

A Simple Debiasing Framework for Out-of-Distribution Detection in Human Action Recognition

Minho Sim¹, Young-Jun Lee¹, Dongkun Lee¹, Jongwhoa Lee¹ and Ho-Jin Choi^{1,*}

¹Korea Advanced Institute of Science and Technology, Daejeon, South Korea

Abstract. In real-world scenarios, detecting out-of-distribution (OOD) action is important when deploying a deep learning-based human action recognition (HAR) model. However, HAR models are easily biased to static information in the video (e.g., background), which can lead to performance degradation of OOD detection methods. In this paper, we propose a simple debiasing framework for out-of-distribution detection in human action recognition. Specifically, our framework eliminates patches with static bias in video using attention maps extracted from the video vision transformer model. Experimental results show that our framework achieves consistent performance improvement on multiple OOD action detection methods and challenging benchmarks. Furthermore, we introduce two new OOD action detection tasks, Kinetics-400 vs. Kinetics-600 exclusive and Kinetics-400 vs. Kinetics-700 exclusive, to validate our method in a setting close to the real-world scenario. With extensive experiments, we demonstrate the effectiveness of our attention-based masking, and in-depth analysis validates the effect of static bias on OOD action detection. The source code and supplementary materials are available at: <https://github.com/Simcs/attention-masking>

1 Introduction

Recent advances in deep learning-based human action recognition (HAR) models [17, 31, 9, 3] have shown considerable performance in challenging action recognition datasets [27, 18, 36], which require models to understand the characteristics of hundred-scale action labels of videos. However, in real-world scenarios, people can perform actions that do not exist in the training dataset. From the model perspective, such actions are considered as *out-of-distribution* (OOD) and should be rejected at test time. Therefore, determining whether an input action belongs to OOD is necessary for the safe deployment of HAR models.

The task of OOD detection was initially formalized in [14], and existing studies have proposed various methods [28, 26, 6, 4] to improve the OOD detection performance in HAR. However, as pointed out in [4], the *static biased* cues (e.g., background or scene context) in the video can cause the degradation of OOD detection performance in the HAR. For example, as shown in Figure 1, a model trained on Kinetics-400 [18] can classify the video from Mimetics [36] with the same ground-truth label as OOD, since the model is biased towards the static information (e.g., grass or sky). In order to improve the capability of recognizing human action labels correctly, recent studies have attempted to alleviate the static bias problem by introducing a debiasing algorithm [8], a simple video rep-

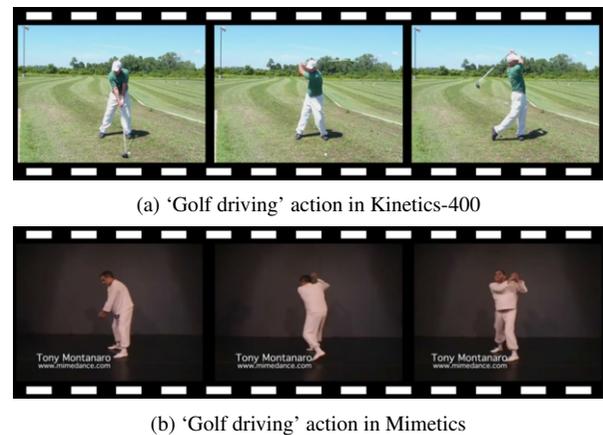


Figure 1. Static bias in human action recognition. A HAR model trained on Kinetics-400 cannot recognize the action label of ‘Golf driving’ in Mimetics because the model is biased toward static parts of the action, such as grass or sky. Such biases can induce degraded OOD detection performance.

resentation learning [34], synthesized sets of benchmarks [23], or a spatio-temporal augmentation [19].

However, the effect of static bias on OOD action detection is underexplored. Therefore, in this paper, we aim to improve the OOD action detection performance by explicitly masking the parts of the video that are likely to be subject to static bias. In a previous work [38], data augmentation by masking non-object patches (e.g., background) based on the attention map obtained from transformer-based image encoder [13] has improved the performance on various image classification benchmarks. Inspired by this, we believe that attention-based masking can also improve OOD detection performance in HAR by removing patches unrelated to the action, which corresponds to the static bias in the video.

To this end, we propose a simple debiasing framework for OOD in HAR to encourage OOD detection methods to concentrate on the non-static biased features of the video input by leveraging an attention map extracted from a pre-trained video encoder model. As illustrated in Figure 2, the core part of our framework is an OOD adapter¹, which consists of two main steps; (1) frame selection and (2) patch masking. These two main steps eliminate patches irrelevant to the action label from the given video clip by utilizing the extracted attention map. Experimental results show that our method consis-

* Corresponding Author. Email: hojinc@kaist.ac.kr.

¹ The ‘adapter’ term used in this paper is different from the term widely used for the parameter-efficient transfer learning [16, 33]

