Evidential Deep Learning for Open Set Action Recognition Supplementary Materials

Wentao Bao, Qi Yu, Yu Kong
Golisano College of Computing and Information Sciences, Rochester Institute of Technology
{wb6219, qi.yu, yu.kong}@rit.edu

In this document, additional materials are provided to supplement our main paper. In section 1, the preliminary knowledge about the evidential deep learning and model calibration are described in detail, which are helpful to understand the methodology of our main paper. In section 2, additional implementation details are provided, which are useful to reproduce our proposed method. Sections 3 and 4 provide additional experimental results to complement the ones presented in our main paper.

1. Detailed Methodology

1.1. Preliminaries of Evidential Deep Learning

Existing video action recognition models typically use softmax on top of deep neural networks (DNN) for classification. However, the softmax function is heavily limited in the following aspects. First, the predicted categorical probabilities have been squashed by the denominator of softmax. This is known to result in an over-confident prediction for the unknown data, which is even more detrimental to open set recognition problem than the closed set recognition. Second, the softmax output is essentially a point estimate of the multinomial distribution over the categorical probabilities so that softmax cannot capture the uncertainty of categorical probabilities, i.e., second-order uncertainty.

To overcome these limitations, recent evidential deep learning (EDL) [8] is developed from the evidence framework of Dempster-Shafer Theory (DST) [9] and the subjective logic (SL) [5]. For a K-class classification problem, the EDL treats the input $\mathbf{x}$ as a proposition and regards the classification task as to give a multinomial subjective opinion in a K-dimensional domain $\{1, \ldots, K\}$. The subjective opinion is expressed as a triplet $\omega = (\mathbf{b}, u, \mathbf{a})$, where $\mathbf{b} = \{b_1, \ldots, b_K\}$ is the belief mass, u represents the uncertainty, and $\mathbf{a} = \{a_1, \ldots, a_K\}$ is the base rate distribution. For any $k \in [1, 2, \ldots, K]$, the probability mass of a multinomial opinion is defined as

$$p_k = b_k + a_k u, \quad \forall y \in \mathbb{Y} \tag{1}$$

To enable the probability meaning of p_k, i.e., $\sum_k p_k = 1$, the base rate a_k is typically set to $1/K$ and the subjective opinion is constrained by

$$u + \sum_{k=1}^{K} b_k = 1 \tag{2}$$

Besides, for a K-class setting, the probability mass $\mathbf{p} = [p_1, p_2, \ldots, p_K]$ is assumed to follow a Dirichlet distribution parameterised by a K-dimensional Dirichlet strength vector $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_K\}$:

$$\text{Dir}(\mathbf{p}|\boldsymbol{\alpha}) = \begin{cases} \dfrac{1}{B(\boldsymbol{\alpha})} \displaystyle\prod_{k=1}^{K} p_k^{\alpha_k - 1}, & \text{for } \mathbf{p} \in \mathcal{S}_K, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $B(\boldsymbol{\alpha})$ is a K-dimensional Beta function, $\mathcal{S}_K$ is a K-dimensional unit simplex. The total strength of the Dirichlet is defined as $S = \sum_{k=1}^{K} \alpha_k$. Note that for the special case when $K = 2$, the Dirichlet distribution reduces to a Beta distribution and a binomial subjective opinion will be formulated in this case.

According to the evidence theory, the term *evidence* is introduced to describe the amount of supporting observations for classifying the data $\mathbf{x}$ into a class. Let $\mathbf{e} = \{e_1, \ldots, e_K\}$ be the evidence for K classes. Each entry $e_k \geq 0$ and the Dirichlet strength $\boldsymbol{\alpha}$ are linked according to the evidence theory by the following identity:

$$\boldsymbol{\alpha} = \mathbf{e} + \mathbf{a}W \tag{4}$$

where W is the weight of uncertain evidence. With the Dirichlet assumption, the expectation of the multinomial probability $\mathbf{p}$ is given by

$$\mathbb{E}(p_k) = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k} = \frac{e_k + a_k W}{W + \sum_{k=1}^{K} e_k} \tag{5}$$

With loss of generality, the weight W is set to K and considering the assumption of the subjective opinion constraint in Eq. (2) that $a_k = 1/K$, we have the Dirichlet strength

$\alpha_k = e_k + 1$ according to Eq. (4). In this way, the Dirichlet evidence can be mapped to the subjective opinion by setting the following equality's:

$$b_k = \frac{e_k}{S} \quad \text{and} \quad u = \frac{K}{S} \tag{6}$$

Therefore, we can see that if the evidence e_k for the k-th class is predicted, the corresponding expected class probability in Eq. (1) (or Eq. (5)) can be rewritten as $p_k = \alpha_k / S$. From Eq. (6), it is clear that the predictive uncertainty u can be determined after α_k is obtained.

Inspired by this idea, the EDL leverages deep neural networks (DNN) to directly predict the evidence $\mathbf{e}$ from the given data $\mathbf{x}$ for a K-class classification problem. In particular, the output of the DNN is activated by a non-negative evidence function. Considering the Dirichlet prior, the DNN is trained by minimizing the negative log-likelihood:

$$\begin{aligned}\mathcal{L}^{(i)}_{EDL}(\mathbf{y}, \mathbf{e}; \theta) &= -\log\left(\int \prod_{k=1}^{K} p_{ik}^{y_{ik}} \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{k=1}^{K} p_{ik}^{\alpha_{ik}-1} d\mathbf{p}_i\right) \\ &= \sum_{k=1}^{K} y_{ik} \left(\log(S_i) - \log(e_{ik}+1)\right)\end{aligned} \tag{7}$$

where $\mathbf{y}_i = \{y_{i1}, \ldots, y_{iK}\}$ is an one-hot K-dimensional label for sample i and $\mathbf{e}_i$ can be expressed as $\mathbf{e}_i = g\left(f(\mathbf{x}_i; \theta)\right)$. Here, f is the DNN parameterized by θ and g is the evidence function such as exp, softplus, or ReLU. Note that in [8], there are two other forms of EDL loss function. In our main paper, we found the Eq. (7) achieves better training empirical performance.

