Appendices of "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

A. Datasets

We follow the datasets setting in [1]. The training InD dataset is UCF101, which contains 101 classes with 9537 training samples and 3783 test samples. The OoD datasets for open-set evaluation are HMDB51 and MiT-v2. We use the test sets of them which contain 1530 samples and 30500 samples respectively. For UCF101 and HMDB51, we follow the MMAction [2] to use the split 1 for training and evaluation, which is the same with [1]. Note that in [1], they find some classes in HMDB51 overlap with those in UCF101 but they do not clean them. We remove the overlapping classes in UCF101 and HMDB51 so that OoD data does not contain any samples of InD classes. The classes we remove in HMDB51 and the corresponding same classes in UCF101 are in Table 1.

HMDB51	35, Shoot bow	29, Push up	15, Golf	26, Pull up
UCF101	2, Archery	71, PushUps	32, GolfSwing	69, PullUps
HMDB51	30, Ride bike	34, Shoot ball	43, Swing baseball	31, Ride horse
UCF101	10, Biking	7, Basketball	6, BaseballPitch	41, HorseRiding

Table 1. Overlapping classes in HMDB51 and UCF101.

B. Evaluation protocols

Based on codes provided by [1], we find that their evaluation metrics including Open maF1 and AUORC are both calculated under a specific certain threshold, *i.e.*, a sample whose uncertainty is larger than the threshold will be considered as an OoD sample. The threshold is determined by top 5% uncertainty in the training set. This is contradictory with the classical metrics in the open-set image recognition, in which common metrics including AUROC and AUPR [3, 4] both consider all thresholds. Each point on the ROC and PR curve is based on one specific threshold, and the area under ROC and PR curve is regarded as the comprehensive result of all thresholds. After discussing with authors in [1], they admit that the AUROC, AUPR and FPR95 which are served as the classical metrics in the open-set image recognition are more suitable for the OSAR problem. So they modify the corresponding code and we provide the correct results in the Table 1 in our paper. We provide a comparison between the result of considering only one threshold and all thresholds in Table 2. The results show that no matter for only considering one threshold or all thresholds, our PSL method can both outperform all methods.

When we use MiT-v2 as the OoD dataset, we find the imbalance problem, which is also mentioned in [1]. The MiT-v2 test set contains 30500 samples while UCF101 test set only contains 3783 samples. This will cause the AUPR to be close to 100% if we regard all samples in MiT-v2 as OoD samples during evaluation. Therefore, we divide the MiT-v2 test set into 10 splits, and evaluate the open-set metrics for 10 times and calculate the mean as the final result. A comparison between the results of evaluating 10 times and 1 time is shown in Table 3. The results illustrate that when we use all samples in MiT-v2 for

*Work done as in intern at Alibaba DAMO Academy.

Models	Methods	One threshold [1]				All thresholds (ours)			
		AUROC↑	AUPR↑	FPR95↓	Acc.↑	AUROC↑	AUPR↑	FPR95↓	Acc.↑
TSM	OpenMax	84.18	76.52	100	95.32	90.89	73.16	38.77	95.32
	MC Dropout	78.50	71.11	37.80	95.06	88.23	67.62	38.12	95.06
	BNN SVI	77.77	71.00	41.13	94.71	91.81	79.65	31.43	94.71
	SoftMax	82.77	74.33	29.58	95.03	91.75	77.69	28.60	95.03
	RPL	77.75	70.93	40.87	95.59	90.53	77.86	37.09	95.59
	DEAR	82.73	74.79	100	94.48	84.16	75.54	89.40	94.48
	PSL(ours)	**87.53**	**79.92**	**14.98**	**95.62**	**94.05**	**86.55**	**23.18**	**95.62**
	Δ	**(+3.35)**	**(+3.10)**	**(-14.60)**	**(+0.03)**	**(+2.24)**	**(+6.90)**	**(-5.42)**	**(+0.03)**

Table 2. Comparison of different evaluation metrics on HMDB51 (OoD) with K400 pretrained.

open-set evaluation, the AUPR will be close to 100%, although our method still achieves the best performance. The AUROC and FPR95 are not sensitive to the OoD sample numbers.

Models	Methods	1 time				10 times			
		AUROC↑	AUPR↑	FPR95↓	Acc.↑	AUROC↑	AUPR↑	FPR95↓	Acc.↑
TSM	OpenMax	93.34	98.46	29.20	95.32	93.34	88.14	28.95	95.32
	MC Dropout	88.71	97.92	39.46	95.06	88.71	83.36	39.46	95.06
	BNN SVI	91.86	98.75	36.21	94.71	91.86	90.12	36.21	94.71
	SoftMax	91.95	98.68	32.00	95.03	91.95	89.16	32.00	95.03
	RPL	90.64	98.57	38.43	95.59	90.64	88.79	38.43	95.59
	DEAR	86.04	98.08	87.66	94.48	86.04	87.38	87.40	94.48
	PSL(ours)	**95.75**	**99.39**	**19.00**	**95.90**	**95.75**	**94.96**	**18.96**	**95.90**
	Δ	**(+2.41)**	**(+0.64)**	**(-10.20)**	**(+0.31)**	**(+2.41)**	**(+4.84)**	**(-9.99)**	**(+0.31)**

Table 3. Comparison of different evaluation methods on MiT-v2 (OoD) with K400 pretrained.

