How Do You Do It? Fine-Grained Action Understanding with Pseudo-Adverbs Supplementary Material

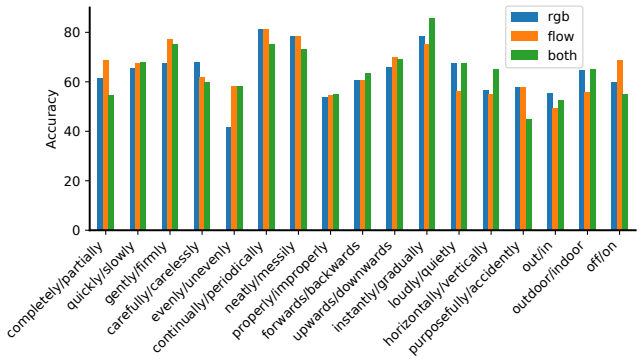

Figure 1. Results of each modality for per adverb-antonym pair.

A. Per Adverb Results

We show results for each adverb-antonym pair in Fig. 1. We first note that our model is capable of learning each of these adverbs since all pairs perform above random performance which is 50%. Our model performs best on 'instantly/gradually', 'continually/periodically' and 'neatly/messily', despite the high imbalance between the number of instances in this adverb-antonym pairs. The most challenging adverb-antonym pairs are 'evenly/unevenly', 'properly/improperly' and 'purposefully/accidentally', since the latter adverb in each pair has very few labelled samples to learn from. While our multi-adverb pseudo-labelling learns well from imbalanced data, generalizing from few samples remains a challenge.

Figure 1 also displays the results of different modalities for each adverb-antonym pair. Different adverbs benefit from different modalities. For instance, 'gently' and 'firmly' are better distinguished with flow features, while recognition of 'carefully' vs. 'carelessly' benefits from the RGB modality. Overall the results of the different modalities are quite comparable, with RGB achieving 63.7, Flow 64.5 and the combination of both 63.9. The fusion of RGB and Flow not always being beneficial highlights that better fusion of modalities is needed for adverbs. This is a challenge since not only are different modalities useful for different adverbs, but the adverbs appearance is highly dependent on the action to which is applies and different actions are also better recognized with different modalities.

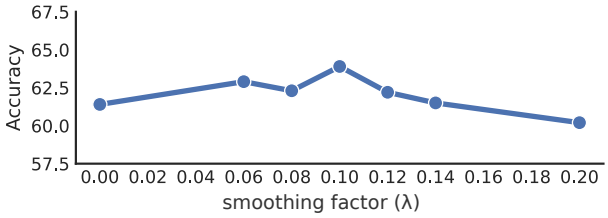

Figure 2. Effect of smoothing factor λ.

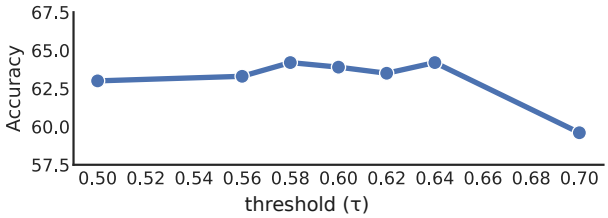

Figure 3. Effect of base threshold τ.

B. Adaptive Thresholding Hyperparameters

Effect of λ. In Fig. 2 we show the effect of smoothing factor λ used in the adaptive thresholding (Eq. 8 in the main paper). The parameter determines how much the adverb-specific thresholds adapt and thus how much the model focuses on underrepresented adverbs. When λ=0 the original threshold τ is used for all adverbs. Fig. 2 shows that best results are obtained with λ=0.1, although any value $0.04 \leq \lambda \leq 0.14$ improves results.

Effect of τ The effect of base threshold for the pseudo-labeled adverbs used is shown in Fig. 3. The model is relatively insensitive to this parameter with any value $0.5 \leq \tau \leq 0.64$ being suitable.

C. Long-Tail Results

In Table 1 we investigate the ability of our method to recognize adverbs in the long tail distributions of adverb, actions and their compositions. For each we report adverb recognition results for the head, middle and tail of the distributions. For adverbs the head is defined as $>$500 instances and the tail is $<$100 instances. For actions the head is $>$100 instances and $<$20 instances. For the action-adverb pairs the head and tail are $>$50 and $<$10 instances respectively.