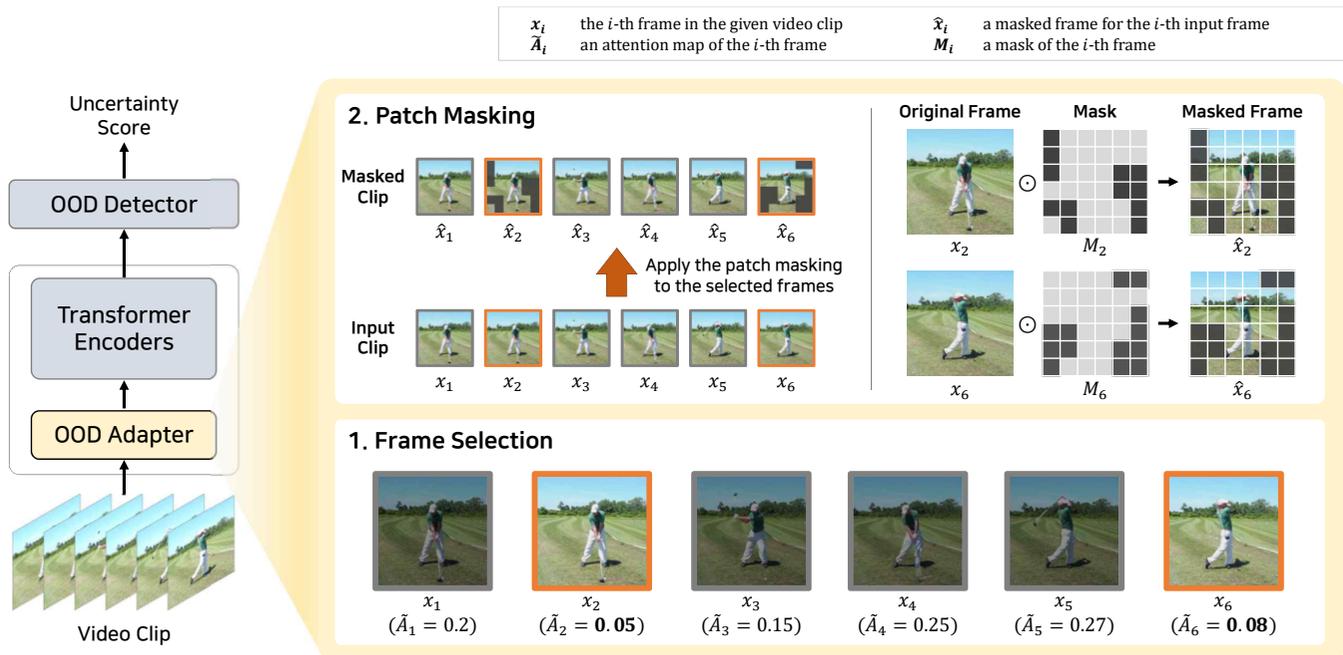


Figure 2. An overview of the simple debiasing framework for out-of-distribution detection in human action recognition. Our framework consists of three main components; OOD adapter, transformer encoders, and OOD detector. In the OOD adapter, we first select temporally redundant frames based on the attention values. Next, with the assumption that less attended patches would contain unrelated features to the action, we mask patches with attention values lower than a certain threshold for each selected frame. In this paper, we leverage ViViT as the transformer-based video encoder.

tently improves the performance of various OOD detection methods on multiple benchmarks while achieving state-of-the-art results. In addition, to verify the effectiveness of our framework in a setting close to real-world scenarios, we further validate our method on OOD action detection tasks using Kinetics [18] dataset. With extensive experiments and analyses, we demonstrate that attention-based masking clearly improves the OOD detection performance in HAR.

In summary, our main contributions are as follows:

- We propose a simple framework for out-of-distribution in human action recognition to enhance the OOD action detection performance by alleviating the static bias problem.
- With attention-based masking, our framework consistently boosts the performance of various OOD detection methods while achieving state-of-the-art results on challenging benchmarks.
- Extensive experiments and analyses demonstrate the validity of our framework and the effect of static bias on OOD detection in HAR.

2 Related Work

Human Action Recognition. The goal of HAR is to predict the action label of an individual or a group of people from a video observation. The convolutional neural network (CNN) based approaches (e.g., 3D-CNN [17]) have been popular choices for action recognition. Motivated by the recent success of the transformer architecture, video transformer networks are proposed to capture the long-range spatiotemporal dependencies (e.g., video vision transformer (ViViT) [3]). In addition, another line of work focused on mitigating the static bias in the video data. [8] adopted adversarial loss on human-masked videos so one cannot infer the scene types based on the learned representations, and [19] proposed FreqAug, which

stochastically filters frequency components from videos to encourage the model to capture essential features. Moreover, various action recognition datasets were proposed to alleviate representation bias caused by background and static objects in the video [24, 36].

Out-of-Distribution Detection. The overconfidence problem in deep learning models motivates the need for the robust detection of OOD samples [2]. Most previous studies are divided into two categories depending on whether the OOD samples are utilized in the training procedure [39]. The OOD detection methods that do not utilize OOD samples typically use the representations of the model trained on the in-distribution (ID) dataset. The pioneering work on OOD detection [14] proposed to use the maximum softmax probability (MSP) to measure the ID-ness of the input based on the assumption that the model will produce flat distribution over the OOD sample. [22] proposed to use the minimum Mahalanobis distance to all class centroids using a Gaussian distribution fitted to the class conditional embeddings. Recently, [25] proposed a unified framework for OOD detection using an energy score which is applicable for any pre-trained neural classifier. Another branch of OOD detection utilizes a set of OOD samples, referred to as *outlier exposure*. [15] proposed outlier exposure loss which encourages the model to produce a flat softmax distribution to the OOD sample. [10] experimented with a few-shot outlier exposure setting where only a handful of known OOD samples were given and showed that a transformer-based model could achieve almost full AUROC-level performance in the OOD image detection domain.

OOD Detection in HAR. In visual OOD detection, most works have covered OOD image detection, which considers image classification as a downstream task, while only a few studies have explored

OOD detection in HAR. [28] proposed to use the Pearson correlation coefficient to measure the deviation of real data from predicted data on the model trained with human activities, and [6] compared various OOD detection algorithms in HAR with inertial data obtained from a smartwatch. [26] trained an additional OOD detector module from synthesized OOD video features for generalized zero-shot action recognition. In addition, open set action recognition (OSAR) essentially tackles the same problem of semantic shift detection in action classification. [30] proposed ODN that detects open set action by incrementally adding new classes to the action recognition head, and [4] proposed DEAR to mitigate overconfident predictions and static bias problems by uncertainty estimation using evidential deep learning. To the best of our knowledge, our work is the first to mitigate the static bias problem by masking videos based on the attention map and analyzing its effect on OOD action detection.

3 Proposed Method

In this section, we propose an adapter-based framework for out-of-distribution detection in human action recognition. As illustrated in Figure 2, the key idea of our framework is to mitigate the effects of static bias of the video clip through two main debiasing steps: (1) frame selection and (2) patch masking. Importantly, all steps are entirely based on the video attention map obtained from the large-scale pre-trained video encoder, such as ViViT [3].

