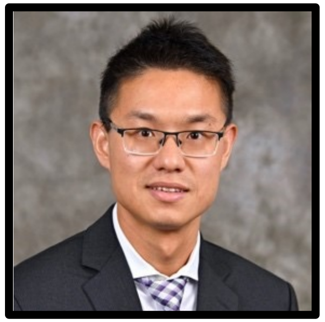
TimeBalance: Temporally-Invariant and Temporally-Distinctive Video Representations for Semi-Supervised Action Recognition



Ishan Dave

Mamshad Rizve

Chen Chen

Mubarak Shah

The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2023

Poster: TUE-AM-224

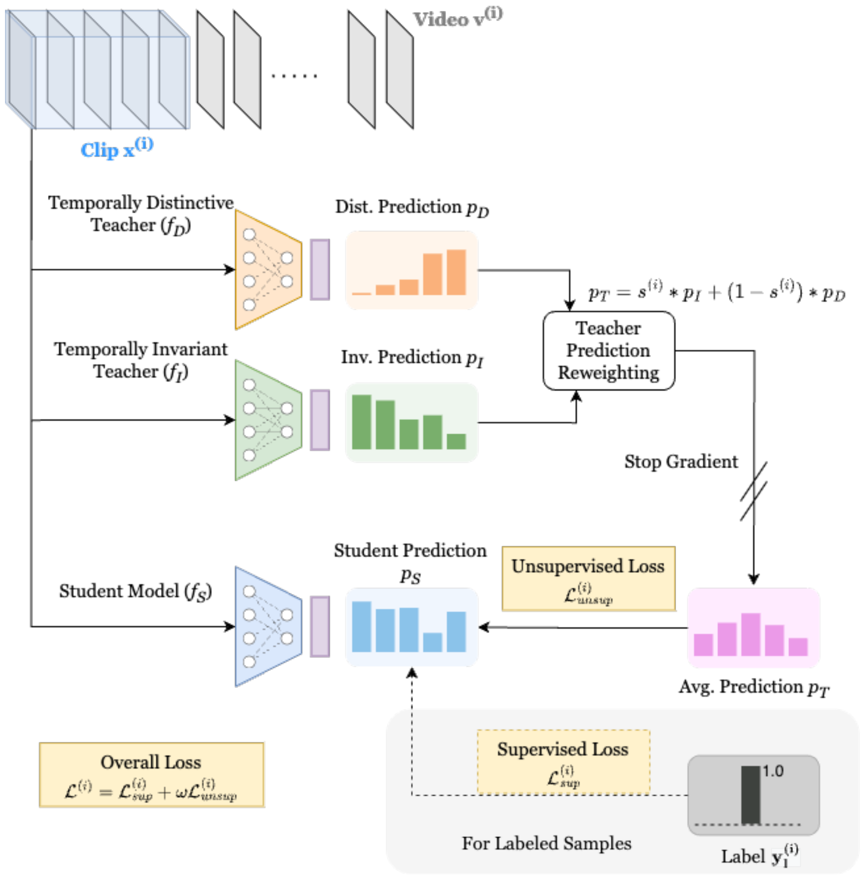
Center for Research in Computer Vision

University of Central Florida, Orlando, USA

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Quick preview-TimeBalance Framework


Video $\mathbf{v}^{(\mathbf{i})}$
Clip $\mathbf{x}^{(\mathbf{i})}$
Temporally Distinctive Teacher (f_D)
Dist. Prediction p_D
$p_T = s^{(i)} * p_I + (1 - s^{(i)}) * p_D$
Teacher Prediction Reweighting
Temporally Invariant Teacher (f_I)
Inv. Prediction p_I
Stop Gradient
Student Prediction p_S
Unsupervised Loss $\mathcal{L}^{(i)}_{unsup}$
Student Model (f_S)
Avg. Prediction p_T
Supervised Loss $\mathcal{L}^{(i)}_{sup}$
1.0
Overall Loss $\mathcal{L}^{(i)} = \mathcal{L}^{(i)}_{sup} + \omega\mathcal{L}^{(i)}_{unsup}$
For Labeled Samples
Label $\mathbf{y}^{(\mathbf{i})}_{\mathbf{l}}$

Dual teacher- Student framework for Semi-supervised Action Recognition

Teachers are pretrained with different types of *self-supervised video representations*: *temporally-invariant* and *temporally-distinctive* representations

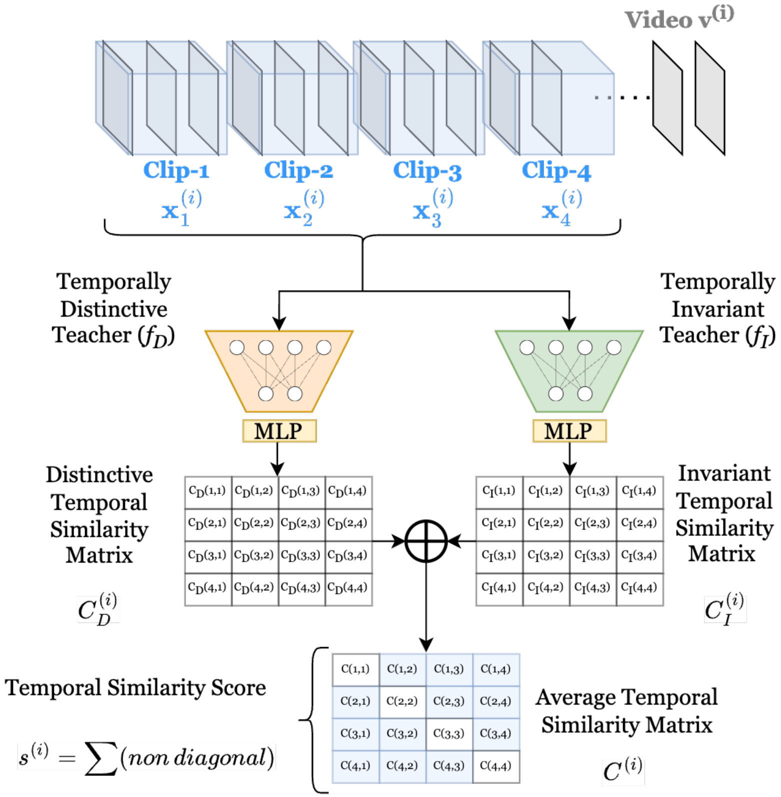
Input video clip: processed by both teachers

