Enlarging Instance-specific and Class-specific Information for Open-set Action Recognition

Jun Cen[1,2*] Shiwei Zhang[2] Xiang Wang[3] Yixuan Pei[4]
Zhiwu Qing[3] Yingya Zhang[2] Qifeng Chen[1]
[1]The Hong Kong University of Science and Technology [2]Alibaba Group
[3]Huazhong University of Science and Technology [4]Xi'an Jiaotong University
jcenaa@connect.ust.hk, {zhangjin.zsw,yingya.zyy}@alibaba-inc.com,
{wxiang,qzw}@hust.edu.cn, peiyixuan@stu.xjtu.edu.cn, cqf@ust.hk

Abstract

Open-set action recognition is to reject unknown human action cases which are out of the distribution of the training set. Existing methods mainly focus on learning better uncertainty scores but dismiss the importance of feature representations. We find that features with richer semantic diversity can significantly improve the open-set performance under the same uncertainty scores. In this paper, we begin with analyzing the feature representation behavior in the open-set action recognition (OSAR) problem based on the information bottleneck (IB) theory, and propose to enlarge the instance-specific (IS) and class-specific (CS) information contained in the feature for better performance. To this end, a novel Prototypical Similarity Learning (PSL) framework is proposed to keep the instance variance within the same class to retain more IS information. Besides, we notice that unknown samples sharing similar appearances to known samples are easily misclassified as known classes. To alleviate this issue, video shuffling is further introduced in our PSL to learn distinct temporal information between original and shuffled samples, which we find enlarges the CS information. Extensive experiments demonstrate that the proposed PSL can significantly boost both the open-set and closed-set performance and achieves state-of-the-art results on multiple benchmarks. Code is available at https://github.com/Jun-CEN/PSL.

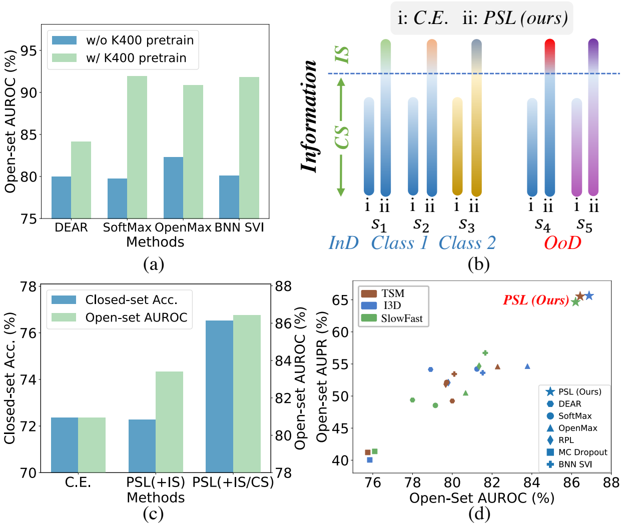


Figure 1. (a) Richer semantic features brought by the pretraining can significantly improve the open-set performance. (b) Information in the feature is divided into IS and CS information. s_4 can be identified as OoD since it has distinct IS information (IS bars in different colors) with s_1 and s_2, while s_5 has distinct CS information (CS bars in different colors) with all InD samples so it may be OoD. Our PSL aims to learn more IS and CS information (bars in longer lengths) than Cross-Entropy (C.E.). (c) Both enlarged IS and CS information boosts the open-set performance. (d) Our PSL achieves the best OSAR performance.

1. Introduction

Deep learning methods for video action recognition have developed very fast and achieved remarkable performance in recent years [1–4]. However, these methods operate under the *closed-set* condition, *i.e.*, to classify all videos into one of the classes encountered during training. This closed-set condition is not practical in the real-world scenario, as videos whose classes are beyond the range of the training set will be misclassified as one of the known classes. Therefore, *open-set action recognition* (OSAR) is proposed to require the network to correctly classify in-distribution (InD) samples and identify out-of-distribution (OoD) samples. InD and OoD classes refer to classes involved and not involved in the training set, respectively.

[*] Work done as an intern at Alibaba DAMO Academy.

Open-set video action recognition is systematically studied in the recent work [5], in which they transfer the existing methods for open-set image recognition into the video domain [6–9] as the baselines, and propose their own method to introduce deep evidential learning [10] to calculate the uncertainty and propose a contrastive evidential debiasing module to alleviate the appearance bias issue in the video domain. All of these methods tend to improve the OSAR performance by calculating a better uncertainty score, based on the feature representations extracted by the neural network (NN). However, the main purpose of training in these methods is still to classify InD samples, which determines the learned feature representations are merely sufficient for InD classification. We find that almost all methods have a significantly better open-set performance when the NN is pretrained with a large dataset (Fig. 1 (a)), so we argue that the diversity of feature representation is extremely important for the OSAR task. Therefore, we propose to boost the open-set ability from the feature representation perspective rather than finding a better uncertainty score.

We first analyze the feature representation behavior in the open-set problem based on the information bottleneck (IB) theory [11, 12]. We divide the information of the feature into *Instance-Specific (IS)* and *Class-Specific (CS) information*. CS information is used for inter-class recognition, so it is similar for samples within the same class but different for samples from other classes. IS information is the special information of each sample within the same class, as two samples cannot be exactly the same even if they belong to the same class. Both CS and IS information are crucial for the open-set task, as illustrated in Fig. 1 (b), where s_4 and s_5 can be identified as OoD samples based on the IS and CS information, respectively. We find that the closed-set classification setting tends to eliminate IS information during training, and cannot fully extract the minimum sufficient CS information for the classification task, so we aim to enlarge IS and CS information in learned feature representations for better OSAR performance.

