Supplementary Material for WACV 2023 ID 0849
Avoiding Lingering in Learning Active Recognition by Adversarial Disturbance

Lei Fan and Ying Wu
Northwestern University
2145 Sheridan Road, Evanston, IL 60208
leifan@u.northwestern.edu, yingwu@northwestern.edu

Abstract

This document is the supplementary material of our WACV submission with ID 0849. We provide the curves during training, the effect of disturbances, the classifier activations over steps, and the video result on the ShapeNet dataset [1].

1. The derivation of Equation 2

We denote the active recognition agent as $f_{\theta,\phi}$ containing two groups of parameters, *i.e.*, the recognizer and the recognition policy. Recall $\mathcal{P}(.)$ is the projection function from a 3D instance to a 2D image. We have the following loss of a two-step recognition process on the object instance x^i, which is

$$\begin{aligned} l^i &= |y^i - f_{\theta,\phi}(x^i)| \\ &= |y^i - \arg\max_y \mathbb{P}[y|v_0, \mathcal{P}(c_0 + \arg\max_a \pi_\phi(a_1|v_0))]| \\ &= |y^i - \arg\max_y \frac{\mathbb{P}[y, \mathcal{P}(c_0 + \arg\max_a \pi_\phi(a_1|v_0))|v_0]}{\mathbb{P}[\mathcal{P}(c_0 + \arg\max_a \pi_\phi(a_1|v_0))|v_0]}| \\ &= |y^i - \arg\max_y \frac{\mathbb{P}(y, \hat{v}_{a_1}|v_0)}{\mathbb{P}[\arg\max_a \pi_\phi(a_1|v_0)|v_0]}| \\ &= |y^i - \arg\max_y \frac{q_\theta(y, \hat{v}_{a_1}|v_0)}{\pi_\phi(a_1|v_0)}|. \end{aligned} \tag{1}$$

As $q_\theta(y, \hat{v}_{a_1}|v_0)$ and $\pi_\phi(a_1|v_0)$ are our recognizer and recognition policy, respectively, we factorize the active recognition process into a multiplication of two modules.

2. The curves during training

We here demonstrate the rewards of both the recognition policy and the adversarial policy, the accuracy of the training set, and the accuracy of the validation set in Fig. 1. As the recognizer gets better during training, the recognition policy could obtain rewards about which view benefits the recognition process. On the other hand, the adversarial policy gradually could not find views that the recognizer fails as the training process converges.

During training, the two policies compete with each other on the recognition performance, which forms a zero-sum game. The convergence of our method could be comprehended as a Nash equilibrium by the min-max training procedure, *i.e.*, we want to obtain the highest recognition reward as there is no way to increase the reward achieved by the adversarial policy.

3. The effect of disturbances

We provide more examples on the view visiting frequencies during the training of ours and the baseline method [2]. The accuracy of each view is therefore marked on each grid. All heatmaps are normalized independently, *i.e.*, each heatmap covers the full-color range. During the agent exploration, the elevation of view grids is not connected end-to-end. In other words, the agent stays at the same position when it attempts to go downwards at the bottom line of the view grid. It explains why the visiting heatmaps of our method are bright at the two ends of elevations.

We could notice that in Fig. 2 (b), the policy collapsed to a monotonous mode. The policy constantly visit views that could provide positive reward as the recognition accuracy of other views are unsatisfactory. We named this phenomenon during training as *lingering* as the agent being reluctant to explore challenging views. On the contrary, the adversarial policy disturbs our agent during training by discovering its current deficiencies. Therefore, with similar training epochs, the proposed method could obtain enough training for each view instead of overfitting to limited views.

4. The classifier activations over steps

We demonstrate the average classifier activations of the correct class at each step in Fig. 3. The increase in classifier activations reflects that the agent could progressively

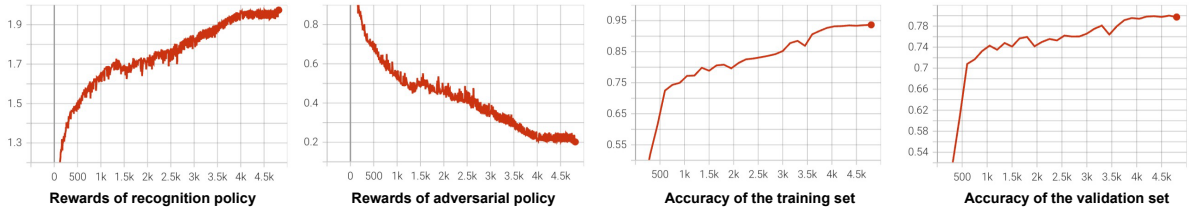


Figure 1. We show the statistics of one training trial of our method on the ShapeNet dataset [1]. In each training epoch, we sample a $T = 5$ trajectory on each object instance and then update our model. The accuracy of the validation set stabilizes after about 4000 epochs.

disambiguate its predictions by taking more movements.