Predictions combined: reweighting strategy

Knowledge distillation: Provides unlabeled supervision to the student

Labeled supervision: Provided when labels are available

Quick preview- Reweighting strategy


Video $\mathbf{v}^{(i)}$
Clip-1
$\mathbf{x}_1^{(i)}$
Clip-2
$\mathbf{x}_2^{(i)}$
Clip-3
$\mathbf{x}_3^{(i)}$
Clip-4
$\mathbf{x}_4^{(i)}$
Temporally Distinctive Teacher (f_D)
Temporally Invariant Teacher (f_I)
MLP
MLP
Distinctive Temporal Similarity Matrix
$C_D^{(i)}$
Invariant Temporal Similarity Matrix
$C_I^{(i)}$
Temporal Similarity Score
$s^{(i)} = \sum(non\ diagonal)$
Average Temporal Similarity Matrix
$C^{(i)}$

$C_D(1,1)$	$C_D(1,2)$	$C_D(1,3)$	$C_D(1,4)$
$C_D(2,1)$	$C_D(2,2)$	$C_D(2,3)$	$C_D(2,4)$
$C_D(3,1)$	$C_D(3,2)$	$C_D(3,3)$	$C_D(3,4)$
$C_D(4,1)$	$C_D(4,2)$	$C_D(4,3)$	$C_D(4,4)$

$C_I(1,1)$	$C_I(1,2)$	$C_I(1,3)$	$C_I(1,4)$
$C_I(2,1)$	$C_I(2,2)$	$C_I(2,3)$	$C_I(2,4)$
$C_I(3,1)$	$C_I(3,2)$	$C_I(3,3)$	$C_I(3,4)$
$C_I(4,1)$	$C_I(4,2)$	$C_I(4,3)$	$C_I(4,4)$

$C(1,1)$	$C(1,2)$	$C(1,3)$	$C(1,4)$
$C(2,1)$	$C(2,2)$	$C(2,3)$	$C(2,4)$
$C(3,1)$	$C(3,2)$	$C(3,3)$	$C(3,4)$
$C(4,1)$	$C(4,2)$	$C(4,3)$	$C(4,4)$

- Clips representation similarity (↑) → f_I(↑)
- Clips representation similarity (↓) → f_D (↑)

Quick preview- Results

Method	Venue	Backbone	Params (M)	Input	#F	UCF101					HMDB51			Kinetics400	
						1%	5%	10%	20%	50%	40%	50%	60%	1%	10%
PL	ICML'13	3D-ResNet18	13.5	V	16	-	17.6	24.7	37	47.5	27.3	32.4	33.5	-	-
MT	NeuRIPS'17	3D-ResNet18	13.5	V	16	-	17.5	25.6	36.3	45.8	27.2	30.4	32.2	-	-
S4L	ICCV'19	3D-ResNet18	13.5	V	16	-	22.7	29.1	37.7	47.9	29.8	31	35.6	-	-
UPS	ICLR'21	3D-ResNet18	13.5	V	16	-	-	-	39.4	50.2	-	-	-	-	-
SD	ICCV'19	3D-ResNet18	13.5	V	16	-	31.2	40.7	45.4	53.9	32.6	35.1	36.3	-	-
MT+SD	WACV'21	3D-ResNet18	13.5	V	16	-	30.3	40.5	45.5	53	32.3	33.6	35.7	-	-
VideoSemi	WACV'21	3D-ResNet18	13.5	V	16	-	32.4	42	48.7	54.3	32.7	36.2	37	-	-
TG-FixMatch	CVPR'21	3D-ResNet18	13.5	V	8	-	44.8	62.4	76.1	79.3	46.5	48.4	49.7	9.8	43.8
TCL	CVPR'21	TSM-R18	-	V	8	-	-	-	-	-	-	-	-	11.6	-
MvPL	ICCV'21	3D-ResNet18	13.5	VFG	8	-	41.2	55.5	64.7	65.6	30.5	33.9	35.8	5	36.9
CMPL	CVPR'22	3D-ResNet18	13.5	V	8	23.8	-	67.6	-	-	-	-	-	16.5	53.7
TACL	TSVT'22	3D-ResNet18	13.5	V	16	-	35.6	50.9	56.1	65.8	34.6	37.2	39.5	-	-
TACL	TSVT'22	3D-ResNet18	13.5	V	16	-	43.7	55.6	59.2	67.2	38.7	40.2	41.7	-	-
3DRotNet	Arxiv'19	3D-ResNet18	13.5	V	16	15	31.5	40.4	47.1	-	-	-	-	-	-
MemDPC	ECCV'20	3D-ResNet18	13.5	V	16	-	-	44.2	50.9	62.3	-	-	-	-	-
MotionFit	ICCV'21	3D-ResNet18	13.5	VF	16	-	-	-	57.7	59	-	-	-	-	-
TCLR	CVIU'22	3D-ResNet18	13.5	V	16	26.9	-	66.1	73.4	76.7	-	-	-	-	-
TimeBalance	CVPR'23	3D-ResNet18	13.5	V	8	**29.1**	**47.9**	**69.8**	**79.1**	**83.3**	**49.8**	**51.4**	**53.1**	**17.1**	**54.9**
ActorCM	Arxiv'21	R(2+1)D-34	33.3	V	8	-	27	40.2	51.7	59.9	32.9	38.2	38.9	-	-
ActorCM	Arxiv'21	R(2+1)D-34	33.3	V	8	-	45.1	53	57.4	64.7	35.7	39.5	40.8	-	-
FixMatch	NeuRIPS'20	SlowFast-R50	60	V	8	16.1	-	55.1	-	-	-	-	-	10.1	49.4
MvPL	ICCV'21	3D-ResNet50	31.8	VFG	8	22.8	-	80.5	-	-	-	-	-	17	58.2
CMPL	CVPR'22	3D-ResNet50	31.8	V	8	25.1	-	79.1	-	-	-	-	-	17.6	58.4
TimeBalance	CVPR'23	3D-ResNet50	31.8	V	8	**30.1**	**53.5**	**81.1**	**83.3**	**85**	**52.6**	**53.9**	**54.5**	**19.6**	**61.2**