C. Implementation details

When we use K400 pretrained model, the only method we need to fulfill is our PSL method, and we follow [1] to set the base learning rate as 0.001 and step-wisely decayed every 20 epochs with total 50 epochs. When we train the model from scratch, we need to conduct experiments on all methods in our Table 1. For our PSL method, we use the LARS optimizer [5] and set the base learning rate and momentum as 0.6 and 0.9 with totally 400 epochs. The reason we use this strategy is inspired by the contrastive learning SimCLR [6]. For other baselines, we find the above learning rate strategy cannot achieve good enough closed-set performance, and we find that setting the base learning rate as 0.05 and step-wisely decayed every 160 epochs with totally 400 epochs can achieve comparable closed-set performance. The batch size for all methods is 256, and we use 16 NVIDIA V100 GPUs to train the model.

D. OSAR performance under I3D and SlowFast backbone

We provide the OSAR results under TSM [7] backbone in Table 1 of the paper. Here, we further provide the OSAR results under I3D [8] and SlowFast [9] backbones in Table 4 and 5. We can see our PSL method still achieves state-of-the-art performance under these two backbones. The performance gain under Slowfast when MiTv2 is OoD dataset is marginal, as baselines already have high performance.

E. Representation analysis through singular value spectrum

To deeply understand the feature representations learned by our method, we analyze the representation through singular value spectrum. We first compute the covariance matrix $C \in \mathbb{R}^{d\times d}$ of the embedding matrix:

$$C = \frac{1}{M}\sum_{i=1}^{M}(z_i - \bar{z})(z_i - \bar{z})^T, \tag{1}$$

where z_i and $\bar{z}_i$ denote the feature representation of a sample and mean representation of all samples respectively. M is the total number of samples. Then we conduct singular value decomposition on the matrix $C = USV^T$, $S = diag(\sigma^k)$, and plot the singular values in sorted order and logarithmic scale $log(\sigma^k)$. We provide the singular value spectrum in Fig. 1.

Datasets	Methods	w/o K400 Pretrain				w/ K400 Pretrain			
		AUROC↑	AUPR↑	FPR95↓	Acc.↑	AUROC↑	AUPR↑	FPR95↓	Acc.↑
UCF101 HMDB51	OpenMax	83.78	54.65	47.60	74.42	92.03	77.72	41.02	95.01
	MC Dropout	75.85	40.04	50.34	74.39	91.66	78.87	33.60	94.11
	BNN SVI	81.53	53.62	49.18	73.15	91.57	78.65	34.60	93.89
	SoftMax	81.24	54.21	48.20	74.42	91.28	79.73	34.18	94.11
	RPL	79.80	52.09	54.07	71.62	92.49	81.72	28.89	94.26
	DEAR	78.91	54.14	81.96	74.42	89.80	80.86	75.63	93.89
	PSL(ours)	**86.88**	**65.63**	**39.85**	**78.85**	**93.62**	**85.54**	**28.38**	**95.46**
	Δ	**(+3.10)**	**(+10.98)**	**(-7.75)**	**(+4.43)**	**(+1.13)**	**(+3.82)**	**(-0.51)**	**(+0.45)**
UCF101 MiTv2	OpenMax	86.33	77.49	44.40	74.63	93.29	90.17	29.84	94.90
	MC Dropout	76.61	62.32	48.43	74.24	93.53	90.97	25.21	94.11
	BNN SVI	83.13	76.20	48.63	73.15	93.52	91.24	25.34	93.89
	SoftMax	82.58	74.91	46.39	74.63	92.62	90.87	30.55	94.11
	RPL	81.47	73.98	49.62	71.89	93.69	92.04	25.97	94.26
	DEAR	81.48	77.03	77.58	74.42	90.88	90.55	60.28	93.89
	PSL(ours)	**88.88**	**83.30**	**34.91**	**78.69**	**95.70**	**95.06**	**20.03**	**95.51**
	Δ	**(+2.55)**	**(+5.81)**	**(-9.49)**	**(+4.06)**	**(+2.01)**	**(+3.02)**	**(-5.18)**	**(+1.25)**

Table 4. OSAR performance under I3D backbone.

