TCLR: Temporal Contrastive Learning for Video Representation

Ishan Dave

Rohit Gupta

Mamshad Rizve

Mubarak Shah

Center for Research in Computer Vision
University of Central Florida, Orlando, USA

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Overview of the presentation

- Motivation for SSL in video
- Instance Contrastive Self-sup learning for Videos
- Inspiration for TCLR
- Temporal Contrastive Learning for video Representation (TCLR)
- Method
- Experiments
- Analysis

Motivation for Self-supervised learning

- 3D CNN can learn spatio-temporal features and outperform 2D CNNs
- Data hungry!
- Requirement of pretraining weights in video models:

  On UCF101, from scratch, RGB modality:
  - Standard models: ~60%
  - SOTA architectures (including recent Video transformers) with lot of data augmentations: ~70%

  With Kinetics-400 pretraining: ~96%
- Annotating video is very costly compared to the annotating image
  - Kinetics-400 has data of 28 days!
- Solution: Learning from unlabeled data!

Some video examples from UCF101 Action Recognition Dataset:

Pushups

Long Jump

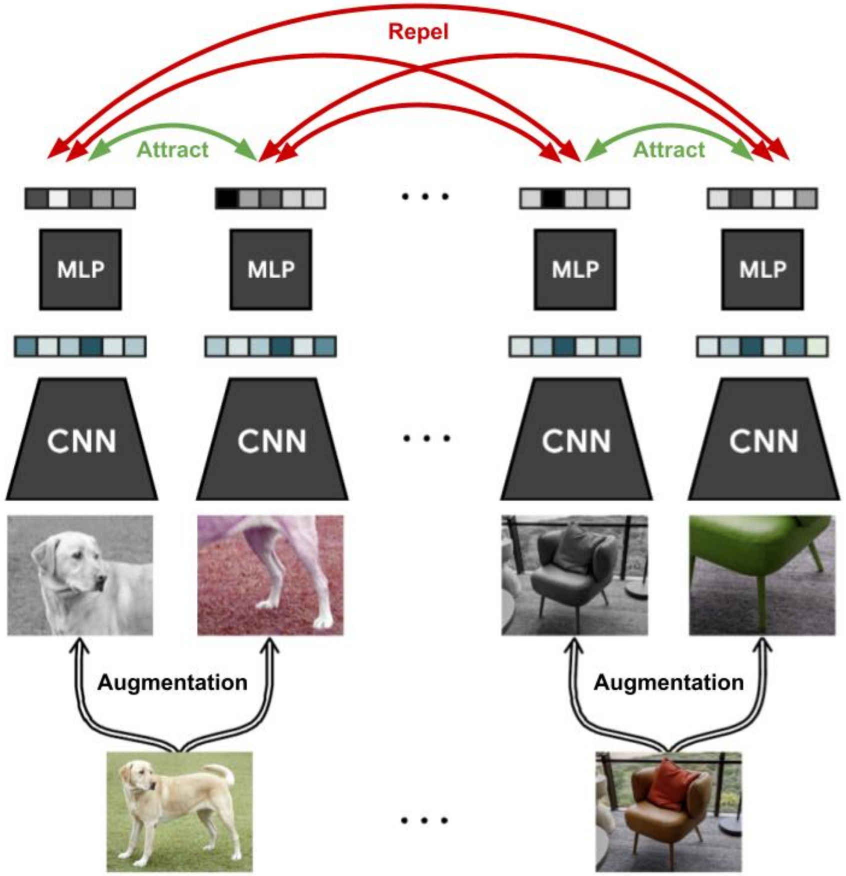
Ice-dancing

Pizza-tossing

Contrastive Self-supervised Learning (CSL)