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

UCF Center for Research in Computer Vision

Overview of the details

- Motivation
- Temporal-Invariance vs Distinctiveness in video representations
- TimeBalance framework
- Temporal Similarity Based Reweighting
- Experiments
- Conclusion

Semi-supervised Action Recognition

- Semi-supervised learning deals with enhancing performance on smaller labeled set leveraging large amount of unlabeled data.
- Semi supervised learning is more crucial for video data due to **higher cost** of video annotations
- Compared to images, videos provide additional **temporal dimension** to exploit to leverage unlabeled data

Prior work

Prior work have been relied on:

- Two-stream networks: RGB + Optical Flow/ Temporal Gradient
- Two-stream skip rate: Fast + slow playback

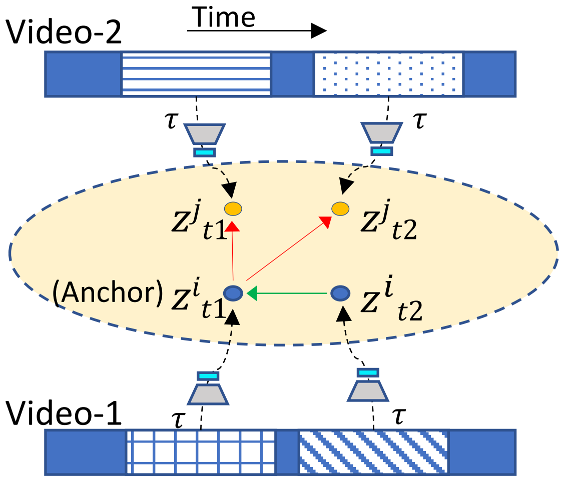
Prior work have input specific inductive-bias

- Requires precomputation of datasets. e.g., computing flow takes about a week for Kinetics400

Prior work have not systematically leveraged SSL video representations

Self-supervised Representations for Videos

Temporally Invariant (f_I)



▭ MLP Projection Head

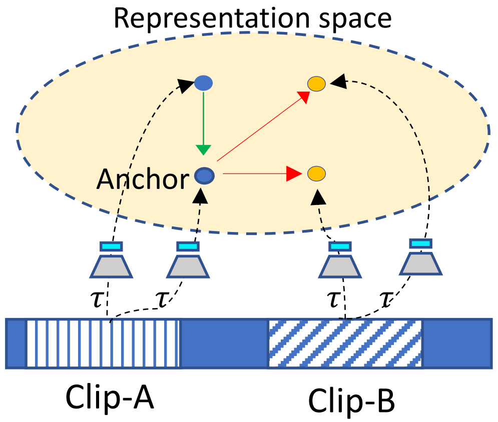
⏢ 3D-CNN backbone

→ Attract

→ Repel

τ Random Augmentation

Temporally Distinctive (f_D)