3.1 Extracting Video Attention Map

In order to extract a video attention map from a given video clip, we need a pure-transformer-based video model $f(\cdot)$. In this work, we leverage ViViT [3] model as $f(\cdot)$. Specifically, among several versions of transformer architectures introduced in the original ViViT paper, we utilize a *factorized encoder* with separate spatial and temporal transformer encoders and fine-tune it on the ID training dataset such as Kinetics-400. To fit the input shape of ViViT, we map the input video clip $\mathbf{V} \in \mathbb{R}^{T \times H \times W \times C}$ into a sequence of tokens $\mathbf{z} \in \mathbb{R}^{T \times \frac{H}{p} \times \frac{W}{p} \times d}$, where (T, H, W, C) denote time, height, width, and channel of the video clip, and p and d indicate the patch size and the length of patch embedding, respectively. The intermediate attention weights for each transformer layer l_i is computed as:

$$A(l_i) = Q(l_i) \cdot K(l_i)^T \quad (1)$$

where $Q(l_i)$ and $K(l_i)$ denote query and key of the transformer layer, respectively. The attention weights indicate how each token attends to every other token, i.e., $A(l_i) = \{a_j | j = 1, 2, \dots, (L + 1)^2\}$, where L can be varied to either the number of patches or the number of frames in the video clip, depending on the type of $f(\cdot)$. Note that the attention map ranges from 0 to 1 due to the normalization of attention weights in each transformer layer. Then, following the similar scheme of attention rollout [1], we multiply every attention weights of transformer layers in the $f(\cdot)$ as:

$$\tilde{A} = \prod_{i=1}^{n_l} A(l_i) \quad (2)$$

where n_l denotes the number of layers in the $f(\cdot)$. By multiplying every attention weights, we expect only patches or frames with high attention values to be activated. Finally, the attention map of the transformer encoder $f(\cdot)$ is obtained by taking only the parts related to the `cls` token from the rollout attention weights \tilde{A} . We called the

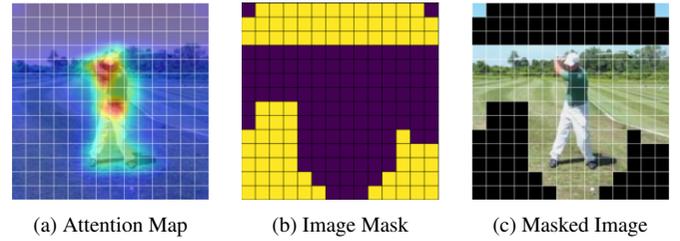


Figure 3. An example of patch masking. We illustrate each step of the patch masking process. (a) visualizes the importance of patches in terms of the model perspective. (b) presents the image mask, which is generated by selecting patches based on the thresholding of attention values. (c) shows the final masked image. The yellow patch in (b) is the patch to be masked in (c).

attention maps extracted from the spatial and temporal transformer encoder of ViViT as spatial attention map and \tilde{A}_s temporal attention map \tilde{A}_t , respectively.

3.2 Adapter-based Video Masking

To reduce the static bias of the given video clip, we adopt a two-stage video masking adapter consisting of frame selection and patch masking. The adapter works in a coarse-to-fine manner, first selecting frames from the entire video clip and then performing fine-grained patch masking on each selected frame.

Frame Selection. Based on the existing literature on *temporal redundancy* [12] that not every video frame equally contributes to video recognition, we selectively perform masking on a few frames. For this, we utilize the temporal attention map \tilde{A}_t . To select the frames with static bias, we exploit the inductive bias that less attended frames would have a low possibility of containing action-related features. Therefore, we select frames with attention values lower than a certain threshold, which we call *lt-threshold* strategy. Consequently, we obtain an index set of masked frames F as:

$$F = \text{lt-threshold}(\tilde{A}_t, \gamma_t) \quad (3)$$

where γ_t represents the temporal threshold parameter for the lt-threshold method.

Patch Masking. After selecting which frames to mask, we apply a fine-grained patch masking based on the attention map over the spatial axis. An example of the patch masking process is shown in Figure 3. For each selected frame, we compute the spatial attention map \tilde{A}_s from the spatial transformer encoder of the ViViT. With the assumption that image patches with the static bias adversely affect OOD detection performance, we masked less attended video patches using the lt-threshold strategy again. As a result, we obtain an index set of masked patches J_t and binary mask $M_t \in \mathbb{R}^{\frac{H}{p} \times \frac{W}{p}}$ for each selected t -th video frame as:

$$J_t = \text{lt-threshold}(\tilde{A}_s, \gamma_s) \quad (4)$$

$$M_t \left[\lfloor \frac{J_{ti}}{p} \rfloor, \text{mod}(J_{ti}, p) \right] = 1 \quad (5)$$

where γ_s represents spatial threshold parameter for the lt-threshold strategy and $\lfloor \cdot \rfloor$ and $\text{mod}(\cdot)$ denote floor operation and the modulo operation, respectively. Finally, we apply the resulting binary

video masks to the original video clip. Here, we followed the masking scheme similar to [37] and replaced the patch embedding corresponding to the target mask patch with a zero vector. Consequently, by applying masks to each selected frame embedding, we obtain a masked sequence of tokens \tilde{z} .

3.3 OOD Detection with Masked Video Clip

After masking is done, we use the masked video clip \tilde{z} to determine whether a given input \mathbf{x} belongs to ID or OOD. For this, we compute the uncertainty score $U(\mathbf{x})$, which indicates the degree to which a given input belongs to OOD, using the OOD detector introduced in Figure 2. Also, it is worth noting that our framework is applicable to arbitrary OOD detection methods that produce an uncertainty score. Finally, following the OOD detection framework introduced in [25], we consider examples with higher uncertainty scores as OOD inputs:

$$G(\mathbf{x}; \tau, f) = \begin{cases} 0 & \text{if } U(\mathbf{x}; f) \leq \tau, \\ 1 & \text{if } U(\mathbf{x}; f) > \tau, \end{cases} \quad (6)$$

where f denotes the proposed framework and τ indicates the OOD detection threshold. In practice, the threshold parameter can be selected using the statistics obtained from ID data so that the majority of the ID samples can be determined as ID.