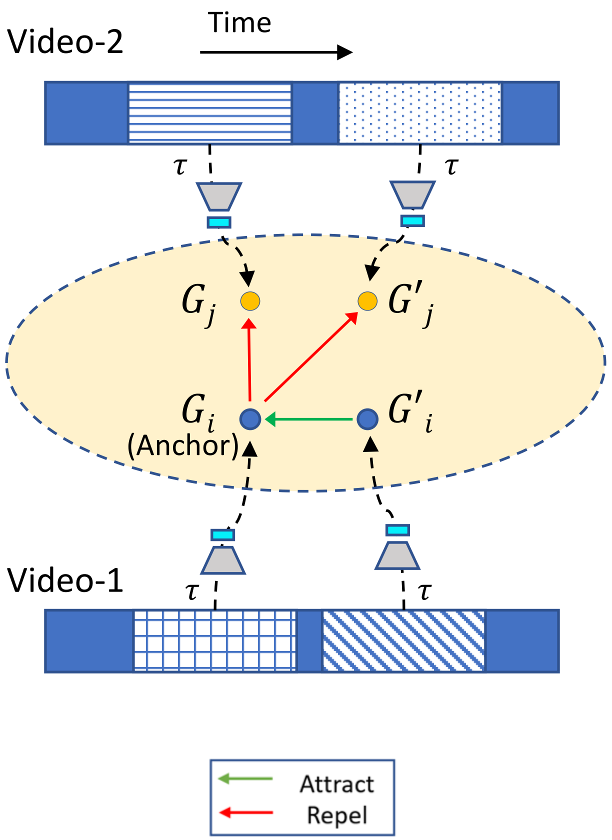
Maximize the agreement between different view (augmented version) of an image while maximizing the disagreement between views of different images

SimCLR



https://github.com/google-research/simclr

CSL in Videos



- SimCLR like extension is simple-yet-effective for VideoSSL
- *Instance discrimination* by maximizing agreement b/w clips of the *same video*

$$\mathcal{L}_{IC}^{i} = -\log \frac{h\left(G_i, G_i'\right)}{\sum_{j=1}^{N}[\mathbb{1}_{[j\neq i]} h(G_i, G_j) + h(G_i, G_j')]}$$

Where,

h is softmax of cosine similarity between embedding *u* and *v*, with temperature τ

$$h(u,v) \ = \ \exp\left(u^T v / (\|u\| \|v\| \tau)\right)$$

$\mathbb{1}_{[j\neq i]}$ is indicator function which is 0 iff i=j, else 1

Temporal Invariance in *Instance Contrastive* Loss

In inference of any video understanding task, we take average prediction of multiple clips of a video

- Instance Contrastive Loss attracts all clips from the same video to similar representations, i.e., it enforces ***temporal invariance***
- Due to temporal invariance IC does not gain from multiple clips
- Hence, to enforce *temporal diversity* in the learned features we introduce *Temporal Contrastive Learning framework*