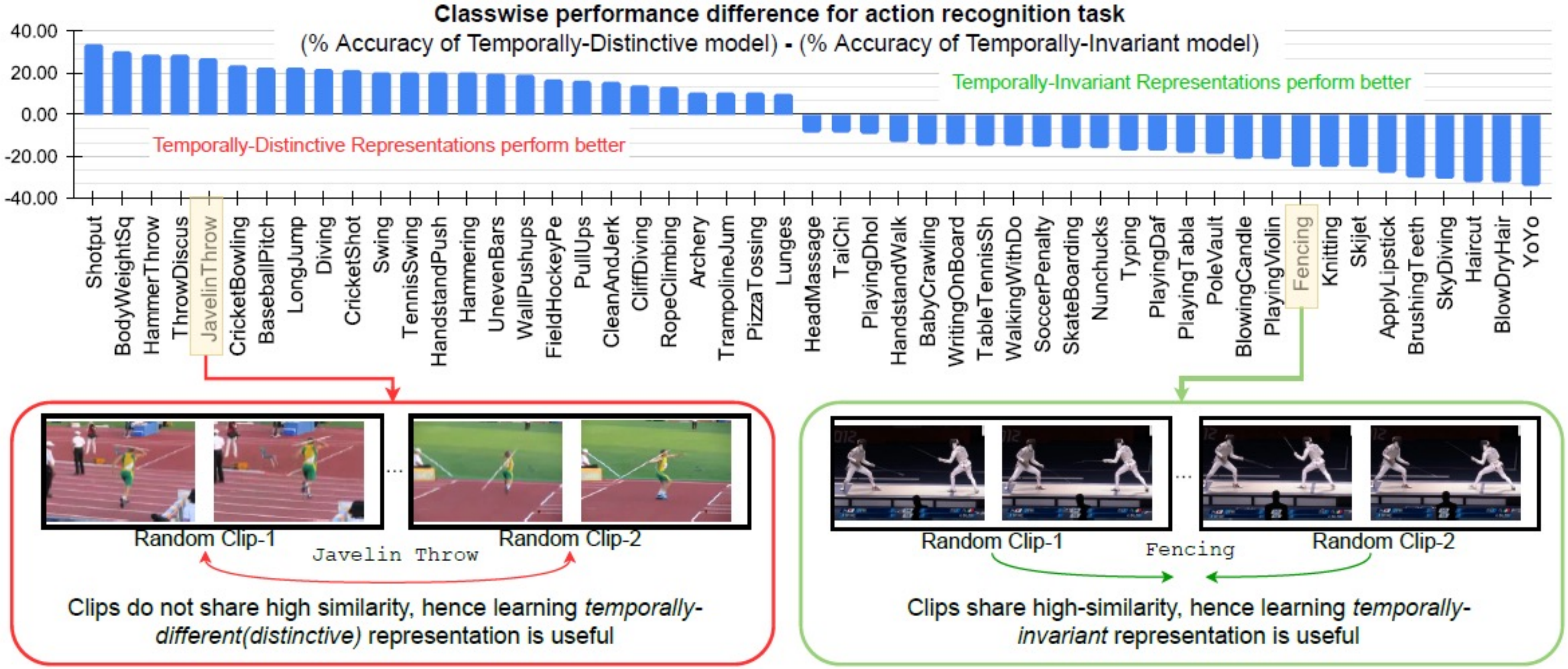
Contrastive loss/ SSL loss encourages **two clip** of video to have **same** representation

Contrastive loss/ SSL loss encourages **two clip** of video to have **different** representation

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

UCF Center for Research in Computer Vision

Complementary behavior of f_I and f_D



Clips do not share high similarity, hence learning *temporally-different(distinctive)* representation is useful

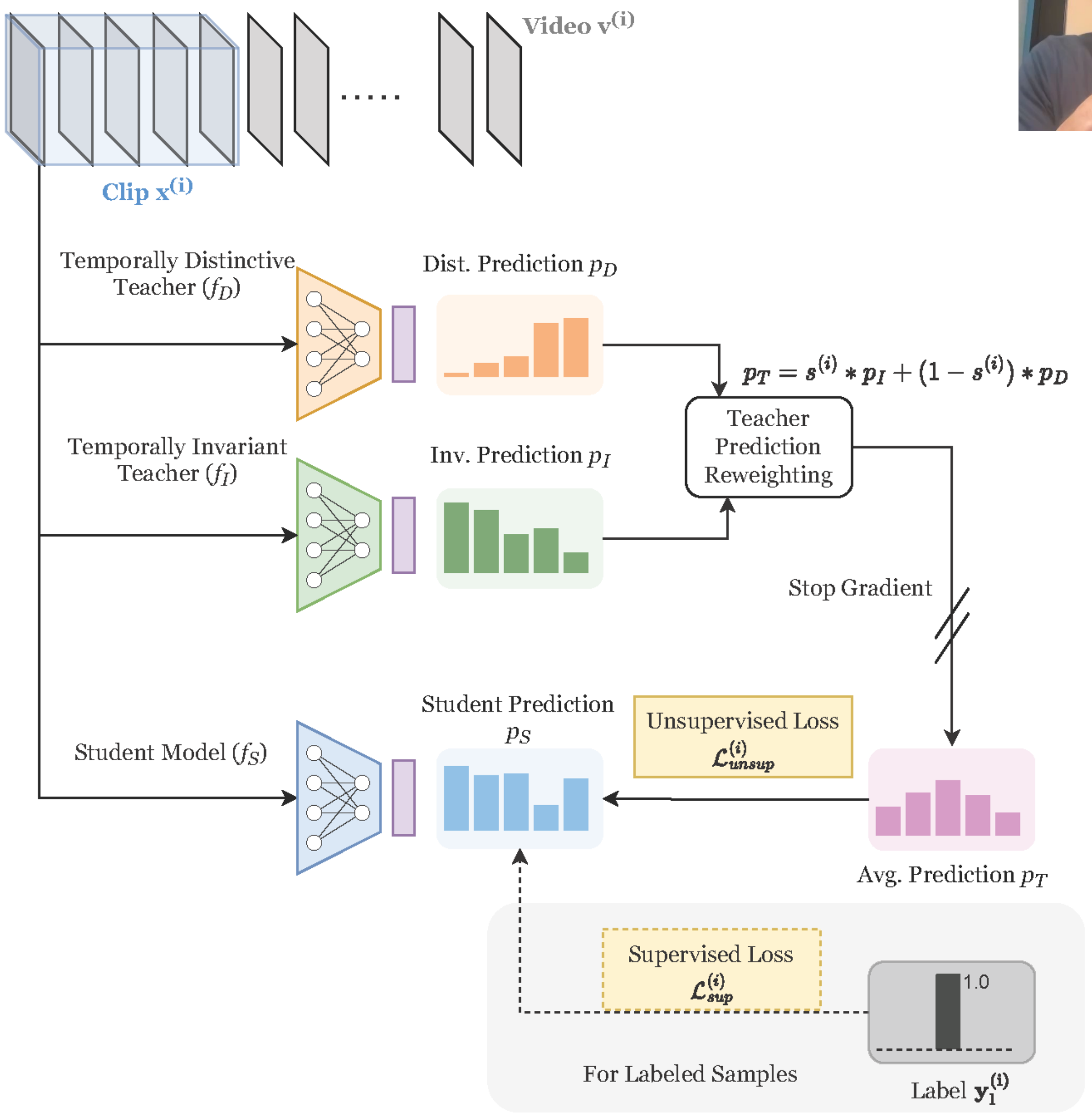
Clips share high-similarity, hence learning *temporally-invariant* representation is useful

