Paper ID mmfp0387

ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

Unsupervised Video Hashing with Multi-granularity Contextualization and Multi-structure Preservation

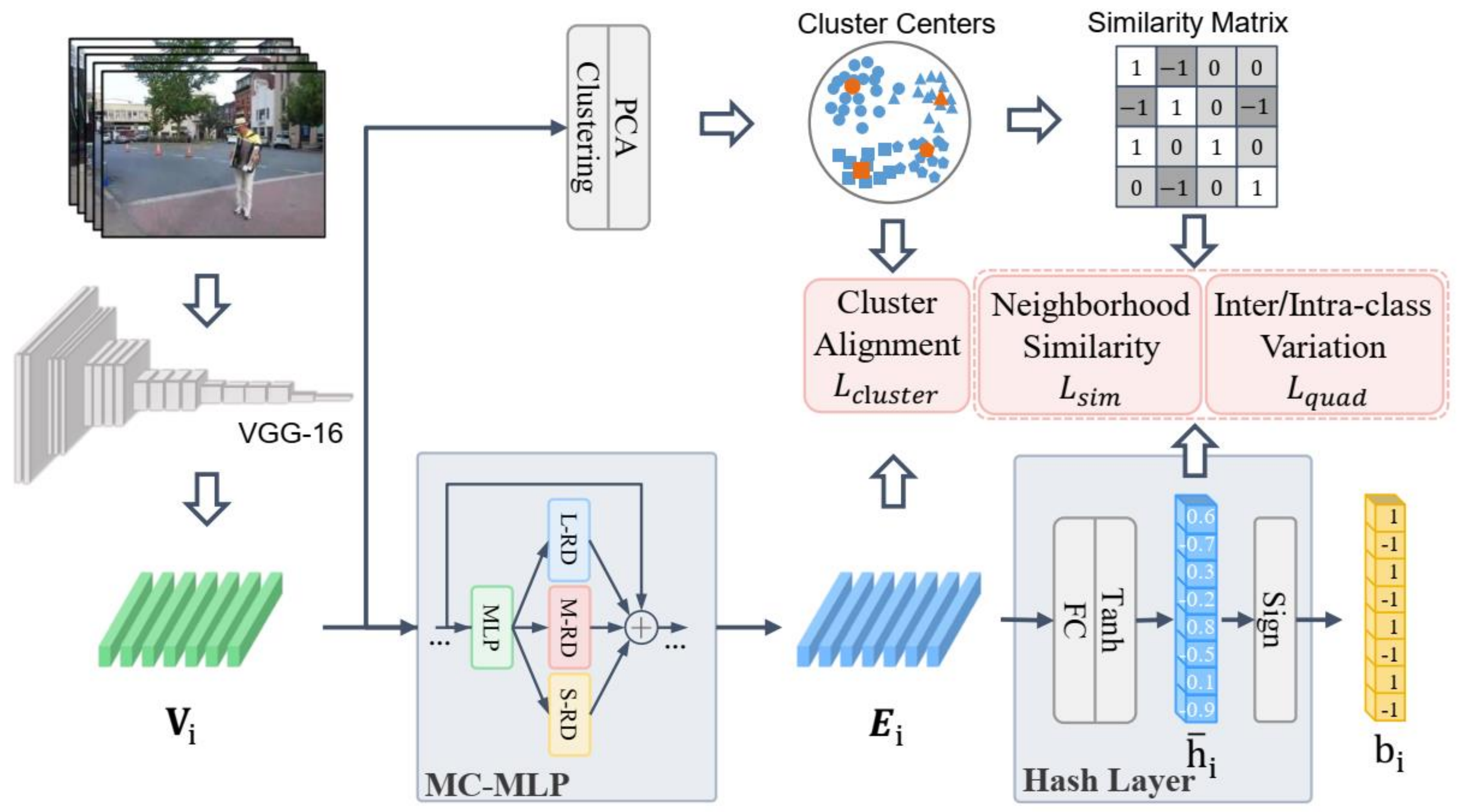
Yanbin Hao, Jingru Duan, Hao Zhang*, Bin Zhu, Pengyuan Zhou, Xiangnan He