4 Experiments

4.1 Experimental Setting

OOD Action Detection Tasks. Inspired by similar open set action recognition tasks proposed by [4], we evaluate the proposed framework using three commonly used video classification datasets: UCF-101 [32], HMDB-51 [21], and MiT-v2 [27]. Following the experimental setting in [4], we conduct experiments on two OOD action detection tasks which set UCF-101 as ID and HMDB-51 and MiT-v2 as OOD, which we referred to as UCF-101 (in) vs. HMDB-51 (out) and UCF-101 (in) vs. MiT-v2 (out) detection, respectively. In addition, to further validate our method in a setting close to the real-world scenario, we adopt the Kinetics [18] dataset, which contains hundreds of action labels. Following the convention of mainstream OOD detection tasks that focus on semantic shifts where ID labels and OOD labels do not overlap [39, 10], we create two new datasets: *Kinetics-600 exclusive* and *Kinetics-700 exclusive* from Kinetics-600 and Kinetics-700 by removing samples with the same label as Kinetics-400. The statistics for each dataset are shown in Table 1, and the details are specified in the appendix. Subsequently, we validate our framework on two OOD action detection tasks: Kinetics-400 (in) vs. Kinetics-600 exclusive (out) and Kinetics-400 (in) vs. Kinetics-700 exclusive (out), which set Kinetics-400 as ID as Kinetics-600 exclusive and Kinetics-700 exclusive as OOD, respectively.

Evaluation Metrics. For each experiment, we measured three frequently used metrics in the OOD detection domain: AUROC, AUPR, and FPR@TPR x . The Area Under the ROC curve (AUROC) and Area Under the PR curve (AUPR) are threshold-independent performance evaluation metrics for binary classification tasks and can be interpreted as the probability of a positive sample having a superior detector score than a negative sample. Another commonly used metric, FPR@TPR x , measures the false-positive rate (FPR) at the true-positive rate (TPR) x . The lower value indicates that the method can maintain a low level of FPR even under a high TPR situation,

Table 1. Statistics of datasets. We include the statistics of the original Kinetics-400, 600, and 700 and created Kinetics-600 exclusive and Kinetics-700 exclusive datasets for the OOD detection task. Note that K400, K600, and K700 denotes Kinetics-400, Kinetics-600, and Kinetics-700, respectively.

Datasets	Classes	Train Videos	Validation Videos	Validation Clips
K400	400	246,245	20,000	88,540
K600	600	392,622	30,000	135,318
K600 excl.	213	138,896	10,576	49,315
K700	700	545,317	35,000	142,604
K700 excl.	309	236,768	14,943	63,399

which is important for safety-critical applications. In this paper, we measured FPR95, which is a shorthand for FPR@TPR95.

Implementation Details. We used a large-scale pre-trained video vision transformer (ViViT) as our backbone model throughout the entire experiment. Specifically, we used a publicly available implementation of the ViViT-Base model and selected a factorized encoder type that comprises a separate spatial and temporal transformer encoder. Following the concept of late fusion of the factorized encoder model, we initialized the weights of the spatial transformer encoder from the ImageNet-21K pre-trained ViT-Base model and randomly initialized the weights of the temporal transformer encoder. Then, for each task, we fine-tuned the pre-trained ViViT model on the ID training set. For example, for OOD action detection tasks that set UCF-101 as ID, we pre-trained our ViViT model using Kinetics-400 training set for 30 epochs and fine-tuned it on the UCF-101 training set with 8 batch size and SGD optimizer with a cosine learning rate scheduler and 0.005 base learning rate. Here, we used a system with one Intel Xeon Gold 6230R CPU and two NVIDIA A100 GPUs. Note that the frames per clip and frame size of the ViViT-Base are set to 16 and 224x224, respectively. After training is done, we compute uncertainty scores on both ID and OOD test samples using multiple OOD detectors and evaluate the performance of the proposed framework. In addition, we used grid search to select the best-performing parameters of our framework. As a result, we set *lt-threshold* parameters $\gamma_s = 0.01$ and $\gamma_t = 0.05$ throughout the entire experiment, which were the combinations obtained from the Kinetics dataset.

4.2 Out-of-Distribution Action Detection

In Table 2, we report the results of our framework on two OOD action detection tasks: UCF-101 (in) vs. HMDB-51 (out) and UCF-101 (in) vs. MiT-v2 (out) detection. We adopted three commonly used OOD detection methods as our baseline OOD detectors: MSP [14], Energy [25], and Mahalanobis distance [22], and observed performance before and after applying our masking scheme. Note that for the energy-based OOD detector, we selected temperature parameter $T = 1.5$ since it showed the best performance. The results show that our method consistently improves the baseline performance of each OOD detector. In UCF-101 vs. HMDB-51 detection task, our method was the most effective for the Mahalanobis distance-based OOD detector, improving its baseline AUROC score by more than 4%. Also, our method significantly improves the FPR95 score of the energy-based OOD detector by about 11.5%. In addition, in Figure 4, we illustrate the changes in the distributions of ID and OOD samples as we apply the proposed framework using histogram statistics. The consistent increase in KL-divergence and the figure show that

Table 2. Out-of-distribution detection in human action recognition results. We trained the ViViT model on the UCF-101 training set and tested it on two different OOD datasets, HMDB-51 and MiT-v2. Performances of three widely used OOD detection methods (MSP, Energy-OOD, and Mahalanobis) before and after applying our framework are computed for each OOD action detection task. For every OOD detection metric, we report the mean and standard deviation of 10 random trials, as well as the difference in the mean score.