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

UCF Center for Research in Computer Vision

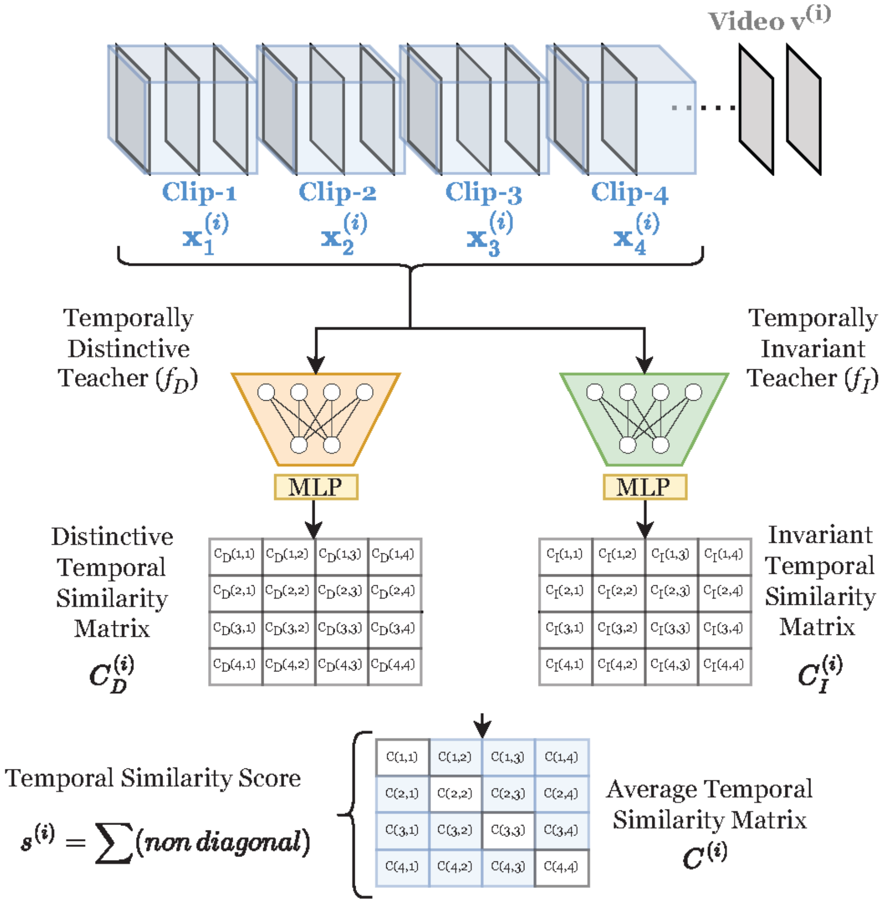
TimeBalance Framework

- Dual teacher- Student framework
- Both teachers are pretrained in self-supervised way and finetuned with the limited labeled data
- Prediction of both teachers are combined with proposed temporal-similarity based reweighting scheme


Video $\mathbf{v}^{(i)}$
Clip $\mathbf{x}^{(i)}$
.....
Temporally Distinctive Teacher (f_D)
Dist. Prediction p_D
Temporally Invariant Teacher (f_I)
Inv. Prediction p_I
$p_T = s^{(i)} * p_I + (1 - s^{(i)}) * p_D$
Teacher Prediction Reweighting
Stop Gradient
Student Model (f_S)
Student Prediction p_S
Unsupervised Loss $\mathcal{L}^{(i)}_{unsup}$
Avg. Prediction p_T
Supervised Loss $\mathcal{L}^{(i)}_{sup}$
1.0
For Labeled Samples
Label $\mathbf{y}^{(i)}_l$

Temporal Similarity Based Reweighting

- Consecutive clips are sampled from a video
- Clips are passed through the f_I and f_D models to get their SSL representation
- Temporal self-similarity matrix is computed using cosine distance
- Obtain a temporal self-similarity score
- Clips representation similarity (↑) → f_I(↑)
- Clips representation similarity (↓) → f_D (↑)


Video $\mathbf{v}^{(i)}$
Clip-1
$\mathbf{x}_1^{(i)}$
Clip-2
$\mathbf{x}_2^{(i)}$
Clip-3
$\mathbf{x}_3^{(i)}$
Clip-4
$\mathbf{x}_4^{(i)}$
Temporally Distinctive Teacher (f_D)
Temporally Invariant Teacher (f_I)
MLP
MLP
Distinctive Temporal Similarity Matrix $C_D^{(i)}$
Invariant Temporal Similarity Matrix $C_I^{(i)}$
Temporal Similarity Score
Average Temporal Similarity Matrix $C^{(i)}$

