Supplementary Material for
SOAR: Scene-debiasing Open-set Action Recognition

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

This supplementary material is organized as follows.

- Sec. 1 provides additional experiments. We highlight the conclusions here:
  - The scene bias problem is a general problem that exists in four different backbones (*i.e.*, I3D [3], SlowFast [5], TSM [9], and TPN [12]).
  - Our SOAR successfully mitigates the scene bias, and achieves state-of-the-art OSAR performance with three additional backbones (*i.e.*, SlowFast [5], TSM [9], TPN [12]), showing the generalization ability and effectiveness of our method.
  - The ablation study on UCF101 [11] + HMDB51 [8] reveals that all designs contribute to the final performance, demonstrating the generalization ability of our method.
- Sec. 2 introduces additional implementation details.

1. Additional experimental results

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

In the main paper, we have shown that our SOAR outperforms previous methods in terms of lower scene bias and higher OSAR performance with the I3D backbone [3]. To validate that the superiority of SOAR is not tied to a specific backbone, we carry out experiments with different backbones, *i.e.*, SlowFast [5], TSM [9], and TPN [12]. Furthermore, to analyze how the scene distance affects the overall OSAR and closed-set classification performance, we carry out the scene bias analysis experiments using Open maF1 as the metric.

Scene bias analysis with different backbones. Fig. 1, Fig. 2, and Fig. 3 shows the scene bias analysis experiments with TPN [12], TSM [9], and SlowFast [5] backbones, respectively. We make the following observations. (1) All figures show the same tendency: known actions with unfamiliar scenes (right part of the left figures) and unknown actions with familiar scenes (left part of the right figures) are hard to recognize. This conclusion holds for all backbones, indicating that it is a general problem rather than a backbone-specific problem. (2) Our SOAR achieves lower scene bias in both scenarios with all backbones, showing the generalization ability and debias ability of our method. Especially, our SOAR achieves better OSAR performance when the closed-set testing set exhibits dissimilar scene to the training set (*i.e.*, the right part of Fig. 1a, Fig. 2a, and Fig. 3a), and when the open-set testing set exhibits similar scene to the training set (*i.e.*, the left part of Fig. 1b, Fig. 2b, and Fig. 3b). Such a performance advantage shows that our method successfully avoids the misleading of scene information, and further demonstrates its debias ability. We note that this ability is critical when the testing environment is different from the training environment.

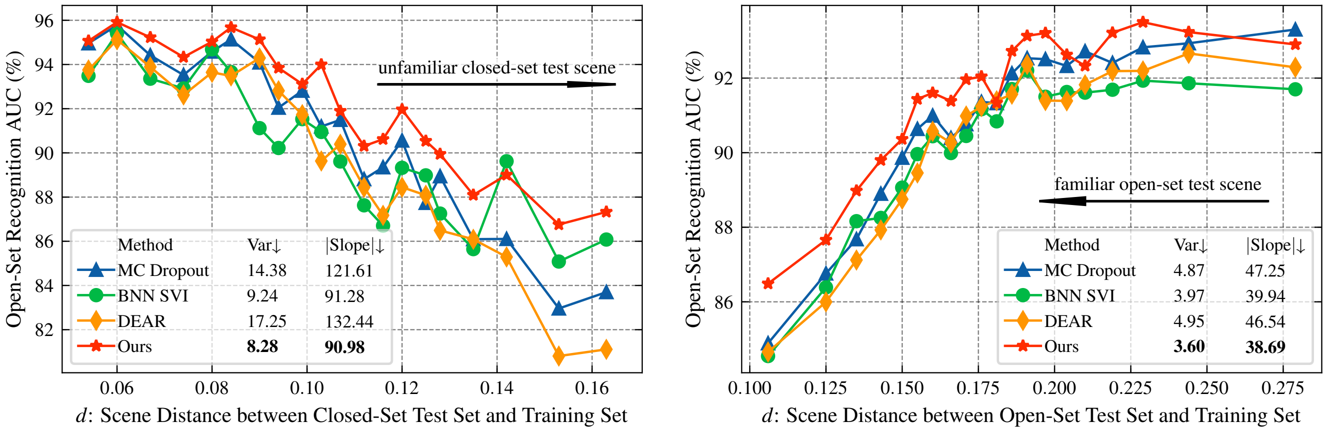
OSAR performance comparison with different backbones. Tab. 1 lists the performance comparison with previous OSAR methods in different backbones. The results show that our SOAR achieves state-of-the-art OSAR performance and outperforms all previous methods in terms of AUC and open macro F1 with all backbones, demonstrating the effectiveness of our method.