OOD Detection Methods	Metrics	UCF-101 (in) + HMDB-51 (out)			UCF-101 (in) + MiT-v2 (out)		
		Original	Ours	Diff.	Original	Ours	Diff.
MSP [14]	AUROC	83.625 ± 0.0005	85.807 ± 0.0034	2.181	91.280 ± 0.0011	93.695 ± 0.0008	2.415
	AUPR	81.561 ± 0.0005	84.188 ± 0.0044	2.626	90.125 ± 0.0010	92.660 ± 0.0008	2.535
	FPR95	58.187 ± 0.0214	47.354 ± 0.0100	10.833	34.843 ± 0.0064	26.182 ± 0.0069	8.660
Energy [25]	AUROC	83.937 ± 0.0052	86.013 ± 0.0037	2.075	91.731 ± 0.0011	94.345 ± 0.0008	2.614
	AUPR	82.609 ± 0.0070	85.197 ± 0.0050	2.588	91.355 ± 0.0010	93.950 ± 0.0008	2.595
	FPR95	59.049 ± 0.0126	47.536 ± 0.0108	11.512	34.501 ± 0.0076	25.194 ± 0.0064	9.306
Mahalanobis [22]	AUROC	80.884 ± 0.0058	85.319 ± 0.0032	4.435	90.560 ± 0.0011	93.539 ± 0.0009	2.979
	AUPR	78.845 ± 0.0083	85.582 ± 0.0051	6.736	89.866 ± 0.0012	92.630 ± 0.0009	2.764
	FPR95	77.490 ± 0.0109	68.073 ± 0.0088	9.416	37.497 ± 0.0040	26.965 ± 0.0051	10.532

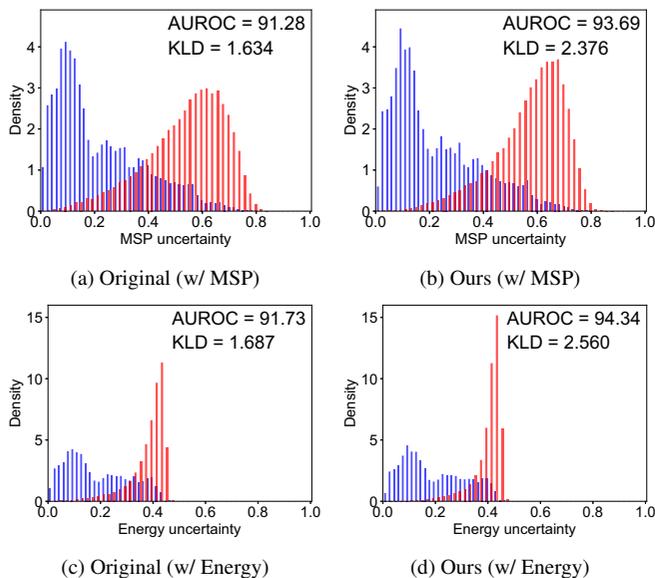


Figure 4. Histogram Statistics. For each OOD detector, we show the change in the distribution before and after applying our method, along with AUROC and KL-divergence scores. We used UCF-101 as ID (blue) and MiT-v2 as OOD (red). Uncertainty values are normalized to [0, 1].

our attention-based masking effectively enhances the ID/OOD separability. More results can be found in the appendix. In addition, we compared our method with existing studies on OOD detection and open set recognition in Table 3. When combined with the energy-based OOD detector, our method achieved the highest AUROC score of 86.10% and 94.34% on each OOD action detection task, which pushes the state-of-the-art results by 0.53% and 2.49%, respectively. Also, note that the ID classification accuracy on the UCF-101 test split is changed from 93.49% to 92.81% after applying our framework, which is less than a 1% performance decrease.

To further validate the proposed method in a setting close to real-world scenarios, we measured the performance of our framework on Kinetics-based OOD action detection tasks where both ID and OOD dataset consist of hundreds of action labels, as shown in Table 4. Note that we used the Mahalanobis distance-based OOD detector as a baseline, of which our method was the most effective in the previous tasks. Here, we can see that our method achieved the highest improvement in terms of FPR95. Before applying our framework to the ViViT fine-tuned on Kinetics-400, we obtained baseline

Table 3. Comparison with the existing studies. We report AUROC scores on two different OOD action detection tasks where the OOD samples are drawn from HMDB-51 and MiT-v2, respectively.

Methods	AUROC (%)	
	UCF-101 + HMDB-51	UCF-101 + MiT-v2
OpenMax [5]	78.76	80.62
MC Dropout [11]	75.41	78.49
BNN SVI [20]	74.78	77.39
RPL [7]	74.23	77.42
DEAR [4]	82.94	86.99
InternVideo [35]	85.48	91.85
Ours (w/ Energy)	86.01	94.34

FPR95 scores of 76.26% and 73.76% in Kinetics-400 vs. Kinetics-600 exclusive and Kinetics-400 vs. Kinetics-700 exclusive, respectively. After applying a mask generated from our two-stage masking scheme, we achieved the FPR95 scores of 71.92% and 69.70%, which is 4.33% and 4.05% higher than the baseline. In addition, we can observe consistent improvements in AUROC and AUPR as well. In Kinetics-400 vs. Kinetics-600 exclusive detection task, our framework achieves the AUROC and AUPR score of 74.28% and 71.92%, which is 2.11% and 1.97% higher than the baseline, respectively.

Table 4. OOD action detection results using Kinetics dataset. We report the performances of our framework on Kinetics-based OOD action detection tasks equipped with the Mahalanobis distance-based OOD detector.

Metrics	K400 + K600 excl.			K400 + K700 excl.		
	Original	Ours	Diff.	Original	Ours	Diff.
AUROC	72.172	74.287	2.114	74.633	76.335	1.702
AUPR	69.943	71.921	1.977	72.225	73.668	1.442
FPR95	76.250	71.920	4.330	73.761	69.701	4.059

4.3 Ablation Studies

Effect of Frame Selection Strategies. When selecting frames to apply patch masking, we leveraged the inductive bias that less attended frames are less likely to be related to the action. To show the effectiveness of our attention-based frame selection, we compared the OOD detection performances between various frame selection strategies while patch masking threshold γ_s is fixed to 0.01. We compared three rule-based frame selection strategies along with our

method, and the details of each strategy are as follows—(1) *All*: applies patch masking to every frame in the video clip, (2) *Skip*: selects frame by frame, (3) *Random*: randomly selects half of the video clip. In Table 5, we report OOD detection performances on each frame selection strategy. The results show that the attention-based frame selection strategy was the most effective, and applying patch masking improves OOD action detection performance regardless of the frame selection strategy. Among the three rule-based strategies, selecting all frames was the most effective.

Table 5. Effect of frame selection strategies. We compare various frame selection strategies on Kinetics-400 vs. Kinetics-600 exclusive detection. Ours indicates selecting less attended frames using an attention map. We mark the best and second performances in bold and underlined, respectively.