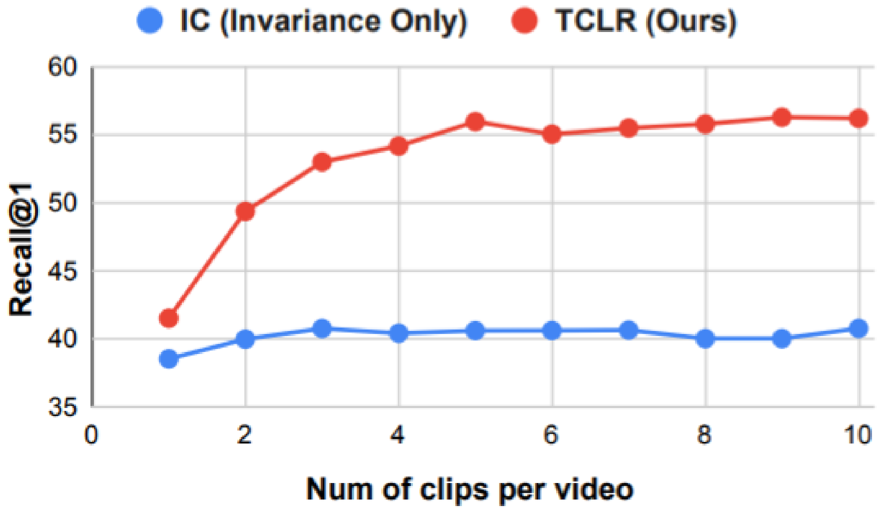
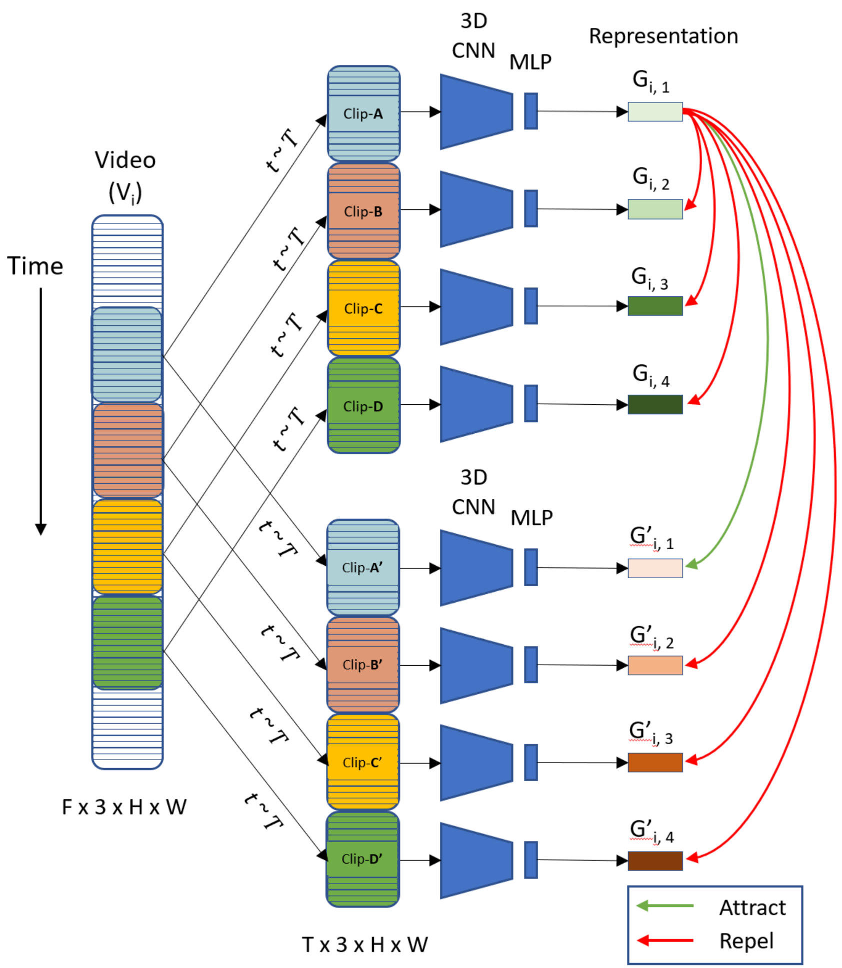


Fig 1: NN Retrieval with increasing number of clips per video

TCLR Framework

Goal: Encourage temporal diversity at 2 *temporal aggregation* steps:
1. Clip level Averaging
2. Temporal Pooling of Feature Map

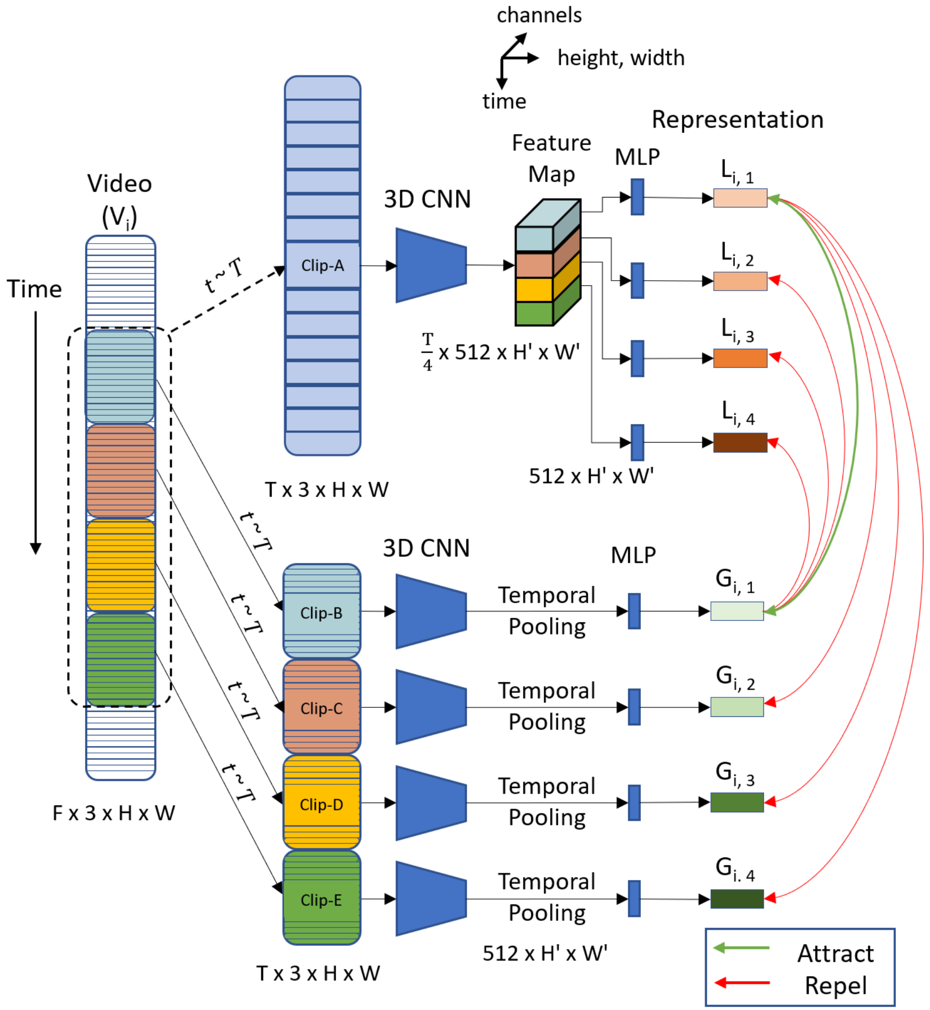
Local-Local Temporal Contrastive Loss ($\mathcal{L}_{LL}$)