To enlarge the IS information, we propose the *Prototypical Similarity Learning* (PSL) framework, in which the representation of an instance is encouraged to have less than 1 similarity with the corresponding prototype. In this way, we encourage the IS information to be retained and not eliminated. In addition, [5] finds that OoD videos can be easily classified as InD videos in a similar appearance. To alleviate this issue, we introduce the shuffled video into PSL and make it have less than 1 similarity with the original sample. As the shuffled video almost shares the same appearance information with the original one, we encourage the similarity to be less than 1 so that the network can extract the distinct temporal information among them. We find this technique actually enlarges the CS information in the feature representation. Fig. 1 (c) shows that enlarging the IS information is helpful for the open-set performance, and more CS information can further benefit the open-set and closed-set performance. To summarize, our contributions include:

- We provide a novel perspective to analyze the open-set recognition task based on the information bottleneck theory, and find that the classical closed-set cross-entropy tends to eliminate the IS information which is helpful to identify OoD samples.
- We propose to enlarge the IS and CS information for better OSAR performance. Specifically, PSL is designed to retain the IS information in the features, and we involve video shuffling in PSL to learn more CS information.
- Experiments on multiple datasets and backbones show our PSL's superiority over a large margin compared to other state-of-the-art counterparts, as shown in Fig. 1 (d).

2. Related Work

Action Recognition. Most recent approaches for action recognition are to exploit appearance and motion cues jointly and achieve remarkable success [1–3, 13–16]. Typically, two-stream networks [17–19] consist of two branches that explore spatial information and temporal dynamics, respectively. Some attempts [1, 20, 21] introduce additional temporal mining operations to overcome the limited temporal information extraction ability of 2D CNN. 3D CNN-based methods [2, 3, 22] inflated 2D kernels for joint spatio-temporal modeling. [23] proposes the prototype similarity learning which pushes the learned representation to the corresponding prototype as close as possible, while our PSL keeps the differences among the same class.

Open-set Action Recognition. The related work of OSAR is limited [5, 24–26]. Recently, [5] systematically studies the OSAR problem and transfers several open-set image recognition methods to the video domain, including SoftMax [6], MC Dropout [7], OpenMax [8], and RPL [27]. In the benchmark of [5], the only two methods designed specifically for the video domain are BNN SVI [24] and their proposed DEAR. BNN SVI is a Bayesian NN application in the OSAR, while DEAR adopts the deep evidential learning [10] to calculate the uncertainty, and utilizes two modules to alleviate the over-confidence prediction and appearance bias problem, respectively. Existing methods pursue better uncertainty scores, while the objective of our PSL is to learn more diverse feature representations for better open-set distinguishability.

Information Bottleneck Theory. Based on the IB theory [11, 28], the NN intends to extract minimum sufficient information of the inputs for the current task. More recent [12, 29, 30] adopt the IB theory on unsupervised contrastive learning to analyze the representation learning behavior under the corresponding tasks. In this work, we provide a new view to analyze the OSAR problem based on the IB theory.

3. Information Analysis in OSAR

3.1. Prototypical Learning

Let f be the encoder to extract the information for an input video sample x and output the feature representation $z = f(x), z \in \mathbb{R}^d$. We first define a prototypical learning (PL) loss [31], which is a general version of the cross-entropy (C.E.) loss:

$$\mathcal{L}_{PL} = -\log \frac{\exp(\frac{z^T k_i}{\tau})}{\exp(\frac{z^T k_i}{\tau}) + \sum\limits_{n \in K_i^-} \exp(\frac{z^T n}{\tau})}, \quad (1)$$

where i is the ground truth label of x, $k_i \in \mathbb{R}^d$ is the prototype for class i, τ is a temperature parameter, $K_i^- = \{k_j | j \in \{1, 2, ..., N\}, j \neq i\}$ is the negative prototype set, and N is the number of InD classes. Note that z and k_i are normalized by L2 norm, so that $z^T k_i$ is the cosine similarity. If we regard prototypes as the row vector of the linear classifier $W \in \mathbb{R}^{N \times d}$, and do not normalize z and k as well as remove τ, $\mathcal{L}_{PL}$ will degenerate to the C.E. loss. We introduce the $\mathcal{L}_{PL}$ so that we can directly manipulate the feature representation z.

3.2. Information Analysis of OSAR

Let x_{InD}, z_{InD}, and Y be the random variables of InD sample, extracted representation of InD sample, and the task to predict the label of x_{InD}, where $z_{InD} = f(x_{InD})$. Given the joint distribution of $p(x_{InD}, Y)$, the relevant information between x_{InD} and Y is defined as $I(x_{InD}, Y)$, where I denotes the mutual information [28]. The learned representation z_{InD} satisfies:

$$I(x_{InD}; z_{InD}) = \underbrace{I(x_{InD}; z_{InD}|Y)}_{IS} + \underbrace{I(z_{InD}; Y)}_{CS}, \quad (2)$$

in which $I(x_{InD}; z_{InD}|Y)$ and $I(z_{InD}; Y)$ denote the *Instance-Specific (IS)* and *Class-Specific (CS) information* respectively. In Fig. 2, IS information is blue and orange areas, and CS information is yellow and green areas. CS information is for the closed-set label prediction task Y, while IS information is the special information of each sample that is not related to Y.

To analyze the information about OSAR, we let T be a random variable that represents the task to distinguish OoD samples from InD samples, then we divide the information contained in z_{InD} about T into two parts [12]:

$$I(z_{InD}; T) = \underbrace{I(z_{InD}|Y; T)}_{IS \text{ about } T} + \underbrace{I(z_{InD}; Y; T)}_{CS \text{ about } T}, \quad (3)$$

where $I(z_{InD}|Y; T)$ and $I(z_{InD}; Y; T)$ are the information about the OoD detection task T in IS and CS information (orange and green areas in Fig. 2 respectively). We can see that larger IS and CS information are helpful for OSAR.

In this paper, we aim to enlarge the information about T contained in CS and IS information for better OSAR performance, as illustrated in Fig. 1 (b) and the enlarged green and orange areas in Fig. 2. We first analyze the CS and IS information behaviors under the classical C.E. loss, and find that CS information is encouraged to be maximized but IS information tends to be eliminated in Sec. 3.3. Then we explain this conclusion from the IB theory view in Sec. 3.4.

