Occlusions in Video Action Detection: Benchmark Datasets And Training Recipes (Supplementary Material)

Rajat Modi[1]*, **Vibhav Vineet**[2], **Yogesh Singh Rawat**[1]
CRCV, University of Central Florida[1], and Microsoft Research[2]

1 Broader Impact Statement

A decade ago[2], [28] showed that successive filters of a convnet could act as general forms of edge, shape and texture detectors. In Fig 6, we now illustrate that higher layers of a neural net can learn to group pixels into objects at a semantic level without any explicit localization supervision/ multi-modal alignment . This shows that we could learn non-parametric object-queries at the highest levels in the architecture (via unsupervised-clustering), instead of directly injecting learnable parameters/proposals into lower decoder-layers[2]. In nature, there can be a large number of objects present in the retinal frame (eg, leaves of a tree). Depending on the granularity of fixation (eg, separation between multiple trees), we might care only about a subset of those classes[21]. However, the objects we don't care about still exist, even though an object query might not bind to them. Therefore, the encoding process in a neural net should *not* prevent separate islands of finer objects (eg, leaves) from getting formed [3].

The representational collapse issue observed in Sec 5 presents a memory-scaling bottleneck. One way to get around it is to consider the behaviour of a self-replicating cellular-automaton [24] like GLOM[6]: unfolding a singular embedding creates dynamic connectionist hardware on-the-fly[10]. This would also allow us to spiritually succeed the computationally-expensive distillation setup, i.e. simply copying a singular embedding lesser number of times on a weaker hardware would allow the student model to exist. A key question that thus remains unanswered for our community is how to compress the entire knowledge of a neural-net into this singular entity (eg, biological seed) and the mechanism behind its unfolding[13](which is an inverse of the protein-folding problem).

2 Mortal computation requires self-replicable embeddings

On the other hand if we want to pack similar levels of intelligence in a brain-like interface which consumes less than 25 watts daily, we have to take an alternate biologically-plausible route[8]: intelligence can be encoded in a single cell (embedding) much like how the genetic code of a human gets encoded in the DNA. In nature, multiple copies of a zygote (cell) lead to emergence of an animal's internal organs. Similarly, in our computers multiple copies of this single replicable embedding shall lead to networks whose structure gets discovered 'on-the-fly' according to local hardware constraints.

It has been well established in distillation literature that [9] a higher-parameterized teacher network can teach a lower-parameter student network to obtain similar performance. If all the knowledge of the teacher could be compressed into a singular embedding, then we won't need distillation at all. Simply copying the single embedding many times on a weaker student hardware would be enough to allow the student model to exist. Although the student model would be weak due to lesser copying,

*Corresponding Author, email: rajatmodi@ucf.edu

[2]This section is meant to be philosophical and optimistic in nature.

[3]And instead encode the complete part-whole hierarchy of a scene[4]. This shall resolve the issue of oversegmentation/ inability to distinguish between part-wholes that SAM[16] still faces when given a grid of input point prompts at a very-fine granularity.[16]

36th Conference on Neural Information Processing Systems (NeurIPS 2023).

the replication step for both student/teacher stems from a common learnt singular representation[4]. These 'connectionist-networks'[10] which start their existence from a single embedding and only remain in the memory as long as the hardware is powered on (hence the name mortal) would require learning algorithms other than backpropogation: where the parameters of each layer could be updated without knowing the precise feed-forward mathematical-functions[7] and allow dynamic growing of the network on a smaller hardware[12]. We remain optimistic for the future that such mortal computation offers to humanity[7].

3 Additional Supplementary Material

This manuscript discusses the supplementary materials in addition to our submission. The supplementary material contains seven sections:

- §A1. shows how our simple VCAPS-Mvitv2 model achieves *a new state-of-the-art* in Video Action Detection, *along* with being *robust* to occlusions.
- §A2. proposes a *new task* called action-segmentation involving *instance-level localization* of actor in a video and presents a benchmark to streamline further research in the field.
- §A3. explores the *importance of background context* in action detection.
- §A4. presents the full benchmark and analyzes the background bias property in existing detectors.
- §4 analyzes our results for *synthetic* occluder motions on O-UCF & O-JHMDB datasets.
- §5 discusses more detail about the video-collection and annotation process of our curated Real-OUCF dataset.
- §A1. explores some plausible ways to solve the representational collapse problem in capsules.
- §A2. presents some qualitative samples from the proposed three Benchmark datasets, along with UCF-101 instance-level annotations. All the datasets, benchmarks and codes for this work will be released for free public usage at `https://anonymous.4open.science/r/OccludedActionBenchmark-B9E2`.
- §A3. provides the NeurIPS recommended datasheet explaining the dataset collection mechanism and other important details.