5. Ablation studies on auxiliary loss terms

In this section, we study the impact of $\mathcal{L}_{entropy}$ and $\mathcal{L}_{forecast}$ in our loss function. The $\mathcal{L}_{forecast}$ works as approximating the state transition function, which predicts the next-step feature based on the current state. It motivates our model to establish the relation between actions and different views. The $\mathcal{L}_{entropy}$ promotes exploratory behaviors. Table 1 compares the final recognition accuracy over both ShapeNet [1] and SUN360 [3] datasets.

Datasets/Method	w/o $\mathcal{L}_{forecast}$	w/o $\mathcal{L}_{entropy}$	Ours
ShapeNet	75.6±.3	75.4±.3	**76.4±.3**
SUN360	69.2±.2	69.0±.3	**69.6±.2**

Table 1. Ablation studies on the ShapeNet and SUN360 datasets.

6. The video result on the ShapeNet dataset

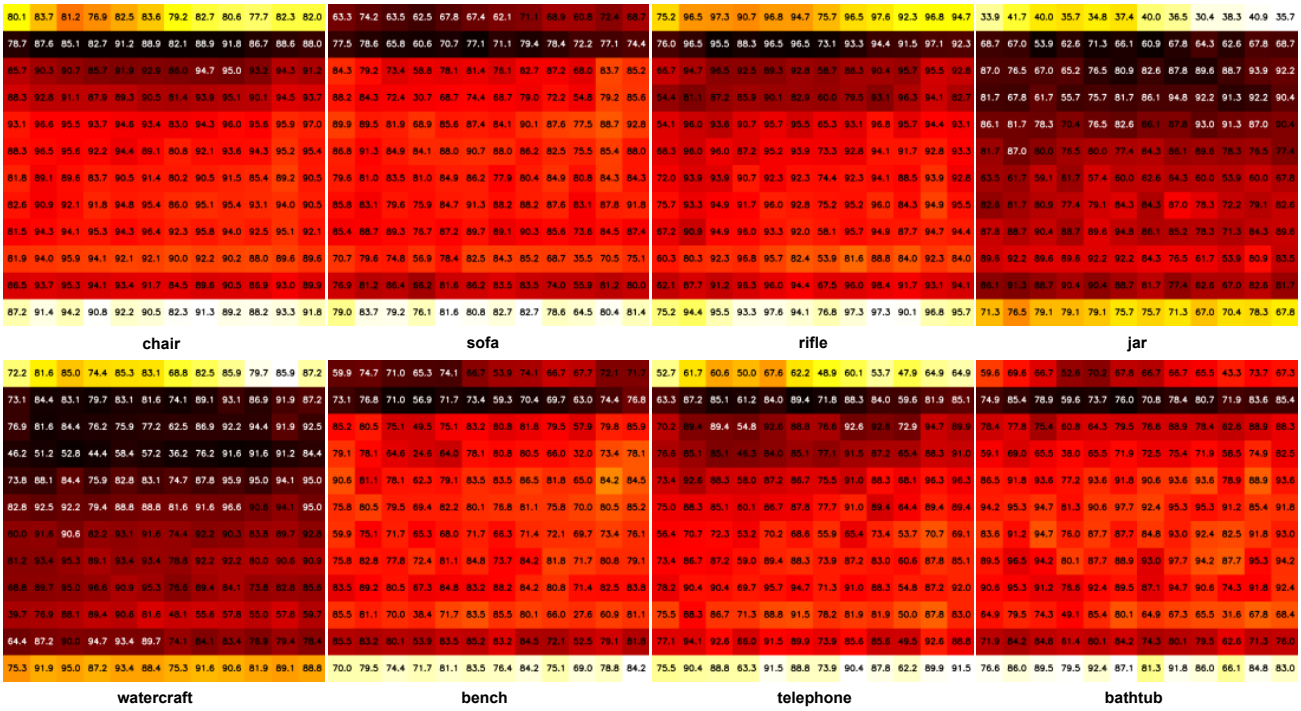
More qualitative results of the proposed method on the ShapeNet dataset [1] are included in the demonstrative video. We show the top-3 guesses at each step to show the advantage brought by active recognition.