Table 1 shows the results on the 5% split of VATEX Ad-

Method	All	Adverbs			Actions			Pairs		
		Head	Mid	Tail	Head	Mid	Tail	Head	Mid	Tail
Supervised only	60.3	68.6	69.0	50.9	62.3	56.5	57.3	78.8	56.5	57.3
Pseudo-Label	60.4	65.5	65.0	55.1	63.1	54.8	63.0	78.2	54.8	62.3
FixMatch	61.2	65.8	63.4	57.4	60.5	63.9	55.0	79.2	63.9	55.0
TCL	58.3	64.9	74.6	55.3	66.5	56.5	58.7	77.8	56.5	58.7
Ours	63.9	68.9	69.1	58.2	65.3	60.2	59.7	84.2	60.3	59.7

Table 1. Results of the long-tail. We show adverb recognition results over the long-tails of adverbs, actions and their pairs. We split each into three categories of head, middle and tail. Our model successfully combats the long-tail of adverbs by increasing tail results over the supervised only baseline and maintaining a similar on the adverbs in the head and middle of the distribution.

verbs. We can see that our method improves results of the tail adverbs significantly over the supervised only baseline (50.9 to 58.2) while maintaining a similar performance for the head and middle of the adverb distribution. Other methods have a smaller improvement over the adverb tail while causing a decrease in performance at the head and middle of the distribution. For the action and pair distributions, our method increases results over the supervised only baseline for the head, middle and tail of the distributions. Other methods are better at certain parts of the distribution, for instance FixMatch obtains best performance at the middle of the action and pair distribution. However, their overall improvement is lower. From these results we can conclude that the success of our method is due to its improvement on the long-tail of adverbs.

D. Baseline Implementation Details

For all models we use the same backbone, video-text embedding functions (f and g) and loss functions (L_{act} and L_{adv}) as in our proposed method. The shared hyper-parameters are also common between our method and baselines. We outline the specific details for each below.

Pseudo Label [5]. For this baseline we take the adverb in the closest embedded action-adverb text representation to be the hard pseudo-label, as defined in Eq. 3. All of these pseudo-labels are used in training, this approach does not use thresholding. The action-only labeled data and adverb pseudo-labeling is introduced at epoch 300, up until then the model is trained with the supervised data only. The loss functions for supervised and pseudo-labeled action-only data are given equal weighting. We experimented with different weightings and introducing the pseudo-labeling at different epochs as in [5], but empirically found these settings to perform best. This baseline uses both RGB and Optical flow modalities.

FixMatch [7]. Fixmatch also uses hard pseudo-labels. An action-only video is first weakly augmented and the adverb in the closest embedded action-adverb text representation is taken as the pseudo-label (Eq. 3). A strongly augmented version of this video is then trained to predict this pseudo-label. We use the same augmentations as in the original paper [7]. The weak augmentations are randomly flipping the video with 50% probably and randomly translating the video by up to 12.5% vertically and horizontally. The strong augmentations are those used in RandAugment [3]: auto-contrast, adjusting brightness, adjusting color balance, adjusting contrast, equalizing the video frame histogram, the identity function, posterizing, rotating, adjusting the sharpness, shearing along the x or y-axis, solarizing and translating the video along the x or y-axis. We randomly select two augmentations for each video in a batch, each with random magnitudes. Full details can be found here [7]. The same augmentation with the same parameters are applied to all frames in a video. Fixmatch uses fixed thresholding where we use a threshold of $\tau = 0.6$. This baseline uses the RGB modality.

TCL [6]. TCL optimizes the consistency in predictions between a normal video and an augmented version played at twice the speed. The agreement is maximized through an instance contrastive loss which encourages the two speeds of the video to have the same prediction for all adverbs. This is done with class logits in the original paper, since we use a video-text embedding space to learn adverbs we use the distance to each of the action-adverb compositions with the ground-truth action. There is also a group contrastive loss which optimizes agreement between the average predictions for groups of videos with the same pseudo-label. This grouping is done with the hard pseudo-label predicted from Eq. 3. As in the original paper we use a weighting of 9 for the instance contrastive loss and 1 for the group contrastive loss. This baseline uses the RGB modality.

E. Dataset Licensing

Our adverb newly proposed adverb datasets are based on three existing video-text datasets: VATEX [8], MSR-VTT [9] and ActivityNet Captions [4]. All three of these datasets use videos from YouTube, as such all of the videos in these datasets and our adverb datasets use either the YouTube Standard License [2] or the Creative Commons BY License [1].

