Revealing the unseen: Benchmarking video action recognition under occlusion (Supplementary Material)

A Appendix

A.1 Overview

All three benchmark datasets and code will be made publicly available at this [link]. We include the following results and details in this supplementary,

1. We provide classwise performance of CTx-Net on UCF-101 and UCF-101-O dataset.
2. We provide additional results for data augmented transformer.
3. We provide a detailed network architecture of the proposed CTx-Net.
4. Qualitative Results for CTx-Net.
5. We provide tsne visualizations for effect of occlusion.
6. We further provide the confusion matrix on UCF-101-Y-OCC dataset for other models as well.
7. Limitation and Social Impact.
8. Datasheet for the proposed datasets.

A.2 Class wise Performance

In figure 1 we provide the classwise performance for the proposed CTx-Net on UCF-101 and UCF-101-O dataset. The result shows the relative robustness score for each class. As expected for most of the classes, the relative robustness score is less than 1 which shows the negative effect of occlusion on the task of action recognition. Further, a few outlying classes have robustness score more than 1 which shows that the model performs better on occluded frame as compared to clean frames. We can also notice from the plots that for most of the classes the score is close to 1 which shows the robustness of the proposed CTx-Net irrespective of classes.

A.3 Additional Results

In Figure 7 performance of MVit+data augmentation on out of distribution occluders is shown. It is performed by including a specific number of occluders during test time which were present during training as well. From the Figure we can clearly see that the proposed CTx-Net has minimal effects on performance while varying the distribution of occluders whereas the performance of data augmented model falls as the number of out of distribution occluders increase. This shows data augmentations inability to generalize well.

A.4 UCF-19-Y-OCC dataset

The UCF-19-Y-OCC is composed of 19 classes. The classes included are - Band Marching, Bench-Press, Biking, Playing Cello, Baby Crawling, Walking with dog, Drumming, Playing Flute, Hand

