MoLo: Motion-augmented Long-short Contrastive Learning for Few-shot Action Recognition

Xiang Wang[1], Shiwei Zhang[2], Zhiwu Qing[1], Changxin Gao[1], Yingya Zhang[2], Deli Zhao[2], Nong Sang[1]

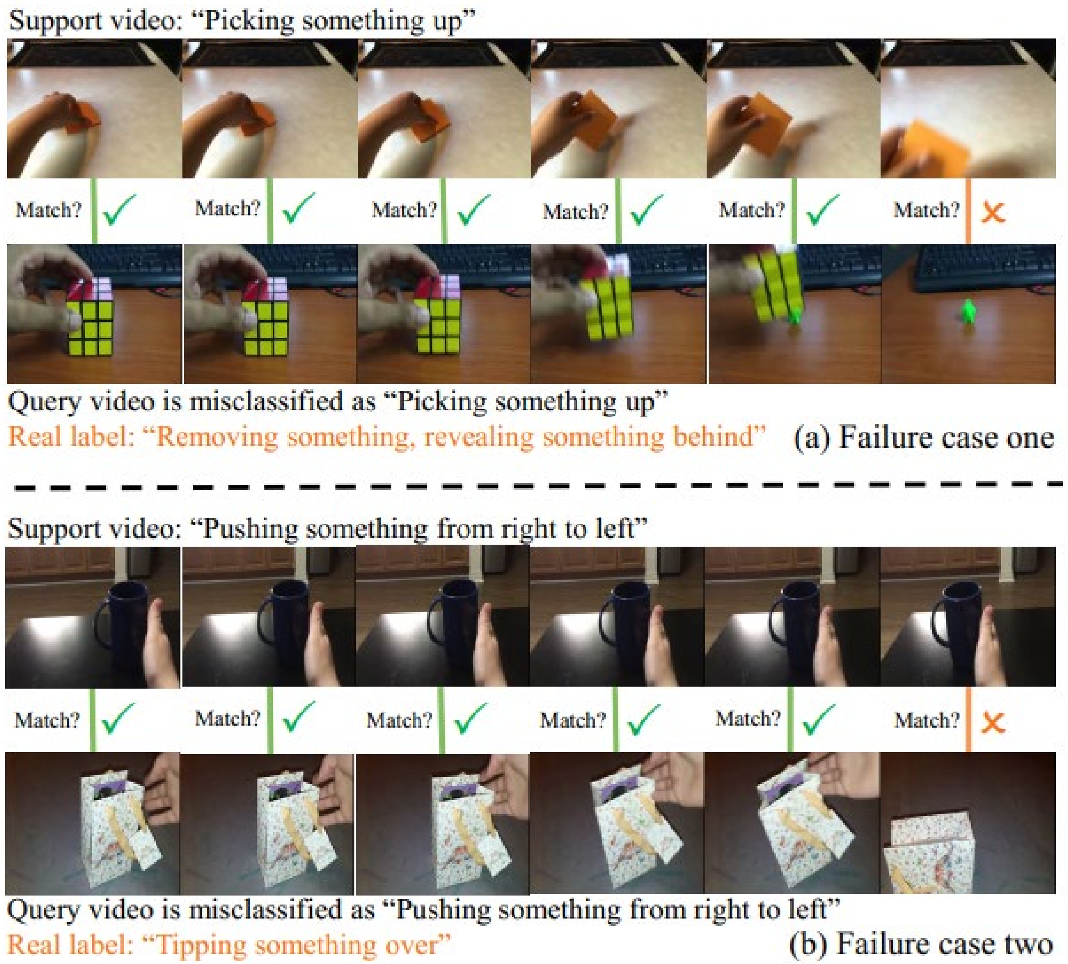
Code: https://github.com/alibaba-mmai-research/MoLo