1.2. EDL for Open Set Action Recognition

To implement the EDL method on video action recognition tasks, we removed the Kullback–Leibler (KL) divergence regularizer term defined in [8], because the digamma function involved in the KL divergence is not numerically stable for large-scale video data. Instead, to compensate for the over-fitting risk, we propose the Evidential Uncertainty Calibration (EUC) as a new regularization. Together with the Contrastive Evidence Debiasing module, the complete training objective of our DEAR method can be expressed as

$$\mathcal{L} = \sum_{i} \mathcal{L}^{(i)}_{EDL} + w_1 \mathcal{L}_{EUC} + w_2 \mathcal{L}_{CED} \tag{8}$$

where $\mathcal{L}_{EUC}$ is defined in Eq. (3) in our main paper, and $\mathcal{L}_{CED}$ is the sum of (or one of for alternative training) $\mathcal{L}(\theta_f, \phi_f)$ and $\mathcal{L}(\theta_h, \phi_h)$ defined in Eq. (4) and Eq. (5) respectively in our main paper. The hyperparameters w_1 and w_2 are set to 1.0 and 0.1, respectively.

During the training process, the DEAR model aims to accurately construct the Dirichlet parameters $\boldsymbol{\alpha}$ by collecting the *evidence* from human action video training set. In the inference phase, the probability of each action class is predicted as $\hat{p}_k = \alpha_k / S$ while the predictive uncertainty is simultaneously computed as $u = K/S$. If an input action video is assigned with high uncertainty, which means a vacuity of evidence to support for closed-set classification, the action is likely to be unknown from the open testing set.

Compared with existing DNN-based uncertainty estimation method such as Bayesian neural networks (BNN) or deep Gaussian process (DGP), the advantage of EDL is that the predictive uncertainty is deterministically learned without inexact posterior approximation and computationally expensive sampling. These merits enable the EDL method to be efficient for training recognition models from large-scale vision data such as the human action videos.

1.3. Hilbert-Schmidt Independence Criterion

Hilbert-Schmidt Independence Criterion (HSIC) is a commonly-used dependency measurement of two high-dimensional variables. In practice, we used the unbiased HSIC estimator in [10] with m samples:

$$\text{HSIC}^{k,l}(U, V) = \frac{1}{m(m-3)}\left[\text{tr}(\tilde{U}\tilde{V}^T) + \frac{\mathbf{1}^T\tilde{U}\mathbf{1}\mathbf{1}^T\tilde{V}\mathbf{1}}{(m-1)(m-2)} - \frac{2}{m-2}\mathbf{1}^T\tilde{U}\tilde{V}^T\mathbf{1}\right], \tag{9}$$

where $\tilde{U}$ is the kernelized matrix of U with RBF kernel k by $\tilde{U}_{ij} = (1-\delta_{ij})k(u_i, u_j)$, $\{u_i\} \sim U$ and the $(1-\delta_{ij})$ sets the diagonal of $\tilde{U}$ to zeros. $\tilde{V}$ is defined similarly with kernel l, and $\mathbf{1}$ is an all-one vector. The HSIC value is equal to zero if and only if the two variables are independent.

1.4. Evaluation of Model Calibration

In our main paper, we used the expected calibration error (ECE) to quantitatively evaluate the model calibration performance of our proposed EUC method. According to [7, 4], the basic idea of model calibration is that, if the confidence estimation $\hat{p}$ (probability of correctness) is well calibrated, we hope $\hat{p}$ represent the true probability of the case when the predicted label $\hat{y}$ is correct. Formally, this can be expressed as

$$\mathbb{P}(\hat{y} = y | \hat{p} = p) = p \tag{10}$$

Since perfect calibration is infeasible due to the finite sample space, a practical way is to group all predicted confidence $\hat{p}$ into M bins in the range of [0,1] such that the width of each bin is $1/M$. Therefore, for the m-th bin, the accuracy can be estimated by

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{I}(\hat{y}_i = y_i) \tag{11}$$

where B_m is the set of indices of prediction $\hat{p}$ when it falls into the m-th bin. $\hat{y}_i$ and y_i are predicted and ground truth labels. Besides, the average confidence for the m-th bin can

be expressed as

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \tag{12}$$

To evaluate the mis-calibration error, the ECE is defined as the expectation of the gap between the accuracy and confidence in M bins for all N samples:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{13}$$

A perfect calibrated model means that ECE=0 and higher ECE value indicates that the model is less calibrated.

2. Implementation Details

Network Architecture. As presented in our main paper, the proposed DEAR method as well as all other baselines are implemented on top of the four recent video action recognition models, i.e., I3D, TSM, SlowFast, and TPN. For simplicity, these models use ResNet-50 as the backbone architecture and the network weights are initialized with the pre-trained model from the Kinetics-400 benchmark. To avoid the impact of the validation experiments on the Kinetics and Mimetics datasets, the pre-trained model is not used and we train the model from scratch using the same hyperparameters.

Specifically, for the **I3D** model, it is straightforward to implement our method by replacing the cross-entropy loss with the proposed EUC regularized EDL loss, and inserting the proposed CED module before the recognition head (fully-connected layers). For the **TSM** model, since the architecture of TSM is based on 2D convolution where the output feature embedding is with the size (B, MC, H, W), we recover the number of video segments M as the temporal dimension such that the 5-dimensional tensor with size (B, C, M, H, W) could be compatible with our proposed CED module for contrastive debiasing. For the **SlowFast** model, our CED module is inserted after the *slow* pathway because the feature embedding of slow pathway is more likely to be biased since it captures the static cues of video content. For the **TPN** model, we used the ResNet-50-like SlowOnly model as the recognition backbone and the auxiliary cross-entropy loss in the TPN head is kept unchanged.

Training and Inference. In the training phase, we choose the $\exp$ function as the evidence function because we empirically found $\exp$ is numerically more stable when using the proposed EDL loss $\mathcal{L}_{EDL}$. We set the hyperparameter λ_0 to 0.01 in EUC loss $\mathcal{L}_{EUC}$ and set λ to 1.0 in the two CED losses. The weight of $\mathcal{L}_{EUC}$ is set to 1.0 and the weight of the sum of the two CED losses is empirically set to 0.1. In practice, we found the model performance is robust to these hyperparameters. We used mini-batch SGD with nesterov strategy to train all the 3D convolution models. For all models, weight decay is set to 0.0001 and momentum factor is set to 0.9 by default. Our experiments are supported by two GeoForce RTX 3090 and two Tesla A100 GPUs. Since no additional parameters are introduced during inference, the inference speed of existing action recognition models is not affected.