Scene bias analysis using Open maF1. In Fig. 2 of the main paper, we show a strong correlation between OSAR performance and the scene distance. We further illustrate how the scene distance affects the overall OSAR and closed-set classification performance (*i.e.*, the C + 1 way classification performance) by conducting the scene bias analysis using Open maF1 in Fig. 4. The results reveal a similar trend that the scene distance and Open maF1 is highly correlated, and our SOAR achieves the best performance as well as the lowest scene bias, demonstrating its effectiveness.

1.2. Ablation study on UCF101 [11]+HMDB51 [8]

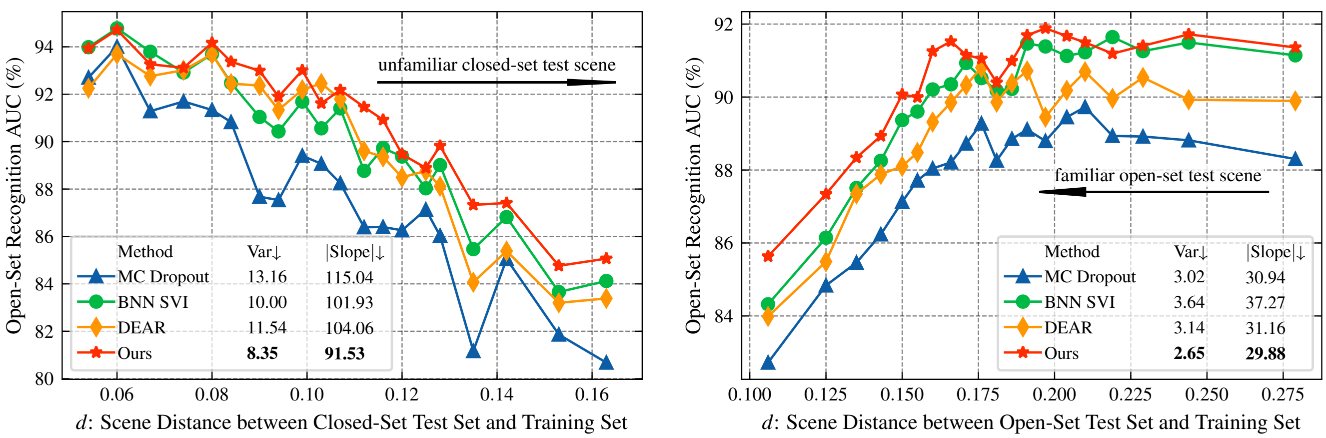
To demonstrate the generalization ability of our proposed AdRecon and AdaScls, we further show the results of the

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



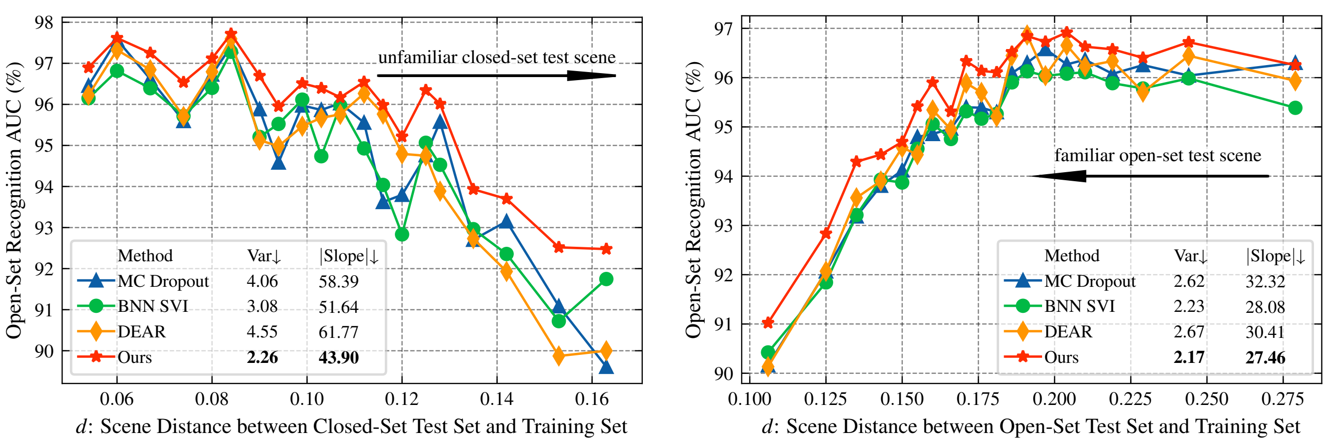
(a) Analysis on the known action in unfamiliar scene scenario. (b) Analysis on the unknown action in familiar scene scenario.