Motivation:

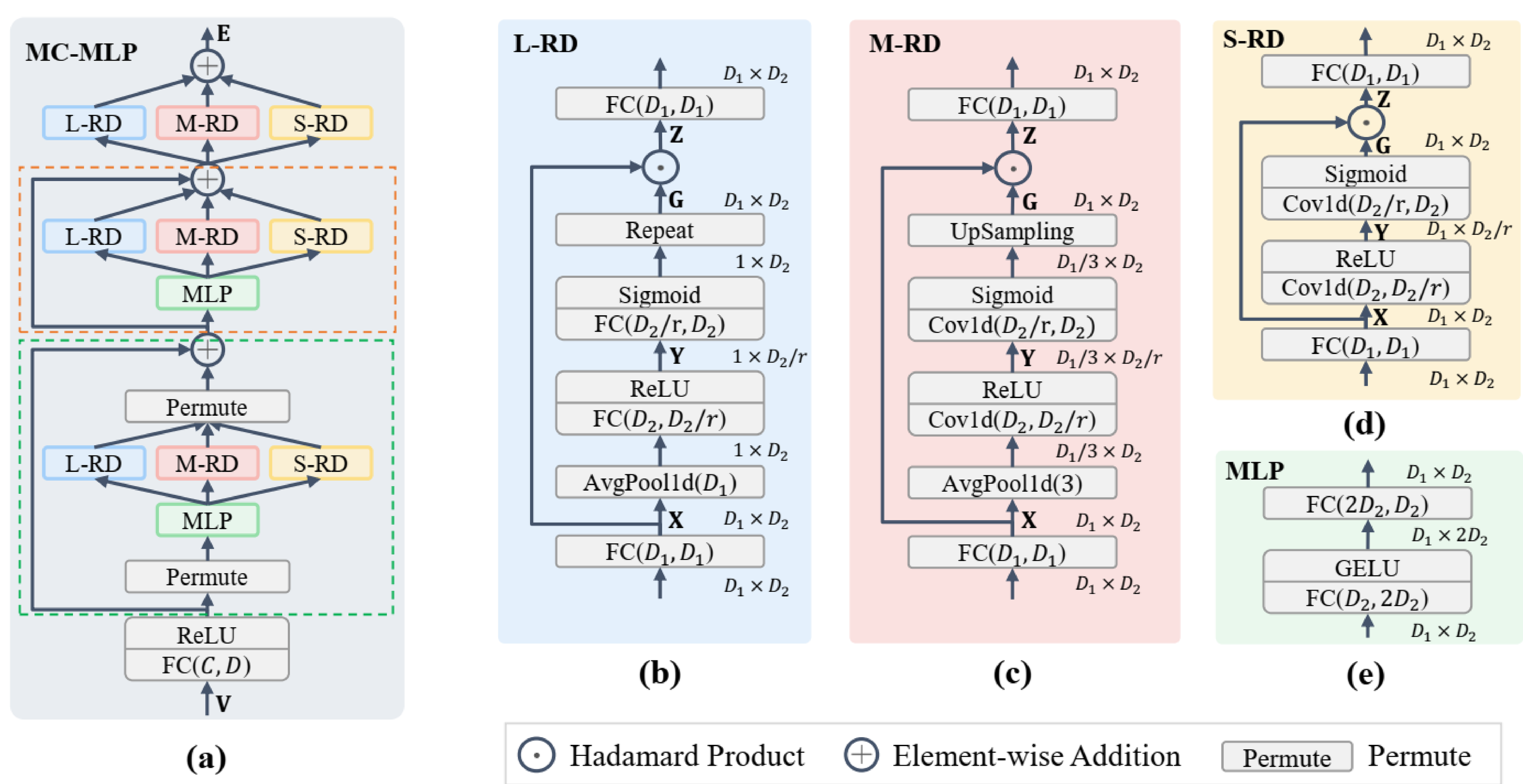
- Existing unsupervised hashing methods generally suffer from incomplete exploration of various perspective dependencies and data structures that exist in visual contents.
- MLP-Mixer achieve comparable performance with the advanced CNNs and Transformers but require a lower computational cost.