Dataset Information. For the UCF-101 and HMDB-51 datasets, we used the *split1* for all experiments. For the MiT-v2 dataset, we only use the testing set for evaluation. To validate the proposed CED module, we refer to [1] and select 10 action categories which are included in both Kinetics and Mimetics dataset. These categories are *canoeing or kayaking*, *climbing a rope*, *driving car*, *golf driving*, *opening bottle*, *playing piano*, *playing volleyball*, *shooting goal (soccer)*, *surfing water*, and *writing*. The recognition model is trained from scratch on the 10 categories of Kinetics training set, and tested on these categories of both Kinetics and Mimetics testing set.

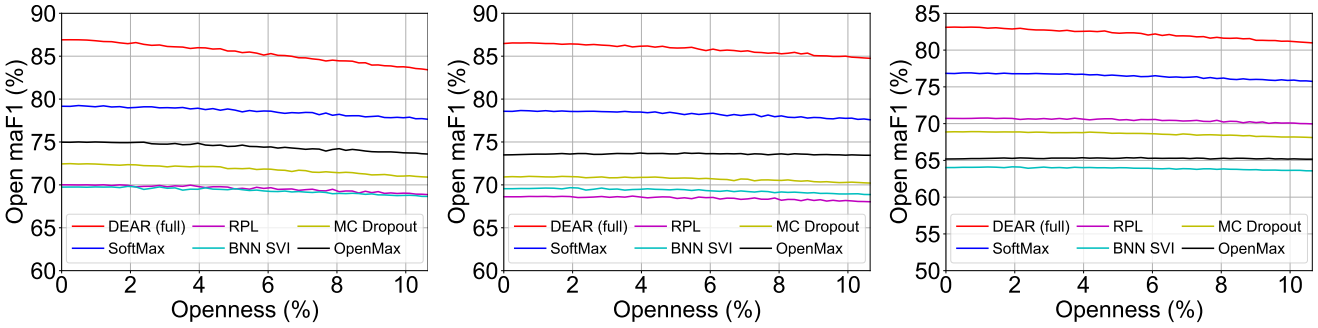
3. Quantitative Results

Open Set Action Recognition. In addition to the I3D-based curves of Open maF1 scores against varying openness in our main paper, we also provide the curves for other action recognition models, including TSM, SlowFast, and TPN in Fig. 1 and Fig. 2. The figures show that when HMDB-51 testing set is used as the unknown, the proposed DEAR method significantly outperforms other baselines with large margins. When MiT-v2 testing set is used as the unknown, the DEAR method could achieve the best performance with relatively low openness.

Out-of-Distribution Detection. From Fig. 3 to Fig. 10, we provide the out-of-distribution detection results to compare our performance with all baselines listed in the main paper. Results on both HMDB-51 and MiT-v2 datasets with I3D, TSM, SlowFast, and TPN are provided. Note that OpenMax, SoftMax, and RPL are not predicting the uncertainty score of input sample, we instead use the confidence score (the maximum score of categorical probabilities) to show the OOD detection performance. These figures show that the uncertainties estimated by the proposed DEAR method exhibit a more long-tailed and flatten distribution than those estimated by MC Dropout and BNN SVI.

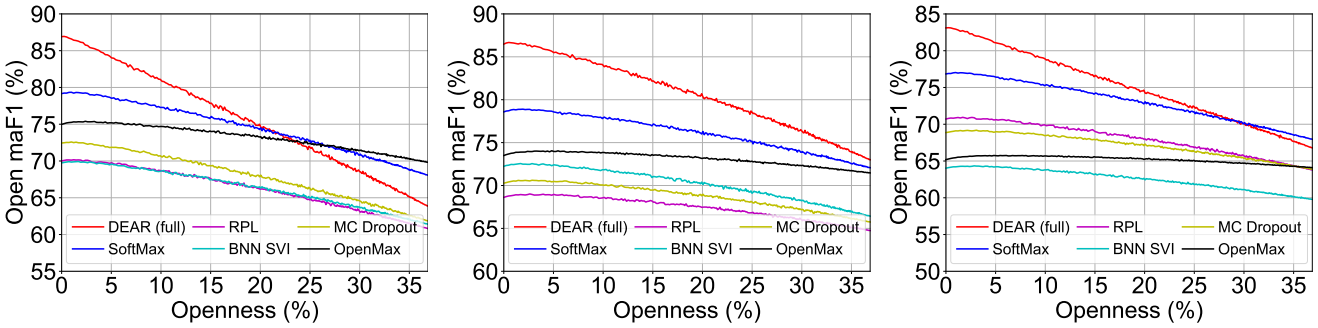
4. Qualitative Results

Open Set Confusion Matrix. In Fig. 11 and Fig. 12, we provide the confusion matrix results. These figures show that when HMDB-51 dataset is used as the unknown, the ratio of mis-classification that classifying the samples from known classes into unknown (see the bottom-left region in each sub-figure) is less on TSM and SlowFast models than that on I3D and TPN models. When MiT-v2 dataset is used



(a) TSM (b) SlowFast (c) TPN

Figure 1: **Open macro-F1 scores against varying Openness**. The HMDB-51 testing set is used as the unknown.



(a) TSM (b) SlowFast (c) TPN

Figure 2: **Open macro-F1 scores against varying Openness**. The MiT-v2 testing set is used as unknown.

as the unknown, the unknown classes are the dominant testing case and from the bottom-right region, we see that the proposed method on I3D and SlowFast models shows significant advantage (brighter red color) over the method on TSM and TPN.

Representation Debiasing Examples. In Fig. 13, we provide examples of three classes, i.e., *playing piano*, *writing*, and *golf driving* from both the biased dataset Kinetics and the unbiased (out-of-context) dataset Mimetics. We compare the recognition results of the variants of our proposed DEAR method with and without CED. These examples show that the CED module could help the DEAR method to recognize human actions on both the biased and unbiased datasets. For example, without the CED module, the model falsely recognizes the *golf driving* as *shooting soccer goal*. The reason could be conjectured that these video samples of the two classes are similar in the static background, i.e., large area of green grassland.