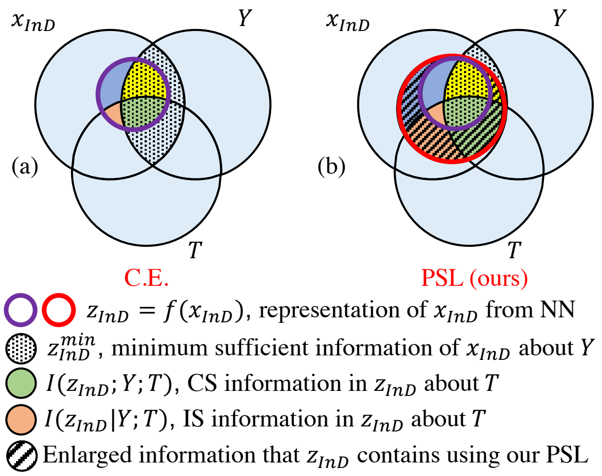

x_{InD}
Y
(a)
T
C.E.
x_{InD}
Y
(b)
T
PSL (ours)
$z_{InD} = f(x_{InD})$, representation of x_{InD} from NN
z_{InD}^{min}, minimum sufficient information of x_{InD} about Y
$I(z_{InD}; Y; T)$, CS information in z_{InD} about T
$I(z_{InD}|Y; T)$, IS information in z_{InD} about T
Enlarged information that z_{InD} contains using our PSL

Figure 2. The neural network (NN) can only extract limited representations z_{InD} of the InD sample x_{InD} for the current task Y (predict the closed-set label), which is not diverse enough for the task T (distinguish OoD samples), as green and orange areas are small in (a). In our PSL, we encourage the NN to learn a more diverse representation so that more IS and CS information about T are contained.

3.3. CS and IS Information Behavior under C.E.

CS information is for closed-set classification task Y, so it is similar for the same class sample, but distinct for the different class sample ($s_1, s_2/s_3$ in Fig. 1). In contrast, IS information is not related to Y and it is distinct for samples in the same class (s_1, s_2 in Fig. 1). Therefore, we have the following proposition which describe the relation between CS/IS information and feature representation similarity.

Proposition 1 *For two feature representations of samples in the same class, more CS information means these two feature representations are more similar, and more IS information decreases their feature similarity.*

CS information is for the closed-set label prediction task Y, which is fully supervised by C.E. loss, so it is maximized during training. In contrast, Eq. 1 shows that C.E. encourages representations of the same class to be exactly same with the corresponding prototype, and such high similarity eliminates the IS information according to Proposition 1. Therefore, **C.E. loss tends to maximize the CS information and eliminate the IS information in the feature representation**. We analyze this conclusion based on Information Bottleneck (IB) theory in next Sec. 3.4.

3.4. IB Theory Analysis for CS and IS Information

Applying the Data Processing Inequality [32] to the Markov chain $Y \rightarrow x_{inD} \rightarrow z_{InD}$, we have

$$I(z_{InD};Y) \leq I(x_{InD};Y). \tag{4}$$

It means that the compressed representation z_{InD} cannot contain more information of Y compared to the original data x_{InD}.

According to the IB theory [11, 28], the NN is to find the optimal solution of z_{InD} with minimizing the following Lagrange:

$$\mathcal{L}[p(z_{InD}|x_{InD})] = I(z_{InD};x_{InD}) - \beta I(z_{InD};Y), \tag{5}$$

where β is the Lagrange multiplier attached to the constrained meaningful condition. Eq. 5 demonstrates the NN is solving a trade-off problem, as the first term tends to keep the information of x_{InD} as less as possible while the second term tends to maximize the information of Y.

Inspired by [12, 33], the sufficient and minimum sufficient representation of x_{InD} about Y can be defined as:

Definition 1 *(Sufficient Representation) A feature representation z^{suf}_{InD} of x_{InD} is sufficient for Y if and only if $I(z^{suf}_{InD};Y) = I(x_{InD};Y)$.*

Definition 2 *(Minimum Sufficient Representation) A sufficient representation z^{min}_{InD} of x_{InD} is minimum if and only if $I(z^{min}_{InD};x_{InD}) \leq I(z^{suf}_{InD};x_{InD})$, $\forall z^{suf}_{InD}$ that is sufficient for Y.*

CS Information Maximization. The goal of training is to optimize f so that $I(z_{InD};Y)$ (CS information) can approximate $I(x_{InD};Y)$, which stays unchanged as data distribution is fixed during training. Therefore, CS information is supposed to be maximized to the upper bound $I(x_{InD};Y)$ because of Eq. 4. In this way, the closed-set classification task pushes the NN to learn the sufficient representation z^{suf}_{InD} according to definition 1 [30].

IS Information Elimination. When z_{InD} is close to the sufficient representation z^{suf}_{InD}, the second term in Eq. 5 will be the fix value $I(x_{InD};Y)$ based on the definition 1. So the key to minimize Eq. 5 is to minimize the first term $I(z^{suf}_{InD};x_{InD})$. Based on the definition 2, the lower bound of $I(z^{suf}_{InD};x_{InD})$ is $I(z^{min}_{InD};x_{InD})$, so we can conclude that the learned representation is supposed to be the minimum sufficient representation z^{min}_{InD} [12]. We substitute $I(z^{suf}_{InD};x_{InD})$ and $I(z^{min}_{InD};x_{InD})$ in definition 2 with Eq. 2 and we have

$$\begin{aligned} &I(x_{InD};z^{min}_{InD}|Y) + I(z^{min}_{InD};Y) \\ \leq &I(x_{InD};z^{suf}_{InD}|Y) + I(z^{suf}_{InD};Y). \end{aligned} \tag{6}$$

As both z^{min}_{InD} and z^{suf}_{InD} are sufficient, the second term of both sides in Eq. 6 is $I(x_{InD};Y)$, so we have

$$0 \leq I(x_{InD};z^{min}_{InD}|Y) \leq I(x_{InD};z^{suf}_{InD}|Y). \tag{7}$$

Therefore, the learned IS information in z^{min}_{InD} is smaller than any IS information in z^{suf}_{InD}, which could be eliminated to 0 [12] (no blue and orange areas in z^{min}_{InD} in Fig. 2).

