Semi-Supervised Action Recognition with Temporal Contrastive Learning

Ankit Singh[1*], Omprakash Chakraborty[2*], Ashutosh Varshney[2], Rameswar Panda[3], Rogerio Feris[3], Kate Saenko[3,4] and Abir Das[2]

[1] IIT Madras, [2] IIT Kharagpur, [3] MIT-IBM Watson AI Lab, [4] Boston University

Problem Definition and Contribution

Goal: Leverage freely available unlabeled videos along with limited labeled videos for action recognition

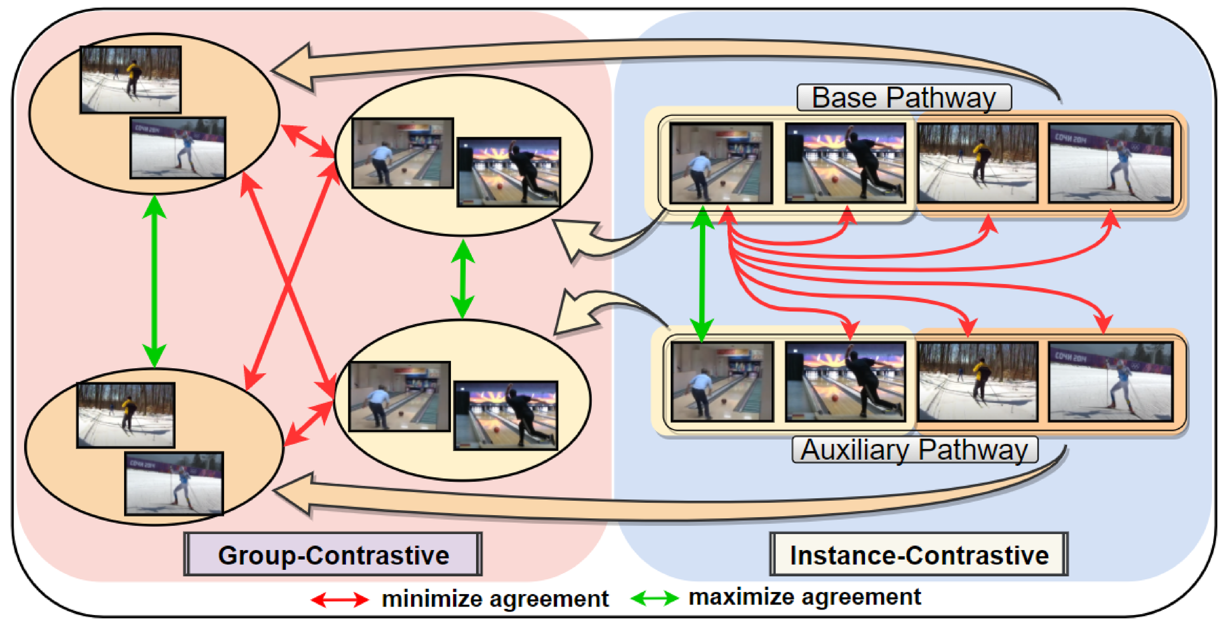
Motivations:

- Annotating videos is expensive and time consuming
- Semi Supervised Action Recognition in Videos is still under explored
- Naive extensions of image based approaches to videos yield sub-optimal performance

Key Contributions:

- we treat the time axis in unlabeled videos specially, by processing them at two different speeds and propose a **T**emporal **C**ontrastive **L**earning (TCL) framework for semi-supervised action recognition
- Introduction of novel group contrastive loss
- State of the art results on large scale video datasets