Frame Selection Strategy	AUROC	AUPR	FPR95
No Masking	72.172	69.943	76.250
All	<u>73.429</u>	<u>71.205</u>	73.454
Skip	73.224	70.758	<u>73.043</u>
Random	73.297	70.605	73.448
Ours	74.287	71.921	71.920

Effect of Patch Masking Thresholds. Our proposed framework depends on the threshold parameter used at *lt-threshold* strategy when masking patches. We analyzed the effect of the threshold by comparing the OOD detection performance while adjusting the patch masking parameter γ_s . For this, we measured OOD detection performances on Kinetics-400 (in) vs. Kinetics-600 exclusive (out) detection with various γ_s while fixing the frame selection strategy to *All*. The visualization of patch masking on various thresholds and the corresponding graph on OOD action detection results are shown in Figure 5 and Figure 6, respectively. The graph shows that there exists an optimal point of patch masking threshold in terms of OOD action detection. As we adjusted the spatial masking threshold from 0 to 0.01, AUROC rose from 72.22% to 73.17%, and then the performance steadily decreased as the threshold increased. We remark that this crossover point in the graph indicates that removing the static bias to an appropriate level helps detect OOD samples. Finally, further performance improvement achieved by our video attention masking tells us that explicitly providing the model with the debiased video clip with considering temporal dependency was the most effective for OOD action detection.

4.4 Groupwise Analysis

For a conceptual analysis of the proposed method, we adopted parent-child groupings of the Kinetics-400 dataset introduced in the original paper [18]. Figure 7 shows the top/bottom five action groups for which our framework was successful. To measure the effectiveness of our method, we calculated the classwise median of the difference in Mahalanobis distance after applying masking and computed the per-group average. Here, a negative value indicates that the average Mahalanobis distance within the group decreased, representing that our method is effective in the case of the ID dataset. Therefore, our method was effective for the green group (e.g., golf and body motions) and not for the red group (e.g., snow ice and swimming). A detailed description of each selected group is included in the appendix. Here, we found a similar tendency among the group where our method was effective. The top 5 groups (golf, makeup, body motions, gymnastics, athletics – jumping) mainly consist of actions where temporal dynamics of the human motion are essential

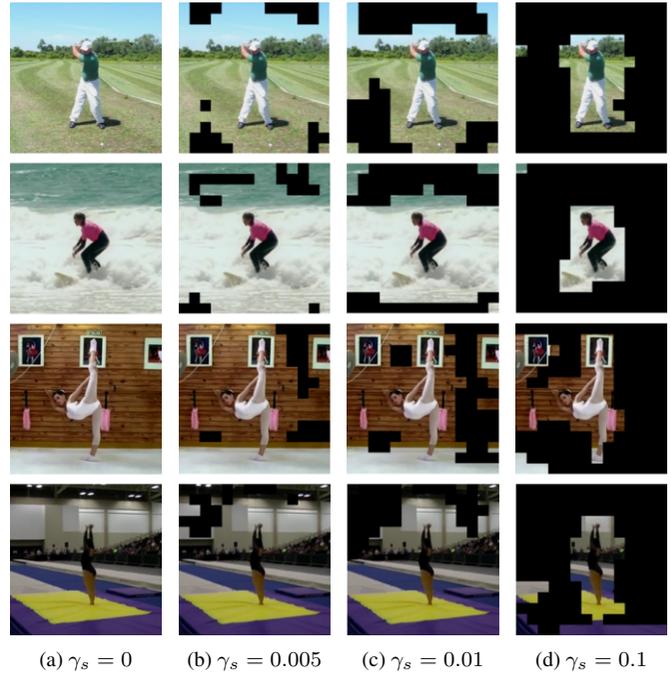


Figure 5. Comparison examples of various patch masking thresholds. We show examples of masked frames containing different actions in Kinetics-400 while adjusting the patch masking threshold γ_s from 0 to 0.1.

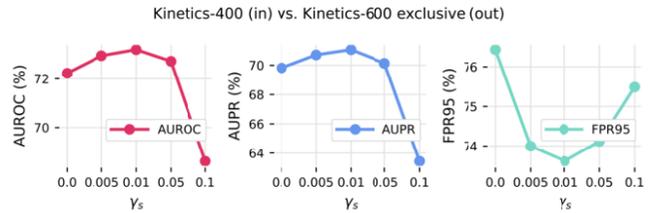


Figure 6. Effect of patch masking thresholds. We show the performance of OOD detection corresponding to the value of patch masking parameters γ_s on Kinetics-400 (in) vs. Kinetics-600 exclusive (out) across all evaluation metrics. When the γ_s is 0.01, we achieve the best performance.

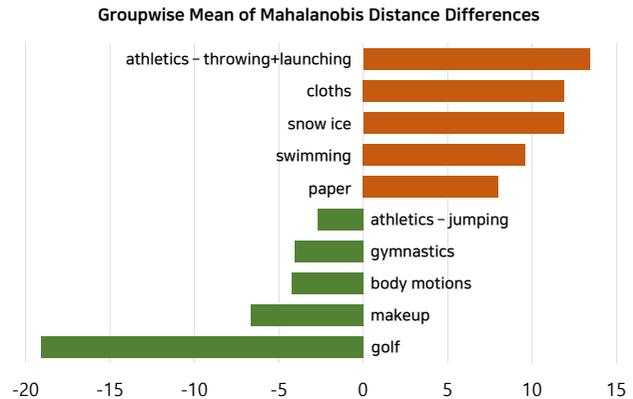


Figure 7. Groupwise analysis on the effectiveness of our framework. We show the top/bottom five action groups in the Kinetics-400 for which our method is effective. Since Kinetics-400 is set as ID in our experiments, a smaller value of difference indicates that our method is more effective.