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

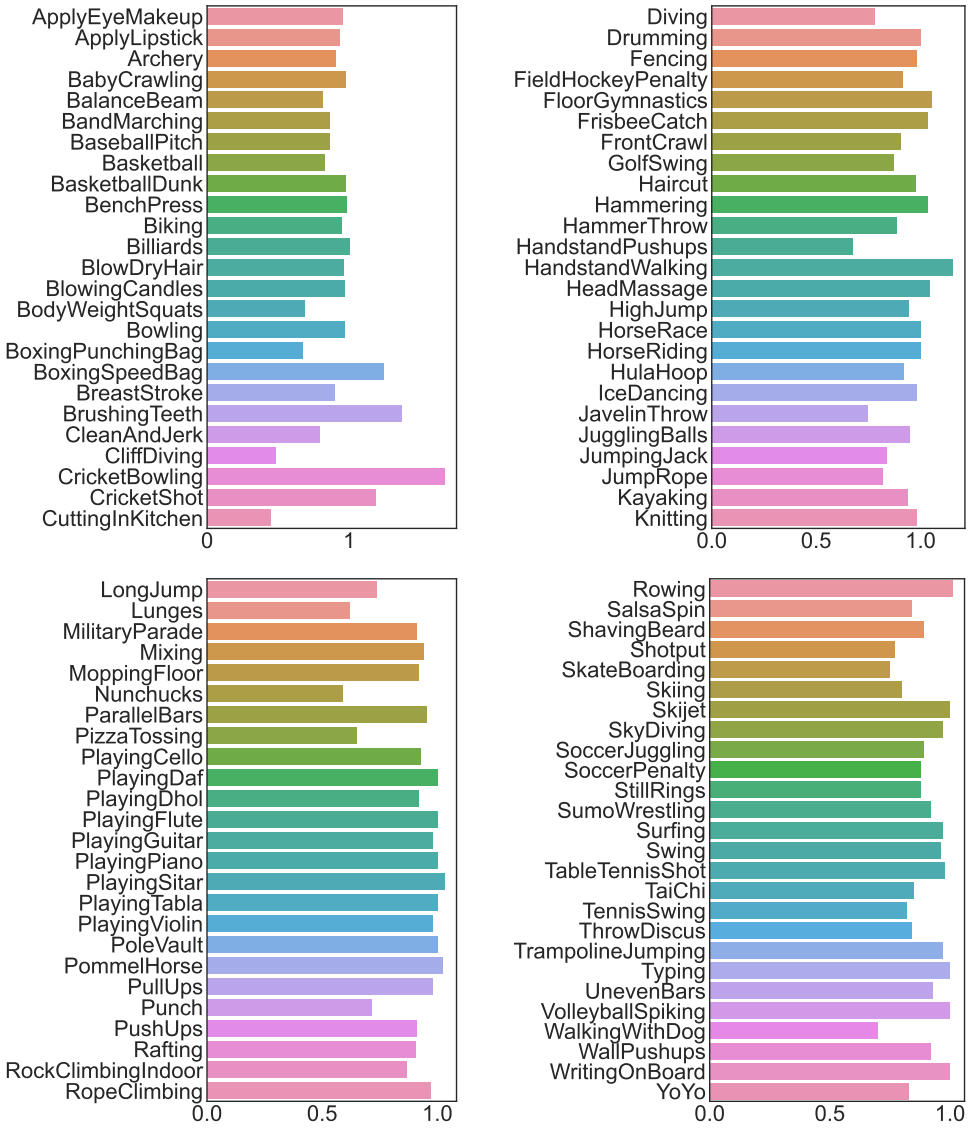

Figure 1: Class wise performance of CTx-Net on UCF-101 dataset using relative robustness as the metric.

stand Pushups, Kayaking, Mopping Floor, Nunchucks, Pizza Tossing, Pushups, Skateboarding, Skiing, Soccer Juggling, Soccer Penalty, Surfing. All of these actions are also present in UCF-101 dataset. Figure 8 shows the distribution of number of clips per class in the dataset. Figure 7 shows the confidence matrix plot for I3D and X3D on UCF-19-Y-OCC dataset. It can be clearly seen that these methods do not have enough discriminative properties in case of natural occlusion.

A.5 Effects of Occlusion on feature representation

From Figure 6 we can see that the feature representations learned by X3D and I3D are quite discriminative in case of clean frames, as they are able to provide distinct cluster in the tsne plot, whereas for UCF-101-O we can see that the feature representation does not have enough discriminative power given the lack of distinct clusters.

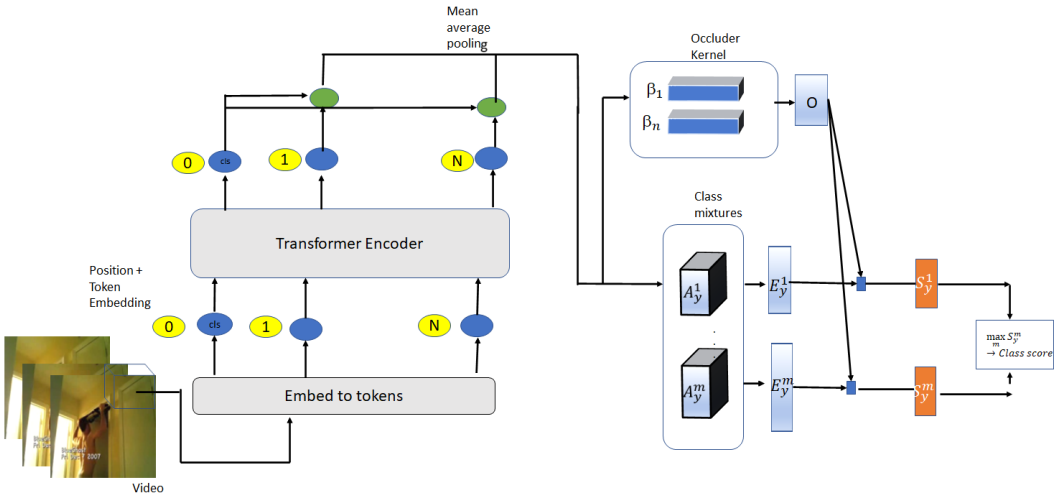

Figure 2: Overview of the CTx-Net architecture. A transformer network used to extract features .Class token is then aggregated with feature tokens, followed by feature pooling using average pooling, followed by this a convolution with vMF kernels μ_n followed by non-linear vMF activation $\mathcal{N}(.)$.The resulting vMF likelihood L is used to compute the occlusion likelihood O using the occluder kernels β.Furthermore, L is used to compute the mixture likelihoods E_y^m using mixture models A_y^m. O and E_y^m compete in explaining L the orange box and are combined to obtain the final class score.