[1]Huazhong University of Science and Technology

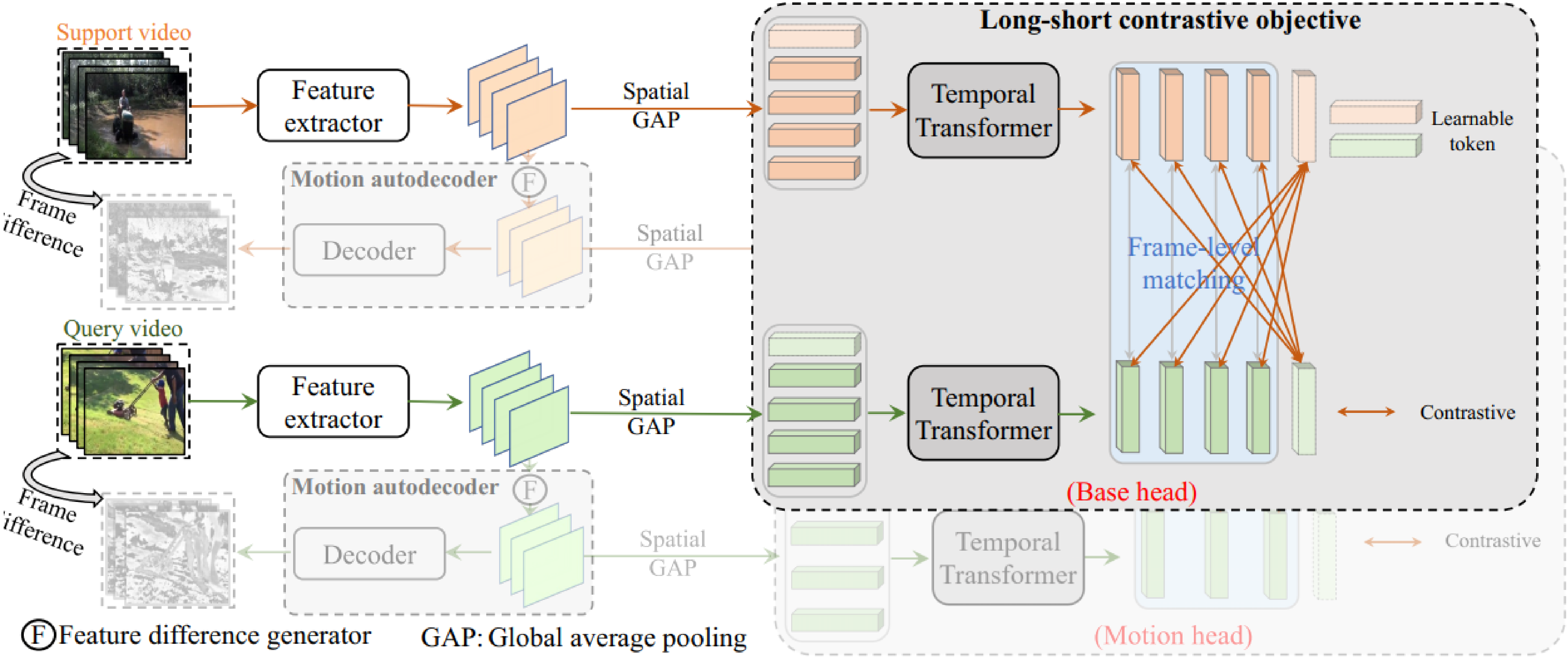
[2]Alibaba Group

Limitations of metrics-based meta-learning frameworks

Limitations of the previous approaches:
(1) the matching procedure between local frames tends to be inaccurate due to the lack of guidance to force long-range temporal perception;
(2) explicit motion learning is usually ignored, leading to partial information loss