References

[1] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *ICML*, 2020. 3

[2] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *CVPR*, 2016. 5, 6, 7, 8

[3] 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*, 2020. 5, 6, 7, 8

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 2

[5] Audun Jøsang. *Subjective logic*. Springer, 2016. 1

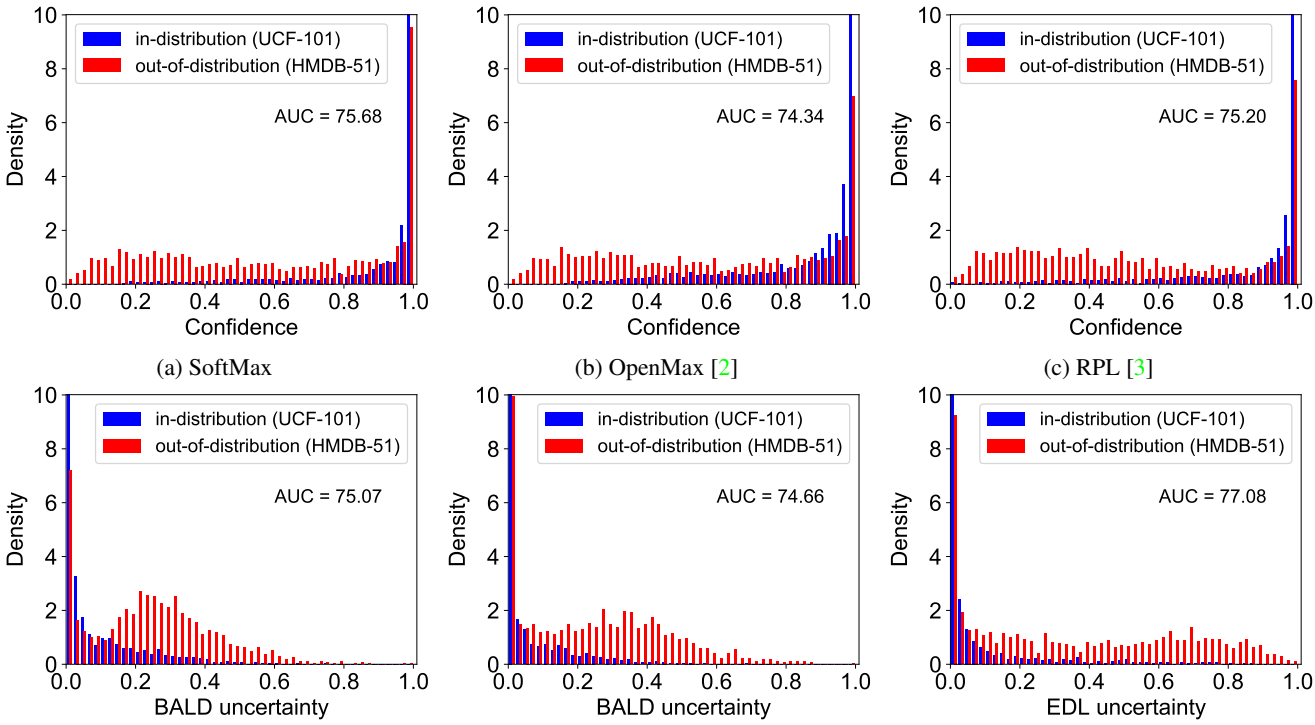
[6] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. BAR: Bayesian activity recognition using variational inference. In *NeurIPS*, 2018. 5, 6, 7, 8

[7] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, 2015. 2

[8] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *NeurIPS*, 2018. 1, 2

[9] Kari Sentz, Scott Ferson, et al. *Combination of evidence in Dempster-Shafer theory*, volume 4015. Sandia National Laboratories Albuquerque, 2002. 1

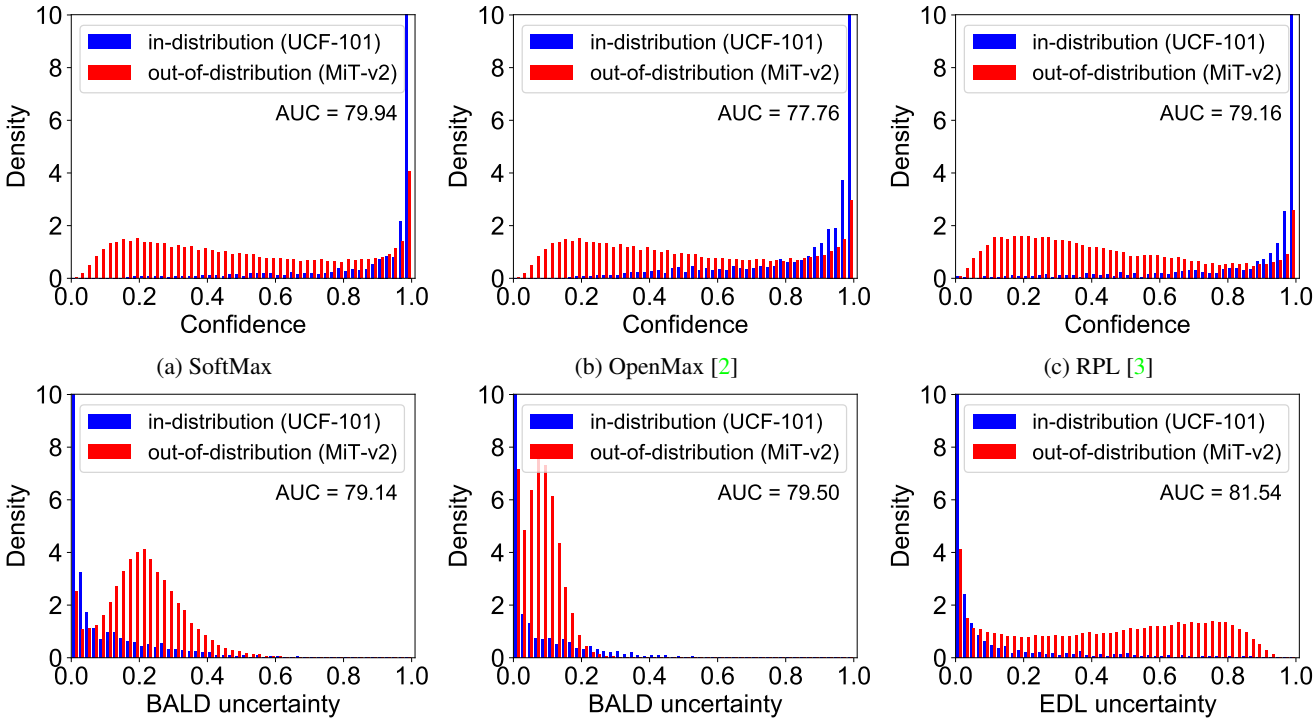
[10] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(5), 2012. 2



(a) SoftMax (b) OpenMax [2] (c) RPL [3]

(d) MC Dropout (e) BNN SVI [6] (f) DEAR (full)