c

Figure 3: Visualization for the patches in a video activated by three different vmf kernel. Left column: represents actions which comprises hand movements closer to upper body, right column: represents action which comprises hand movements closer to lower body

A.6 Qualitative Results

Figure 9 shows qualitatively the localization of occluders in a video. Each row represents frames in a video, followed by localization of occluders performed by CTx-Net. We observe that CTx-Net is able to localize occluders exhibiting a variety of shapes and motions effectively.Additionally, we present qualitative findings pertaining to the CTx-Net model trained on data augmented with synthetic occlusions, as shown in Figure 10. Notably, we observe an interesting phenomenon: the model

Figure 4: Visualization for the patches in a video activated by different components of mixture model for two different classes.

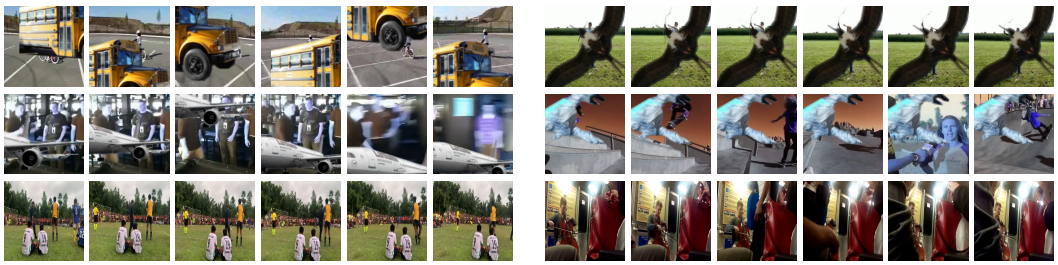

Figure 5: Samples from occluded benchmark datasets. First row: UCF101-O, second row: K-400-O, and third row UCF-19-Y-OCC.

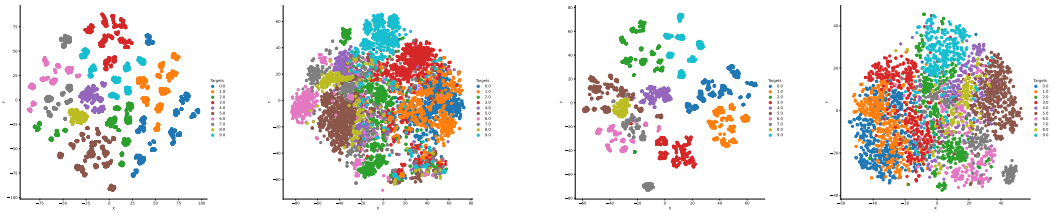

Figure 6: T-sne analysis of features under occlusion comparing I3D and X3D. Starting from left, (i) I3D features on UCF-101, (ii) I3D features on UCF-101-O, (iii), X3D features on UCF-101 and (iv) X3D features on UCF-101-O.

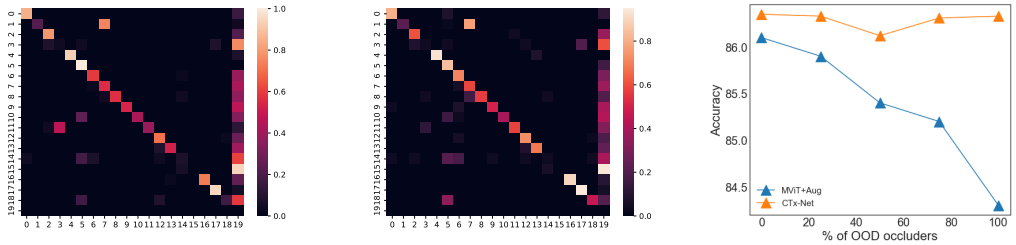

Figure 7: Class-wise performance analysis. Confusion matrix for (a) I3d, and (b) X3d. (c)Performance of the proposed CTx-Net and MViT+data aug on out of distribution occluders.

tends to identify certain attributes as occlusions even within video segments that remain unoccluded throughout. This behavior is rooted in the utilization of augmented data to train a class mixture model within the CTx-Net architecture. Consequently, this affects the model's ability to generalize effectively to non-occluded scenarios. Our study also extends to the comparative analysis detailed in Table 3 and Table 4. Particularly, we analyze instances where VideoMAE's performance is lower in contrast to MViTv2 and MViT. Visualizations of videos that exhibit correct classification by MViT and MViTv2, yet are misclassified by VideoMAE, are shown in Figure 11. These observations contribute to a comprehensive understanding of the strengths and limitations exhibited by the models under study.

