, 5, 6, 7

[14] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. `https://github.com/open-mmlab/mmaction2`, 2020. 6

[15] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. *NeurIPS*, 32, 2019. 4

[16] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *CVPR*, pages 7882–7891, 2019. 1, 2

[17] Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968. 4, 5

[18] Luke Ditria, Benjamin J Meyer, and Tom Drummond. Opengan: Open set generative adversarial networks. In *ACCV*, 2020. 2

[19] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*, 2018. 3

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 1

[21] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 4, 7

[22] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 5

[23] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017. 2

[24] 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, 3

[25] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE TPAMI*, pages 3614–3631, 2020. 2

[26] Hongji Guo, Zhou Ren, Yi Wu, Gang Hua, and Qiang Ji. Uncertainty-based spatial-temporal attention for online action detection. In *European Conference on Computer Vision*, pages 69–86. Springer, 2022. 1

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6, 7

[28] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, pages 771–787, 2018. 3

[29] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 2

[30] Xin Hu, Zhenyu Wu, Hao-Yu Miao, Siqi Fan, Taiyu Long, Zhenyu Hu, Pengcheng Pi, Yi Wu, Zhou Ren, Zhangyang Wang, et al. Eˆ 2tad: An energy-efficient tracking-based action detector. *arXiv preprint arXiv:2204.04416*, 2022. 3

[31] Lalit P Jain, Walter J Scheirer, and Terrance E Boult. Multi-class open set recognition using probability of inclusion. In *ECCV*, pages 393–409, 2014. 1, 2

[32] Audun Jøsang. *Subjective logic*, volume 3. Springer, 2016. 4

[33] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *CVPR*, pages 9012–9020, 2019. 3

[34] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *ICCV*, pages 813–822, 2021. 2

[35] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, pages 3519–3529, 2019. 8

[36] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Bar: Bayesian activity recognition using variational inference. *NeurIPS Workshop*, 2018. 3, 4, 7

[37] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Specifying weight priors in bayesian deep neural networks with empirical bayes. In *AAAI*, pages 4477–4484, 2020. 3

[38] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. 2, 6, 7, 8

[39] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *CVPR*, pages 6970–6979, 2022. 2, 3, 4

[40] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, pages 513–528, 2018. 1, 3, 6

[41] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019. 1

[42] Wei Liu, Yuanzheng Cai, Miaohui Zhang, Hui Li, and Hejin Gu. Scene background estimation based on temporal median filter with gaussian filtering. In *ICPR*, pages 132–136, 2016. 5

[43] Murari Mandal, Lav Kush Kumar, Mahipal Singh Saran, et al. Motionrec: A unified deep framework for moving object recognition. In *WACV*, pages 2734–2743, 2020. 5

[44] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *NeurIPS*, 34:12251–12264, 2021. 1, 3, 7

[45] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE TPAMI*, pages 1–8, 2019. 2, 3, 4, 6, 7, 8

[46] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *ICML*, pages 7034–7044, 2020. 4

[47] Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2307–2316, 2019. 2

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035. 2019. 6

[49] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014. 5

[50] Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *CVPR*, pages 11814–11823, 2020. 2

[51] Massimo Piccardi. Background subtraction techniques: a review. In *2004 IEEE international conference on systems, man and cybernetics (IEEE Cat. No. 04CH37583)*, volume 4, pages 3099–3104, 2004. 5

[52] AJ Piergiovanni and Michael S Ryoo. Representation flow for action recognition. In *CVPR*, pages 9945–9953, 2019. 1, 2

[53] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, pages 1713–1721, 2015. 5

[54] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE TPAMI*, 2020. 5

[55] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *NeurIPS*, 34:23386–23400, 2021. 5

[56] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE TPAMI*, pages 1757–1772, 2012. 2, 4, 6

[57] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE TPAMI*, pages 2317–2324, 2014. 1, 2

[58] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *NeurIPS*, 31, 2018. 2, 3, 4, 5

[59] Zheng Shou, Xudong Lin, Yannis Kalantidis, Laura Sevilla-Lara, Marcus Rohrbach, Shih-Fu Chang, and Zhicheng Yan. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. In *CVPR*, pages 1268–1277, 2019. 1, 2

[60] Lei Shu, Hu Xu, and Bing Liu. Doc: Deep open classification of text documents. *arXiv preprint arXiv:1709.08716*, 2017. 2

[61] Yu Shu, Yemin Shi, Yaowei Wang, Yixiong Zou, Qingsheng Yuan, and Yonghong Tian. Odn: Opening the deep network for open-set action recognition. In *ICME*, pages 1–6, 2018. 2

[62] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NeurIPS*, 27, 2014. 2

[63] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 3, 4, 6, 7, 8

[64] Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *ICCV*, pages 6301–6310, 2019. 3

[65] Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *CVPR*, pages 13480–13489, 2020. 2

[66] Ozan Tezcan, Prakash Ishwar, and Janusz Konrad. Bsuv-net: A fully-convolutional neural network for background subtraction of unseen videos. In *WACV*, pages 2774–2783, 2020. 5

[67] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICML*, pages 9690–9700, 2020. 4

[68] Lei Wang, Piotr Koniusz, and Du Q Huynh. Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns. In *ICCV*, pages 8698–8708, 2019. 1, 2

[69] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE TPAMI*, (11):2740–2755, 2018. 1

[70] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 1

[71] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *IJCV*, 129(5):1675–1690, 2021. 8

[72] Zhenyu Wu, Zhou Ren, Yi Wu, Zhangyang Wang, and Gang Hua. Txvad: Improved video action detection by transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4605–4613, 2022. 1

[73] Zhenyu Wu, Karthik Suresh, Priya Narayanan, Hongyu Xu, Heesung Kwon, and Zhangyang Wang. Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1201–1210, 2019. 3

[74] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE TPAMI*, 2020. 3

[75] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *ECCV*, pages 606–624, 2018. 3

[76] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. *NeurIPS*, 30, 2017. 3

[77] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR*, pages 591–600, 2020. 1

[78] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *CVPR*, pages 4016–4025, 2019. 2

[79] Tan Yu, Zhou Ren, Yuncheng Li, Enxu Yan, Ning Xu, and Junsong Yuan. Temporal structure mining for weakly supervised action detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5522–5531, 2019. 1

[80] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, pages 15404–15414, 2021. 2

[81] 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

[82] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*, 2018. 3

[83] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Recognize actions by disentangling components of dynamics. In *CVPR*, pages 6566–6575, 2018. 1, 2

[84] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 5

[85] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017. 7

[86] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *CVPR*, pages 4401–4410, 2021. 2