3.5. Enlarge CS and IS Information for OSAR

Based on the analysis in Sec. 3.3 and Sec. 3.4, we show that C.E. tends to maximize the CS information and eliminate the IS information in the feature representation. Both larger IS and CS information are crucial for OSAR according to Eq. 3, but C.E. does not bring the optimal information. On the one hand, IS information is eliminated so we lose a part of information which is beneficial for the OSAR. On the other hand, the learned representation is not sufficient and does not contain enough CS information in practice due to the model capacity and data distribution shift between training and test sets, which can be supported by the fact that test accuracy cannot reach 100%. Therefore, we propose our method to enlarge the CS and IS information for better OSAR performance in next Sec. 4.

4. Methods

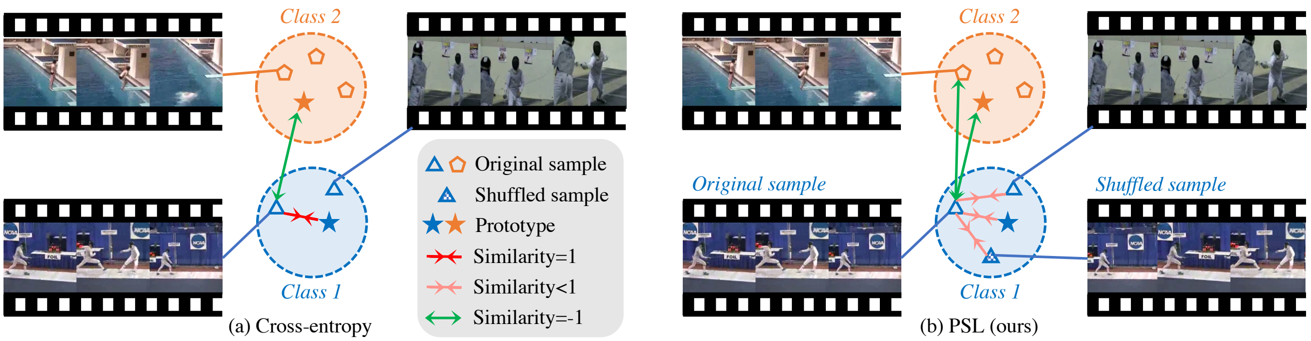
4.1. Prototypical Similarity Learning

According to Sec. 3.3, we notice that IS information is suppressed by the C.E. loss and a key reason is C.E. encourages feature representations of the same class to be exactly same. Therefore, we argue that the feature representation of the same class samples should have a similarity $s < 1$. In other words, we aim to **keep the intra-class variance which prevents intra-class collapse to retain IS information**. Based on the classical PL loss Eq. 1, we develop prototypical similarity learning (PSL):

$$\mathcal{L}_{PSL} = -\log \frac{\exp(\frac{1-\left|z^T k_i - s\right|}{\tau})}{\exp(\frac{1-|z^T k_i - s|}{\tau}) + \sum\limits_{n \in K_i^-} \exp(\frac{z^T n}{\tau})}, \tag{8}$$

where s and τ are fixed hyperparameters. In this way, we expect the prototype k_i to act as the CS information for the InD class i, which is used to predict the label, and the dissimilarity between the z and k_i represents the IS information. Traditional PL loss (or C.E. loss) encourages the features of samples in the same classes to be as tight as possible, while our PSL aims to keep the variance within the same class.

However, we find Eq. 8 will converge to the trivial solution, where the z converges to the training result of Eq. 1 and only k_i shifts. To solve this problem, we introduce the similarity between different samples within a mini-batch into the denominator of Eq. 8. In this way, we directly constrain the relationship between sample features instead of only supervising the similarity between the sample feature and its prototype. We name the modified loss as PSL with contrastive terms (CT):



(a) Cross-entropy

(b) PSL (ours)

Figure 3. (a) C.E. encourages the sample feature z to be exactly same with the corresponding prototype k_i. (b) Our PSL encourages the similarity between z and k_i, features of shuffled sample Q_{shuf} and other samples in the same class Q_{sc} to have a similarity less than 1.

$$\mathcal{L}_{PSL}^{CT} = \frac{\exp(\frac{1-\left|z^T k_i - s\right|}{\tau})}{\exp(\frac{1-|z^T k_i - s|}{\tau}) + \sum\limits_{n \in Q_n} \exp(\frac{z^T n}{\tau}) + \sum\limits_{p \in Q_{sp}} \exp(\frac{|z^T p - s|}{\tau})}, \tag{9}$$

where $Q_n = K_i^- \cup Q_{ns}$. Q_{ns} refers to the negative samples, *i.e.*, samples in other classes, and Q_{sp} refers to the soft positive samples which contains samples in the same class Q_{sc} here. The reason we call soft positive samples is that we think samples in the same class share CS information but have distinct IS information.

4.2. Video Shuffling for PSL

PSL aims to keep IS information during training, and in this section we introduce how to enlarge CS information through video shuffling. The appearance bias is a significant problem in the OSAR. For instance, the OoD classes *Smile* and *Chew* are easily classified as InD classes *ApplyEyeMakeup* and *ApplyLipstick*, as the majority area of all these classes are occupied by a face, as shown in Fig. 7. The NN is confused by the extremely similar spatial information and neglects the minor different temporal information. This phenomenon encourages us to strengthen the temporal information extraction ability of the NN to distinguish classes with very similar appearances but different actions. We find that introducing a simple yet effective way, *i.e.*, to regard the shuffled video Q_{shuf} as the soft positive sample in Eq. 9, is extremely suitable and useful in our PSL framework. In this case, $Q_{sp} = Q_{sc} \cup Q_{shuf}$. Shuffled video means shuffling the frames within a single video. As the appearance information of the shuffled video is almost the same as the original video, a smaller than 1 similarity forces the NN to learn the distinct temporal information between them. Unlike existing works which predict the sequence or the type of the shuffled video [34–37], we regard the shuffle video as a whole sample and directly compare its feature representation with the original video in our PSL. **We find this technique can improve the closed-set accuracy which indicates more CS information is learned**. We summarize the difference between our PSL and classical C.E. in Fig. 3.