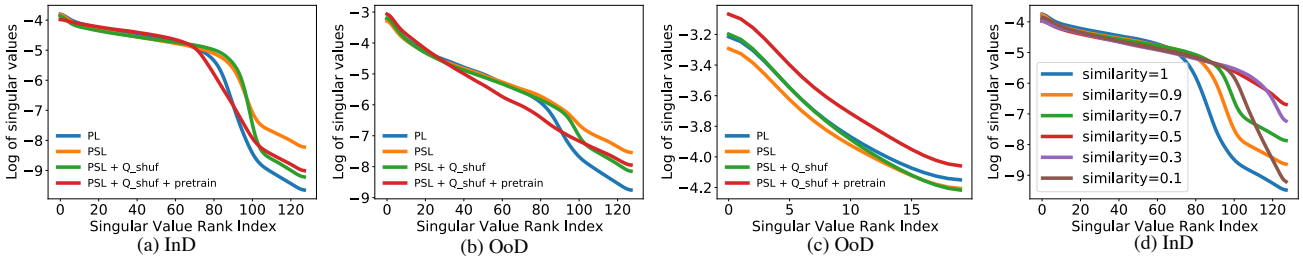
Datasets	Methods	w/o K400 Pretrain				w/ K400 Pretrain			
		AUROC↑	AUPR↑	FPR95↓	Acc.↑	AUROC↑	AUPR↑	FPR95↓	Acc.↑
UCF101 HMDB51	OpenMax	80.67	50.49	52.46	75.40	92.49	78.27	35.65	96.30
	MC Dropout	76.10	41.37	50.82	75.16	91.83	77.71	29.82	96.70
	BNN SVI	81.66	56.72	49.66	76.58	93.34	85.57	27.89	96.56
	SoftMax	79.15	48.54	48.79	75.63	93.82	85.56	24.74	96.70
	RPL	81.35	54.65	51.64	78.36	93.81	85.41	24.06	**96.93**
	DEAR	78.00	49.38	68.49	76.21	92.28	87.09	62.99	96.48
	PSL(ours)	**86.20**	**64.65**	**42.48**	**79.40**	**95.24**	**89.76**	**18.72**	96.52
	Δ	**(+4.54)**	**(+7.93)**	**(-6.31)**	**(+1.04)**	**(+1.42)**	**(+2.67)**	**(-5.34)**	**(-0.49)**
UCF101 MiTv2	OpenMax	79.60	70.05	51.08	75.63	94.34	89.90	25.42	96.30
	MC Dropout	75.88	63.12	51.40	75.63	93.43	90.43	24.52	96.70
	BNN SVI	82.89	76.13	46.88	76.58	93.53	92.34	28.81	96.56
	SoftMax	51.08	75.63	79.60	70.05	94.67	93.34	22.14	96.70
	RPL	81.42	73.07	49.13	78.36	94.76	93.39	21.99	**96.93**
	DEAR	78.21	69.30	62.02	76.21	92.60	93.09	59.98	96.48
	PSL(ours)	**85.00**	**77.08**	**43.16**	**79.40**	**96.81**	**96.22**	**14.52**	96.52
	Δ	**(+2.11)**	**(+0.95)**	**(-3.72)**	**(+1.04)**	**(+2.05)**	**(+2.83)**	**(-7.47)**	**(-0.49)**

Table 5. OSAR performance under SlowFast backbone.

PSL has larger singular values than the PL in the larger rank index, illustrating that more information is contained in the not significant dimensions, which is reasonable as PSL keeps the IS information with no direct supervision signal, but these IS information does help for better OSAR performance according to Table 2 in the paper. PSL with shuffled samples Q_{shuf} has larger singular values than PSL in the small rank index, indicating more diverse information is learned in the important dimensions, which are supposed to refer to CS information as CS information is learned by the explicit supervision signal. The closed-set accuracy with Q_{shuf} is higher than without Q_{shuf} in Table 2 further testifies our conclusion. In Tabel 2 we see that the representations of the same class are tighter with more CS information. Therefore, learning the distinct temporal information from shuffled videos can enlarge the open-set task related CS information while PSL can enlarge the IS information, which fulfills the goal to enlarge Eq. 3 for better OSAR performance.

F. Open-set performance *w.r.t.* s with Q_{shuf}

We provide extension results of Table 4 in the paper. The results are based on HMDB51 (OoD) from scratch. s for Q_{sc} is set as 0.7, and we change the value of s for Q_{shuf} in Table 6. We can see that the performance is optimal when s for Q_{shuf} is 0.8, but the same s with Q_{sc} which is 0.7 also achieves the good performance. So to reduce the number of hyper-parameters, we pick up the same s for Q_{sc} and Q_{shuf} by default. In addition, we can see that the closed-set accuracy is lower when $s = 1$ compared to $s = 0.8$. This is because we set the similarity between the original video and the shuffled video as 1, which is not reasonable as the temporal information is totally lost in the shuffled video.