Proposed framework:

- ***MCMSH:*** overall structure of the proposed Multi-granularity Contextualized and Multi-Structure preserved Hashing



- ***MC-MLP:*** densely integrate the three self-gating modules L/M/S-RD into MLP-Mixer to build the Multi-granularity Contextualized MLP(MC-MLP).



Experiments

- ***Ablation study:***

1. Without L/M/S-RD modules *VS* With L/M/S-RD modules

Model	32 bits					64 bits				
	k=20	k=40	k=60	k=80	k=100	k=20	k=40	k=60	k=80	k=100
MLP-Mixer	0.288	0.244	0.223	0.208	0.195	0.323	0.277	0.254	0.237	0.223
+L-RD	0.298	0.253	0.229	0.212	0.198	0.330	0.283	0.258	0.240	0.225
+M-RD	0.298	0.253	0.230	0.213	0.199	0.332	0.284	0.258	0.240	0.225
+S-RD	0.295	0.250	0.226	0.209	0.195	0.329	0.282	0.257	0.239	0.224
MC-MLP	**0.302**	**0.258**	**0.235**	**0.217**	**0.202**	**0.335**	**0.288**	**0.263**	**0.245**	**0.230**

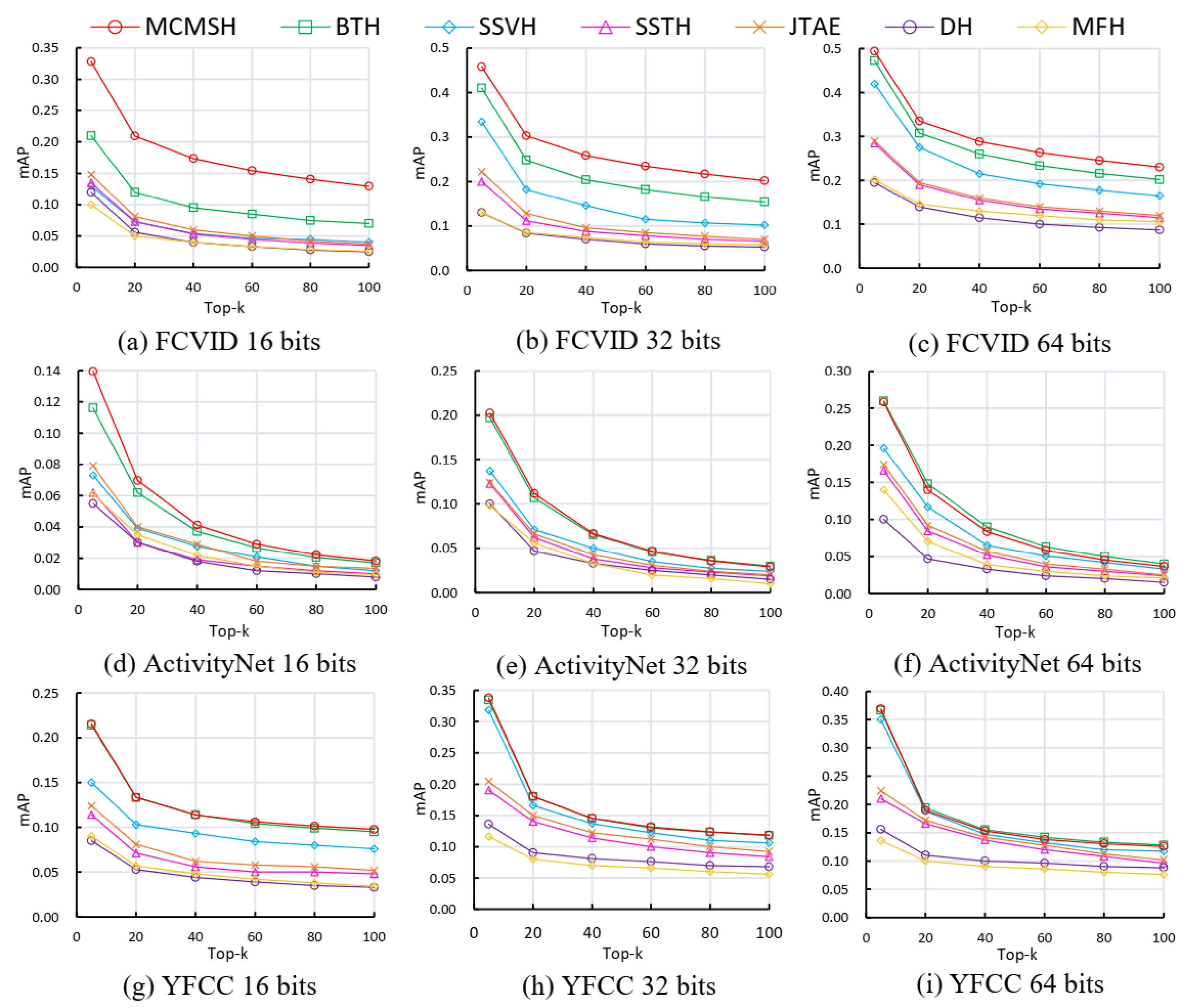
Table1: Performance (mAP@k) comparison with different MC-MLP variants on FCVID with 32-bits and 64-bits hash code lengths.