Enforce diversity at <u>clip level</u> by contrasting clips from the same video

$$\mathcal{L}_{LL}^{i} = -\sum_{p=1}^{N_T} \log \frac{h\left(G_{i,p}, G'_{i,p}\right)}{\sum_{q=1}^{N_T}[\mathbb{1}_{[q\neq p]} h(G_{i,p}, G_{i,q}) + h(G_{i,p}, G'_{i,q})]}$$

Global-Local Temporal Contrastive Loss ($\mathcal{L}_{GL}$)



Enforce temporal diversity at <u>feature level</u> by contrasting global clips feature map with pooled local features

$$\mathcal{L}^i_{GL_k} = \log \frac{h\left(L_{i,k}, G_{i,k}\right)}{\sum_{q=1}^{N_T} h(L_{i,k}, G_{i,q})} + \log \frac{h\left(G_{i,k}, L_{i,k}\right)}{\sum_{q=1}^{N_T} h(G_{i,k}, L_{i,q})}$$

Summing over all timestamps N_T

$$\mathcal{L}^i_{GL} = -\sum_{k=1}^{N_T} \mathcal{L}^i_{GL_k}$$

Experimental Setting

- TCLR self-supervised pre-training
  - No Labels used
  - UCF-101 or Kinetics-400 videos
- Architectures Tested: 3D ResNet, R-(2+1)-D, C3D
- Downstream Tasks:

1. Full Finetuning on Downstream Action Recognition Task
    - UCF101, HMDB51, Diving48
2. Nearest Neighbor Retrieval
    - Downstream task with no finetuning
3. Finetuning with Limited Labels: 1%, 10%, 20%, 50%
4. Linear Evaluation
    - Only Finetune Final Classifier Layer

Results on Action Recognition (R3D-18)

Method	Publication Venue	Frames × Resolution	Pretraining → Finetuning Datasets			
			UCF101 → UCF101	UCF101 → HMDB51	Kinetics → UCF101	Kinetics → HMDB51
Pace Pred [58]	ECCV 20	16 × 112	65.0	-	-	-
VCP [37]	AAAI 20	16 × 112	66.0	31.5	-	-
PRP [65]	CVPR 20	16 × 112	66.5	29.7	-	-
MemDPC [22]	ECCV 20	40 × 224	69.2	-	-	-
TCP [36]	WACV 20	- × 224	64.8	34.7	70.5	41.1
VIE [68]	CVPR 20	16 × 112	-	-	72.3	44.8
UnsupIDT [51]	ECCVw 20	16 × 112	-	-	73.0	41.6
CSJ [5]	-	16 × 224	70.4	36.0	76.2	46.7
BFP [8]	WACV 21	40 × 128	63.6	-	66.4	45.3
IIC (RGB) [49]	ACMMM 20	16 × 112	61.6	-	-	-
CVRL (Reproduced) [44]	CVPR 21	16 × 112	75.8	44.6	-	-
SSTL [45]	-	16 × 112	-	-	79.1	*49.7*
VTHCL [63]	-	8 × 224	-	-	80.6	48.6
VideoMoCo [42]	CVPR 21	16 × 112	-	-	74.1	43.6
RSPNet [13]	AAAI 20	16 × 112	-	-	74.3	41.8
Temp Trans [26]	ECCV 20	16 × 112	*77.3*	*47.5*	79.3*	49.8*
TaCo [7]	-	16 × 224	-	-	*81.4*	45.4
MFO	ICCV-21	16 x 112	-	-	*79.1*	47.6
TCLR	**-**	**16 × 112**	**82.4**	**52.9**	**84.1**	**53.6**