A1. A Robust Video-Level State Of The Art

We present a new state-of-the-art in spatio-temporal video action detection specifically on UCF-24 and JHMDB-21 datasets in Tab1. Note that we achieve 83.1% on UCF-24 and 98.1% on JHMDB-21 in terms of the widely accepted[14] v-mAP metric at the 0.5ioU threshold. One of the most desirable properties in an action-detector is that it should perform well on *existing* standard datasets[23] as well as be *robust to occlusions* at the same time. Our simple model namely VCAPS-Mvitv2 achieves the best of both worlds, thereby setting a new video-level state of the art for our community.

The robustness of an action-detector can be measured in two ways, 1) the *actual robustness* under occlusions which has been illustrated as absolute value (i.e. 67.3%) in Tab6 of our original manuscript. 2) Measuring the drop in performance of a detector as the ioU threshold during evaluation is swept from 0.2 to 0.5. We define this quantity as $\kappa = 1 - (\frac{vmAP_{0.2} - vmAP_{0.5}}{vmAP_{0.2}})$ which simply measures the relative performance drop from 0.2→0.5ioU. In Tab1, we note that on the much challenging UCF-24 dataset, our method obtains $\kappa = 0.84$, which is greater than all the other methods, thereby indicating more localization robustness. On JHMDB-21, we obtain 92.8% in terms of the absolute v-mAP score, which is significantly better than other existing methods. We acknowledge that TubeR[29] and ST-Mixer [26] are slightly better than our method in terms of f-mAP scores on UCF-24 dataset, although our method is considerably *more* robust (0.84 vs 0.71 on κ score).

A2. A New Instance Level Benchmark

Traditionally[20, 29], spatio-temporal action-detection has relied on predicting *bounding boxes* across an actor for *each* frame. A much harder task instead would be a *finer-grained* localization, i.e.

[4] An undeniable fact of nature is that intelligence/consciousness in humans emerges from self-replication of a singular cell (zygote). It still remains to be seen whether singular prokaryotic organisms like amoeba themselves are conscious [22, 19]

Table 1: **Comparison with existing methods:** Comparison of our method across existing supervised approaches, *: denotes results using a CSN152 backbone. $\kappa = 1 - (\frac{vmAP_{0.2} - vmAP_{0.5}}{vmAP_{0.2}})$. Higher value of κ denotes more robustness.

			UCF-24				JHMDB-21			
	Backbone		f-mAP	v-mAP			f-mAP	v-mAP		
Methods	2D	3D	0.5	0.2	0.5	κ	0.5	0.2	0.5	κ
Yang et al.[15]	✓		75.0	76.6	-	-	-	-	-	-
Li et al.[20]	✓		78.0	82.8	53.8	0.65	70.8	77.3	70.2	0.91
Kopuklu et al.[17]	✓	✓	80.4	75.8	48.8	0.64	75.7	88.3	85.9	0.97
Zhao et al.[29]		✓	**81.3**	85.3	60.2	0.71	82.3*	81.8	80.7	**0.99**
Duarte et al.[5]		✓	78.6	97.1	80.3	0.83	64.6	95.1	-	-
Kumar et al.[18]		✓	69.2	95.3	71.9	0.75	68.1	96.8	68.4	0.71
Tao et al.[26]		✓	**83.7**	-	-	-	86.7	-	-	-
Ours		✓	81.2	**98.6**	**83.1**	**0.84**	**93.0**	**98.1**	**92.8**	0.95

predicting instance-level mask for an actor instead of only bounding boxes. Surprisingly, to the best of our knowledge, only one approach i.e. VideoCapsuleNet[5] is able to solve the much harder task of instance-level action segmentation by adapting the network trivially out-of-the-box.

We argue that research in the field of instance-level action-segmentation has largely been inhibited due to the lack of proper instance-level actor annotations for standard action-detection datasets[23]. While the popular JHMDB[11] dataset consists of instance puppet masks, the larger UCF-24 dataset only has bounding box annotations. To rectify this, we release the instance-level annotations for UCF-24 some of whose samples have been illustrated in Fig56. Our insight is that these annotations will now allow the existing research in action-detection and Video Instance Segmentation [25] to progress concurrently due to inherently similar problem formulations [5]. Finally, we note that the results of VCAPS on instance-level benchmark in Tab2 are significantly lower than on the bounding-box level benchmark in Tab4. This indicates that instance-level localization task is significantly harder task than isolating 'broader level' bounding boxes.