$c_D(1,1)$	$c_D(1,2)$	$c_D(1,3)$	$c_D(1,4)$
$c_D(2,1)$	$c_D(2,2)$	$c_D(2,3)$	$c_D(2,4)$
$c_D(3,1)$	$c_D(3,2)$	$c_D(3,3)$	$c_D(3,4)$
$c_D(4,1)$	$c_D(4,2)$	$c_D(4,3)$	$c_D(4,4)$

$c_I(1,1)$	$c_I(1,2)$	$c_I(1,3)$	$c_I(1,4)$
$c_I(2,1)$	$c_I(2,2)$	$c_I(2,3)$	$c_I(2,4)$
$c_I(3,1)$	$c_I(3,2)$	$c_I(3,3)$	$c_I(3,4)$
$c_I(4,1)$	$c_I(4,2)$	$c_I(4,3)$	$c_I(4,4)$

$c(1,1)$	$c(1,2)$	$c(1,3)$	$c(1,4)$
$c(2,1)$	$c(2,2)$	$c(2,3)$	$c(2,4)$
$c(3,1)$	$c(3,2)$	$c(3,3)$	$c(3,4)$
$c(4,1)$	$c(4,2)$	$c(4,3)$	$c(4,4)$

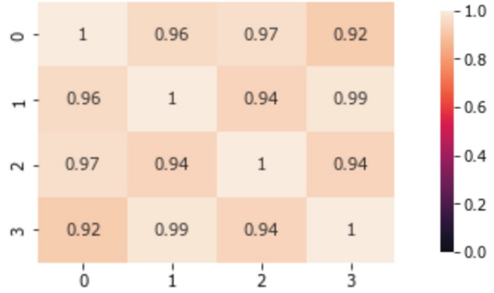
$$s^{(i)} = \sum(non\ diagonal)$$

$$\mathbf{p}_T^{(i)} = s^{(i)}.\mathbf{p}_I^{(i)} + (1 - s^{(i)}).\mathbf{p}_D^{(i)}$$

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

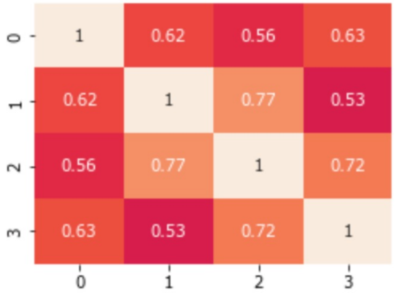
UCF Center for Research in Computer Vision

Illustration of teacher reweighting



	0	1	2	3
0	1	0.96	0.97	0.92
1	0.96	1	0.94	0.99
2	0.97	0.94	1	0.94
3	0.92	0.99	0.94	1

Temporal Similarity Matrix from Invariant teacher C_I



	0	1	2	3
0	1	0.62	0.56	0.63
1	0.62	1	0.77	0.53
2	0.56	0.77	1	0.72
3	0.63	0.53	0.72	1

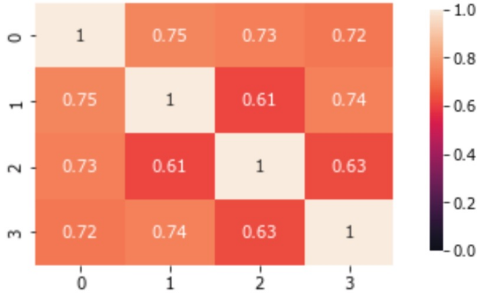
Temporal Similarity Matrix from Distinctive teacher C_D

Instance Similarity score $s = 0.795$

Repetitive action

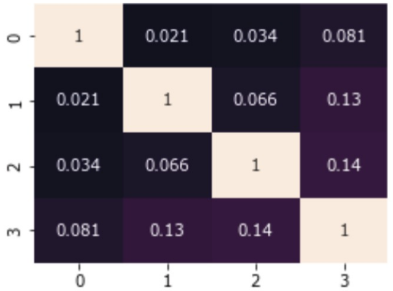
High similarity score

More weight given to f_I



	0	1	2	3
0	1	0.75	0.73	0.72
1	0.75	1	0.61	0.74
2	0.73	0.61	1	0.63
3	0.72	0.74	0.63	1

Temporal Similarity Matrix from Invariant teacher C_I



	0	1	2	3
0	1	0.021	0.034	0.081
1	0.021	1	0.066	0.13
2	0.034	0.066	1	0.14
3	0.081	0.13	0.14	1