(a) InD (b) OoD (c) OoD (d) InD

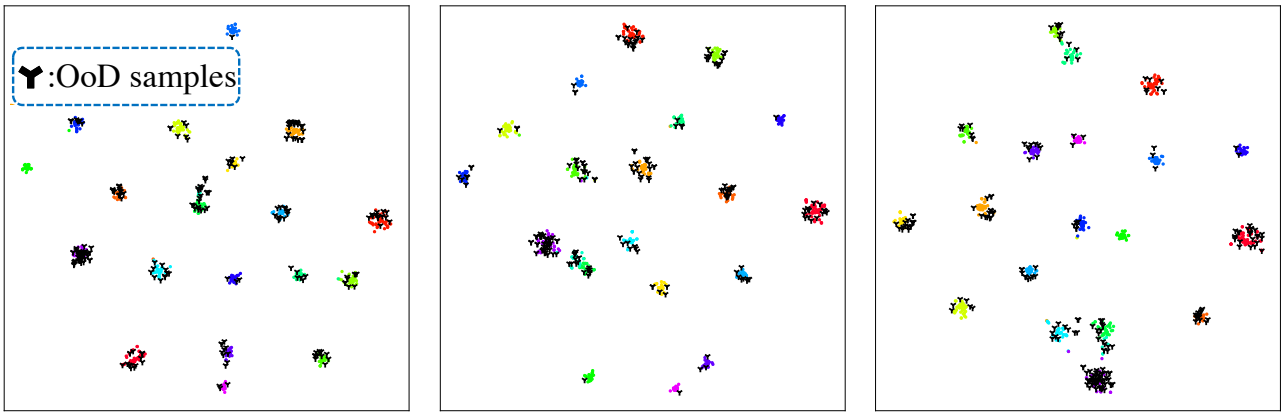
Figure 1. Singular value spectrum on HMDB51 (OoD) under different training conditions (a)-(c) and hyper-parameter s (d). (c) contains the top 20 singular values in (b).

s	AUROC↑	AUPR↑	FPR95↓	Acc.↑
1	82.04	53.82	51.82	72.89
0.9	83.12	57.04	46.84	73.31
0.8	**86.43**	**65.58**	**41.75**	76.53
0.7	85.25	63.91	48.34	**76.98**
0.6	85.26	62.93	46.89	76.77
0.5	84.08	61.76	53.53	75.13
0.4	82.75	59.09	52.72	73.79
0.3	77.34	53.84	68.14	67.67
0.2	73.94	50.63	75.55	60.21
0.1	68.86	41.39	82.15	39.00

Table 6. Ablation results of different s for Q_{shuf}.

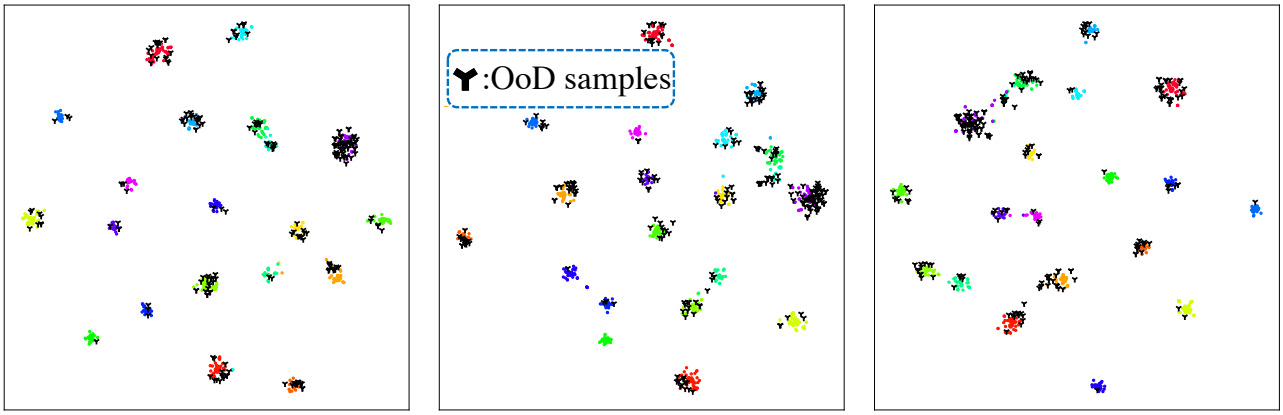
G. t-SNE visulization

To illustrate the variance within a class, we provide the Table 2, Fig. 5 and 6 in the paper, which is enough to show the variance change due to different components in our PSL method. Here, we provide the t-SNE visualization for straight understanding. All results are based on HMDB (OoD) from scratch. We provide the visualization results of PSL, PSL with Q_{ns}, PSL with Q_{ns}, Q_{sc}, and PSL with Q_{ns}, Q_{sc}, Q_{shuf} in Fig. 2, 3, 4, 5 respectively. From Fig. 2 we can see PSL alone cannot keep the intra-class variance when s decreases. Fig. 3 and Fig. 4 tell us that Q_{ns} and Q_{sc} are important for PSL to keep the intra-class variance. Furthermore, Q_{shuf} makes the feature representation tighter if we compare Fig. 4 and Fig. 5, which shows the model learns more CS information with Q_{shuf}.