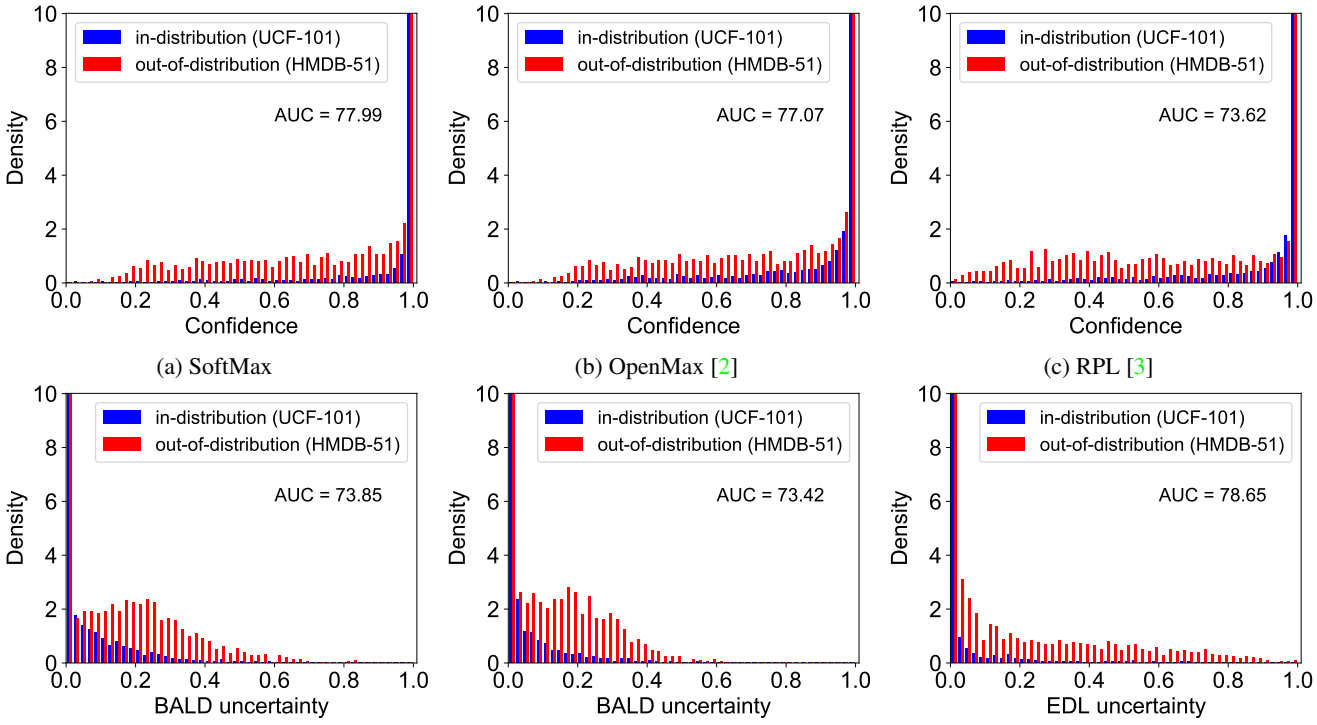
Figure 3: **I3D-based Out-of-distribution Detection with HMDB-51 as Unknown.** Values are normalized to [0,1] within each distribution.



(a) SoftMax (b) OpenMax [2] (c) RPL [3]

(d) MC Dropout (e) BNN SVI [6] (f) DEAR (full)

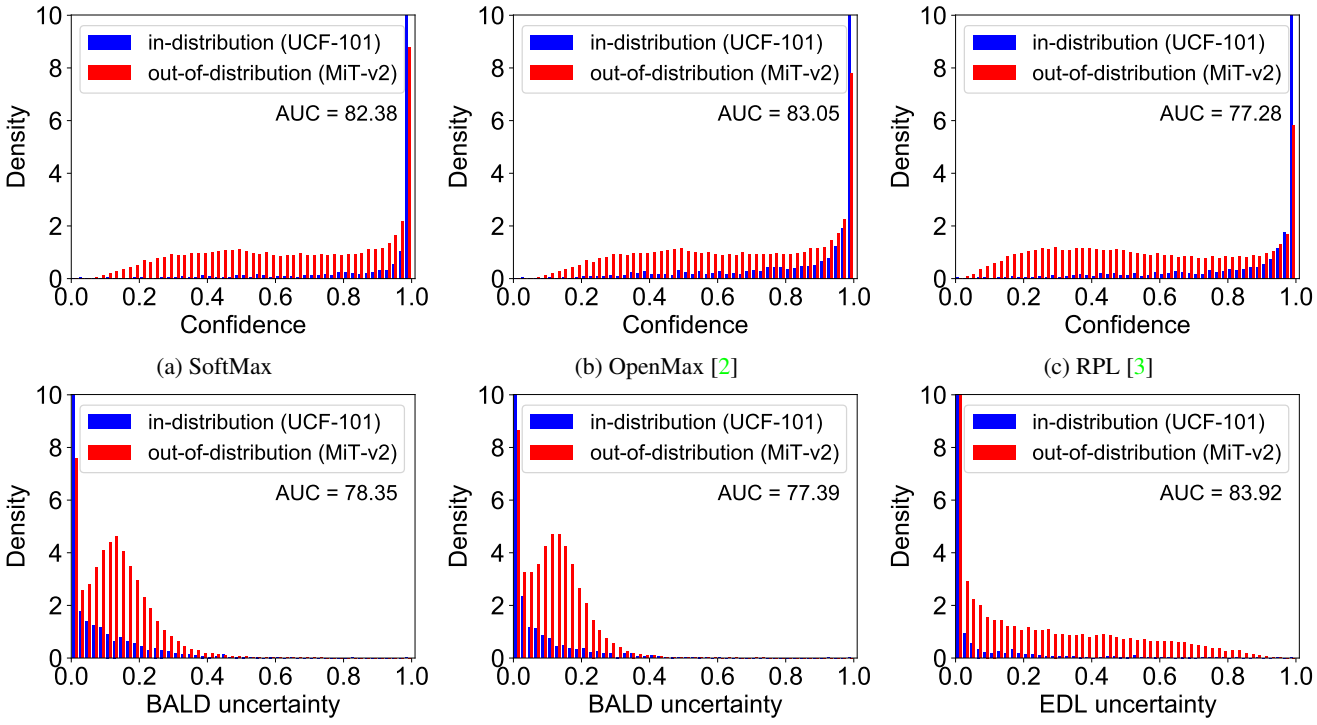
Figure 4: **I3D-based Out-of-distribution Detection with MiT-v2 as Unknown.** Values are normalized to [0,1] within each distribution.