Annotation Procedure for UCF-24: We generate instance-level segmentation masks on UCF101-24 videos leveraging the recent SOTA in Video Instance segmentation [25]. To make sure that our instance-level action tubes are temporally coherent, we perform inference over successive chunk sizes of 100 frames, with a temporal overlap of t = 30 frames. Next, we run a sliding window for smoothening the predicted tube to remove any stray pixel-level artifacts. Finally, we manually refine the obtained segmentation masks using the CVAT tool.[1]

Table 2: **Instance Level Benchmark on O-UCF**: A benchmark showing a VCAPS model trained using our instance-level annotations on UCF-24 datasets and evaluated on clean/occluded test-set.

			BG1			BG2			BG3					
	Occ As Aug	Clean	FG1	FG2	FG3	FG1	FG2	FG3	FG1	FG2	FG3	Circle	Sin	Avg
VCAPS[5]	×	34.7	29.3	21.0	13.7	28.3	19.1	13.5	29.2	20.0	13.1	13.3	18.1	19.9
VCAPS[5]	✓	41.1	36.4	29.5	23.5	36.3	28.1	24.4	34.9	28.2	23.8	22.6	25.8	28.5

A3. Preliminary Experiments

In Fig1 of the main manuscript, we had run a set of preliminary experiments which served as a motivation for this benchmark study. One notable result has been illustrated in Table 3. *Distractor* refers to the setting when there is just one single occluder in the background of a video. *No Context* refers to the setting when all the background pixels of the video have been blackened, thereby making it easier for the network[20, 5] to classify the remaining pixels as an actor. It can be clearly observed that the highest drops (i.e. δ_r for No Context case) are observed when the background is entirely masked. This shows that existing networks are highly *biased* to the background information

[5]The only difference is VIS involves instance-classification which could be solved by just a single-frame object detection. However, action detection requires higher-level action-classification which requires temporal reasoning. The per-frame localization task of both the problems is fundamentally identical.

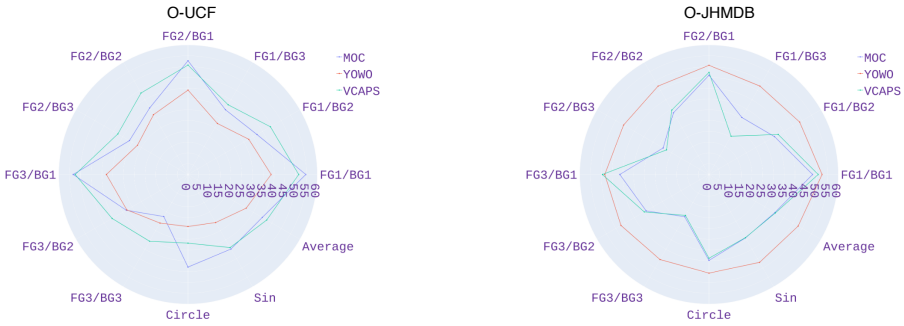

Figure 1: **Occlusion Robustness** across multiple severity levels. The outermost envelope represents most robust model. Radial Axis represents vmAP-0.5 score. For O-UCF, VCAPS is most robust, whereas for O-JHMDB, YOWO is most robust across several categories.

while trying to reason about an actor location. We note that this dependency on background is *counter-intuitive* to the desirable behaviours of learning object(actor)-centric representations[2, 29].

Table 3: **Distractor vs Background-Context Sensitivity**: Highest performance drops are observed when background is masked (no context), whereas presence of distractor objects in background has slightly less effect.

	UCF-24					JHMDB-21				
		Distractor		No Context			Distractor		No Context	
Methods	Clean	Abs	δ_r	Abs	δ_r	Clean	Abs	δ_r	Abs	δ_r
MOC[20]	54.4	38.6	0.71	26.6	0.49	77.2	53.7	0.70	49.5	0.64
YOWO[17]	48.8	43.8	**0.90**	32.2	0.66	**85.7**	**73.5**	0.86	**66.5**	**0.78**
VCAPS[5]	**75.5**	**67.7**	**0.90**	**53.1**	**0.70**	65.7	61.8	**0.94**	47.5	0.72

A4. Full Benchmark Analysis

Table 4: **O-UCF Benchmark:** Illustrates the full benchmark across 9 severity levels of static occlusions and different trajectories in dynamic occlusions.