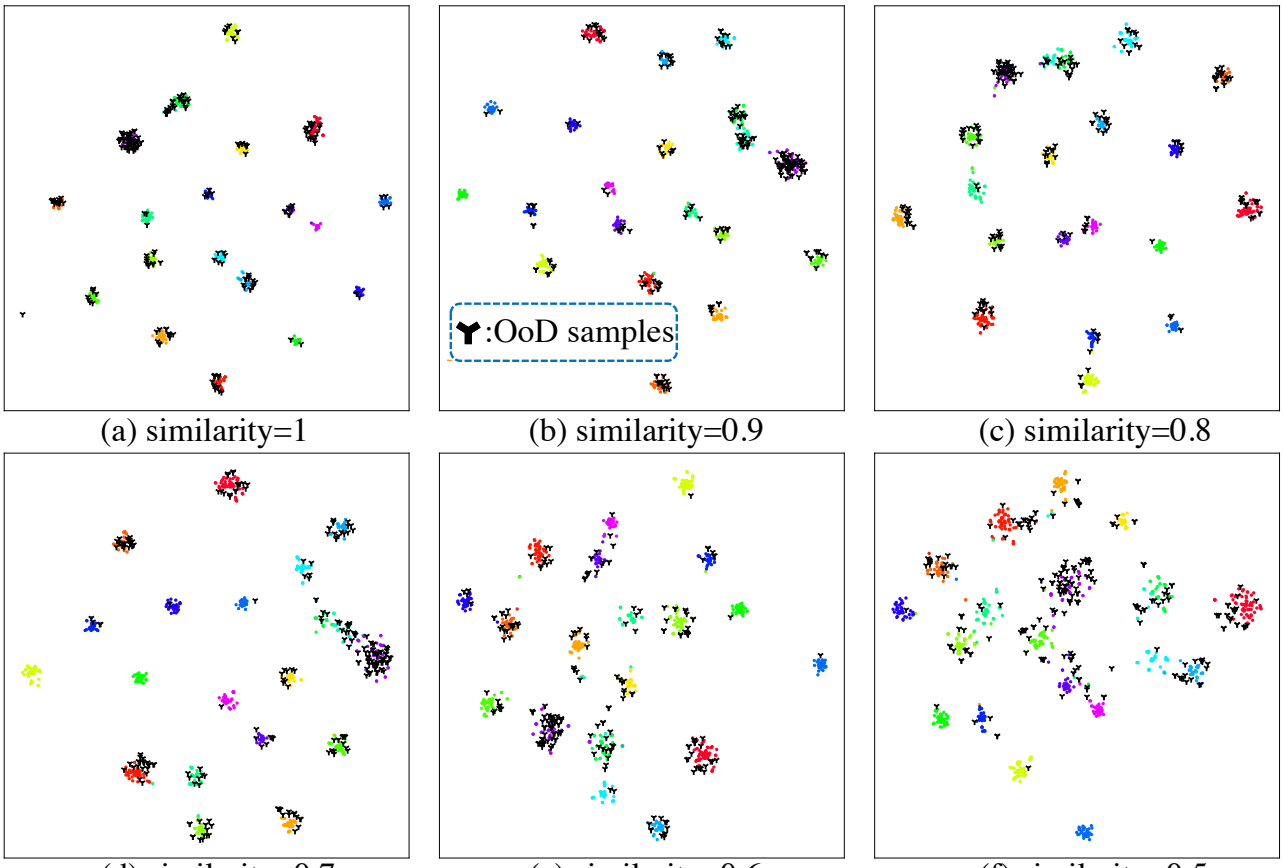
(a) similarity=1 (b) similarity=0.7 (c) similarity=0.5

Figure 2. t-SNE visualization of PSL.



(a) similarity=0.9 (b) similarity=0.8 (c) similarity=0.7

Figure 3. t-SNE visualization of PSL with Q_{ns}.



(a) similarity=1 (b) similarity=0.9 (c) similarity=0.8

(d) similarity=0.7 (e) similarity=0.6 (f) similarity=0.5

Figure 4. t-SNE visualization of PSL with Q_{ns}, Q_{sc}.