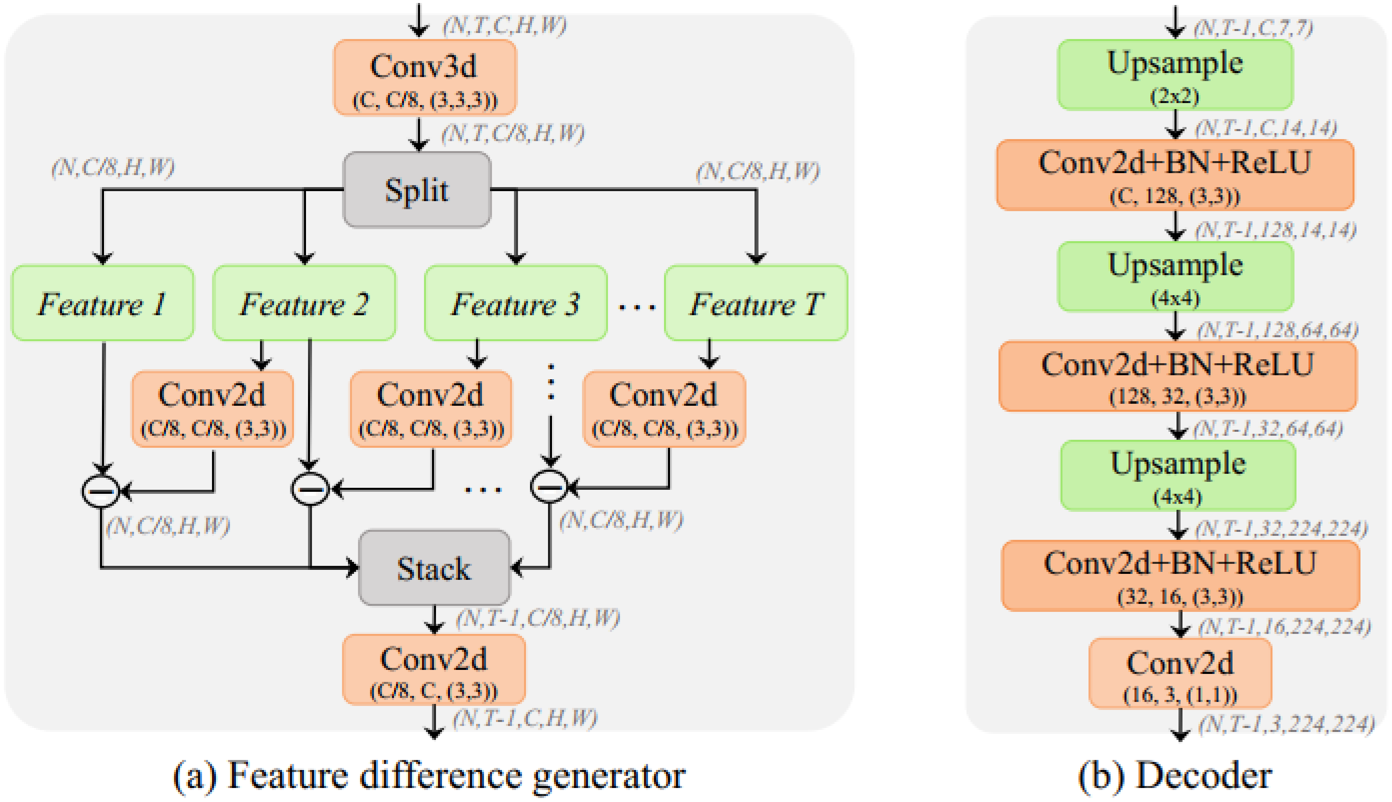


MoLo



Motion-augmented **Lo**ng-short Contrastive Learning (MoLo)

MoLo



Motion-augmented **Lo**ng-short Contrastive Learning (MoLo)