Results on Nearest Neighbor Retrieval

Method	UCF101 / HMDB51 Results			
	R@1	**R@5**	**R@10**	**R@20**
VCOP [61]	14.1 / 7.6	30.3 / 22.9	40.4/34.4	51.1 / 48.8
VCP [37]	18.6 / 7.6	33.6 / 24.4	42.5 / 36.6	53.5 / 53.6
Pace Pred [58]	23.8 / 9.6	38.1 / 26.9	46.4 / 41.1	56.6 / 56.1
Var. PSP [15]	24.6 / 10.3	41.9 / 26.6	51.3 / 38.8	62.7 / 51.6
Temp Trans [26]	26.1 / -	48.5 / -	59.1 / -	69.6 / -
CSJ [5]	21.5 / -	40.5 /-	53.2 / -	64.9 / -
MemDPC [22]	20.2 / 7.7	40.4 / 25.7	52.4 / 40.6	64.7 / 57.7
RSPNet [13]	41.1 / -	59.4 / -	68.4 / -	*77.8* / -
STS [57]	38.3 / 18.0	*59.9 / 37.2*	*68.9 / 50.7*	77.2 / *64.8*
SSTL [45]	*44.5 / 21.8*	57.4 / 35.7	63.5 / 44.2	70.0 / 57.7
TCLR	**56.2 / 22.8**	**72.2 / 45.4**	**79.0 / 57.8**	**85.3 / 73.1**

Results on Limited Label Classification

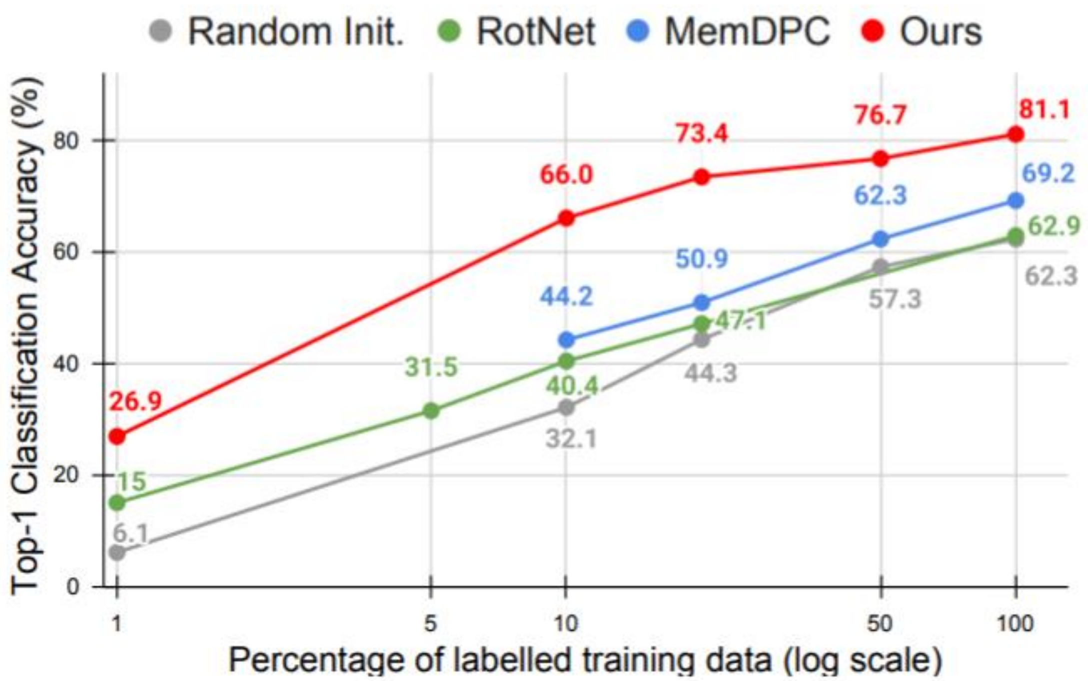


Fig: Evaluating Label Efficiency using Limited Label Learning on UCF101 (split-1) action classification task

