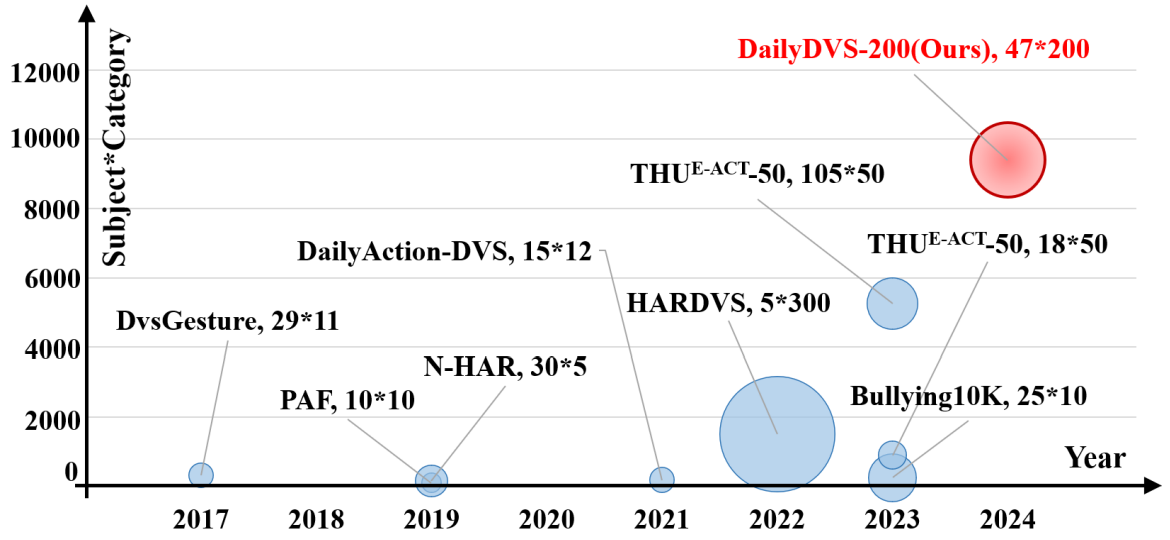
DailyDVS-200: A Comprehensive Benchmark Dataset for Event-Based Action Recognition

Qi Wang[1*], Zhou Xu[1*], Yuming Lin[1], Jingtao Ye[1], Hongsheng Li[1], Guangming Zhu[1],
Syed Afaq Ali Shah[2], Mohammed Bennamoun[3], Liang Zhang[1]✉
[1]Xidian University, [2]Edith Cowan University, [3]University of Western Australia