Generalization performance in different video scenarios

Comparison with state-of-the-art

Method	Reference	SSv2-Full 1-shot	SSv2-Full 2-shot	SSv2-Full 3-shot	SSv2-Full 4-shot	SSv2-Full 5-shot	Kinetics 1-shot	Kinetics 2-shot	Kinetics 3-shot	Kinetics 4-shot	Kinetics 5-shot
MatchingNet [61]	NeurIPS'16	-	-	-	-	-	53.3	64.3	69.2	71.8	74.6
MAML [14]	ICML'17	-	-	-	-	-	54.2	65.5	70.0	72.1	75.3
Plain CMN [88]	ECCV'18	-	-	-	-	-	57.3	67.5	72.5	74.7	76.0
CMN++ [88]	ECCV'18	34.4	-	-	-	43.8	-	-	-	-	-
TRN++ [86]	ECCV'18	38.6	-	-	-	48.9	-	-	-	-	-
TARN [3]	BMVC'19	-	-	-	-	-	64.8	-	-	-	78.5
CMN-J [89]	TPAMI'20	-	-	-	-	-	60.5	70.0	75.6	77.3	78.9
ARN [81]	ECCV'20	-	-	-	-	-	63.7	-	-	-	82.4
OTAM [4]	CVPR'20	42.8	49.1	51.5	52.0	52.3	72.2*	75.9	78.7	81.9	84.2*
ITANet [83]	IJCAI'21	49.2	55.5	59.1	61.0	62.3	73.6	-	-	-	84.3
TRX (Ω={1}) [44]	CVPR'21	38.8	49.7	54.4	58.0	60.6	63.6	75.4	80.1	82.4	85.2
TRX (Ω={2, 3}) [44]	CVPR'21	42.0	53.1	57.6	61.1	64.6	63.6	76.2	81.8	83.4	85.9
TA^{2}N [35]	AAAI'22	47.6	-	-	-	61.0	72.8	-	-	-	85.8
MTFAN [79]	CVPR'22	45.7	-	-	-	60.4	**74.6**	-	-	-	**87.4**
STRM [58]	CVPR'22	43.1	53.3	59.1	61.7	68.1	62.9	76.4	81.1	83.8	86.7
HyRSM [74]	CVPR'22	54.3	62.2	65.1	67.9	69.0	73.7	80.0	83.5	84.6	86.1
Bi-MHM [74]	CVPR'22	44.6*	49.2*	53.1*	54.8*	56.0*	72.3*	77.2*	81.1*	84.1*	84.5*
Nguyen *et al.* [41]	ECCV'22	43.8	-	-	-	61.1	74.3	-	-	-	**87.4**
Huang *et al.* [21]	ECCV'22	49.3	-	-	-	66.7	73.3	-	-	-	86.4
HCL [85]	ECCV'22	47.3	54.5	59.0	62.4	64.9	73.7	79.1	82.4	84.0	85.8
MoLo (OTAM)	-	55.0	61.8	64.8	67.7	69.6	73.8	80.2	83.1	84.2	85.1
MoLo (Bi-MHM)	-	**56.6**	**62.3**	**67.0**	**68.5**	**70.6**	74.0	**80.4**	**83.7**	**84.7**	85.6

Table 1. Comparison with recent state-of-the-art few-shot action recognition methods on the SSv2-Full and Kinetics datasets under the 5-way setting. The experimental results are reported as the shot increases from 1 to 5. ”-” indicates the result is not available in published works. The best results are bolded and the underline means the second best performance. “$*$” stands for the results of our implementation.