Figure 4. Predictions of different adverbs for the same action.

F. Further Qualitative Results

Fig. 4 shows results of our model predicting different adverbs for an action. This demonstrates how adverbs could be applicable to anomaly detection in videos. For instance,the figure shows 'cut horizontally', which we know is anomalous as 'vertically' is predicted for most 'cut' actions.

G. Action-Adverb Distribution

A full size version of the action-adverb distribution from the main paper is shown in Fig. 5. Not only are the individual distributions of actions and adverbs long-tailed, but the action-adverb compositions are also heavily long-tailed.

H. Supplementary Video

We show video versions from Fig. 8 and Fig. 9 from the main paper in the supplementary video.

References

[1] Cc by 3.0 us. https://creativecommons.org/licenses/by/3.0/us/. Accessed: 2021-11-15. 2

[2] Standard youtube license. https://www.youtube.com/static?template=terms. Accessed: 2021-11-15. 2

[3] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 2

[4] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[5] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Internation Conference on Machine Learning (ICML) Workshops*, 2013. 2

[6] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[7] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised

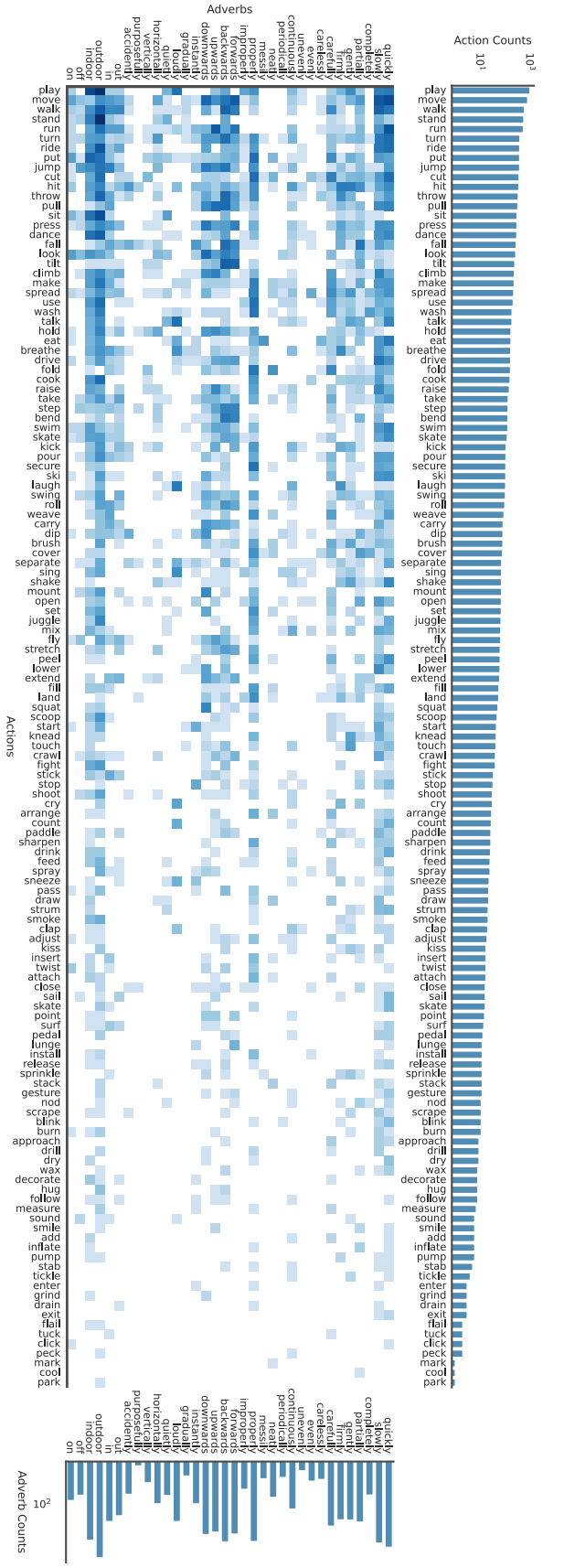

Figure 5. Distribution of action-adverb pairs in VATEX shown on a log-scale.

learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[8] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[9] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2