(a) SoftMax (b) OpenMax [2] (c) RPL [3]

(d) MC Dropout (e) BNN SVI [6] (f) DEAR (full)

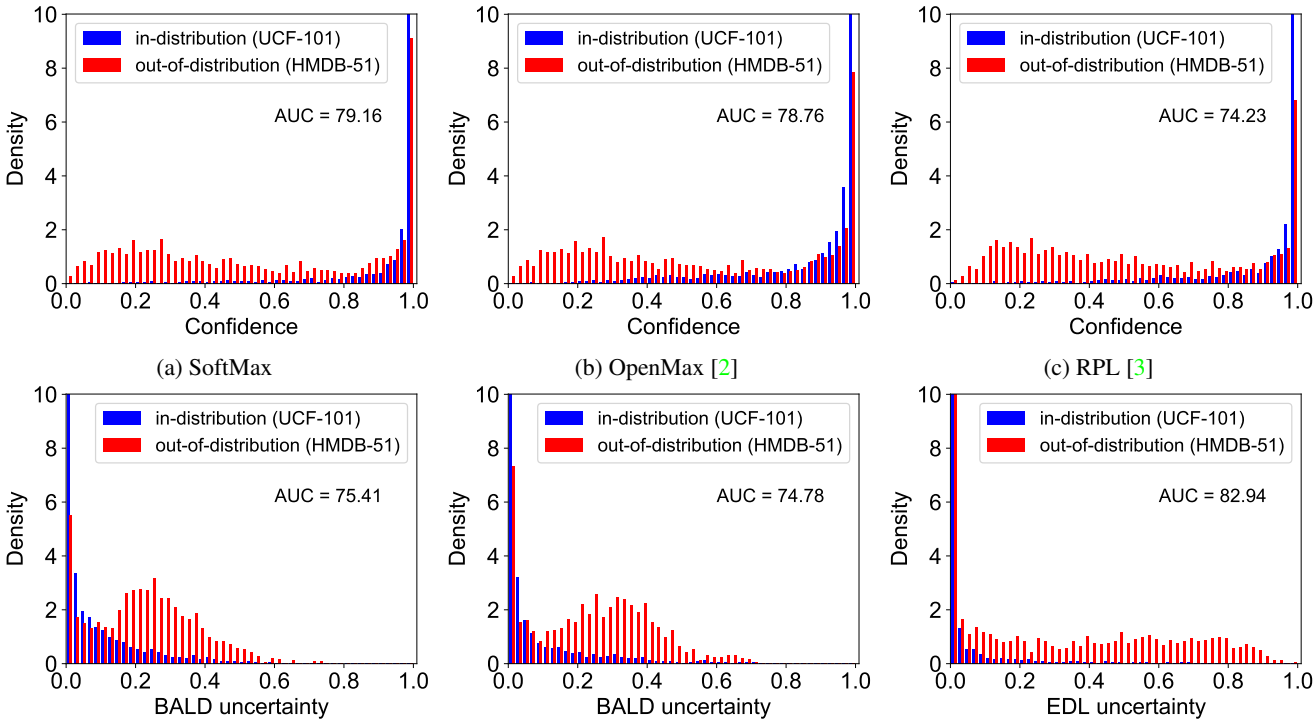
Figure 5: **TSM-based Out-of-distribution Detection with HMDB-51 as Unknown.** Values are normalized to [0,1] within each distribution.



(a) SoftMax (b) OpenMax [2] (c) RPL [3]

(d) MC Dropout (e) BNN SVI [6] (f) DEAR (full)

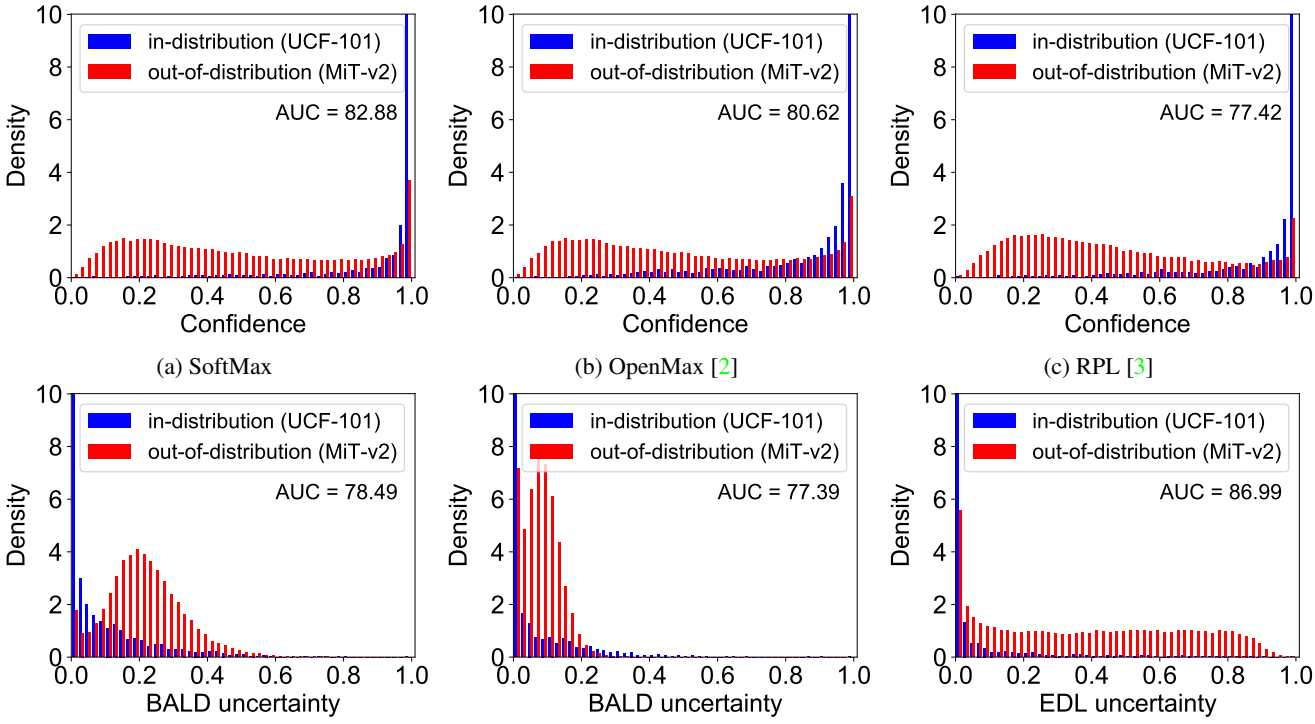
Figure 6: **TSM-based Out-of-distribution Detection with MiT-v2 as Unknown.** Values are normalized to [0,1] within each distribution.