A.7 Model architecture

Figure 2 shows the model architecture of the proposed CTx-Net which uses a transformer backbone. To calculate the class score, first the feature and class tokens are obtained for the given video. Following this, the feature token and class tokens are aggregated. Then vmf likelihood of the obtained features is then calculated. Class models are used to obtain class likelihood for each part detected, Similarly the occluder model is then used to obtain likelihood of which fea- tures are occluded.Both these scores are then combined to obtain the final class score. We also visualize the patches which activate the vmf and class mixture models the most. In Figures 3 and 4 we can see that the vmf kernels learned corresponds to more fundamental movements like moving the hand up, whereas class mixture model capture same actions being performed from differing points of views.

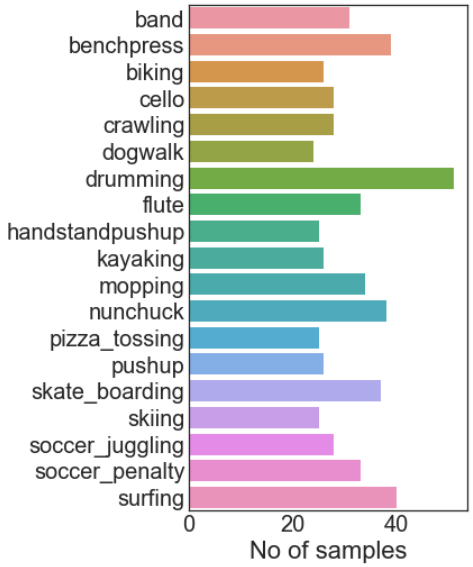

Figure 8: Class wise distribution of samples in UCF-19-Y-OCC dataset.

A.8 Occluders

From Figure 12 we can see some of the images that were used for training the occluder kernel. These are a randomly selected, out of distribution images in which no action seems to take place. Hence, this helps in separating out the random occlusion that occur in the video. Figure 5 also shows the occluded samples from the proposed datasets.UCF-101-O and K-400-O showing different severity of occlusions used

A.9 Limitations and Societal Impact

Benchmarking computer vision models for occlusion-aware video action recognition can lead to significant advancements in various application domains like development of enhanced surveillance systems which take occlusion into account, autonomous driving among others. These enhanced surveillance systems might cause some privacy concerns. For synthetically occluded datasets since, we sample objects randomly from the PASCAL VOC dataset the occluder can often present an unrealistic appearance both since the texture of occluder is significantly different from that of rest of the scene but also, the motion of exhibited by the objects is relatively simple as opposed to the complicated motions followed real life occluders.

A.10 Datasheet for Dataset

Motivation The core motivation behind this dataset was to study the effect of natural occlusion on video action recognition models, which could further help in moving the field forward by studying the performance of models systematically and in a real world setting.

Composition **Content and Composition** The instances in the dataset consists of videos in which actions are completely or partially occluded. The UCF-101-Y-OCC dataset consists of 19 classes, each class consists of 30 clips each lasting 5 second. The UCF-101-O and K-400-O consists of all the videos in test split of UCF and Kinetics dataset occluded synthetically.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? Yes for K-400-O and UCF-101-O, No for UCF-101-Y-OCC.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? No.