Temporal Similarity Matrix from Distinctive teacher C_D

Instance Similarity score $s = 0.387$

Non-Repetitive action

Low similarity score

More weight given to f_D

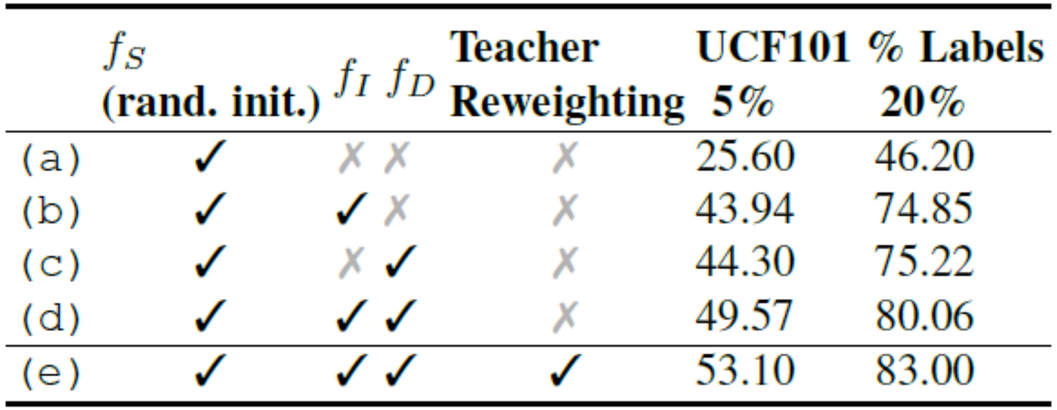
Results

Method	Venue	Backbone	Params	Input	#F	UCF101					HMDB51			Kinetics400	
			(M)			1%	5%	10%	20%	50%	40%	50%	60%	1%	10%
PL	ICML'13	3D-ResNet18	13.5	V	16	-	17.6	24.7	37	47.5	27.3	32.4	33.5	-	-
MT	NeuRIPS'17	3D-ResNet18	13.5	V	16	-	17.5	25.6	36.3	45.8	27.2	30.4	32.2	-	-
S4L	ICCV'19	3D-ResNet18	13.5	V	16	-	22.7	29.1	37.7	47.9	29.8	31	35.6	-	-
UPS	ICLR'21	3D-ResNet18	13.5	V	16	-	-	-	39.4	50.2	-	-	-	-	-
SD	ICCV'19	3D-ResNet18	13.5	V	16	-	31.2	40.7	45.4	53.9	32.6	35.1	36.3	-	-
MT+SD	WACV'21	3D-ResNet18	13.5	V	16	-	30.3	40.5	45.5	53	32.3	33.6	35.7	-	-
VideoSemi	WACV'21	3D-ResNet18	13.5	V	16	-	32.4	42	48.7	54.3	32.7	36.2	37	-	-
TG-FixMatch	CVPR'21	3D-ResNet18	13.5	V	8	-	44.8	62.4	76.1	79.3	46.5	48.4	49.7	9.8	43.8
TCL	CVPR'21	TSM-R18	-	V	8	-	-	-	-	-	-	-	-	11.6	-
MvPL	ICCV'21	3D-ResNet18	13.5	VFG	8	-	41.2	55.5	64.7	65.6	30.5	33.9	35.8	5	36.9
CMPL	CVPR'22	3D-ResNet18	13.5	V	8	23.8	-	67.6	-	-	-	-	-	16.5	53.7
TACL	TSVT'22	3D-ResNet18	13.5	V	16	-	35.6	50.9	56.1	65.8	34.6	37.2	39.5	-	-
TACL	TSVT'22	3D-ResNet18	13.5	V	16	-	43.7	55.6	59.2	67.2	38.7	40.2	41.7	-	-
3DRotNet	Arxiv'19	3D-ResNet18	13.5	V	16	15	31.5	40.4	47.1	-	-	-	-	-	-
MemDPC	ECCV'20	3D-ResNet18	13.5	V	16	-	-	44.2	50.9	62.3	-	-	-	-	-
MotionFit	ICCV'21	3D-ResNet18	13.5	VF	16	-	-	-	57.7	59	-	-	-	-	-
TCLR	CVIU'22	3D-ResNet18	13.5	V	16	26.9	-	66.1	73.4	76.7	-	-	-	-	-
TimeBalance	CVPR'23	3D-ResNet18	13.5	V	8	**29.1**	**47.9**	**69.8**	**79.1**	**83.3**	**49.8**	**51.4**	**53.1**	**17.1**	**54.9**
ActorCM	Arxiv'21	R(2+1)D-34	33.3	V	8	-	27	40.2	51.7	59.9	32.9	38.2	38.9	-	-
ActorCM	Arxiv'21	R(2+1)D-34	33.3	V	8	-	45.1	53	57.4	64.7	35.7	39.5	40.8	-	-
FixMatch	NeuRIPS'20	SlowFast-R50	60	V	8	16.1	-	55.1	-	-	-	-	-	10.1	49.4
MvPL	ICCV'21	3D-ResNet50	31.8	VFG	8	22.8	-	80.5	-	-	-	-	-	17	58.2
CMPL	CVPR'22	3D-ResNet50	31.8	V	8	25.1	-	79.1	-	-	-	-	-	17.6	58.4
TimeBalance	CVPR'23	3D-ResNet50	31.8	V	8	**30.1**	**53.5**	**81.1**	**83.3**	**85**	**52.6**	**53.9**	**54.5**	**19.6**	**61.2**

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

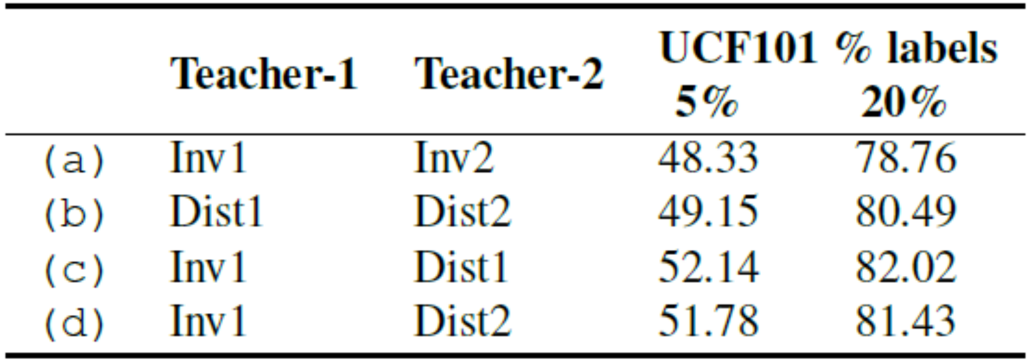
UCF CENTER FOR RESEARCH IN COMPUTER VISION

Ablations



	f_S (rand. init.)	f_I	f_D	Teacher Reweighting	UCF101 % Labels	
					5%	20%
(a)	✓	✗	✗	✗	25.60	46.20
(b)	✓	✓	✗	✗	43.94	74.85
(c)	✓	✗	✓	✗	44.30	75.22
(d)	✓	✓	✓	✗	49.57	80.06
(e)	✓	✓	✓	✓	53.10	83.00