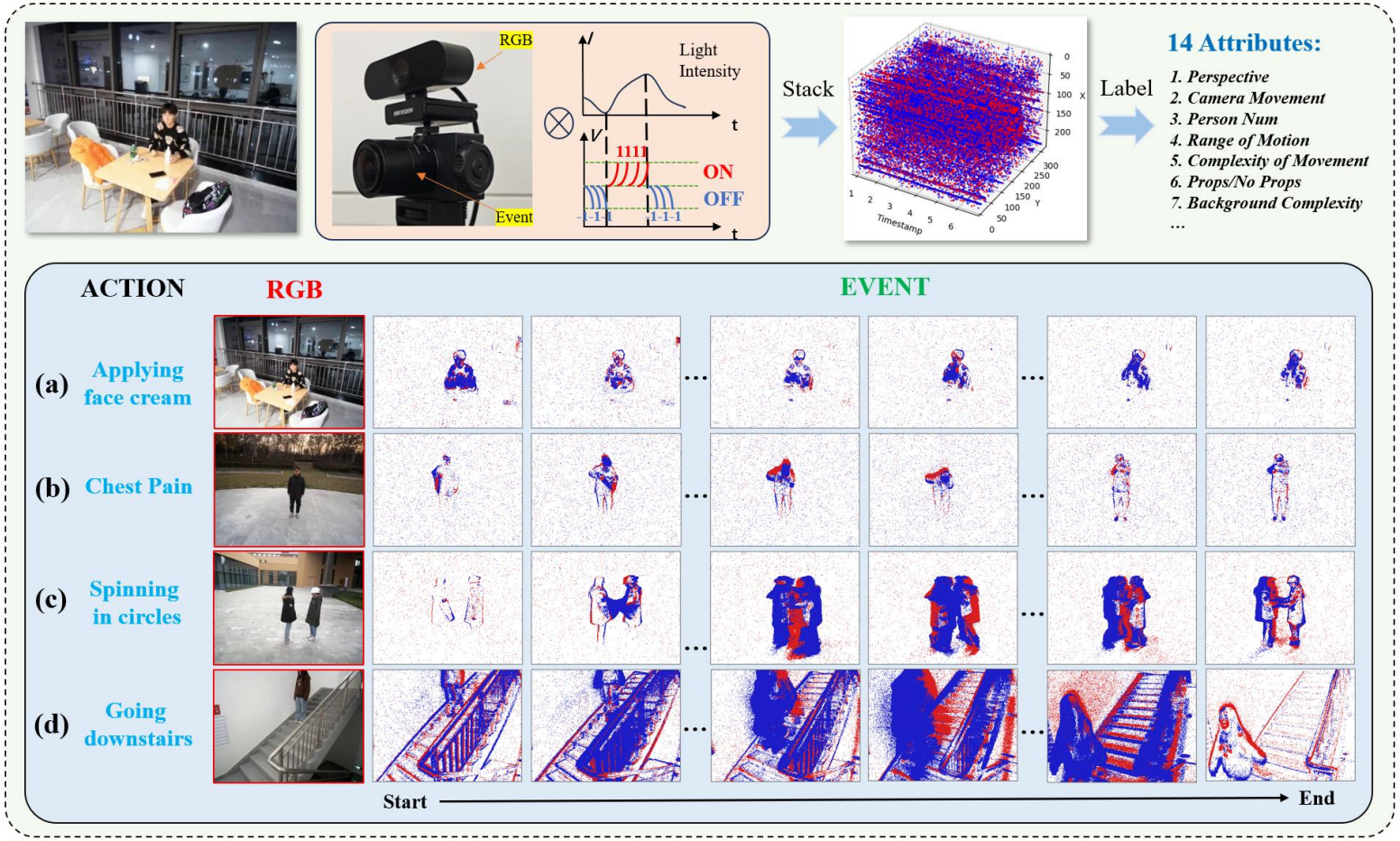
Contributions

- We introduce DailyDVS-200, a pioneering large-scale neuromorphic dataset for action recognition, featuring over 22,000 samples across 200 categories with real-world challenges and detailed attribute annotations.
- We've set up benchmarks for diverse recognition models on this dataset, establishing a solid foundation for future performance comparisons in the field of neuromorphic action recognition.
- Our group evaluation and analysis, based on attribute annotations, not only assesses the influence of various attributes on event-based models but also paves the way for new research directions in neuromorphic datasets.

DailyDVS-200					
Scale	22046	**14 attributes:**			
Class	200	**PR**	Props	**PE**	Perspective
Subject	47	**PO**	Posture	**DI**	Diurnality
Sensors	DVXplorer Lite	**DR**	Duration	**LO**	Location
Resolution	320×240	**AR**	Action Range	**DT**	Distance
Duration	1-20s	**PN**	Person Num	**HE**	Height
EventCount	18 billion	**CM**	Camera Motion	**SH**	Shadow
EventTime	24 hour	**IL**	Illumination Direction	**BC**	Background Complexity



Data acquisition process



- As we move forward, you'll see how these meticulously captured and annotated data points contribute to the robustness of our dataset. They form the backbone of our benchmarks and are instrumental in pushing the boundaries of neuromorphic action recognition.