4.3. Uncertainty Score

As our PSL aims to learn richer CS and IS information in the feature representation, we use the Mahalanobis distance to measure the uncertainty as it can be calculated from the feature representation perspective [38, 39]:

$$u = (z - \mu_m)^T \textstyle\sum_m^{-1} (z - \mu_m), \tag{10}$$

where μ_m and $\sum_m$ denote the mean and covariance of the whole training set features, and z is the test sample feature.

5. Experiments

Datasets. Following [5], we use UCF101 [40] as the InD dataset for training and closed-set evaluation, and use HMDB51 [41] and MiT-v2 [42] as OoD data for open-set evaluation. Different from [5] which does not clean the OoD data that may contains InD classes, we remove the overlapping classes between InD and OoD dataset during evaluation. See Appendix A for more details.

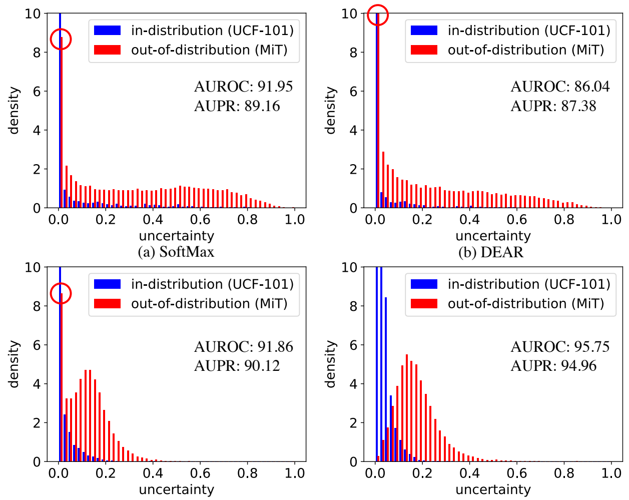
Evaluation protocols. For closed-set performance, we evaluate like the traditional way to calculate the top-1 accuracy Acc. (%). For open-set performance, we follow the classical open-set recognition protocol [6, 43] to use the obtained uncertainty score Eq. 10 to calculate AUROC (%), AUPR (%) and FPR95(%).[1]

Implementation details. For Kinetics400 (K400) [3] pretrained model, our implementation setting is the same with [5]. The base learning rate is 0.001 and step-wisely decayed every 20 epochs with total of 50 epochs. We argue that as K400 is extremely large, the K400 pretrained model may already have seen the OoD data used in inference, so

[1] We find AUROC in [5] only considers one specific threshold based on their code, and after discussion and agreement they provide the modified correct score in our Tab. 1. See Appendix B for details.

		w/o K400 Pretrain				w/ K400 Pretrain			
Datasets	**Methods**	**AUROC↑**	**AUPR↑**	**FPR95↓**	**Acc.↑**	**AUROC↑**	**AUPR↑**	**FPR95↓**	**Acc.↑**
UCF101 (InD) HMDB51 (OoD)	OpenMax [8]	*82.28*	*54.59*	*50.69*	*73.92*	90.89	73.16	38.77	95.32
	MC Dropout [7]	75.75	41.21	54.78	73.63	88.23	67.62	38.12	95.06
	BNN SVI [24]	80.10	53.43	52.33	71.51	*91.81*	*79.65*	31.43	94.71
	SoftMax [6]	79.72	52.13	53.22	*73.92*	91.75	77.69	*28.60*	95.03
	RPL [27]	79.67	51.85	56.40	71.46	90.53	77.86	37.09	*95.59*
	DEAR [5]	80.00	49.23	53.28	71.33	84.16	75.54	89.40	94.48
	PSL(ours)	**86.43**	**65.54**	**41.67**	**76.53**	**94.05**	**86.55**	**23.18**	**95.62**
	Δ	**(+4.15)**	**(+10.95)**	**(-9.02)**	**(+2.61)**	**(+2.24)**	**(+6.90)**	**(-5.42)**	**(+0.03)**
UCF101 (InD) MiTv2 (OoD)	OpenMax [8]	*84.43*	*76.69*	*47.74*	*73.92*	*93.34*	88.14	*28.95*	95.32
	MC Dropout [7]	75.66	62.20	51.57	73.63	88.71	83.36	39.46	95.06
	BNN SVI [24]	79.48	71.73	52.52	71.51	91.86	*90.12*	36.21	94.71
	SoftMax [6]	80.55	73.17	50.49	*73.92*	91.95	89.16	32.00	95.03
	RPL [27]	80.21	72.04	52.83	71.46	90.64	88.79	38.43	*95.59*
	DEAR [5]	79.00	67.10	52.44	71.33	86.04	87.38	87.40	94.48
	PSL(ours)	**86.53**	**79.95**	**40.99**	**76.53**	**95.75**	**94.96**	**18.96**	**95.62**
	Δ	**(+2.10)**	**(+3.26)**	**(-6.75)**	**(+2.61)**	**(+2.41)**	**(+4.84)**	**(-9.99)**	**(+0.03)**

Table 1. Comparison with state-of-the-art methods on **HMDB51 and MiTv2 (OoD)** using TSM backbone. Acc. refers to closed-set accuracy. AUROC, AUPR and FPR95 are open-set metrics. Best results are in **bold** and second best results in *italic*. The gap between best and second best is in **blue**. DEAR and our methods contain video-specific operation.