Generalization performance in different video scenarios

Comparison with state-of-the-art

Method	Reference	UCF101			SSv2-Small			HMDB51		
		1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
MatchingNet [61]	NeurIPS'16	-	-	-	31.3	39.8	45.5	-	-	-
MAML [14]	ICML'17	-	-	-	30.9	38.6	41.9	-	-	-
Plain CMN [88]	ECCV'18	-	-	-	33.4	42.5	46.5	-	-	-
CMN-J [89]	TPAMI'20	-	-	-	36.2	44.6	48.8	-	-	-
ARN [81]	ECCV'20	66.3	-	83.1	-	-	-	45.5	-	60.6
OTAM [4]	CVPR'20	79.9	87.0	88.9	36.4	45.9	48.0	54.5	65.7	68.0
ITANet [83]	IJCAI'21	-	-	-	39.8	49.4	53.7	-	-	-
TRX [44]	CVPR'21	78.2	92.4	<u>96.1</u>	36.0	51.9	<u>56.7</u>*	53.1	66.8	75.6
TA^{2}N [35]	AAAI'22	81.9	-	95.1	-	-	-	59.7	-	73.9
MTFAN [79]	CVPR'22	84.8	-	95.1	-	-	-	59.0	-	74.6
STRM [58]	CVPR'22	80.5	92.7	**96.9**	37.1	49.2	55.3	52.3	67.4	<u>77.3</u>
HyRSM [74]	CVPR'22	83.9	93.0	94.7	40.6	<u>52.3</u>	56.1	<u>60.3</u>	<u>71.7</u>	76.0
Bi-MHM [74]	CVPR'22	81.7*	88.2*	89.3*	38.0*	47.6*	48.9*	58.3*	67.1*	69.0*
Nguyen *et al.* [41]	ECCV'22	84.9	-	95.9	-	-	-	59.6	-	76.9
Huang *et al.* [21]	ECCV'22	71.4	-	91.0	38.9	-	**61.6**	60.1	-	77.0
HCL [85]	ECCV'22	82.5	91.0	93.9	38.7	49.1	55.4	59.1	71.2	76.3
MoLo (OTAM)	-	<u>85.4</u>	<u>93.4</u>	95.1	<u>41.9</u>	50.9	56.2	59.8	71.1	76.1
MoLo (Bi-MHM)	-	**86.0**	**93.5**	95.5	**42.7**	**52.9**	56.4	**60.8**	**72.0**	**77.4**

Table 2. Comparison with state-of-the-art few-shot action recognition methods on UCF101, SSv2-Small, and HMDB51 in terms of 1-shot, 3-shot, and 5-shot classification accuracy. "-" stands for the result is not available in published works. The best results are bolded in black, and the underline represents the second best result. "$*$" indicates the results of our implementation.

Generalization performance in different video scenarios

Ablation study

Long-short contrastive	Autodecoder	Head Base	Head Motion	SSv2-Full 1-shot	SSv2-Full 5-shot
		✓		44.6	56.0
			✓	46.3	60.6
✓		✓		52.2	68.0
✓	✓	✓		53.2	68.1
	✓		✓	47.8	61.8
✓	✓		✓	53.9	69.7
		✓	✓	49.2	63.4
✓		✓	✓	53.3	68.2
	✓	✓	✓	53.2	68.1
✓	✓	✓	✓	**56.6**	**70.6**

Table 3. Ablation study on SSv2-Full under 5-way 1-shot and 5-way 5-shot settings. The top line represents the baseline Bi-MHM. To avoid confusion, note that the "motion head without autodecoder" setting contains the feature difference generator by default.

Each module is complementary to each other

Generalization performance in different video scenarios

Ablation study

Setting	SSv2-Full 1-shot	SSv2-Full 5-shot	Kinetics 1-shot	Kinetics 5-shot
Frame Difference	56.6	70.6	74.0	85.6
RAFT Flow [57]	**56.8**	**71.1**	**74.4**	**85.9**
TRX [44]	42.0	64.6	63.6	85.9
TRX + Motion autodecoder	**45.6**	**66.1**	**64.8**	**86.3**

Different motion reconstruction targets

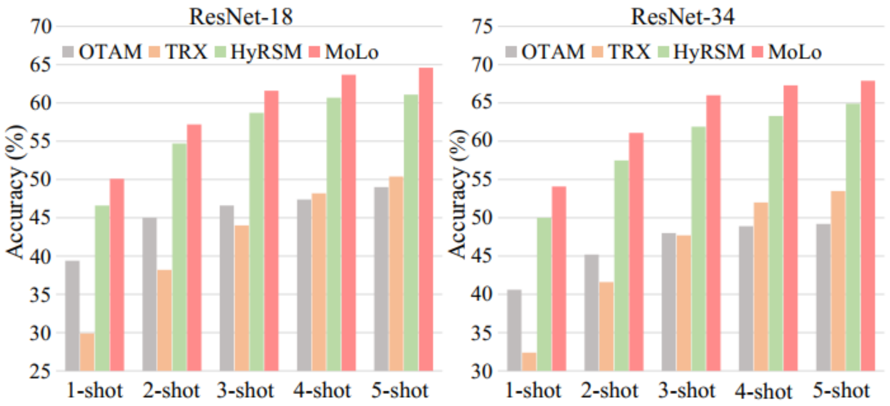


Figure 4. Performance comparison of varying backbone depth on the SSv2-Full dataset under the 5-way K-shot setting. The experiments are carried out with the shot changing from 1 to 5.