References

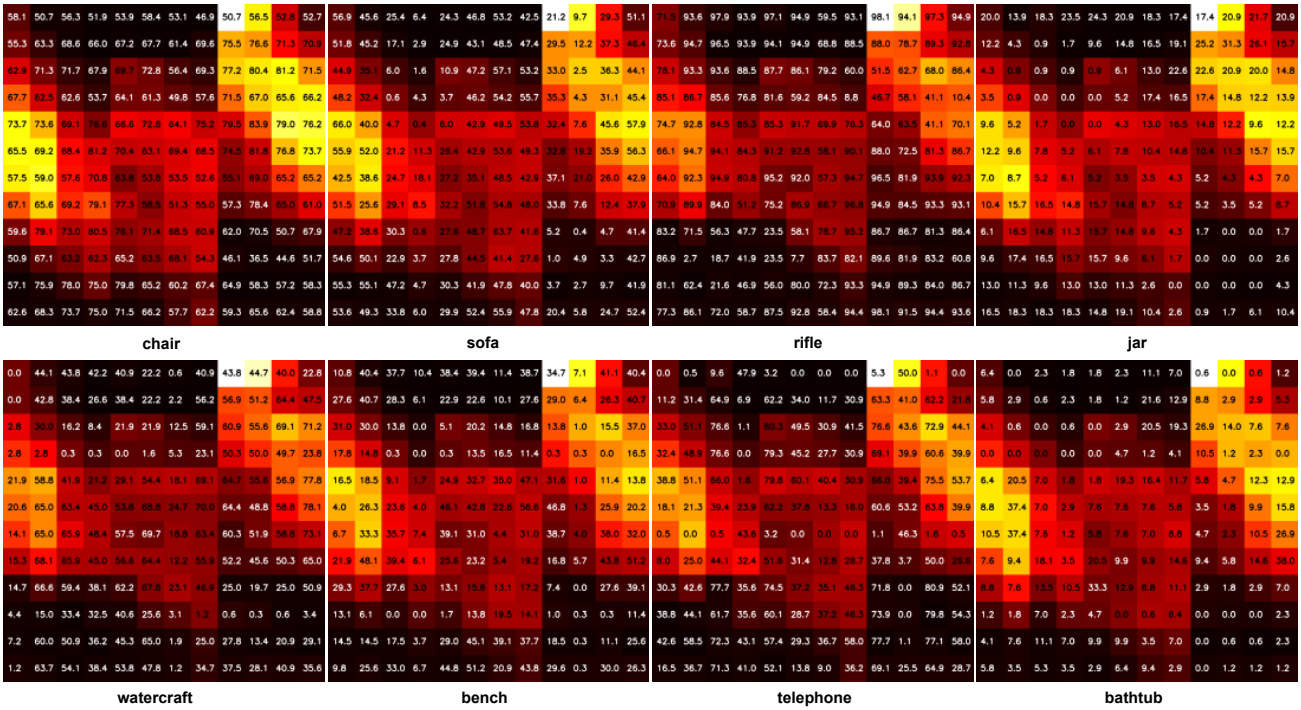
[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[2] Dinesh Jayaraman and Kristen Grauman. Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion. In *European Conference on Computer Vision*, 2016.

[3] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

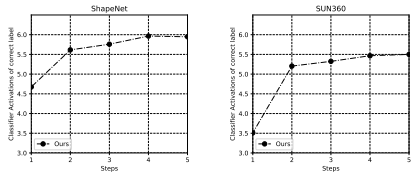


(a) Our view-visiting frequencies and their corresponding single view accuracy during training.



(b) The view-visiting frequencies and their corresponding single view accuracy during training of [2] which does not address the *lingering* problem.

Figure 2. A comparison between the proposed method and the baseline who does not consider the *lingering* problem on the ShapeNet dataset [1]. We discretize the viewing sphere for each object as a 12×12 grid. Then, the visiting frequencies of views by heatmaps on different categories are demonstrated. The training accuracy of views is accordingly marked on each grid. Note that we normalize each heatmap separately.



(a) ShapeNet (b) SUN360

Figure 3. The classifier activations of the correct category over steps on both the ShapeNet [1] and the SUN360 dataset [3].