(a) SoftMax

(b) DEAR

(c) BNN SVI

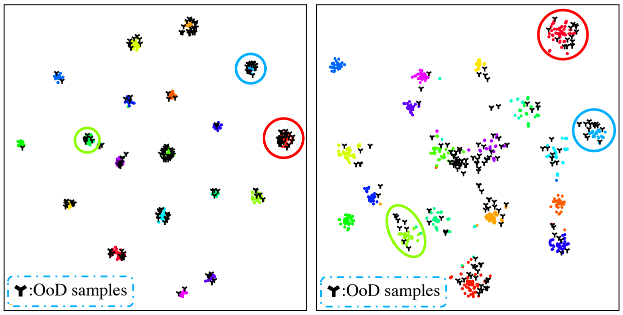
(d) PSL (ours)

Figure 4. The uncertainty distribution of InD and OoD samples of (a) Softmax, (b) DEAR, (c) BNN SVI and (d) our PSL method.

we conduct experiments from scratch (no ImageNet pretrained) to ensure that OoD data is absolutely unavailable during training. We use the LARS optimizer [44] and set the base learning rate and momentum as 0.6 and 0.9 with total of 400 epochs. The experiments are conducted on TSM [1], I3D [3] and SlowFast [2]. The batch size for all methods is 256. More details are in Appendix C.

5.1. Evaluation Results

Comparison with state-of-the-art. We report the results on HMDB51 (OoD) and MiT-v2 (OoD) in Table 1 using TSM backbone [1]. The evaluation results of other backbones including I3D and SlowFast are in the Appendix D. We can see that for w/ or w/o K400 pretrain, our PSL method has significantly better open-set and closed-set performance than all baselines. The uncertainty distribution



(a) Cross-entropy

(b) PSL (ours)

Figure 5. Feature representation visualization of cross-entropy and our PSL method. OoD samples are in black and InD samples are in other colors. In the red, blue and green circles, it is clear that OoD samples distribute at the edge of InD samples in our PSL, while greatly overlap with each other in the cross-entropy method.

of InD and OoD samples are depicted in Fig. 4 for MiT-v2 (OoD) with K400 pretrained. Three baseline methods have a clear over confidence problem, *i.e.*, the far left column is extremely high (red circles in Fig. 4), which means a large number of OoD samples have almost 0 uncertainty, while our method significantly alleviates this problem through the distinct representation of OoD samples, illustrated in Fig. 5. Besides, we can find that the open-set performance w/ K400 pretrain is higher than w/o pretrain for almost all methods in Table 1 and Fig. 1 (a), which can testify the importance of richer semantic representation for OSAR.

Comparison with metric learning methods. Our method concentrates on the feature representation aspect for the OSAR problem, so we also implement several well-known metric learning methods and show the result in Table 3. The evaluation is conducted using TSM model and OoD dataset is HMDB51. We do not use video shuffling in our method for fair comparison. We can see that our method still achieves the best open-set performance. The most im-

					InD		OoD					
	s	Q_{ns}	Q_{sc}	Q_{shuf}	**Mean**	**Variance**	**Mean**	**Variance**	**AUROC↑**	**AUPR↑**	**FPR95↓**	**Acc.↑**
$\mathcal{L}_{PL}$	✗	✗	✗	✗	0.81	0.0015	0.63	0.0029	80.95	52.79	52.51	72.36
$\mathcal{L}_{PSL}$	✓	✗	✗	✗	0.79	0.0016	0.62	0.0028	81.79	54.16	52.33	72.33
$\mathcal{L}_{PSL}^{CT}$	✓	✓	✗	✗	0.71	0.0022	0.61	0.0036	82.60	57.36	50.03	72.17
	✓	✓	✓	✗	0.71	0.0023	0.49	0.0035	83.42	59.05	51.32	72.28
	✓	✓	✓	✓	0.74	0.0016	0.63	0.0029	86.43	65.58	41.75	77.19

Table 2. Abaltion results of different components in $\mathcal{L}_{PSL}^{CT}$.

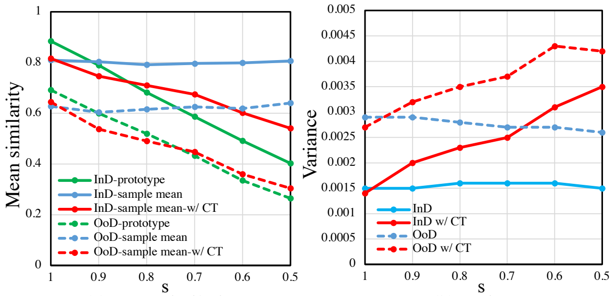
	AUROC↑	AUPR↑	FPR95↓	Acc.↑
SoftMax	80.95	52.79	52.51	72.36
Triplet [45]	81.02	54.75	53.88	75.50
Normface [46]	80.99	54.90	53.19	73.34
Circle [47]	78.76	51.65	55.27	72.15
Arcface [48]	81.23	55.03	53.67	**75.95**
LSoftMax [49]	80.87	54.01	52.29	73.05
PSL($s = 0.8$)	**83.42**	**59.05**	**51.32**	72.28
PSL($s = 0.6$)	82.75	58.57	52.27	73.26

Table 3. Comparison with different metric learning methods.

portant difference between our method and all other metric learning methods is that they aim to push the features of one class as tight as possible like C.E., while our method aims to keep the feature variance within a class to retain IS information. We calculate the mean similarity between the sample feature and the corresponding class center. The mean similarity ranges from 0.77 to 0.82 for other metric learning methods, while mean similarity is 0.71 ($s = 0.8$) and 0.6 ($s = 0.6$) for our PSL. So our method has looser feature distribution within a class, as shown in Fig. 5.