2. Different structures and their combinations

Loss	k=5	k=20	k=40	k=60	k=80	k=100
$L_{cluster}\,(\alpha=1)$	0.466	0.304	0.256	0.230	0.211	0.197
$L_{sim}\,(\beta=1)$	0.430	0.270	0.228	0.206	0.190	0.176
$L_{quad}\,(\gamma=1)$	0.441	0.262	0.211	0.185	0.167	0.154
$0.8L_{cluster}+0.1L_{sim}$	0.490	0.332	0.285	0.260	0.241	0.225
$0.8L_{cluster}+0.01L_{quad}$	0.486	0.328	0.282	0.257	0.239	0.224
$0.1L_{sim}+0.01L_{quad}$	0.464	0.290	0.239	0.213	0.195	0.181
MCMSH	**0.494**	**0.335**	**0.288**	**0.263**	**0.245**	**0.230**

Table2: Performance (mAP@k) comparison with a single data structure and their combination using FCVID with 64-bits hash codes.

- ***Comparison with SOTAs:***



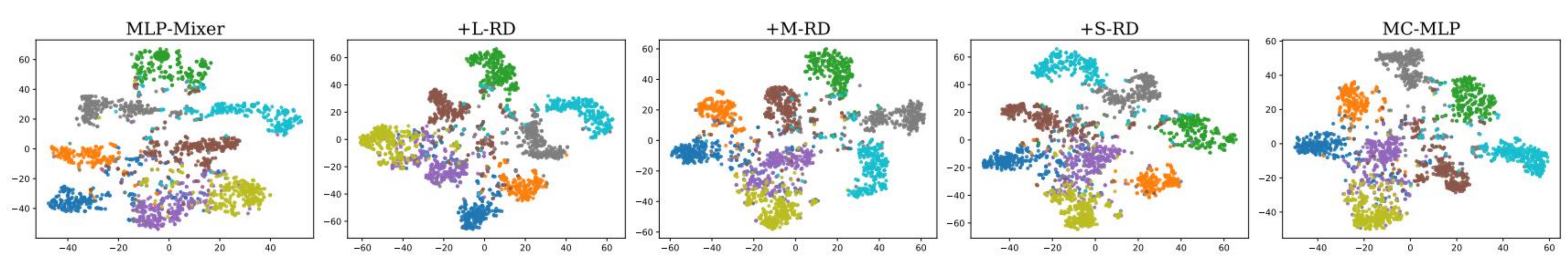
- ***Model Complexity Comparison:*** compare the most competitive method BTH and our MCMSH.