		BG1			BG2			BG3					
Methods	Occ As Aug	FG1	FG2	FG3	FG1	FG2	FG3	FG1	FG2	FG3	Circle	Sin	Avg
MOC[20]	×	54.7	37.0	34.7	52.7	35.6	31.4	53.3	33.1	22.5	42.9	39.9	39.8
YOWO[17]	×	38.6	32.6	27.3	39.1	32	27.1	37.9	32.8	26	24.1	25.7	31.2
VCAPS[5]	×	51.4	44.2	37.3	50.7	43.6	37.6	52.4	40.7	35.7	31.8	39.1	42.2
MOC[20]	✓	48.3	43.6	39.6	47.3	43.1	39.1	47.2	44.6	38.5	48.1	45.6	44.1
YOWO[17]	✓	46.5	45.3	43.4	46.1	45.3	43.4	46.5	45.7	44.3	43.8	43.6	44.9
VCAPS[5]	✓	54.7	51.1	48.7	54.8	50.4	47.6	54.2	50.7	47.8	43.4	45.3	49.9

In Tables 4, 5, we present the full benchmark of the proposed O-UCF and O-JHMDB datasets across the 9 severity levels of static occlusions and circular/sinusoidal dynamic motions.

Background Bias: For an ideal action-detector, it is expected that occlusions in the background wont impact the localization performance in the actor region. We observe this trend in Fig2, where the performance of an ideal detector would be a black line parallel to x-axis. In the graph, most of the methods suffer negligible drops on increasing the occlusion severity in the background. Note that MOC[20] suffers most drop across different severity levels. Our VCAPS-Mvitv2 obtains the highest robustness across all occlusion severity levels for both O-UCF & O-JHMDB datasets.

Table 5: **O-JHMDB Benchmark:** Illustrates the full benchmark across 9 severity levels of static occlusions and different trajectories in dynamic occlusions.

	Occ As Aug	BG1			BG2			BG3			Circle	Sin	Avg
		FG1	FG2	FG3	FG1	FG2	FG3	FG1	FG2	FG3			
MOC[20]	×	48.1	35.1	30.5	46	33	24.6	41.4	33.7	22.6	39.8	33.9	35.3
YOWO[17]	×	52.5	48.6	47.3	50.6	47.3	45.8	48.4	47.2	45.5	45.7	47	47.8
VCAPS[5]	×	50.7	37.1	20.5	47.3	34.5	22.8	49.5	34.7	21.9	38.8	33.8	35.6
MOC[20]	✓	59.6	55.9	53.3	58.9	54.3	52.9	58.5	54.8	52.2	57.2	57.2	55.9
YOWO[17]	✓	71.1	70	69.3	69.2	70.3	66.9	68.4	67	65	65.7	65.1	68
VCAPS[5]	✓	60.8	59	54	61.2	56.9	53.9	61.5	57.8	54.4	55.3	58.2	59.7

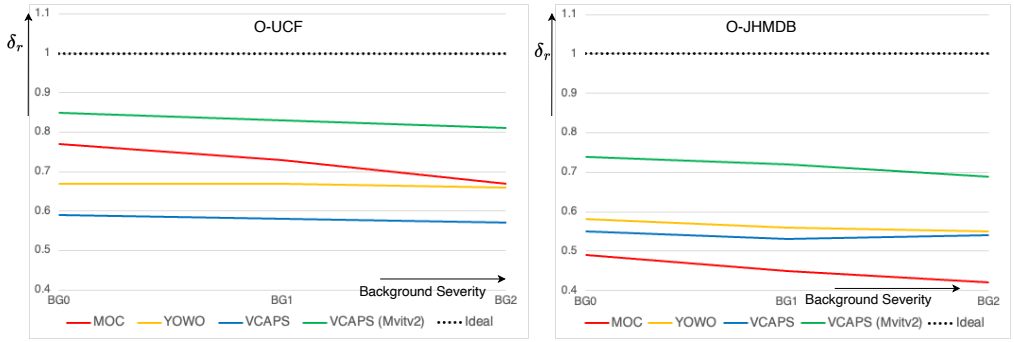

Figure 2: **Background Bias:** As the occlusion severity over background is increased, the performance of an ideal detector should not drop (dashed line). Our VCAPS-Mvitv2 shows highest robustness of all studied models. (X axis): Different severity levels of background occlusions. (Y Axis): Relative Robustness (δ_r)