(a) SoftMax (b) OpenMax [2] (c) RPL [3]

(d) MC Dropout (e) BNN SVI [6] (f) DEAR (full)

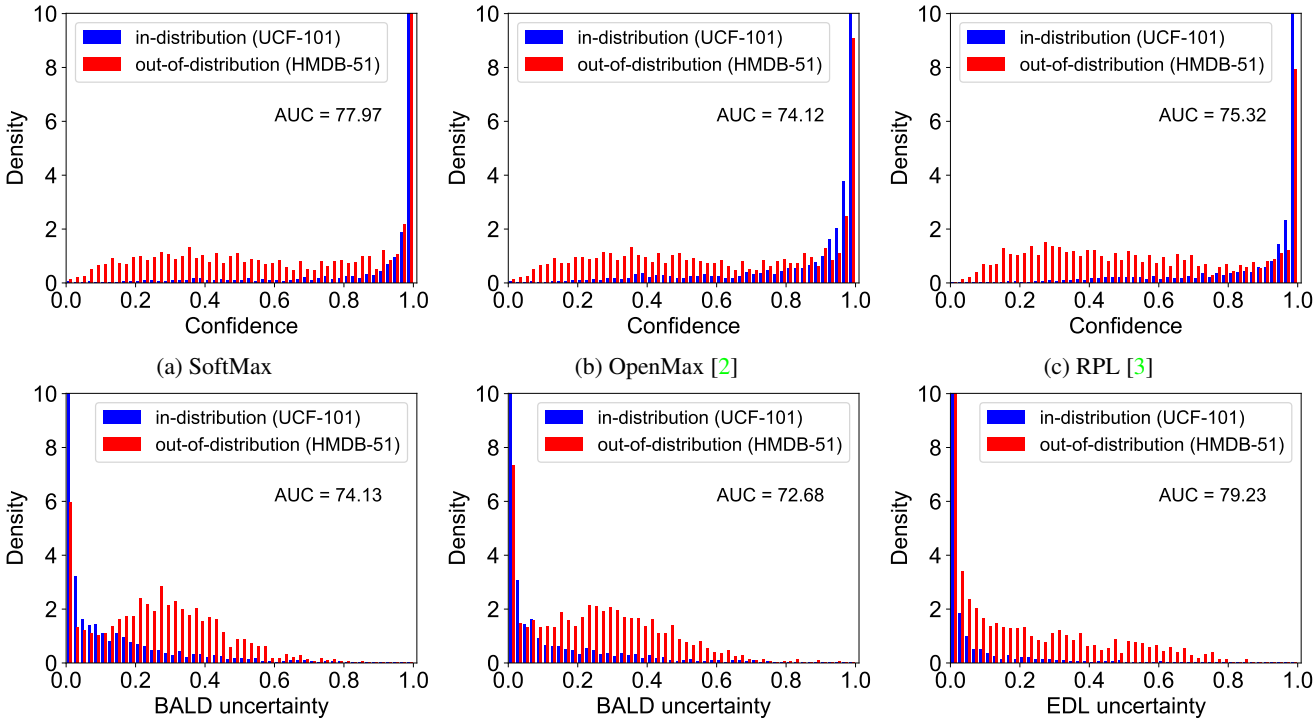
Figure 7: **SlowFast-based Out-of-distribution Detection with HMDB-51 as Unknown.** Values are normalized to [0,1] within each distribution.



(a) SoftMax (b) OpenMax [2] (c) RPL [3]

(d) MC Dropout (e) BNN SVI [6] (f) DEAR (full)

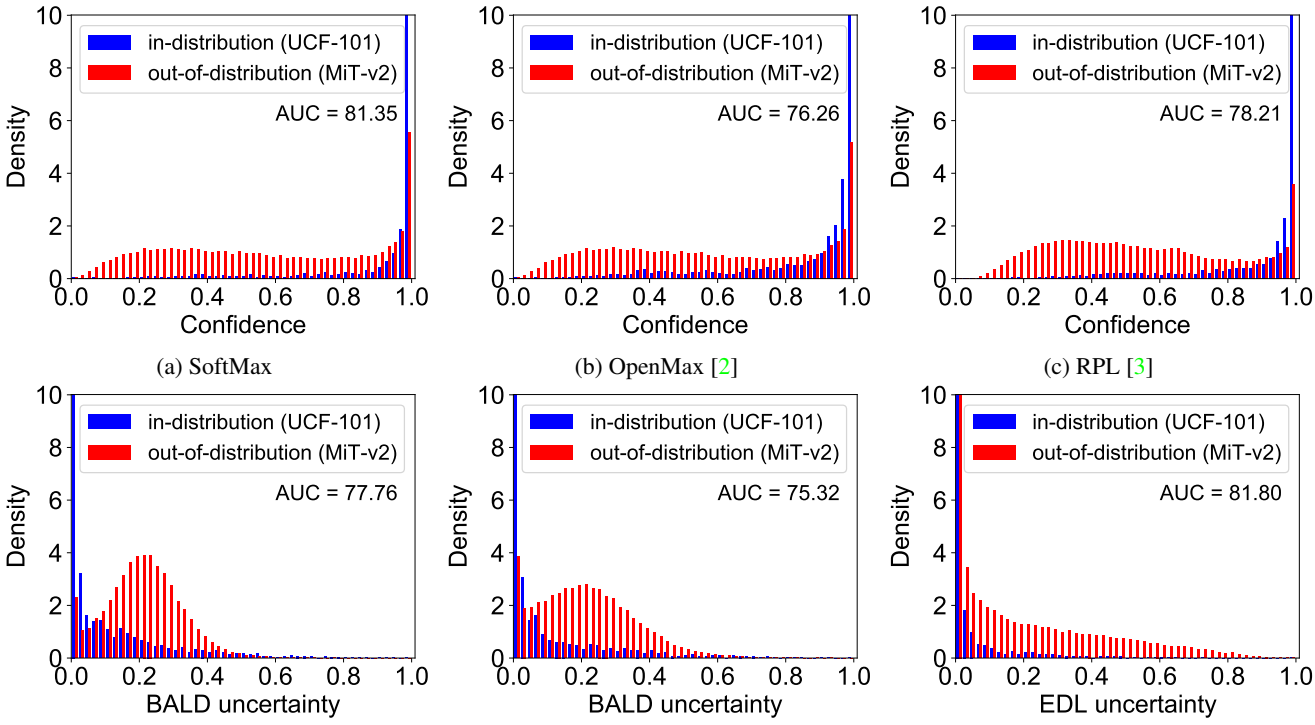
Figure 8: **SlowFast-based Out-of-distribution Detection with MiT-v2 as Unknown.** Values are normalized to [0,1] within each distribution.