Figure 1. Quantitative scene bias analysis using UCF101 [11] as known and MiTv2 [10] as unknown with the *TPN* backbone [12].



(a) Analysis on the known action in unfamiliar scene scenario. (b) Analysis on the unknown action in familiar scene scenario.

Figure 2. Quantitative scene bias analysis using UCF101 [11] as known and MiTv2 [10] as unknown with the *TSM* backbone [9].

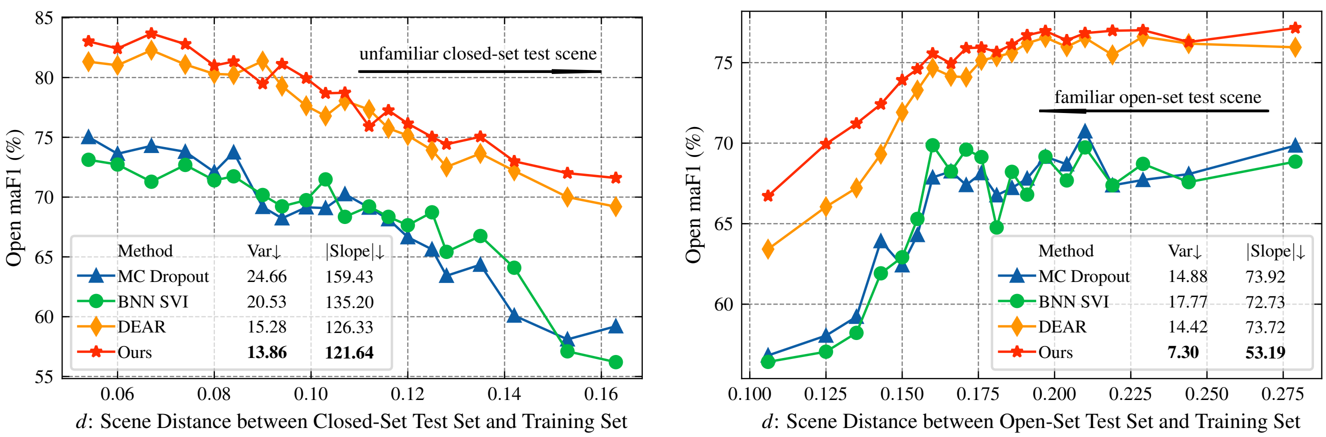


(a) Analysis on the known action in unfamiliar scene scenario. (b) Analysis on the unknown action in familiar scene scenario.

Figure 3. Quantitative scene bias analysis using UCF101 [11] as known and MiTv2 [10] as unknown with the *SlowFast* backbone [5].

Backbone	Methods	UCF101 [11]+MiTv2 [10]				UCF101 [11]+HMDB51 [8]				Closed-set Accuracy
		AUC ↑	FAR@95 ↓	TPR@10 ↑	Open maF1 ↑	AUC ↑	FAR@95 ↓	TPR@10 ↑	Open maF1 ↑	
TPN [12]	SoftMax	43.36	97.82	8.89	55.01 ± 0.32	44.92	97.33	6.42	72.31 ± 0.12	92.00
	OpenMax [2]	60.02	73.93	23.02	65.31 ± 0.19	62.65	64.23	19.30	65.32 ± 0.12	55.37
	MC Dropout [6]	90.86	32.59	72.51	71.96 ± 0.19	84.89	64.76	57.19	77.47 ± 0.14	91.28
	BNN SVI [7]	90.23	32.23	67.86	69.57 ± 0.19	84.93	66.82	58.82	75.38 ± 0.15	90.11
	DEAR [1]	90.31	33.67	68.32	73.57 ± 0.19	85.16	62.72	57.14	84.82 ± 0.14	92.02
	SOAR (Ours)	**91.45**	**30.96**	**74.37**	$\mathbf{74.48 \pm 0.21}$	**86.67**	**61.02**	**60.62**	$\mathbf{85.43 \pm 0.14}$	**92.63**
TSM [9]	SoftMax	46.39	94.45	9.35	54.29 ± 0.34	44.58	98.44	9.32	76.29 ± 0.19	92.11
	OpenMax [2]	61.49	58.90	12.49	64.30 ± 0.25	60.97	63.83	10.46	64.39 ± 0.17	53.48
	MC Dropout [6]	87.87	41.69	61.22	65.67 ± 0.26	84.82	63.67	63.53	75.68 ± 0.20	92.15
	BNN SVI [7]	89.92	40.42	72.66	65.94 ± 0.25	83.28	65.96	54.31	77.63 ± 0.19	91.83
	DEAR [1]	89.12	38.98	68.07	67.33 ± 0.36	84.26	**57.79**	62.16	86.05 ± 0.17	91.94
	SOAR (Ours)	**90.47**	**37.17**	**69.69**	$\mathbf{69.33 \pm 0.21}$	**85.96**	60.62	**65.98**	$\mathbf{87.67 \pm 0.17}$	**92.49**
SlowFast [5]	SoftMax	56.02	89.33	15.54	61.12 ± 0.26	55.39	91.58	20.57	75.02 ± 0.15	96.17
	OpenMax [2]	68.49	39.38	10.48	69.74 ± 0.17	67.00	77.54	25.35	66.46 ± 0.16	60.33
	MC Dropout [6]	95.01	**18.21**	88.99	71.12 ± 0.16	89.52	53.95	75.82	73.35 ± 0.15	96.24
	BNN SVI [7]	94.83	20.51	87.37	68.92 ± 0.19	88.68	60.88	74.05	71.14 ± 0.16	96.01
	DEAR [1]	95.12	20.35	87.63	75.51 ± 0.17	89.33	58.78	75.95	89.71 ± 0.17	**96.56**
	SOAR (Ours)	**95.72**	18.84	**90.68**	$\mathbf{76.47 \pm 0.14}$	**90.72**	**52.32**	**76.93**	$\mathbf{90.64 \pm 0.19}$	96.53