4 Motion Analysis

In the main manuscript, we had experimented with *realistic* motion on the OVIS-dataset. We had also experimented with *synthetic* motions on O-UCF and O-JHMDB datasets, and show those results in Table 6. We observe that in case of MOC using the 2D DLA34 backbone, the performance is better in case of static occlusions, (i.e. 28.7 vs 27.3 on O-UCF). This shows that 2D backbones work well if occluder is static. On the other hand, YOWO uses a combination of 2D backbone (YOLO) and a 3D backbone (ResNext). This shows a slight positive improvement of 1.2% in UCF and 0.6% in JHMDB, thereby indicating that 3D backbone seems to be helping. The best improvement is evident by purely switching to a 3D based backbone as in the case of VCAPS, where gains as much as 4.1% can be observed. From this, it can be seen that 3D backbones *might* help networks reason about actor locations even under the challenging scenarios when the occluders might be moving. However, we note that these differences are close within a small margin of error. Also, our experiments with different type of trajectories in Tab45, i.e. circle and sinusoids show inconsistent- trends thereby indicating that there is no particular motion type to which networks are *more* sensitive.

5 Real-OUCF dataset

We have curated Real-OUCF dataset for realistic occlusion scenarios.

Video Selection Process: The original videos in UCF-24 dataset were mostly of sporting events, but were of significantly lower resolution, i.e. 240 by 320. Furthermore, they consisted of only a single actor. Now, we have significantly upgraded that test set to reflect multiple actors which mutually occlude each other with a high degree of overlap as high as 99%. First, we scraped the videos matching keywords like "riding bike" etc. One curious way we were able to get such overlap ratios was by searching specifically for events like *tandem surfing, tandem diving etc.* In tandem-events, two actors try to synchronize with each other. A lateral captured viewpoint, which captures multiple actors (with one actor behind another), offers very challenging yet realistic conditions for occlusions. Finally, our curated videos are consisting of professional sporting events like Olympics, as well as casual settings, for eg, people just playing cricket in a park.

Table 6: **Static vs Motion Trends**: 2D backbones work well on Static Occlusions whereas 3D backbones work best under Dynamic Occlusions. S- Static Occluders. D- Dynamic Occluders. For analysing clear effects, results are shown with only one occluder over an actor (no bg occlusion).

			UCF-24				JHMDB-21			
Methods	Backbone	Type	Clean	S	D	D-S	Clean	S	D	D-S
MOC[20]	DLA34 [27]	2D	54.4	28.7	27.3	-0.4	77.2	34.4	34.0	-1.1
YOWO[17]	YOLO +ResNext	2D+3D	48.8	24.0	25.2	1.2	85.7	54.6	55.2	0.6
VCAPS[5]	i3D [3]	3D	75.5	42.5	46.6	4.1	65.7	23.0	23.9	0.9

Annotation Criteria: In spatio-temporal action detection, there are two questions that the annotations try to answer 1) what is the time interval during which the action occurs 2) what is the spatial location of the actor in *each* frame of the action-interval. Therefore, we first temporally crop the scraped videos to obtain action start and end times. Finally, for all the frames in between this interval, we spatially localize the actors using the CVAT annotation tool. One subtle thing is that action-detection does not try to label *all* the actors in the video. For eg, if some people are standing , and some people are doing some useful activity like biking, we are only concerned with the biking. This is in line with the annotation procedure of official UCF24.

A1. Representational Collapse in Capsules

In Fig 5 of the main manuscript, we have shown the problem of representational collapse[6], where capsules cannot decode multiple entities (objects), if the number of objects in the scene are greater than the parameters in the network. A classical way to solve this problem would be to to increase the number of parameters by reinitializing and retraining the machine again. It agrees with the evolutionary observation that more parameters mean more intelligence and the only option to do better is to scale up (i.e 85 billion neurons in the brain vs 100 trillion in ChatGPT4). This is a promising direction if we want to continue to pay exorbitant amounts of money for specialist hardware, expensive electricity or wait for the hardware to become cheaper in future.