(a) SoftMax (b) OpenMax [2] (c) RPL [3]

(d) MC Dropout (e) BNN SVI [6] (f) DEAR (full)

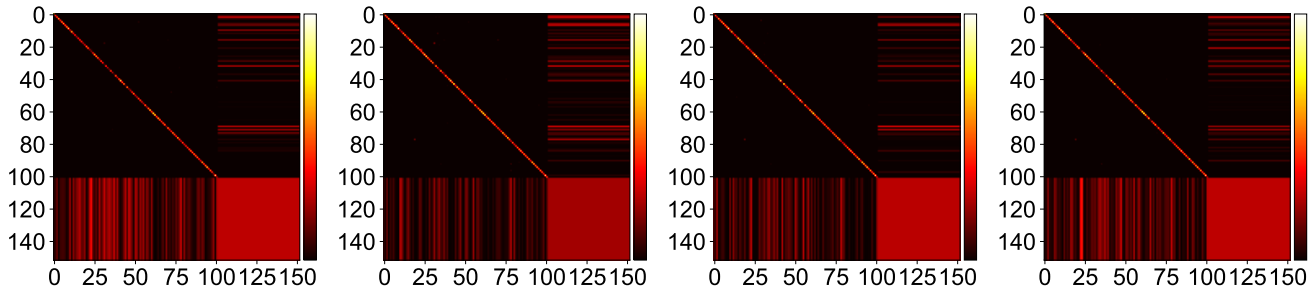
Figure 9: **TPN-based Out-of-distribution Detection with HMDB-51 as Unknown.** Values are normalized to [0,1] within each distribution.



(a) SoftMax (b) OpenMax [2] (c) RPL [3]

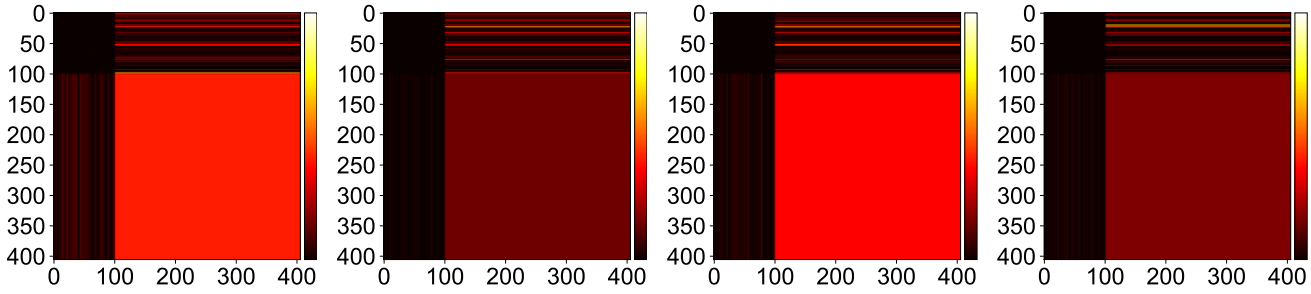
(d) MC Dropout (e) BNN SVI [6] (f) DEAR (full)

Figure 10: **TPN-based Out-of-distribution Detection with MiT-v2 as Unknown.** Values are normalized to [0,1] within each distribution.



(a) I3D (b) TSM (c) SlowFast (d) TPN

Figure 11: **Confusion Matrices of DEAR using HMDB-51 as Unknown.** The x-axis and y-axis represent the ground truth and predicted labels, respectively. The first 101 rows and columns are known classes from UCF-101 while the rest 51 classes are unknown from HMDB-51. Values are uniformly scaled into [0,1] and high value is represented by a lighter color (best viewed in color).



(a) I3D (b) TSM (c) SlowFast (d) TPN

Figure 12: **Confusion Matrices of DEAR using MiT-v2 as Unknown.** The x-axis and y-axis represent the ground truth and predicted labels, respectively. The first 101 rows and columns are known classes from UCF-101 while the rest 305 classes are unknown from MiT-v2. Values are uniformly scaled into [0,1] and high value is represented by a lighter color (best viewed in color).

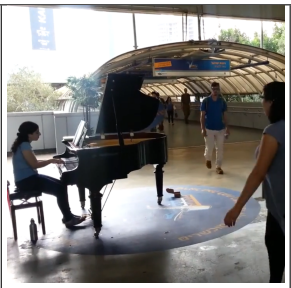
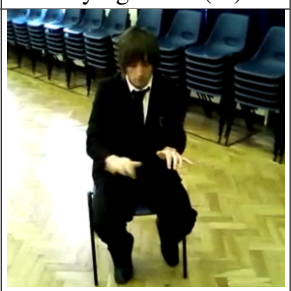
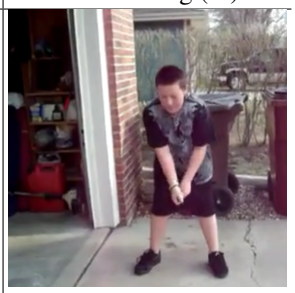
Kinetics (Biased)			
DEAR (w/o CED) DEAR (full)	Playing Volleyball (✗) Playing Piano (✓)	Opening Bottle (✗) Writing (✓)	Shooting Soccer Goal (✗) Golf Driving (✓)
Mimetics (Unbiased)			
DEAR (w/o CED) DEAR (full)	Golf Driving (✗) Playing Piano (✓)	Golf Driving (✗) Writing (✓)	Opening Bottle (✗) Golf Driving (✓)

Figure 13: **Examples of Kinetics and Mimetics.** The check mark (✓) indicates that the predicted label is correct while the cross mark (✗) means that the predicted label is incorrect.