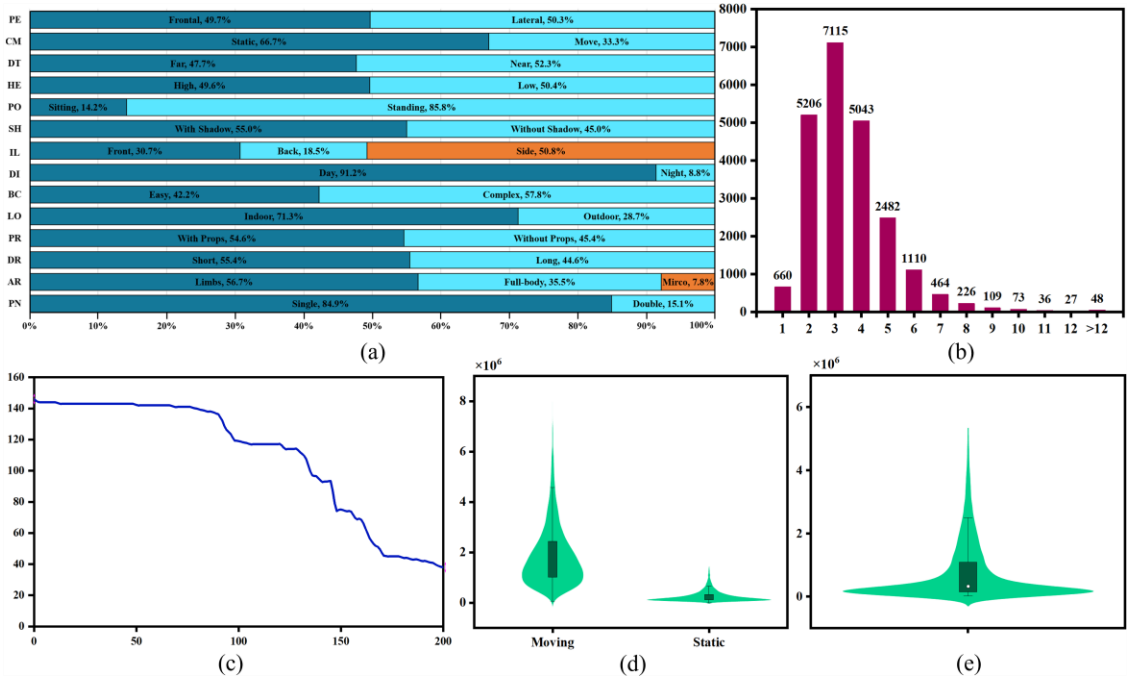
DailyDVS-200: Event-Based Action Recognition Dataset