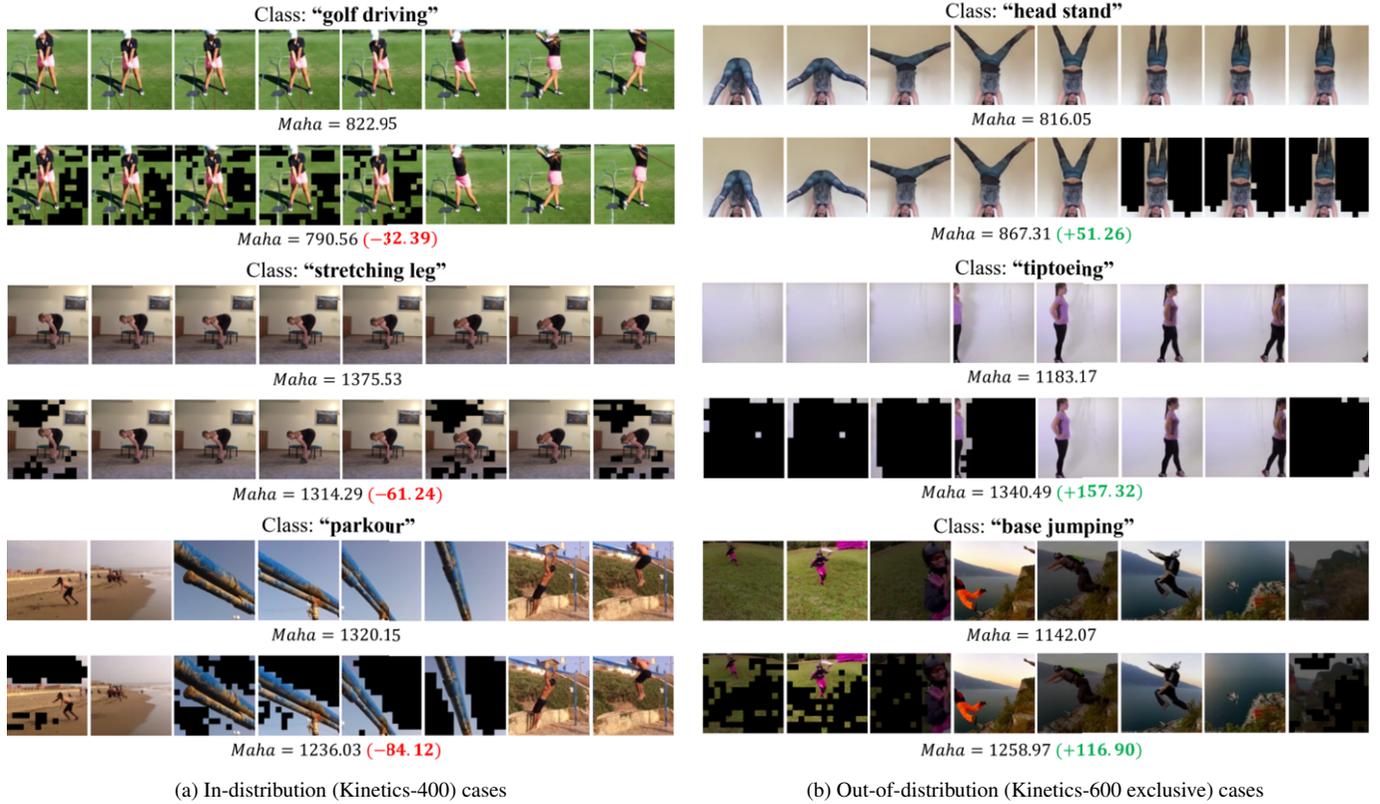


Figure 8. Case study. We present examples of ID (Kinetics-400) and OOD (Kinetics-600 exclusive) with the differences in video clips and Mahalanobis distance before and after applying our framework. The value in parentheses means the difference, **red** and **green** indicate negative and positive differences, respectively. *Maha* below each video clip denotes the computed Mahalanobis distance, and frame selection threshold γ_t and patch masking threshold γ_s are set to 0.05 and 0.01, respectively.

(e.g., golf driving, stretching leg, and parkour). In this case, debiasing video clips using our adapter-based video masking framework significantly improves the OOD action detection performance indicated by the decreased Mahalanobis distances. However, in the case of groups where our method was ineffective, static bias factors were closely related to the action label. For example, actions in the snow ice and swimming groups (e.g., snowboarding and swimming backstroke) are highly dependent on the background, such as snow or the sea, which explains why our method was not effective.

4.5 Case Study

We included the changes in frames and Mahalanobis distances after applying our video attention masking to the video clips in the ID (Kinetics-400) and OOD (Kinetics-600 exclusive) datasets as shown in Figure 8. For ID cases, we selected video clips from golf driving, stretching leg, and parkour, which are actions included in golf, body motion, and athletics – jumping groups, respectively. For OOD cases, we included video clips from head stand, tiptoeing, and base jumping classes where motion features are essential. Interestingly, we found that our frame selection strategy is good at finding static parts in the video. As shown in golf driving and head stand actions in Figure 8, we can see that our frame selection strategy chose frames with relatively low motion changes. In addition, our patch masking clearly recognizes the foreground object and properly masked non-action-related parts of the video clips (e.g., background). After masking was done, the Mahalanobis distances of ID samples consis-

tently decreased, while that of OOD samples increased. In addition, for each pair of video clips in the ID and OOD dataset, we can see the crossover of Mahalanobis distance after masking. We remark that these crossovers significantly affect the OOD detection performance.

5 Conclusion

In this paper, we focused on improving the OOD detection method in the context of HAR by alleviating the effect of static bias. To deal with complicated video data with spatiotemporal dependencies, we proposed an attention map-based video masking framework that consists of two debiasing steps: (1) frame selection and (2) patch masking. Specifically, we first select frames to perform masking and mask out patches that are less likely to be related to the action label with *lt-threshold* strategy. Through extensive experiments and ablation studies, we have validated the effectiveness of our framework in improving the performance of various OOD detection methods on multiple benchmarks. In addition, groupwise analysis and case studies show results supporting our assumption that reducing static bias is helpful for OOD detection in HAR. As future work, comparing our attention-based video masking with various image processing algorithms [29, 40] could provide valuable insights into the field.

Acknowledgements

This research was supported and funded by the Korean National Police Agency. [Project Name: XR Counter-Terrorism Education and Training Test Bed Establishment/Project Number: PR08-04-000-21].