5.2. Ablation Study

Contrastive terms in $\mathcal{L}_{PSL}^{CT}$ for IS information. The intuition of PSL is to keep the intra-class variance to retain the IS information which is helpful for OSAR. We expect that the representation z within a class has a similarity $s < 1$ with the prototype k_i, so each sample can keep its own IS information. However, we find that the loss $\mathcal{L}_{PSL}$ may lead the network to find the trivial representation of samples z which is similar to using loss $\mathcal{L}_{PL}$, where only k_i shifts and z does not. We calculate the mean of similarity $sim(z, \bar{z}_i)$, where $\bar{z}_i$ denotes the mean representation of all samples in the same class i, and the mean of similarity with the corresponding prototype $sim(z, k_i)$, as well as the feature variance in all dimensions. Fig. 6 (a) and (b) show that with the hyper-parameter s decreasing, the $sim(z, k_i)$ decreases as expected by $\mathcal{L}_{PSL}$ (green curves), but the $sim(z, \bar{z}_i)$ and variance stay unchanged (blue curves), meaning that the representation of samples are still similar with using $\mathcal{L}_{PL}$, and only the prototypes are pushed away by the sample representations. In contrast, with CT in $\mathcal{L}_{PSL}^{CT}$, the $sim(z, \bar{z}_i)$ decreases and variance increases with s decreases (red curves), indicating that CT is significantly effective to keep the intra-class variance.



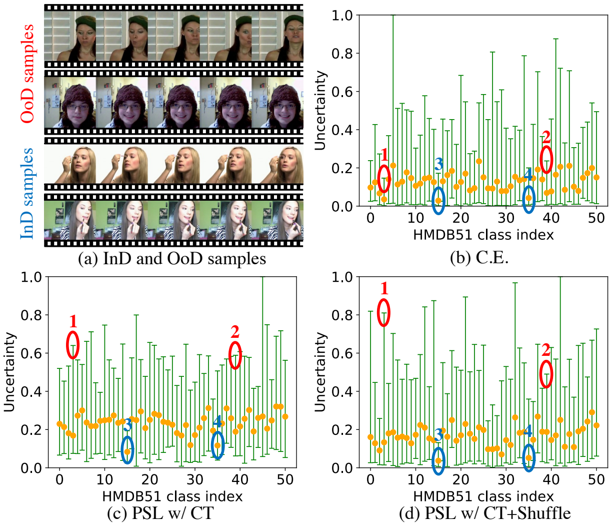
(a) Mean similarity (b) Variance

Figure 6. Mean similarity and variance analysis for CT terms.

To individually study the effectiveness of Q_{ns} and Q_{sc} in $\mathcal{L}_{PSL}^{CT}$, we provide the ablation results in Table 2. For OoD samples, we calculate the similarity with the mean representation of its predicted class. Table 2 shows that using Q_{ns} alone can significantly increase the intra-class variance for both InD and OoD samples, meaning the pushing effect of representations in other classes can implicitly help retain the IS information. On top of that, Q_{sc} can further learn more IS information that is helpful to distinguish OoD samples, as the mean similarity of InD samples stay unchanged, but OoD samples are smaller which means OoD samples are far away from InD samples.

Shuffled videos for CS information. Tab. 2 shows that Q_{shuf} can improve both closed-set and open-set performance, which proves introducing shuffled videos in PSL can enlarge CS information. Smaller intra-class variance brought by Q_{shuf} testify Proposition 1 that more CS information means more similar features within the same class.

We draw the uncertainty of all classes in HMDB51, as shown in Fig. 7. Note that some classes in HMDB51 are actually InD as they appear in the UCF101, like the class 3 *golf* and 4 *shoot bow* in Fig. 7. We find that in C.E. some OoD classes have extremely low uncertainty, such as class 1 *chew* and 2 *smile*, because they are spatially similar to some InD classes like *ApplyEyeMakeup* and *ApplyLipstick* in Fig. 7 (a). Comparing (b) and (c) shows that our PSL can increase the average uncertainty of OoD classes (higher yellow points), and some OoD classes which are similar to InD classes like 1 and 2 have much higher uncertainty in our PSL method. After shuffled samples are involved, some InD classes whose uncertainty are increased in (c) like 3 and 4 have lower uncertainty in (d), and the uncertainty of some



(a) InD and OoD samples
(b) C.E.
(c) PSL w/ CT
(d) PSL w/ CT+Shuffle

Figure 7. (a) *chew* and *smile* are OoD samples from HMDB51, and *ApplyEyeMakeup* and *ApplyLipstick* are InD samples from UCF101. (b-d) Uncertainty distribution of each class in HMDB51. Class 1: *chew*, 2: *smile*, 3: *golf*, 4: *shoot bow*. Classes 1 and 2 are OoD while 3 and 4 are InD.

$s(Q_{shuf})$	$s(Q_{sc})$	AUROC↑	AUPR↑	FPR95↓	Acc.↑
0.7	0.7	85.25	63.91	48.34	76.98
0.5		86.03	64.36	43.70	76.53
0.3		83.80	60.42	48.76	75.50
0		79.54	50.59	54.43	72.59
0.8	0.8	86.43	65.58	41.75	76.53
0.9	0.9	83.12	57.04	46.84	73.31
1	1	82.04	53.82	51.82	72.89

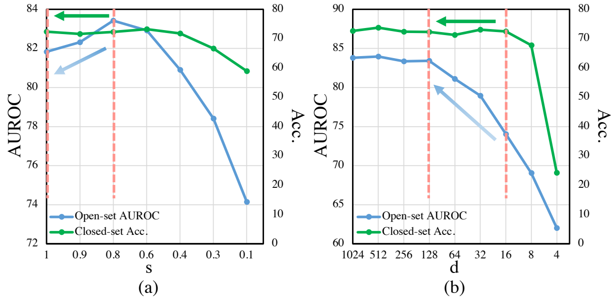
Table 4. Ablation study of similarity s for Q_{shuf} and Q_{sc}.

OoD classes sharing similar appearance with InD classes like class 1 is further improved.

Q_{sp} in Eq. 9 contains Q_{shuf} and Q_{sc}, so we analyze whether should we assign the same s for the shuffled video Q_{shuf} and other videos in the same class Q_{sc}. Tab. 4 shows that the same s have good enough performance. So we set the same s for Q_{shuf} and Q_{sc} in the default setting to reduce the number of hyper-parameters.