Mortal computation requires self-replicable embeddings: On the other hand if we want to pack similar levels of intelligence in a brain-like interface which consumes less than 25 watts daily, we have to take an alternate biologically-plausible route[8]: intelligence can be encoded in a single cell (embedding) much like how the genetic code of a human gets encoded in the DNA. In nature, multiple copies of a zygote (cell) lead to emergence of an animal's internal organs. Similarly, in our computers multiple copies of this single replicable embedding shall lead to networks whose structure gets discovered 'on-the-fly' according to local hardware constraints.

It has been well established in Distillation literature that [9] a higher-parameterized teacher network can teach a lower-parameter student network to obtain similar performance. If all the knowledge of the teacher could be compressed into a singular embedding, then we won't need distillation at all. Simply copying the single embedding many times on a weaker student hardware would be enough to allow the student model to exist. These 'connectionist-networks'[10] which start their existence from a single embedding and only remain in the memory as long as the hardware is powered on (hence the name mortal) would require learning algorithms other than backpropogation: where the parameters of each layer could be updated without knowing the precise feed-forward mathematical-functions[7] and allow dynamic growing of the network on a smaller hardware[12].

We are very excited to see how this direction turns out for our community.

A2. Dataset Samples

In this section, we show some qualitative samples from our proposed-datasets and benchmarks. Specifically, Fig34 shows the realistic occlusion dataset which we have collected for evaluating robustness to real-world occlusions. Next, Fig78contains the synthetic dataset samples from O-UCF and O-JHMDB consisting of controlled occlusions. Fig56 contains the exhaustive instance-level annotations of UCF-24 videos we have released to facilitate our proposed benchmark of action-segmentation.

[6]This section is meant to be philosophical and optimistic in nature. The kind reader is requested to skip it if concrete-results are desired.

Figure 3: **Our Real-OUCF:** contains realistic occlusions with instance-level action annotations.

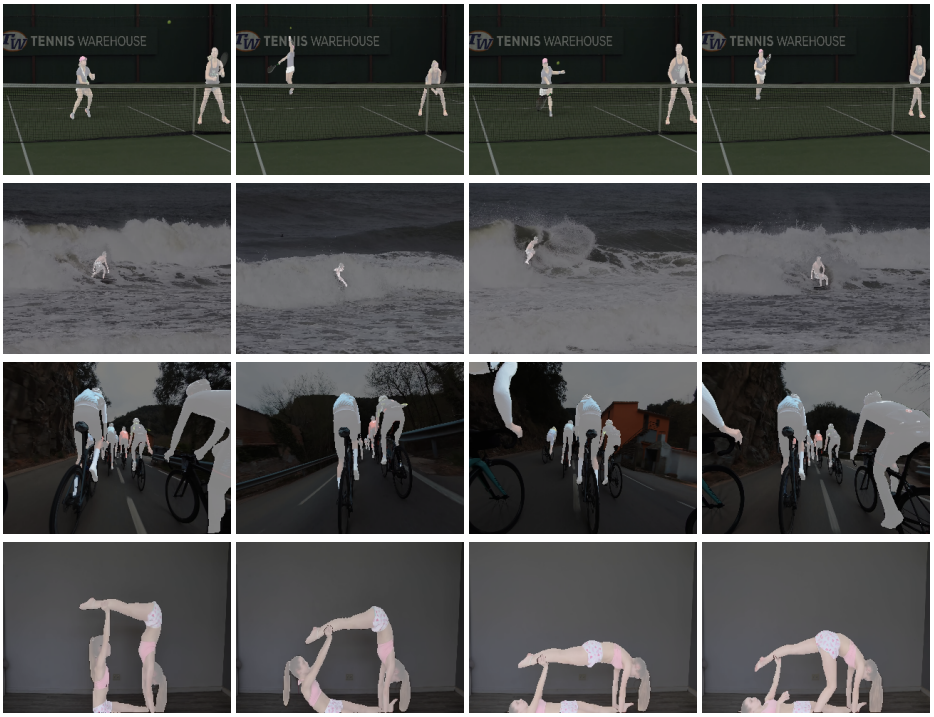

Figure 4: **Our Real-OUCF:** contains realistic occlusions with instance-level action annotations.

Figure 5: **Our instance-level annotations for UCF-24:** We propose a new benchmark of instance-level action segmentation and release official annotations.

Figure 6: **Our instance-level annotations for UCF-24:** We propose a new benchmark of instance-level action segmentation and release official annotations.