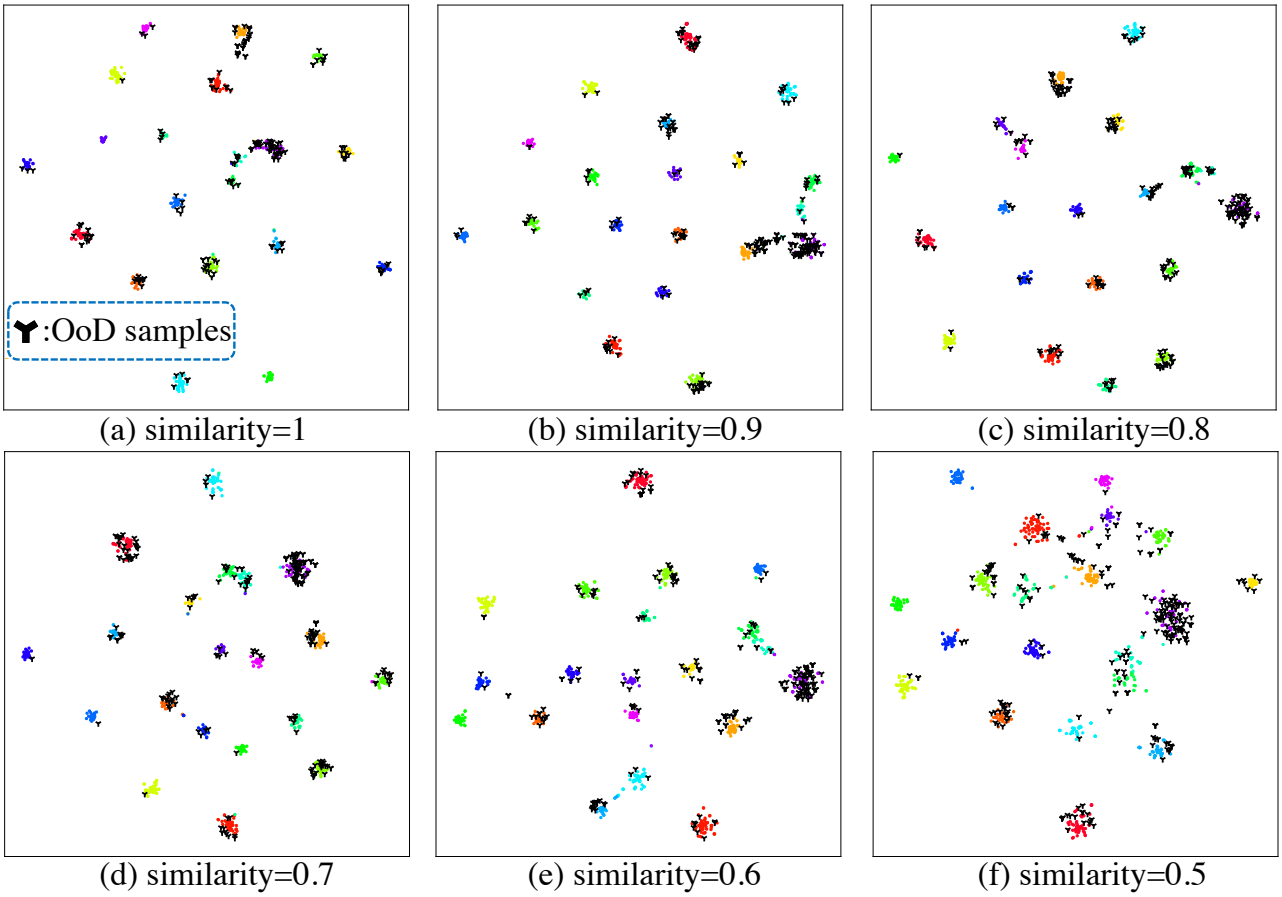


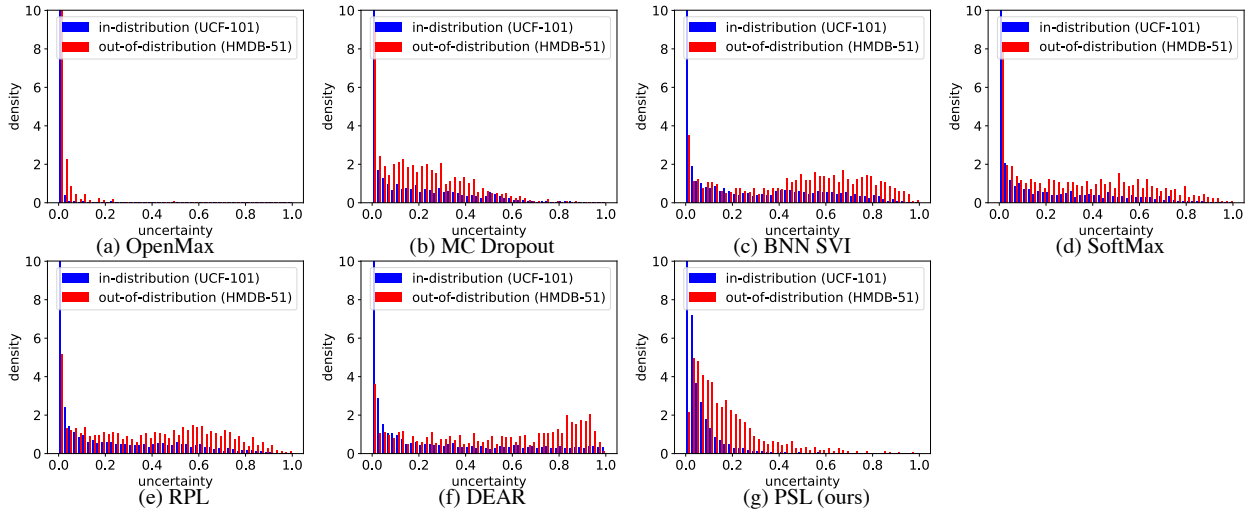
Figure 5. t-SNE visualization of PSL with Q_{ns}, Q_{sc}, Q_{shuf}.