5.3. Discussion

Both CS and IS information are useful. We provide the closed-set and open-set performance under different hyper-parameter s and feature dimension d in Fig. 8. (a) shows that $s = 0.8$ has better open-set performance than $s = 1$ and has comparable closed-set accuracy, which illustrates that retaining the IS information which is eliminated by C.E. ($s = 1$) is beneficial. When $s < 0.8$, the NN cannot learn enough CS information, so both closed-set and open-set performance drops. Therefore, a proper mixture of CS and IS information is ideal. (b) shows that when d grows from 4 to 16, more CS information is contained so that both closed-set and open-set performance improves. When d grows from 16 to 128, the feature does not include more CS information as closed-set accuracy is comparable. However, open-set performance keeps increasing which means more IS information is contained based on more feature dimensions. This interesting experiment shows that enough information for closed-set recognition is not enough for open-set recognition because IS information is not related to the closed-set task but useful for the open-set task.



(a)
(b)

Figure 8. Ablation study of similarity s and feature dimension d.

Epoch	Mean	Variance	AUROC↑	Acc-Test.↑	Acc-Train.↑
200	0.577	3.3e-3	75.08	68.39	99.85
400	0.602	3.1e-3	82.92	73.26	100
800	0.613	3.0e-3	82.54	73.29	100

Table 5. Training process analysis when $s = 0.6$ w/o Q_{shuf}.

Feature variance and open-set performance analysis. Fig. 8 (a) shows that when features get looser ($s = 1-0.8$), the open-set performance is improved, but if features get continually looser ($s = 0.8-0.1$), the open-set performance drops. So there is no strict relation between the feature variance and open-set performance. One may argue that continual training can benefit the open-set performance [50], which is alongside with smaller feature variance [51]. We show that the benefit of continual training comes from better closed-set performance, not tighter features. Tab. 5 shows that when we train the model from 200 to 400 epochs, the closed-set accuracy is higher, and feature is tighter (larger mean similarity and smaller variance), and the open-set performance is better. But from epoch 400 to 800 we find the model is already overfitted to the training set, as the accuracy of test set remains unchanged. So although the features get tighter in the 800 epoch, both the closed-set and open-set performance remain same.

6. Conclusion

We analyze the OSAR problem from the information perspective, and show that cross-entropy tends to eliminate IS information and cannot fully learns CS information which are both useful for the open-set task. So we propose PSL to retain IS information and introduce shuffle videos into PSL to enlarge CS information. Comprehensive experiments demonstrate the effectiveness of our PSL and the importance of IS and CS information in the OSAR task.

Acknowledgements This work is supported by Alibaba Group through Alibaba Research Intern Program.

References

[1] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 1, 2, 6

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *ICCV*, 2019. 1, 2, 6

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 1, 2, 5, 6

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

[5] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *ICCV*, 2021. 2, 5, 6

[6] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 2, 5, 6

[7] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2, 6

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

[9] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *NeurIPS*, 2017. 2

[10] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *NeurIPS*, 2020. 2

[11] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop*, 2015. 2, 4

[12] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *CVPR*, 2022. 2, 3, 4

[13] Ziyuan Huang, Shiwei Zhang, Liang Pan, Zhiwu Qing, Mingqian Tang, Ziwei Liu, and Marcelo H Ang Jr. TAda! Temporally-adaptive convolutions for video understanding. In *ICLR*, 2022. 2

[14] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yi Xu, Xiang Wang, Mingqian Tang, Changxin Gao, Rong Jin, and Nong Sang. Learning from untrimmed videos: Self-supervised video representation learning with hierarchical consistency. In *CVPR*, 2022. 2

[15] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. OadTR: Online action detection with Transformers. In *ICCV*, 2021. 2

[16] Yixuan Pei, Zhiwu Qing, Jun Cen, Xiang Wang, Shiwei Zhang, Yaxiong Wang, Mingqian Tang, Nong Sang, and Xueming Qian. Learning a condensed frame for memory-efficient video class-incremental learning. In *NeurIPS*, 2022. 2

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

[18] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 2

[19] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 2

[20] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 2

[21] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. TDN: Temporal difference networks for efficient action recognition. In *CVPR*, 2021. 2

[22] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015. 2

[23] Lei Bai, Lina Yao, Xianzhi Wang, Salil S Kanhere, and Yang Xiao. Prototype similarity learning for activity recognition. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2020. 2

[24] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. BAR: Bayesian activity recognition using variational inference. In *NeurIPS Workshops*, 2018. 2, 6

[25] 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*, 2018. 2

[26] Yang Yang, Chunping Hou, Yue Lang, Dai Guan, Danyang Huang, and Jinchen Xu. Open-set human activity recognition based on micro-doppler signatures. *Pattern Recognition*, 2019. 2

[27] 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. 2, 6

[28] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 2, 3, 4

[29] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS*, 2020. 2

[30] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2021. 2, 4

[31] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Robust classification with convolutional prototype learning. In *CVPR*, 2018. 3

[32] Thomas M Cover. *Elements of information theory*. 1999. 4

[33] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 2018. 4

[34] Qinghongya Shi, Hong-Bo Zhang, Zhe Li, Ji-Xiang Du, Qing Lei, and Jing-Hua Liu. Shuffle-invariant network for action recognition in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2022. 5

[35] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *ECCV*, 2020. 5

[36] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, 2017. 5

[37] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017. 5

[38] Vikash Sehwag, Mung Chiang, and Prateek Mittal. SSD: A unified framework for self-supervised outlier detection. In *ICLR*, 2021. 5

[39] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018. 5

[40] 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. 5

[41] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, 2011. 5

[42] Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogerio 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*, 2021. 5

[43] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 5

[44] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 6

[45] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 7

[46] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. NormFace: L2 hypersphere embedding for face verification. In *ACM MM*, 2017. 7

[47] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle Loss: A unified perspective of pair similarity optimization. In *CVPR*, 2020. 7

[48] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 7

[49] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, 2016. 7

[50] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *ICLR*, 2022. 8

[51] XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. In *ICLR*, 2022. 8