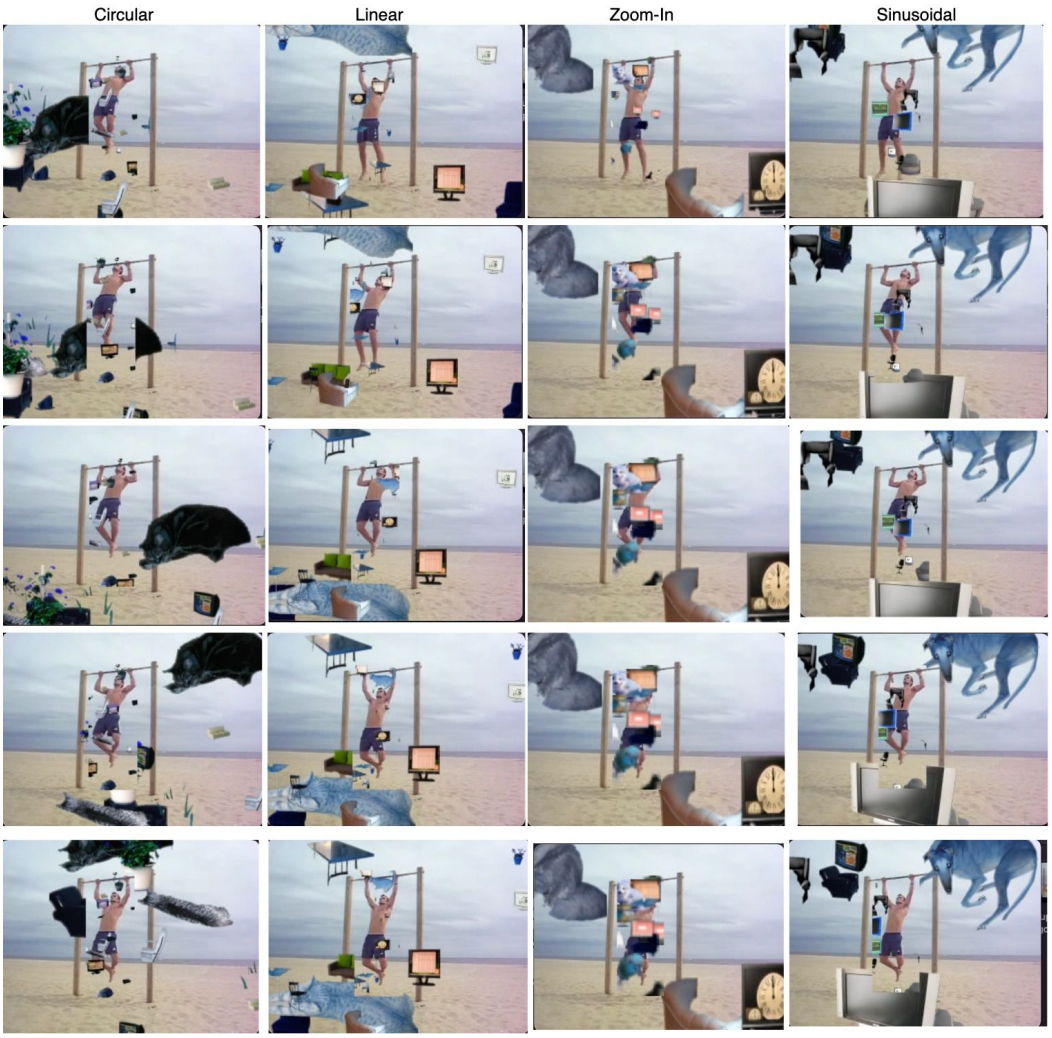

Figure 7: **Different occluder trajectories:** are illustrated for our proposed O-UCF and O-JHMDB datasets.

Figure 8: **Nature of Occluders**: Occluders in the proposed O-UCF & O-JHMDB datasets belong to either indoor/outdoor samples.

References

[1] Computer vision annotation tool. `https://github.com/opencv/cvat`, 2013.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Mingyu Ding, Yikang Shen, Lijie Fan, Zhenfang Chen, Zitian Chen, Ping Luo, Joshua B Tenenbaum, and Chuang Gan. Visual dependency transformers: Dependency tree emerges from reversed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14528–14539, 2023.

[5] Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Videocapsulenet: A simplified network for action detection. *Advances in neural information processing systems*, 31, 2018.

[6] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *arXiv preprint arXiv:2102.12627*, 2021.

[7] Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

[8] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *Neural Computation*, pages 1–40, 2022.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[10] Geoffrey E Hinton. Connectionist learning procedures. In *Machine learning*, pages 555–610. Elsevier, 1990.