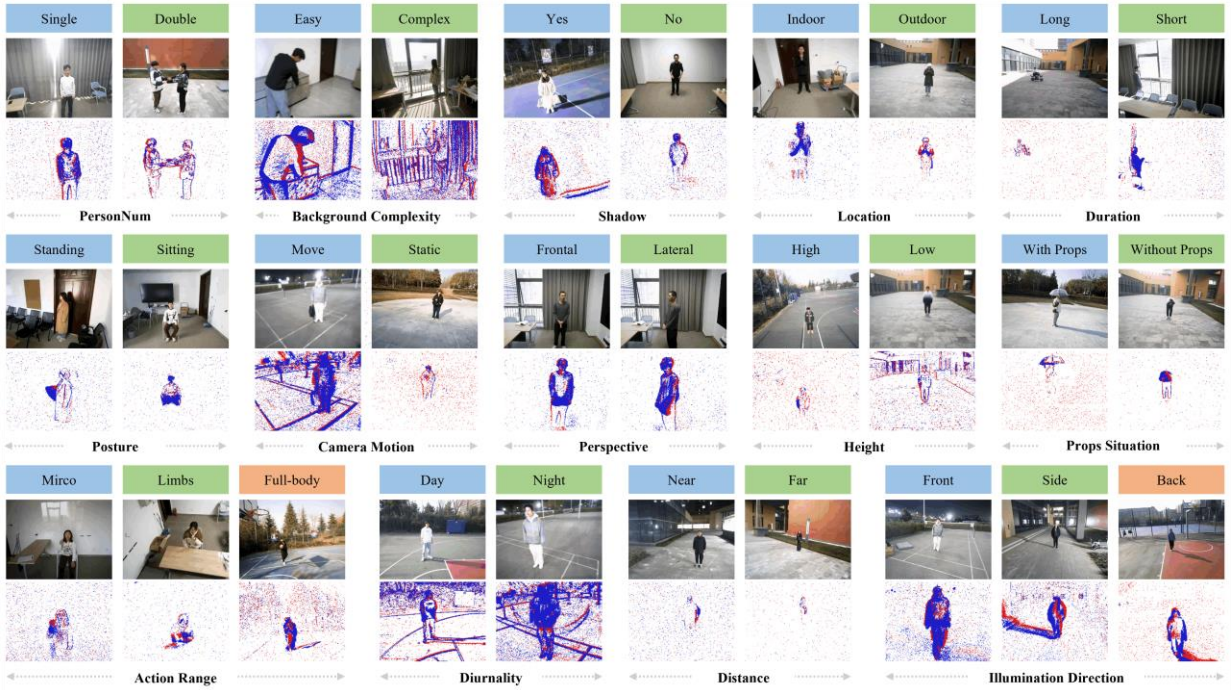
Dataset	Year	Sensors	Object	Scale	Class	Sub	Real	MA	AA	DR
ASLAN-DVS	2011	DAVIS240c	Action	3,697	432	-	✗	-	-	-
MNISTDVS	2013	DAVIS128	Image	30,000	10	-	✗	-	-	-
N-Caltech101	2015	ATIS	Image	8,709	101	-	✗	-	-	0.3s
N-MNIST	2015	ATIS	Image	70,000	10	-	✗	-	-	0.3s
CIFAR10-DVS	2017	DAVIS128	Image	10,000	10	-	✗	-	-	1.2s
HMDB-DVS	2019	DAVIS240c	Action	6,766	51	-	✗	-	-	19s
UCF-DVS	2019	DAVIS240c	Action	13,320	101	-	✗	-	-	25s
N-ImageNet	2021	Samsung-Gen3	Image	1,781,167	1,000	-	✗	-	-	-
ES-lmageNet	2021	-	Image	1,306,916	1,000	-	✗	-	-	-
DvsGesture	2017	DAVIS128	Action	1,342	11	29	✔	✗	✗	6s
N-CARS	2018	ATIS	Car	24,029	2	-	✔	✗	✗	0.1s
ASL-DVS	2019	DAVIS240	Hand	100,800	24	5	✔	✗	✗	0.1s
PAF	2019	DAVIS346	Action	450	10	10	✔	✗	✗	5s
DailyAction	2021	DAVIS346	Action	1,440	12	15	✔	✗	✗	5s
HARDVS	2022	DAVIS346	Action	107,646	300	5	✔	✔	✗	5s
$\text{THU}^{\text{E-ACT}}$-50	2023	CeleX-V	Action	10,500	50	105	✔	✗	✗	2-5s
$\text{THU}^{\text{E--ACT}}$-50-CHL	2023	DAVIS346	Action	2,330	50	18	✔	✗	✗	2-5s
Bullying10K	2023	DAVIS346	Action	10,000	10	25	✔	✗	✗	2-20s
DailyDVS-200 (**Ours**)	2024	DVXplorer Lite	Action	22,046	200	47	✔	✔	✔	1-20s

- Our proposed dataset consists of 200 distinct action categories, collected from 47 subjects. The dataset we provide has the richest number of subject-category combinations. The detailed comparison with existing benchmark datasets can be found in table.

DailyDVS-200: Event-Based Action Recognition Dataset

- A preview of our proposed DailyDVS-200 dataset and examples of our attribute annotations.
- Statistical data and detailed analysis of DailyDVS-200.
- DailyDVS-200 dataset exemplifies diversity in action recognition.





Various scene:

- Household Activities
- Office Tasks
- Sports and Physical Activities
- Health-related Activities
- Interactions
- Bullying and Violence
- Transportation-related Activities

Various setting & Various action characteristics:

- Different Perspectives
- Different Distances
- Lighting Conditions
- Different Camera Movements

- Fine-grained Micro, Limb, and Whole-body Movements
- Short-duration and Long-duration
- Actions with and without Props Interaction