Table 1. Comparison with state-of-the-art methods with different backbones. All methods are trained on UCF101 [11], and evaluated on two different open sets where unknown samples are from HMDB51 [8] and MiTv2 [10], respectively.



(a) Analysis on the known action in unfamiliar scene scenario. (b) Analysis on the unknown action in familiar scene scenario.

Figure 4. Quantitative scene bias analysis using *Open maF1*, which combines the OSAR and closed-set action recognition performances. The experiments are carried out with the I3D backbone [3], using UCF101 [11] as known and MiTv2 [10] as unknown. Our SOAR is least affected by the scene.

AdRecon	Bg. Est.	Unc. Weight	AUC ↑	FAR@95 ↓	TPR@10 ↑	Open maF1 ↑
-	-	-	85.63	78.59	68.10	87.73 ± 0.22
✓	-	-	85.72	82.66	71.63	86.68 ± 0.21
✓	-	✓	87.17	69.82	70.46	88.08 ± 0.20
✓	✓	-	86.95	70.18	71.76	88.01 ± 0.21
✓	✓	✓	**87.49**	**69.41**	**72.31**	$\mathbf{89.52 \pm 0.21}$

Table 2. Ablation study on the adversarial reconstruction on UCF101 [11] + HMDB51 [8] datasets.

$\mathcal{L}_{\text{s\_cls}}$	$\mathcal{L}_{\text{s\_guide}}$	AUC ↑	FAR@95 ↓	TPR@10 ↑	Open maF1 ↑
-	-	85.63	78.59	68.10	$\mathbf{87.73 \pm 0.22}$
✓	-	86.87	73.42	68.48	87.42 ± 0.23
✓	✓	**87.22**	**71.45**	**69.80**	87.47 ± 0.19

Table 3. Ablation study on the adversarial scene classification on UCF101 [11] + HMDB51 [8] datasets.

ablation study in the UCF101 [11]+HMDB51 [8] testing set in Tab. 2 and Tab. 3, respectively. The results reveal that all designs in both modules contribute to the final performance, which aligns with the conclusion made in the main paper, demonstrating the generalization ability of our method.

2. Additional implementation details

For the TPN backbone [12], we follow DEAR [1] to use the slow-only version for feature extraction. For the TSM backbone [9], we use the default setting in MMAction2 [4] for feature extraciton following DEAR [1]. For the SlowFast backbone [5], we interpolate the extracted spatio-temporal features from the slow and fast pathways to the same size, and concatenate them in the channel dimension as the final spatio-temporal feature $\boldsymbol{F}$.

We note that our reported results are different from those reported in DEAR [1] as they use binarized prediction for the AUC prediction, which only has one operating point on the ROC curve, while we use the raw prediction for the AUC computation.

References

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

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, pages 1563–1572, 2016. 3

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1, 3

[4] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. `https://github.com/open-mmlab/mmaction2`, 2020. 3

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 1, 2, 3

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 3

[7] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Bar: Bayesian activity recognition using variational inference. *arXiv preprint arXiv:1811.03305*, 2018. 3

[8] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. 1, 3

[9] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019. 1, 2, 3

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

[11] 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. 1, 2, 3

[12] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR*, pages 591–600, 2020. 1, 2, 3