Group Contrastive Loss

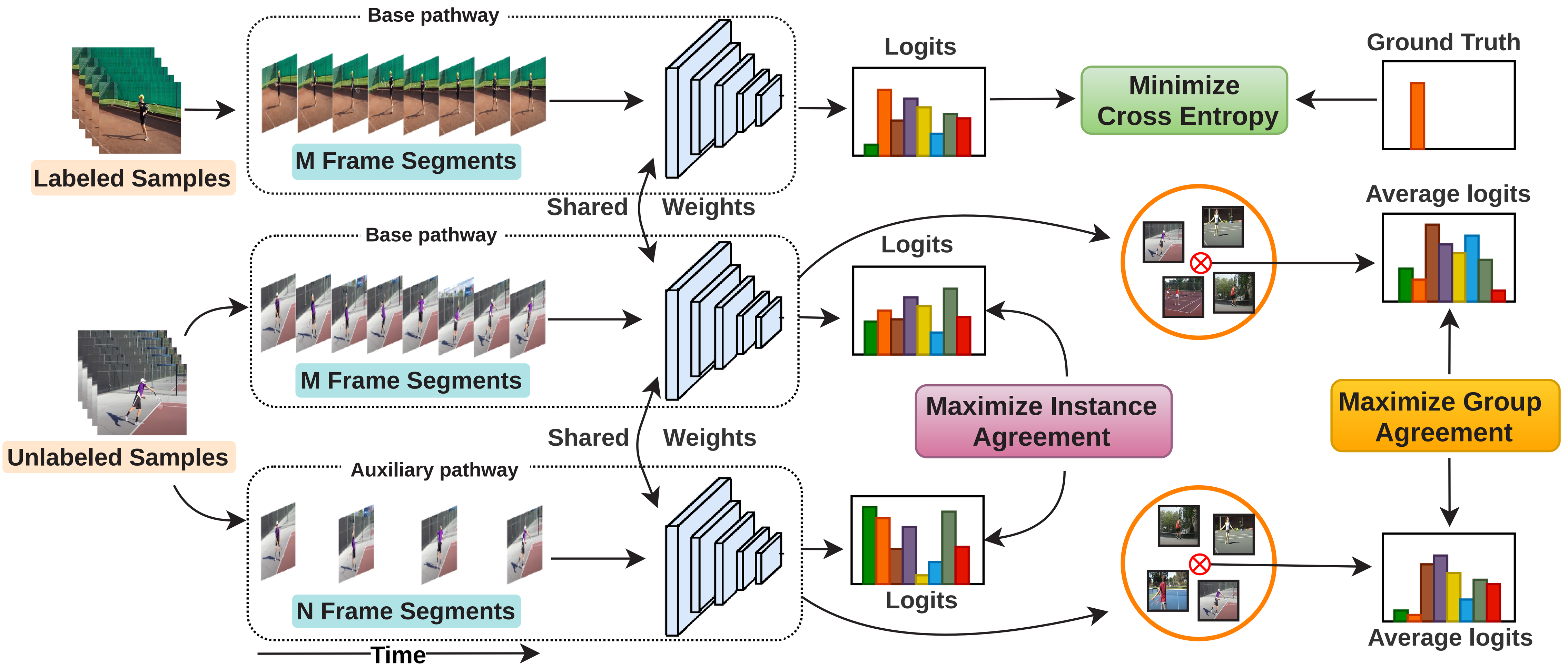


Main idea: Directly applying contrastive loss between different video instances in absence of class-labels does not take the high level action semantics into account.

$$\mathcal{L}_{gc}(R_f^l, R_s^l) = -\log \frac{h(R_f^l, R_s^l)}{h(R_f^l, R_s^l) + \sum\limits_{\substack{m=1 \\ p \in \{s,f\}}}^{C} \mathbb{1}_{\{m \neq l\}} h(R_f^l, R_p^m)}$$

Key Idea

Contrasting video representations between base and auxiliary pathways exploits temporal information in videos to learn rich feature representation



Loss functions for TCL framework:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \gamma * \mathcal{L}_{ic} + \beta * \mathcal{L}_{gc}$$

- $\mathcal{L}_{sup}$: Cross Entropy Loss
- $\mathcal{L}_{gc}$: Group Contrastive Loss
- γ: Weight of Instance Contrastive loss
- β: Weight of Group Contrastive loss

- $\mathcal{L}_{ic}$: Instance Contrastive Loss

$$\mathcal{L}_{ic}(U_f^i, U_s^i) = -\log \frac{h(g(U_f^i), g(U_s^i))}{h(g(U_f^i), g(U_s^i)) + \sum\limits_{\substack{k=1 \\ p \in \{s,f\}}}^{B} \mathbb{1}_{\{k \neq i\}} h(g(U_f^i), g(U_p^k))}$$