Experimental Results

Methods	Year	Input Type	Backbone	top-1 acc.(%)	top-5 acc.(%)
C3D	2015	Frame	3D CNN	21.99	45.81
I3D	2017	Frame	ResNet50	32.30	59.05
R2Plus1D	2018	Frame	ResNet34	36.06	63.67
SlowFast	2019	Frame	ResNet50	**41.49**	68.19
TSM	2019	Frame	ResNet50	40.87	**71.46**
EST	2019	Learnable	ResNet34	**32.23**	**59.66**
TimeSformer	2021	Token	Transformer	44.25	74.03
Swin-T	2022	Token	Transformer	**48.06**	**74.47**
ESTF	2022	Token	ResNet18	24.68	50.18
GET	2023	Token	Transformer	37.28	61.59
Spikformer	2022	Spike	Transformer	**36.94**	**62.37**
SDT	2024	Spike	Transformer	35.43	58.81

- Evaluation of different methods on our dataset. Swin Transformer achieved the highest accuracy in both top-1 and top-5, with 48.06% and 74.47% respectively, and also achieved the highest accuracy among all methods. In the spike-based methods, Spikformer achieved the highest top-1 and top-5 accuracy of 36.94% and 62.37%, respectively.

Experimental Results

Methods	1. BC		2. DR		3. IL			4. DT		5. HE		6. LO		7. CM	
	Complex	Easy	Long	Short	Back	Front	Side	Far	Near	High	Low	Indoor	Outdoor	Move	Static
SlowFast	45.98	39.20	40.94	42.18	44.05	34.52	46.13	41.62	41.64	38.34	45.36	43.21	39.08	21.05	51.81
TSM	45.37	40.84	39.68	44.69	42.91	36.60	47.16	40.71	44.09	38.90	46.51	43.41	40.93	26.88	50.16
EST	35.22	30.59	29.42	34.52	32.89	24.71	38.26	28.93	35.28	27.45	37.64	33.76	29.82	13.52	41.51
Spikformer	38.96	33.45	33.86	36.68	36.77	30.09	39.11	31.57	39.00	31.82	39.52	38.54	30.33	13.74	46.15
Swin-T	**52.86**	**45.37**	**48.57**	**47.64**	**51.61**	**39.11**	**53.39**	**47.16**	**48.89**	**44.46**	**52.14**	**50.15**	**44.70**	**27.84**	**58.05**

Methods	8. PN		9. PE		10. PR		11. AR			12. SH		13. PO		14. DI	
	One	Two	Frontal	Lateral	No	Yes	Full-body	Limbs	Micro	No	Yes	Stand	Sit	Day	Night
SlowFast	39.15	55.57	44.52	39.04	41.81	41.48	52.05	37.59	25.16	37.07	42.73	42.93	35.92	43.10	33.65
TSM	40.07	55.90	46.13	39.17	46.66	38.98	53.31	37.21	33.02	42.75	42.39	43.11	39.61	44.20	33.02
EST	29.88	45.56	33.42	31.15	34.81	30.08	41.18	28.55	19.81	35.18	31.55	32.70	30.13	34.74	18.55
Spikformer	33.88	44.10	37.09	33.94	37.39	33.80	43.86	31.95	23.27	39.47	34.45	35.01	37.24	38.30	19.81
Swin-T	**44.70**	**66.88**	**50.98**	**45.43**	**51.35**	**45.33**	**59.24**	**43.15**	**34.59**	**46.78**	**48.36**	**48.96**	**44.08**	**50.22**	**36.32**

- The fine-grained group testing results on models trained on our dataset.

Methods	0.5s			0.25s			0.125s		
	gap0	gap2	gap4	gap0	gap2	gap4	gap0	gap2	gap4
C3D [53]	21.99	14.98	11.70	31.10	24.82	22.82	44.81	32.62	27.75
I3D [9]	32.30	22.94	20.82	45.39	29.54	29.42	59.10	36.70	37.50
R2Plus1D [54]	36.06	26.39	**24.97**	49.65	36.62	32.10	58.88	48.06	40.29
SlowFast [17]	**41.49**	**26.90**	24.55	**52.16**	33.28	25.43	**64.09**	44.81	44.64
TSM [31]	40.87	23.67	22.97	49.55	**37.94**	**32.37**	61.76	**51.48**	**48.77**

- Evaluation of existing models with different frame settings on our dataset.