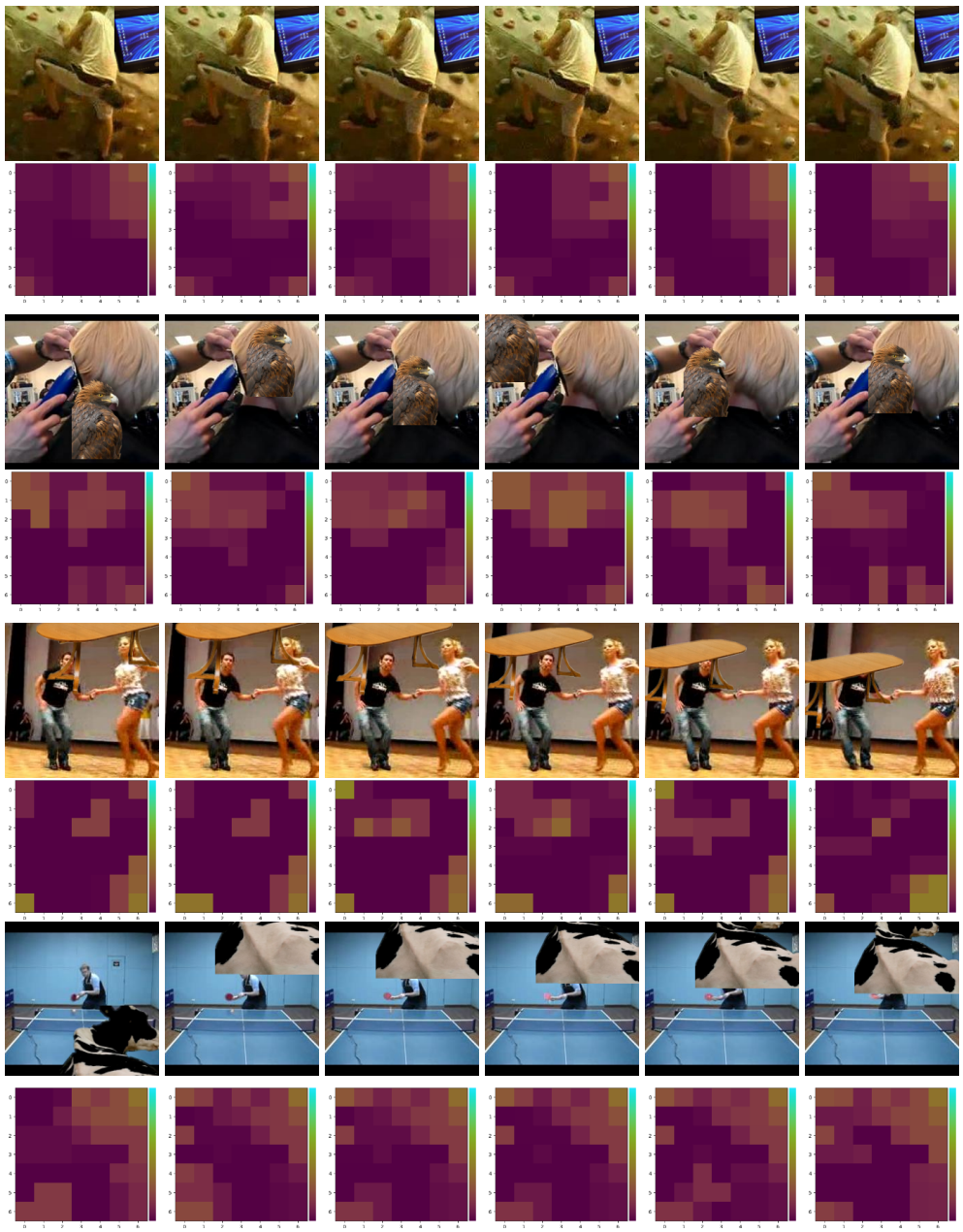

Figure 9: Qualitative results for occlusion localization. Each row represents a video followed by localization of occlusion in each of the above frames.

Are there recommended data splits (e.g., training, development/validation, testing)? No, all the proposed datasets are for evaluation only.

Are there any errors, sources of noise, or redundancies in the dataset? No, The entire UCF-101-Y-OCC dataset was annotated by the authors.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? UCF-101-Y-OCC is composed of YouTube videos. UCF-101-O and K-400-O are self-contained.

Is there a label or target associated with each instance? Yes, a class label representing

