#0200

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

香港中文大學 The Chinese University of Hong Kong

ACM Multimedia 2022 Lisbon, Portugal | 10-14 October

Dual Contrastive Learning for Spatio-temporal Representation

Shuangrui Ding, Rui Qian, Hongkai Xiong*

• Abstract

➢ Task: Video Representation Learning

➢ Problem: Background Bias. As seen in Fig.1, sampling two clips in one video as positive pair leads to the similar scene but distinct motion, thus leading background bias.

➢ Solution: Decouple video into dynamic and static modality and formulate the dual contrastive framework.

• Method

➢ DCLR: **D**ual **C**ontrastive **L**earning for spatio-temporal **R**epresentation.

➢ Static-dynamic decoupling in data input:

- ➢ Static Frame.
- ➢ Frame Difference.
- ➢ Transform standard contrastive objective into dual form.

➢ Static-dynamic decoupling in feature space:

- ➢ Activation alignment constraint.
- ➢ Distill dynamic-/static-related features.

Fig1: An illustration for positive and negative pair in spatio-temporal contrastive learning.

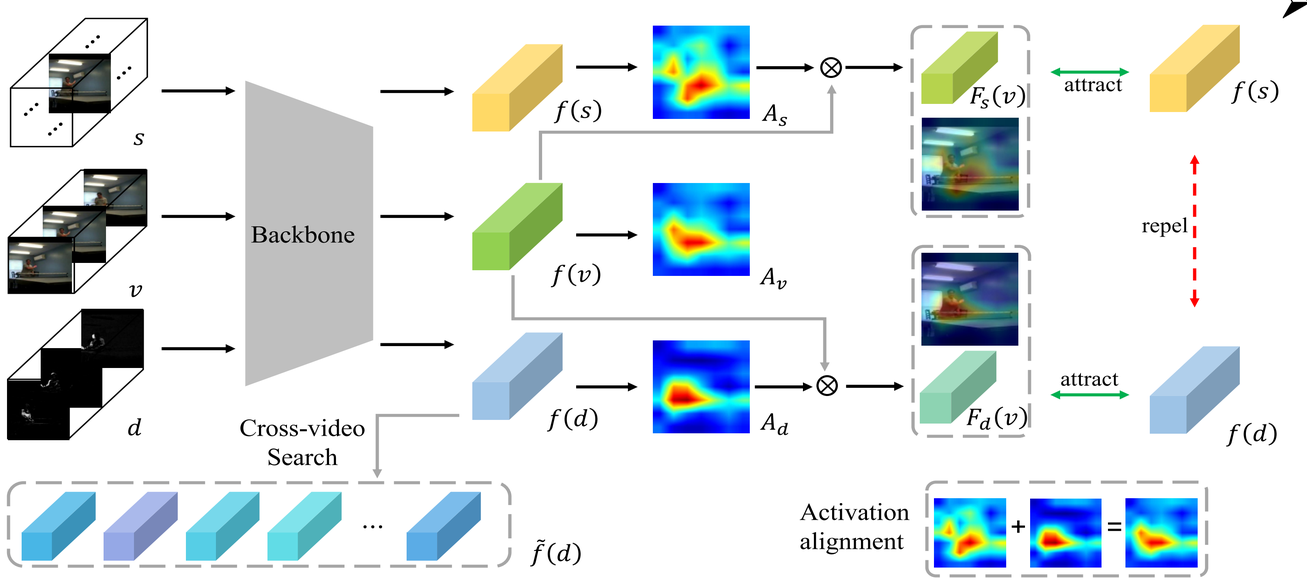

Fig2: An overview of the proposed method.

• Experiments

➢ SOTA results on UCF-101 and HMDB-51.

➢ SOTA results on Diving-48.

Method	Backbone	Pretrain Dataset	Frames	Res.	Freeze	UCF-101	HMDB-51
CCL [28]	R3D-18	Kinetics-400	16	112	✓	52.1	27.8
MemDPC [16]	R3D-34	Kinetics-400	40	224	✓	54.1	30.5
RSPNet [6]	R3D	Kinetics-400	16	112	✓	61.8	42.8
MLRep [41]	R3D	Kinetics-400	16	112	✓	63.2	33.4
FAME [9]	R(2+1)D	Kinetics-400	16	112	✓	72.2	42.2
DCLR(Ours)	R(2+1)D	Kinetics-400	16	112	✓	**72.3**	**46.4**
VCP [35]	R3D	UCF-101	16	112	✗	66.3	32.2
IIC [45]	C3D	UCF-101	16	112	✗	72.7	36.8
MLRep [41]	R3D	UCF-101	16	112	✗	76.2	41.1
TempTrans [24]	R(2+1)D	UCF-101	16	112	✗	81.6	46.4
DCLR(Ours)	R(2+1)D	UCF-101	16	112	✗	**82.3**	**50.1**
3DRotNet [25]	R3D	Kinetics-400	16	112	✗	62.9	33.7
Pace Prediction [53]	R(2+1)D	Kinetics-400	16	112	✗	77.1	36.6
MemDPC [16]	R3D	Kinetics-400	40	224	✗	78.1	41.2
Pace [53]	R(2+1)D	Kinetics-400	16	112	✗	77.1	36.6
VideoMoCo [40]	R(2+1)D	Kinetics-400	32	112	✗	78.7	49.2
MLRep [41]	R3D	Kinetics-400	16	112	✗	79.1	47.6
TempTrans [24]	R3D	Kinetics-400	16	112	✗	79.3	49.8
RSPNet [6]	R(2+1)D	Kinetics-400	16	112	✗	81.1	44.6
ASCNet [21]	R3D	Kinetics-400	16	112	✗	80.5	52.3
SRTC [65]	R(2+1)D	Kinetics-400	16	112	✗	82.0	51.2
DCLR(Ours)	R(2+1)D	Kinetics-400	16	112	✗	**83.3**	**52.7**

Tab1: Results on UCF-101 and HMDB-51.

Method	Pretrain Dataset	Res.	Top-1
Random Init.	-	-	50.7
BE [51]	UCF-101	224	58.8
FAME [9]	UCF-101	224	67.8
DCLR(Ours)	UCF-101	112	72.7
BE [51]	Kinectics-400	224	62.4
FAME [9]	Kinectics-400	224	72.9
DCLR(Ours)	Kinectics-400	112	75.1

Tab2: Results on Diving-48.

• Take-away

➢ Spatio-temporal contrastive learning exists background bias.

➢ Decoupling static and dynamic cues in both data input and feature space can resist the background shortcut.

Want to know more?
Please refer to our paper!