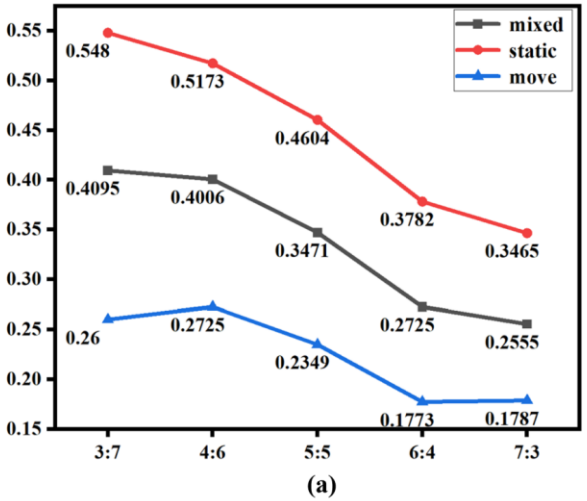
Experimental Results

Methods	Top-10 accurate actions	Top-10 incorrect actions	Methods	Top-10 accurate actions	Top-10 incorrect actions
Slowfast	1. pull out the chair 2. open curtains 3. **hand in hand circling** 4. close the door 5. wash towels 6. fall down 7. turn the light on 8. **wipe the table** 9. cross your legs 10. **push-up**	1. hammer table 2. **play table tennis** 3. open window 4. clean windows 5. **pitch** 6. charge a phone 7. arrang cards 8. V sign 9. use a tablet 10. kick a ball	EST	1. tie shoelaces 2. mutual bow 3. turn the light on 4. **push-up** 5. turn the light off 6. fall down 7. lie on the table 8. close window 9. cheers 10. **hand in hand circling**	1. **play table tennis** 2. open window 3. close curtains 4. slap the table 5. **pitch** 6. OK sign 7. make paper cuttings 8. charge a phone 9. hit people with things 10. headache
Swin-T	1. **hand in hand circling** 2. arm wrestling 3. **push-up** 4. close curtains 5. open curtains 6. close the door 7. sit-up 8. **wipe the table** 9. moves heavy objects 10. cross legs	1. hammer table 2. **pitch** 3. take off headphones 4. write 5. blow nose 6. V sign 7. crush paper into a ball 8. take something from bag 9. chest pain 10. open the bottle	Spikformer	1. **wipe the table** 2. close curtains 3. stand up 4. **hand in hand circling** 5. lie on the table 6. mutual bow 7. cross legs 8. clean the windows 9. **push-up** 10. go upstairs	1. **pitch** 2. **play table tennis** 3. plug in the power strip 4. take off shoes 5. trim nails 6. stomachache 7. play with hair 8. roll up sleeves 9. backache 10. headache

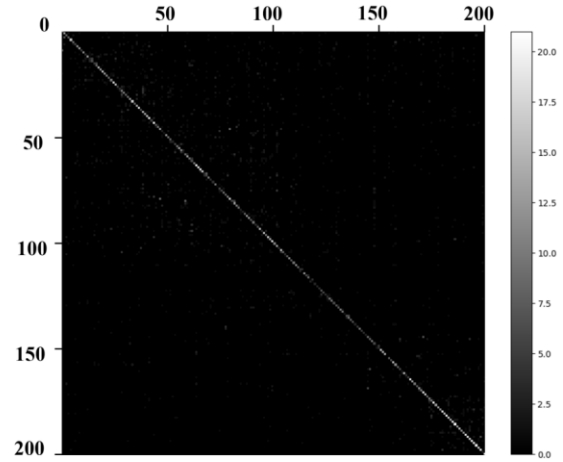
- Top-10 accurate and top-10 incorrect actions of different methods. The same action is marked with the same color.

Experimental Results

- Evaluation of using different sizes of Moving camera set for action recognition and Confusion matrix of swin transformer.



(a)



(b)

- Comparison with other large-scale datasets under the same training configuration.

Methods	TSM	ESTF
THU$^{\text{E-ACT}}$-50	95.60 / 98.75†	95.25
THU$^{\text{E-ACT}}$-50-CHL	49.07 / 83.83†	49.50
Hardvs	97.33 / 98.55†	96.67
Bullying10K	74.22 / 91.90†	84.72
DailyDVS-200(**Ours**)	**36.05** / **65.90**†	**31.29**

Thanks

Code & Dataset:
github.com/QiWang233/DailyDVS-200