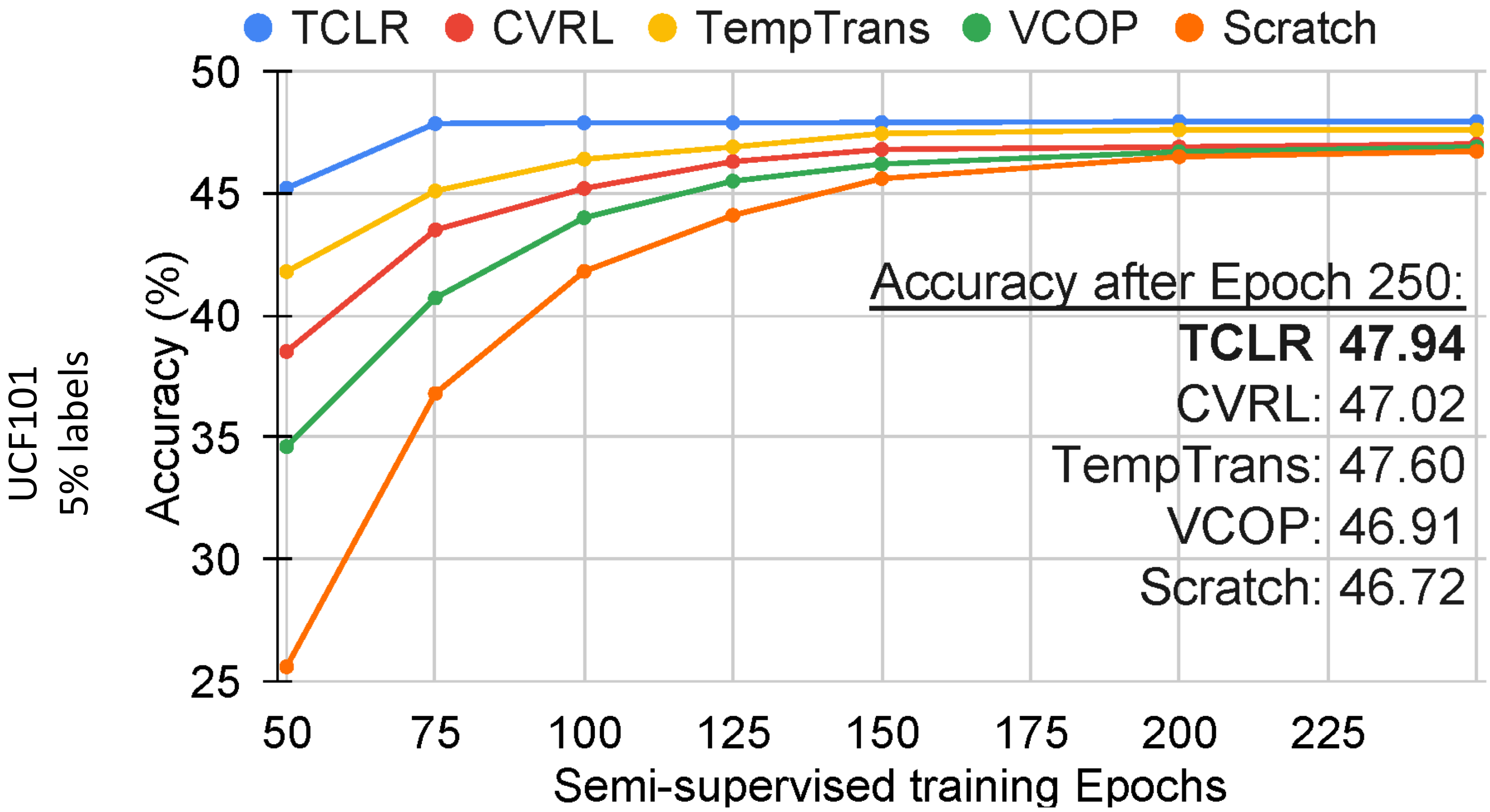
Ablation: Different Teacher Combinations



	Teacher-1	Teacher-2	UCF101 % labels	
			5%	20%
(a)	Inv1	Inv2	48.33	78.76
(b)	Dist1	Dist2	49.15	80.49
(c)	Inv1	Dist1	52.14	82.02
(d)	Inv1	Dist2	51.78	81.43

Different Teacher Combinations

Different Initializations of student



● TCLR ● CVRL ● TempTrans ● VCOP ● Scratch
UCF101
5% labels
Accuracy (%)
50
45
40
35
30
25
50
75
100
125
150
175
200
225
Semi-supervised training Epochs

Accuracy after Epoch 250:

TCLR	**47.94**
CVRL:	47.02
TempTrans:	47.60
VCOP:	46.91
Scratch:	46.72

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

UCF Center for Research in Computer Vision

Conclusion

- TimeBalance, a teacher-student framework for semi-supervised action recognition
- We utilize the complementary strengths of *temporally-invariant* and *temporally-distinctive* representations to leverage unlabeled videos
- State-of-the-art for semi-supervised action recognition: UCF101, HMDB51, Kinetics400