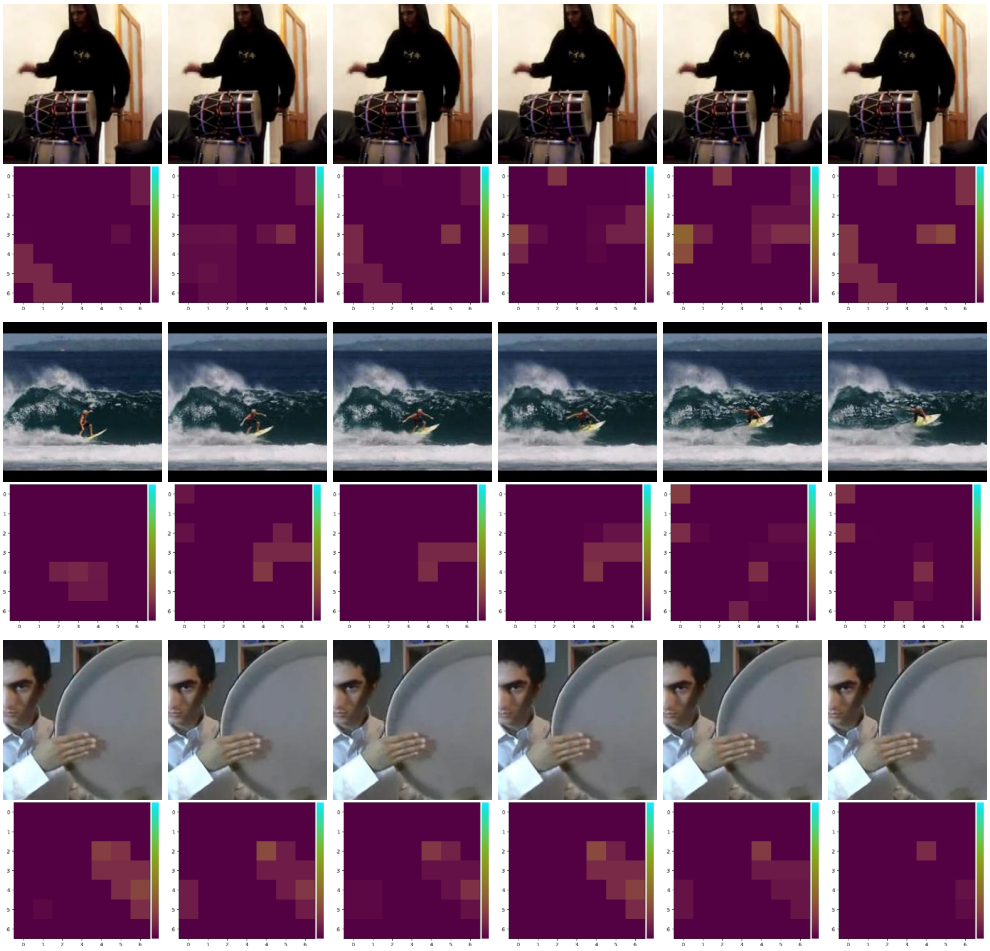

Figure 10: Qualitative results for occlusion localization of CTx-Net (augmented). Each row represents a video followed by localization of occlusion in each of the above frames.

Figure 11: Examples of videos in which Video MAE is unable to classify correctly whereas MViT and MViTv2 are able to correctly classify.

Figure 12: Visualization of images used for training occluder model

the action is associated with each dataset.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? No.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

Does the dataset identify any subpopulations (e.g., by age, gender) No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? Yes, the individual might be identifiable from YouTube.

Collection process Each instance of UCF-101-Y-OCC was extracted from publically available videos on YouTube. The entire video was then viewed to ensure the action occurring is relevant as well as a part of it is occluded. UCF-101-O and K-400-O are composed of test split of UCF-101 and K-400 datasets.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? YouTube API was the only external source of data used for UCF-101-Y-OCC. Annotations for UCF-101-Y-OCC were performed manually by the authors.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Not applicable since UCF-101-O and K-400-O contain entire test split of UCF-101-O and K-400-O.

Over what timeframe was the data collected? 2 Months for UCF-101-Y-OCC.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? YouTube was used for UCF-101-Y-OCC. UCF-101 and K-400 were synthetically occluded to obtain UCF-101-O and K-400-O.

Were the individuals in question notified about the data collection? Data was publically available on YouTube.

Labelling/Preprocessing/cleaning The video collected from YouTube was the segmented temporally into 5 seconds segments which contained occlusion. UCF-101-O and K-400-O have annotations same as UCF-101 and K-400 whereas UCF-101-Y-OCC was annotated manually.

Uses **Has the dataset been used for any tasks already?** For testing the performance of current models when exposed to occluded actions.

Is there a repository that links to any or all papers or systems that use the dataset? NA

What (other) tasks could the dataset be used for? Action Recognition under occlusion is the only intended task for the proposed dataset.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses No.

Are there tasks for which the dataset should not be used? No.

Distribution Will the dataset be distributed to third parties outside the entity (e.g., company, institution, organization) on behalf of which the dataset was created? No

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Will be available along with the code through download link.

When will the dataset be distributed? During the review process.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? Dataset will be distributed under Creative Commons License.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No