References

- [1] Samira Abnar and Willem Zuidema, ‘Quantifying attention flow in transformers’, *arXiv preprint arXiv:2005.00928*, (2020).
- [2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, ‘Concrete problems in ai safety’, *arXiv preprint arXiv:1606.06565*, (2016).
- [3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid, ‘Vivit: A video vision transformer’, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, (2021).
- [4] Wentao Bao, Qi Yu, and Yu Kong, ‘Evidential deep learning for open set action recognition’, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13349–13358, (2021).
- [5] Abhijit Bendale and Terrance E Boulton, ‘Towards open set deep networks’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, (2016).
- [6] Philip Boyer, David Burns, and Cari Whyne, ‘Out-of-distribution detection of human activity recognition with smartwatch inertial sensors’, *Sensors*, **21**(5), 1669, (2021).
- [7] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 507–522. Springer, (2020).
- [8] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang, ‘Why can’t i dance in the mall? learning to mitigate scene bias in action recognition’, *Advances in Neural Information Processing Systems*, **32**, (2019).
- [9] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, ‘Convolutional two-stream network fusion for video action recognition’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941, (2016).
- [10] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan, ‘Exploring the limits of out-of-distribution detection’, *Advances in Neural Information Processing Systems*, **34**, 7068–7081, (2021).
- [11] Yarin Gal and Zoubin Ghahramani, ‘Dropout as a bayesian approximation: Representing model uncertainty in deep learning’, in *international conference on machine learning*, pp. 1050–1059. PMLR, (2016).
- [12] Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibiyan, ‘Frameexit: Conditional early exiting for efficient video recognition’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15608–15618, (2021).
- [13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, ‘Masked autoencoders are scalable vision learners’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, (2022).
- [14] Dan Hendrycks and Kevin Gimpel, ‘A baseline for detecting misclassified and out-of-distribution examples in neural networks’, *arXiv preprint arXiv:1610.02136*, (2016).
- [15] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich, ‘Deep anomaly detection with outlier exposure’, *arXiv preprint arXiv:1812.04606*, (2018).
- [16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, ‘Parameter-efficient transfer learning for nlp’, in *International Conference on Machine Learning*, pp. 2790–2799. PMLR, (2019).
- [17] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, ‘3d convolutional neural networks for human action recognition’, *IEEE transactions on pattern analysis and machine intelligence*, **35**(1), 221–231, (2012).
- [18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., ‘The kinetics human action video dataset’, *arXiv preprint arXiv:1705.06950*, (2017).
- [19] Jinyong Kim, Taehy Kim, Minh Shim, Dongyoon Han, Dongyoon Wee, and Junmo Kim, ‘Spatiotemporal augmentation on selective frequencies for video representation learning’, *arXiv preprint arXiv:2204.03865*, (2022).
- [20] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo, ‘Bar: Bayesian activity recognition using variational inference’, *arXiv preprint arXiv:1811.03305*, (2018).
- [21] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, ‘Hmdb: a large video database for human motion recognition’, in *2011 International conference on computer vision*, pp. 2556–2563. IEEE, (2011).
- [22] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, ‘A simple unified framework for detecting out-of-distribution samples and adversarial attacks’, *Advances in neural information processing systems*, **31**, (2018).
- [23] Haoxin Li, Yue Wu, Yuan Liu, Hanwang Zhang, and Boyang Li, ‘Evaluating and mitigating static bias of action representations in the background and the foreground’, *arXiv preprint arXiv:2211.12883*, (2022).
- [24] Yingwei Li, Yi Li, and Nuno Vasconcelos, ‘Resound: Towards action recognition without representation bias’, in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 513–528, (2018).
- [25] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li, ‘Energy-based out-of-distribution detection’, volume 33, pp. 21464–21475, (2020).
- [26] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao, ‘Out-of-distribution detection for generalized zero-shot action recognition’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9985–9993, (2019).
- [27] Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogério Schmidt Feris, and Aude Oliva, ‘Multi-moments in time: Learning and interpreting models for multi-action video understanding’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**(12), 9434–9445, (2021).
- [28] Mario Munoz-Organero, ‘Outlier detection in wearable sensor data for human activity recognition (har) based on drnns’, *IEEE Access*, **7**, 74422–74436, (2019).
- [29] Fabio Persia, Fabio Bettini, and Sven Helmer, ‘An interactive framework for video surveillance event detection and modeling’, in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 2515–2518, (2017).
- [30] Yu Shu, Yemin Shi, Yaowei Wang, Yixiong Zou, Qingsheng Yuan, and Yonghong Tian, ‘Odn: Opening the deep network for open-set action recognition’, in *2018 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6. IEEE, (2018).
- [31] Karen Simonyan and Andrew Zisserman, ‘Two-stream convolutional networks for action recognition in videos’, *Advances in neural information processing systems*, **27**, (2014).
- [32] 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).
- [33] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal, ‘VI-adaptor: Parameter-efficient transfer learning for vision-and-language tasks’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5227–5237, (2022).
- [34] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun, ‘Removing the background by adding the background: Towards background robust self-supervised video representation learning’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11804–11813, (2021).
- [35] Yi Wang, Kunchang Li, Yizhuo Li, Yanan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al., ‘Internvideo: General video foundation models via generative and discriminative learning’, *arXiv preprint arXiv:2212.03191*, (2022).
- [36] Philippe Weinzaepfel and Grégory Rogez, ‘Mimetics: Towards understanding human actions out of context’, *International Journal of Computer Vision*, **129**(5), 1675–1690, (2021).
- [37] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu, ‘Simmim: A simple framework for masked image modeling’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, (2022).
- [38] Haochang Xu, Shuangrui Ding, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian, ‘Masked autoencoders are robust data augmentors’, *arXiv preprint arXiv:2206.04846*, (2022).
- [39] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu, ‘Generalized out-of-distribution detection: A survey’, *arXiv preprint arXiv:2110.11334*, (2021).
- [40] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov, ‘A-vit: Adaptive tokens for efficient vision transformer’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10809–10818, (2022).