Varying backbone depth

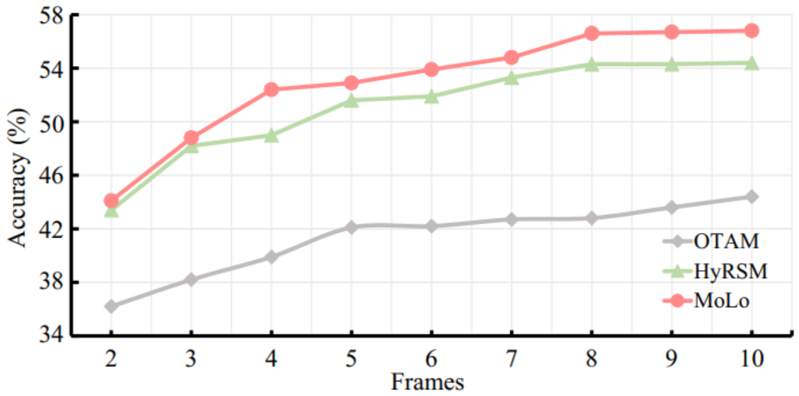


Figure 5. Ablation study on the effect of changing the number of input video frames under the 5-way 1-shot SSv2-Full setting.

Different number of frames

Generalization performance in different video scenarios

Ablation study

Setting	SSv2-Full		Kinetics	
	1-shot	5-shot	1-shot	5-shot
Temporal Transformer×1	**56.6**	70.6	**74.0**	**85.6**
Temporal Transformer×2	56.4	**71.7**	72.5	84.9
Temporal Transformer×3	56.0	71.3	71.6	84.2
Temporal Transformer×4	55.9	69.6	71.1	83.9
Temporal Transformer×5	55.8	69.4	70.5	83.3

Table 5. Ablation study for different number of temporal Transformer layers on the SSv2-Full and Kinetics datasets.

Different number of temporal Transformer layers

Setting	SSv2-Full		Kinetics	
	1-shot	5-shot	1-shot	5-shot
Temporal Transformer-only	53.2	68.1	72.7	84.6
Temporal Transformer w/ TAP	54.8	69.5	73.3	85.2
Temporal Transformer w/ token (MoLo)	**56.6**	**70.6**	**74.0**	**85.6**

Table 6. Comparison experiments on the effect of learnable token and other variants on the SSv2-Full and Kinetics datasets.

Analysis of long-short contrastive objective

Method	SSv2-Full						Kinetics					
	5-way	6-way	7-way	8-way	9-way	10-way	5-way	6-way	7-way	8-way	9-way	10-way
OTAM [4]	42.8	38.6	35.1	32.3	30.0	28.2	72.2	68.7	66.0	63.0	61.9	59.0
TRX [44]	42.0	41.5	36.1	33.6	32.0	30.3	63.6	59.4	56.7	54.6	53.2	51.1
HyRSM [74]	54.3	50.1	45.8	44.3	42.1	40.0	73.7	69.5	66.6	65.5	63.4	61.0
MoLo	**56.6**	**51.6**	**48.1**	**44.8**	**42.5**	**40.3**	**74.0**	**69.7**	**67.4**	**65.8**	**63.5**	**61.3**

Table 8. N-way 1-shot classification accuracy comparison with recent few-shot action recognition methods on the test sets of SSv2-Full and Kinetics datasets. The experimental results are reported as the way increases from 5 to 10.

N-way few-shot classification

Welcome to visit to our poster!

ID: 00299

Thank you!