[11] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *International Conf. on Computer Vision (ICCV)*, pages 3192–3199, December 2013.

[12] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016.

[13] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[14] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017.

[15] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, 2017.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[17] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019.

[18] Akash Kumar and Yogesh Singh Rawat. End-to-end semi-supervised learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14700–14710, 2022.

[19] Mariana Lenharo. Decades-long bet on consciousness ends-and it's philosopher 1, neuroscientist 0. *Nature*, 2023.

[20] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *European Conference on Computer Vision*, pages 68–84. Springer, 2020.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[22] Arthur S Reber, William B Miller, and František Baluška. Consciousness: unicellular organisms know the secret. *Nature*, 620(7972):37–37, 2023.

[23] 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.

[24] Stephen Wolfram et al. *A new kind of science*, volume 5. Wolfram media Champaign, IL, 2002.

[25] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, pages 588–605. Springer, 2022.

[26] Tao Wu, Mengqi Cao, Ziteng Gao, Gangshan Wu, and Limin Wang. Stmixer: A one-stage sparse action detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14720–14729, 2023.

[27] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018.

[28] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[29] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13598–13607, 2022.

A3. Datasheets For Datasets

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

O-UCF, O-JHMDB were created to perform systematic benchmark study of occlusions in video action detection. Real-OUCF was created to evaluate sota action-detectors on real-world occlusions.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

These were created by a research-group whose identity shall be released after the review process.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Dataset contains videos of several people performing actions.

Is there a label or target associated with each instance? If so, please provide a description.

For each instance, we provide an instance-level segmentation mask along with our datasets.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The idea for a real-world occlusion dataset is that it is easy to train an existing network using synthetic occlusions as a data augmentation. However, there is a need of consistent real-world tet set to benchmark all the future approaches against, and that is the test set which our Real-OUCF provides.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

A very detailed care has been taken to make our annotations noise free. In the rare case that a correction is needed, the updated annotations will be posted on our official link (Shared on top of this manuscript).

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

All the videos of our dataset were obtained after hand picking from YOUTUBE, and then scraping them.Due to offline availability of the dataset, our dataset is now self-contained.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, it is spatio-temporal action detection. It primarily considers human actors.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

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? If so, please describe how.

The classes in our Real-OUCF dataset primarily belong to sports such as basketball. Several videos of the dataset are picked from official sporting events like NBA, Olympics where it might be possible to identify famous athletes by face only. However, all the videos we use are already there in the public domain.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Data was raw-video directly observed with human eyes.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Manual human curation

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Graduate students, who were graciously supported by research grants with warm thanks.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Dataset was collected over a period of 6 months from Dec22-May23.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Pre-processing was done by temporally cropping a video into smaller clips which indicate start/end of the action. After auto-labelling spatio-temporal masks via SAM, we manually annotate/refine masks using the CVAT tool.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Link shall be provided soon for the raw data.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, we use normal Segment Anything and Computer Vision Annotation Tool. https://github.com/facebookresearch/segment-anything https://github.com/opencv/cvat

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Yes, for our benchmark study on the impact of occlusions in spatio-temporal video action detection.

What (other) tasks could the dataset be used for?

Amodal mask completion, Occlusion Robustness Testing etc.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

When will the dataset be distributed?

Just before the start of the neurips 2023 conference.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

No, its free for use by all parties. However, the discretion to update more video samples in future, and refine the existing annotations over time lie with the original dataset authors.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No, its free for use by all parties. However, the discretion to update more video samples in future, and refine the existing annotations over time lie with the original dataset authors.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No, its free for use by all parties. However, the discretion to update more video samples in future, and refine the existing annotations over time lie with the original dataset authors.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Dataset will be supported by the authors/research group of this paper.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

By email-address, the details shall be revealed post-review .

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Changes if any , will be posted once a year to our github repo.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes, this shall be useful to give reliable comparisons on different dataset versions in existing literature.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We also release a python script which allows to create on -the-fly semi-realistic occlusions consisting of static occlusions at different severity levels and different motions of the occluder. Anyone could use this to generate infinite occluder variations.

For realistic occlusions, once could always collect more real world samples or expand the count of existing classes. Note that the test set we provide is already significantly larger than traditional action-detection datasets. [23, 11]

Any other comments?

We are grateful to you for taking the time to read this datasheet and reviewing this manuscript.