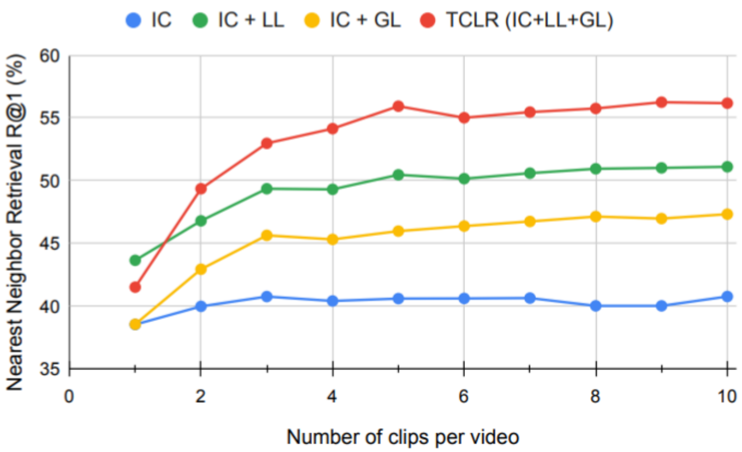
TCLR significantly improves the label efficiency of video representation learning and is able to beat the fully supervised baseline model with only <u>*10% of labelled data*</u>.

Ablation Experiments

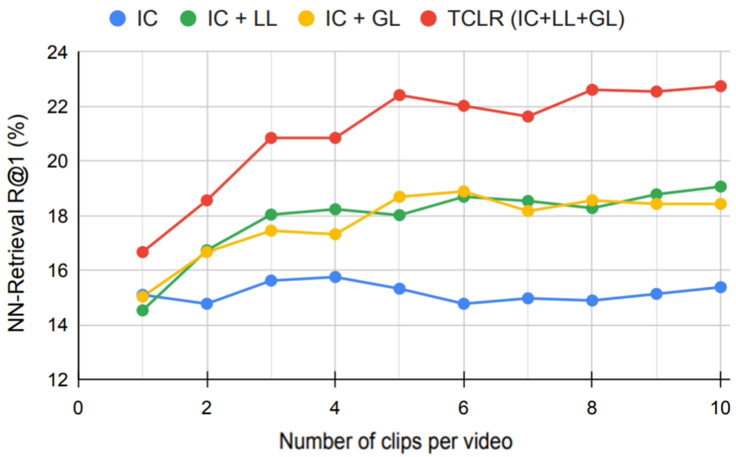
Contrastive Losses			Top-1 Classification Accuracy			R@1 Retrieval
$\mathcal{L}_{IC}$	$\mathcal{L}_{LL}$	$\mathcal{L}_{GL}$	Linear Eval UCF101	Finetune UCF101	Transfer HMDB51	UCF101
Random Init			17.15	62.39	26.95	8.21
	✓	✓	23.39	74.29	47.35	14.17
✓			54.58	71.31	38.32	40.76
✓	✓		62.70 +8%	77.70 +6%	49.77 +11%	51.10 +10%
✓		✓	64.55 +10%	76.30 +5%	47.87 +10%	47.32 +7%
✓	✓	✓	69.91 +15%	82.40 +11%	52.80 +14%	56.17 +15%

Temporal Diversity helps in various downstream tasks

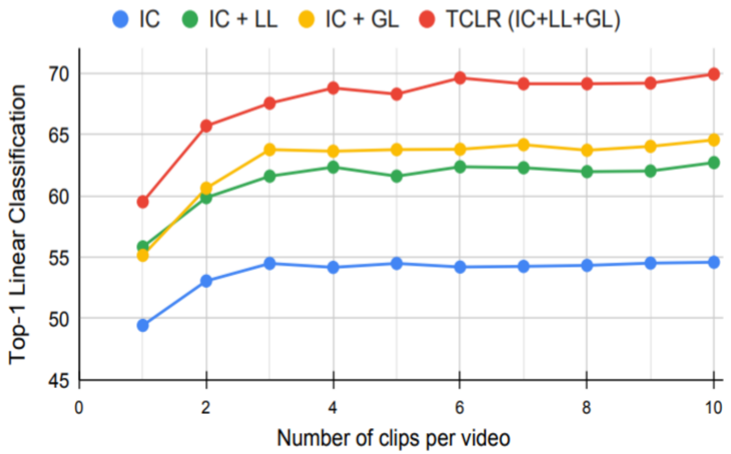
Unlike standard instance contrastive loss, TCLR can benefit from using multiple clips during inference