Experiments & Results

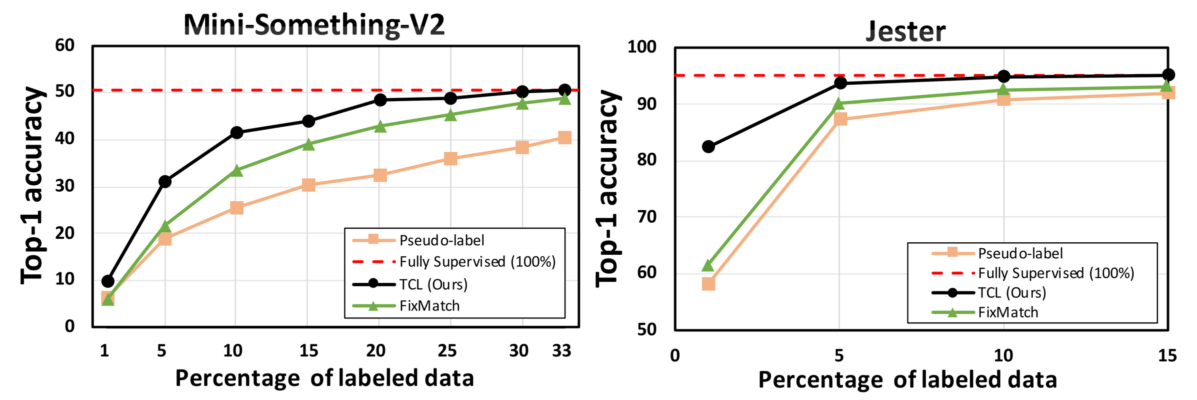
Performance Comparison in Mini-Something-V2:

Approach	ResNet-18 1%	ResNet-18 5%	ResNet-18 10%	ResNet-50 1%	ResNet-50 5%	ResNet-50 10%
Supervised (8f)	5.98±0.68	17.26±1.17	24.67±0.68	5.69±0.51	16.68±0.25	25.92±0.53
Pseudo-Label (`ICMLW'13`)	6.46±0.32	18.76±0.77	25.67±0.45	6.66±0.89	18.77±1.18	28.85±0.91
Mean Teacher (`NeurIPS'17`)	7.33±1.13	20.23±1.59	30.15±0.42	6.82±0.18	21.80±1.54	32.12±2.37
S4L (`ICCV'19`)	7.18±0.97	18.58±1.05	26.04±1.89	6.87±1.29	17.73±0.26	27.84±0.75
MixMatch (`NeurIPS'19`)	7.45±1.01	18.63±0.99	25.78±1.01	6.48±0.83	17.77±0.12	27.03±1.66
FixMatch (`NeurIPS'20`)	6.04±0.44	21.67±0.18	33.38±1.58	6.54±0.71	25.34±2.03	37.44±1.31
TCL (Ours)	7.79±0.57	29.81±0.77	38.61±0.91	7.54±0.32	27.22±1.86	40.70±0.42
TCL w/ Finetuning	8.65±0.76	30.55±1.36	40.06±1.14	8.56±0.31	28.84±1.22	41.68±0.56
TCL w/ Pretraining & Finetuning	9.91±1.84	30.97±0.07	41.55±0.47	9.19±0.43	29.85±1.76	41.33±1.07

Semi-supervised action recognition under domainshift (Charades-Ego) :

Approach	10% ($\rho = 1$)	10% ($\rho = 0.5$)	10% ($\rho = 0$)
Supervised (8f)	17.53 ±0.49		
Pseudo-Label (`ICMLW'13`)	18.00±0.16	17.87±0.14	17.79±0.33
FixMatch (`NeurIPS'20`)	18.02±0.31	18.00±0.29	17.96±0.25
TCL (Ours)	19.13±0.37	18.95±0.17	18.50±0.95
TCL w/ Finetuning	19.68±0.37	19.58±0.31	19.56±0.82