H. InD and OoD uncertainty distribution

We provide the InD and OoD distribution on HMDB51 (OoD) and MiT-v2 (OoD) with K400 pretrain and without K400 pretrain. All results are based on TSM backbone for illustration. The results are shown in Fig. 6, 7, 8, and 9.

From Fig. 6 and 8 we can see that if there is no K400 pretrain, all methods have the overlapping uncertainty between InD and OoD distribution except OpenMax and our PSL. For instance, Fig. 6 (f) DEAR [1] shows the uncertainty of InD and OoD samples both cover the range from 0 to 1. In contrast, Fig. 6 (g) PSL shows that in our method, the InD distribution covers from 0 to 0.3, while the OoD distribution covers from 0 to 0.8. It means our method tends to assign higher uncertainty to OoD samples. For OpenMax, Fig. 6 (a) shows that InD uncertainy distribution is extremely close to 0, which is a good phenomenon, but the OoD uncertainty distribution only covers from 0 to 0.3, and the OoD samples whose uncertainty is larger than 0.3 is too sparse, which means OpenMax tends to assign low uncertainty to both InD and OoD samples, but assigner lower uncertainty to InD samples.

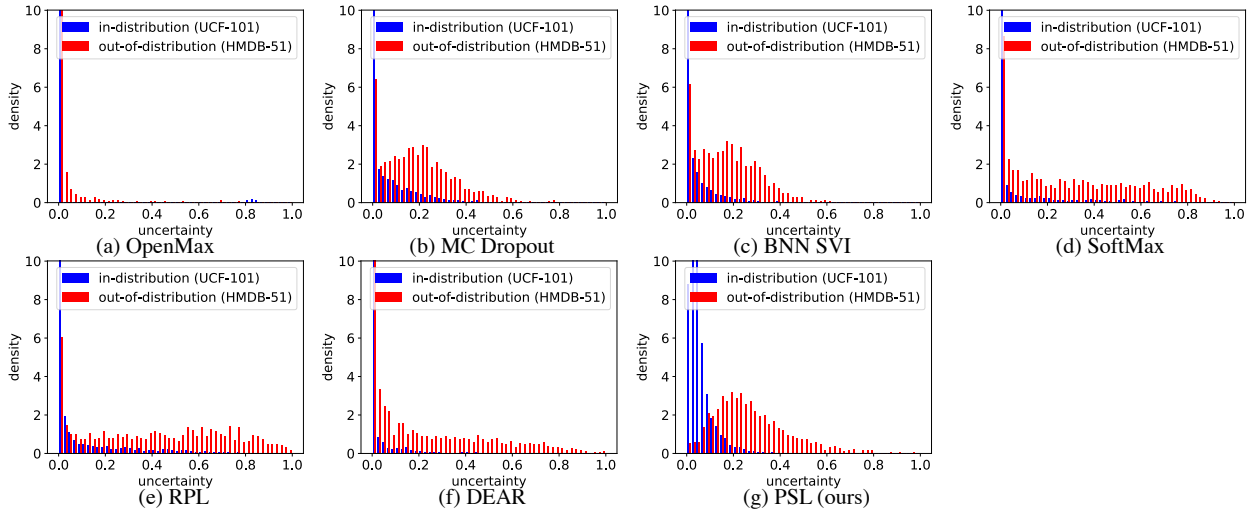
If we compare Fig. 6 to Fig. 7 or compare Fig. 8 to Fig. 9, we can find that the InD distribution of all methods are closer to 0 with K400 pretrain. But all methods except our PSL have a serious over confidence problem, which is illustrated by the fact that the far left column of OoD samples is extremely high, which is also emphasized through the red circles in Fig. 4 of the paper. In contrast, the density of OoD distribution is highest at 0.2 uncertainty in our PSL method, and the density of OoD distribution is almost 0 at 0 uncertainty. Besides, it is very clear that the OoD distribution and InD distribution in our PSL is most distinguishable among all methods.