Method	Param.	FLOPs	Average Encoding Time
BTH	3.17M	0.05G	0.53ms
MCMSH	1.76M	0.05G	0.47ms

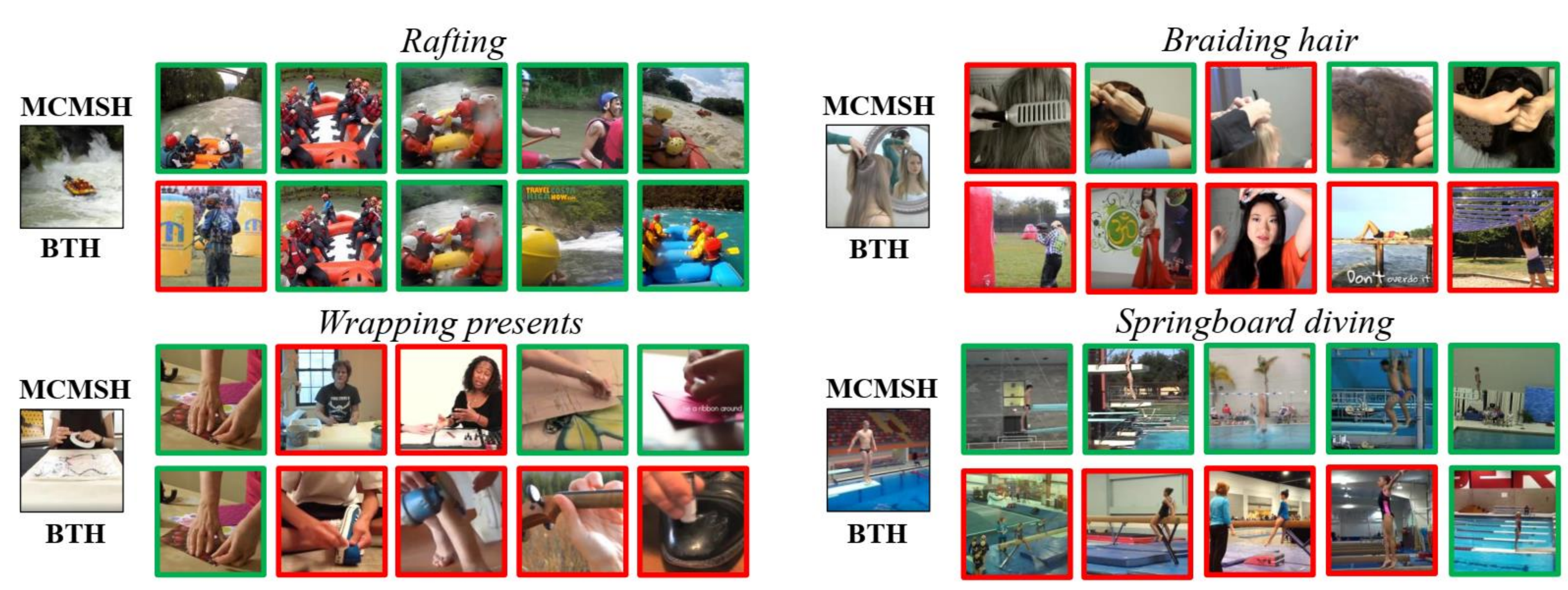
Table3: Comparison of parameters, FLOPs and average encoding time between BTH and MCMSH. The average encoding time is computed in the same platform.

Visualization

- ***Visualization of feature distributions w and w/o L/M/S-RD:***



- ***Retrieved result of MCMSH and BTH on ActivityNet dataset:***



Conclusion and Resources

- ***Conclusion:***

1. The three self-gating modules L/M/S-RD focus on different kinds of axial contexts to model multi-granular spatio-temporal interactions.
2. The three data structures are complementary to each other.
3. MCMSH achieves the effectiveness and efficiency compared with state-of-the-arts.

- ***Contact & Resources:***

haoyanbin@hotmail.com
duanjr@mail.ustc.edu.cn
zhanghaoinf@gmail.com
andrewzhu1216@gmail.com
pyzhou@ustc.edu.cn
xiangnanhe@gmail.com

GitHub

PDF Adobe