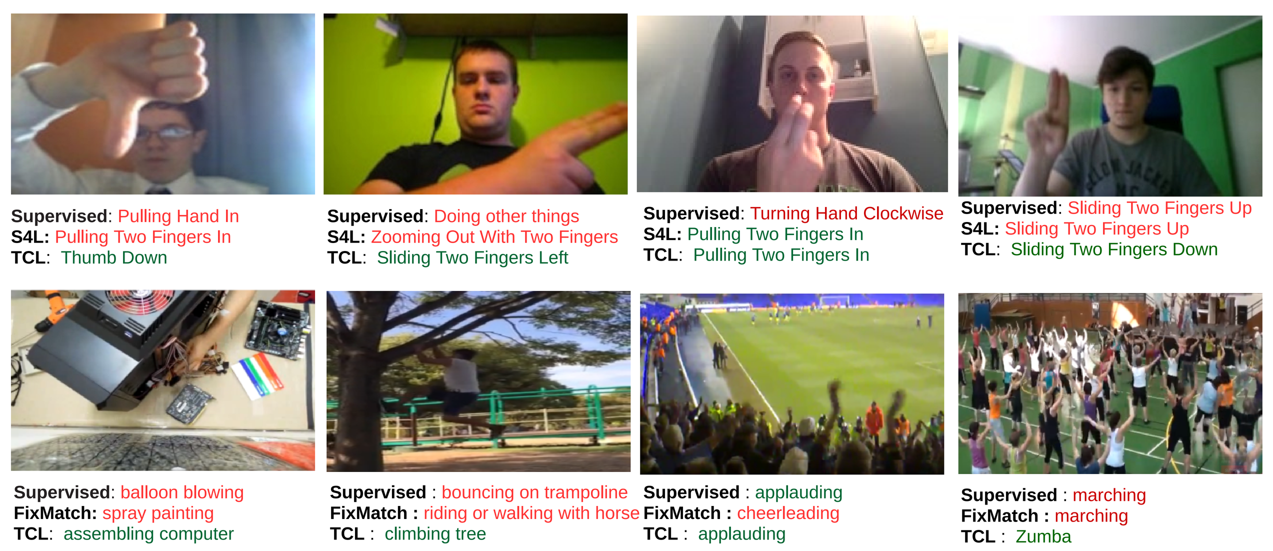
Comparison of TCL with Pseudo-Label and FixMatch



Performance Comparison in Jester and Kinetics-400:

Approach	Jester 1%	Jester 5%	Jester 10%	Kinetics-400 1%	Kinetics-400 5%
Supervised (8f)	52.55±4.36	85.22±0.61	90.45±0.33	6.17±0.32	20.50±0.23
Pseudo-Label (`ICMLW'13`)	57.99±3.70	87.47±0.64	90.96±0.48	6.32±0.19	20.81±0.86
Mean Teacher (`NeurIPS'17`)	56.68±1.46	88.80±0.44	92.07±0.03	6.80±0.42	22.98±0.43
S4L (`ICCV'19`)	64.98±2.70	87.23±0.15	90.81±0.32	6.32±0.38	23.33±0.89
MixMatch (`NeurIPS'19`)	58.46±3.26	89.09±0.21	92.06±0.46	6.97±0.48	21.89±0.22
FixMatch (`NeurIPS'20`)	61.50±0.77	90.20±0.35	92.62±0.60	6.38±0.38	25.65±0.28
TCL (Ours)	75.21±4.48	93.29±0.24	94.64±0.21	7.69±0.21	30.28±0.13
TCL w/ Finetuning	77.25±4.02	93.53±0.15	94.74±0.25	8.45±0.25	31.50±0.23
TCL w/ Pretraining & Finetuning	82.55±1.94	93.73±0.25	94.93±0.02	11.56±0.22	31.91±0.46

Qualitative Examples



Comparison with self-supervised methods:

Self-Supervised Approach	Top-1 Accuracy
Odd-One-Out Networks	19.56
Memory-augmented Dense Predictive Coding	18.67
Video Clip Order Prediction	23.93
TCL (Ours)	29.81±0.77

Project Webpage:
`https://cvir.github.io/TCL/`

Code & Dataset & Model

* denotes equal contribution