NN Retrieval- UCF101

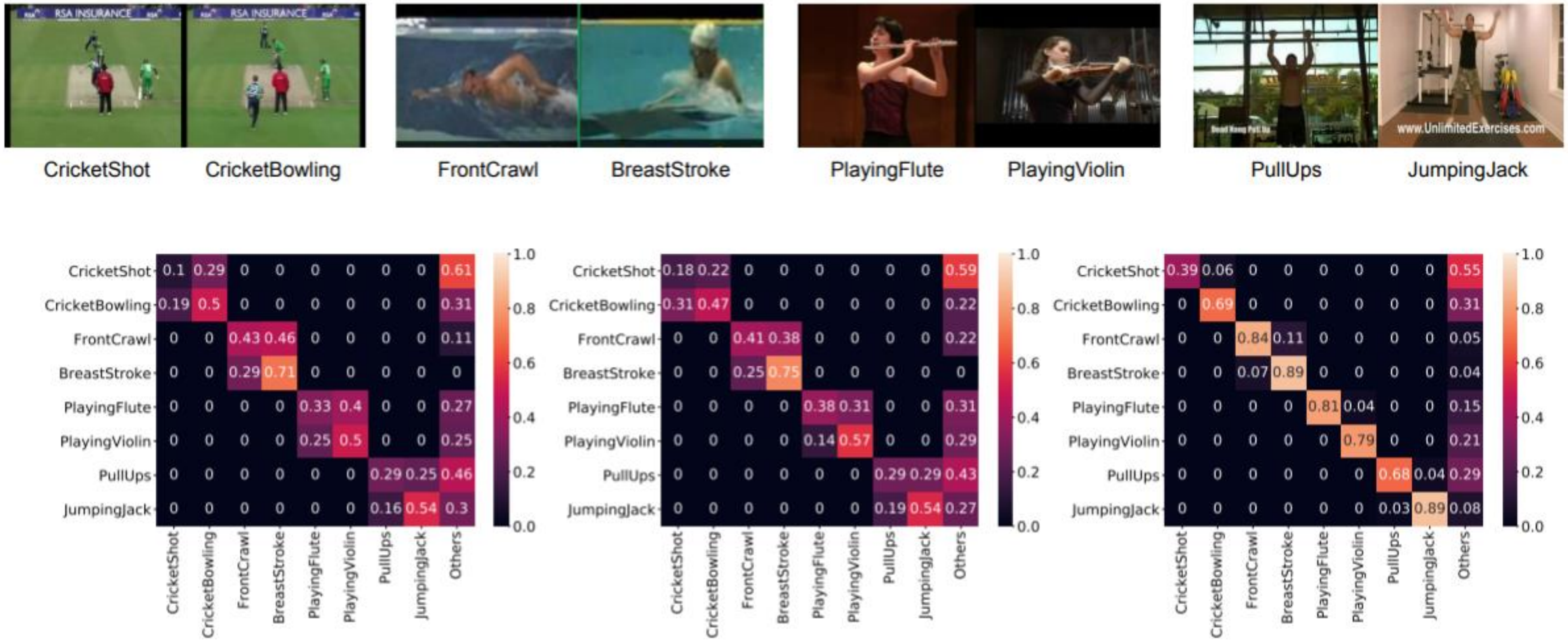


NN Retrieval- HMDB51



Linear Classification

Temporal diversity helps distinguish visually similar classes



(a) Scratch (b) IC pretraining (c) TCLR pretraining

Conclusion

- Proposed 2 novel temporal contrastive losses
- SoTA over various video understanding tasks
- Temporal diversity helps in VideoSSL

Thank you!

Question? ishandave@knights.ucf.edu

New results and link to the repo will be released soon on arxiv:

https://arxiv.org/abs/2101.07974