(a) OpenMax (b) MC Dropout (c) BNN SVI (d) SoftMax

(e) RPL (f) DEAR (g) PSL (ours)

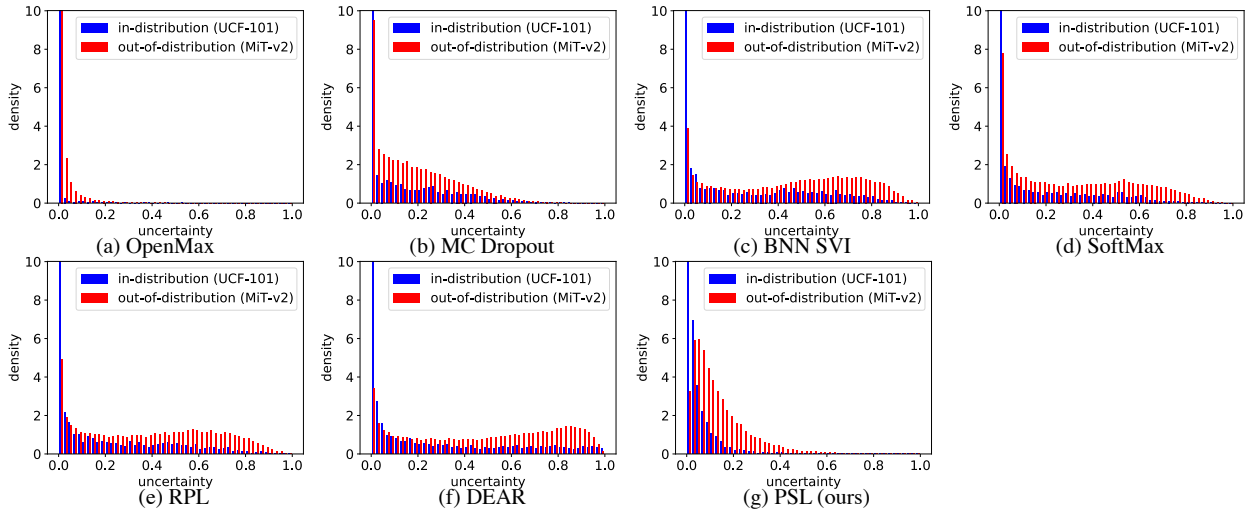
Figure 6. Uncertainty distribution on HMDB51 (OoD) w/o K400 pretrain.



(a) OpenMax (b) MC Dropout (c) BNN SVI (d) SoftMax

(e) RPL (f) DEAR (g) PSL (ours)

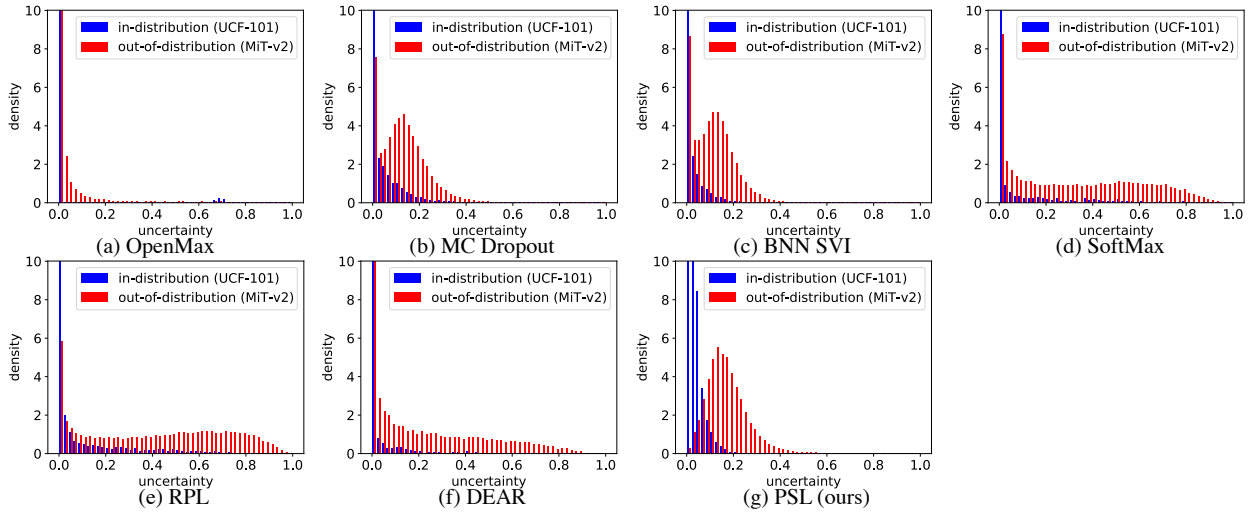
Figure 7. Uncertainty distribution on HMDB51 (OoD) w/ K400 pretrain.



(a) OpenMax (b) MC Dropout (c) BNN SVI (d) SoftMax

(e) RPL (f) DEAR (g) PSL (ours)

Figure 8. Uncertainty distribution on MiT-v2 (OoD) w/o K400 pretrain.



(a) OpenMax (b) MC Dropout (c) BNN SVI (d) SoftMax

(e) RPL (f) DEAR (g) PSL (ours)

Figure 9. Uncertainty distribution on MiT-v2 (OoD) w/ K400 pretrain.

References

[1] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *ICCV*, 2021. 1, 2, 6

[2] Dahua Lin Yue Zhao, Yuanjun Xiong. Mmaction. https://github.com/open-mmlab/mmaction, 2019. 1

[3] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 1

[4] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 1

[5] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

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

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

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