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

Ishan Rajendrakumar Dave, Mamshad Nayeem Rizve, Chen Chen, Mubarak Shah
Center for Research in Computer Vision, University of Central Florida, Orlando, USA
{ishandave, nayeemrizve}@knights.ucf.edu, {chen.chen, shah}@crcv.ucf.edu

A. Overview

- Section B: Implementation details about network architectures and training setup.
- Section C: Ablation study for our framework.
- Section D: Supportive diagrams and explanation for our method.

B. Implementation Details

B.1. Network Architecture

B.1.1 Backbone

For teacher models f_I and f_D, we utilize 3D-ResNet50 model from the implementation of `Slow-R50` [2] of official PyTorchVideo[1]. For experiments with 3D-ResNet18, we utilize its official PyTorch implementation `r3d_18`[2].

B.1.2 Non-Linear Projection Head

We use non-linear projection head $g(\cdot)$ during the self-supervised pretraining of temporally-invariant and temporally-distinctive teachers to reduce the dimensions of the representation. We utilize Multi-layer Perceptron (MLP) as a non-linear projection head to project 2048-dimensional model features to 128-dimensional vectors in normalized representation space. The design of MLP is as follows, where `nn` indicates `torch.nn` PyTorch package:

```
nn.Linear(2048,512, bias = True)
nn.BatchNorm1d(512)
nn.ReLU(inplace=True)
nn.Linear(512, 128, bias = False)
nn.BatchNorm1d(128)
```

[1] https://github.com/facebookresearch/pytorchvideo
[2] https://github.com/pytorch/vision/blob/main/torchvision/models/video

B.2. Training Details

For all weight updates, we utilize Adam Optimizer [3] with default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a base learning rate (α_I, α_D, α_S) of 1e-3. For all training, we utilize a linear warmup of 10 epochs. A patience-based learning rate scheduler is also used, which drops the learning rate to its half value on a loss plateau.

C. Additional Ablations

C.1. Loss function for teacher supervision

In order to distill teacher knowledge, we study three different loss functions as $\mathcal{L}_{unsup}$ and report the results in Table 1. For these experiments, we use 3D-Resnet50 as the student model on the UCF101 dataset [4]. We observe that all three losses perform reasonably while $\mathcal{L}_2$ performs the best, which we use as the default loss in our method.

Unlabeled Supervision	UCF101 % Labels		
	5%	20%	50%
$\mathcal{L}_2$	53.48	83.15	85.02
KL-Divergence	52.62	82.76	84.50
JS-Divergence	50.91	82.10	83.94

Table 1. Ablation of different Teacher Losses. $\mathcal{L}_2$ distillation loss performs the best, which we use in our default setting.

C.2. Student f_S from Scratch

We perform experiments with student from random initialization and compare them with the prior methods in Table 2.

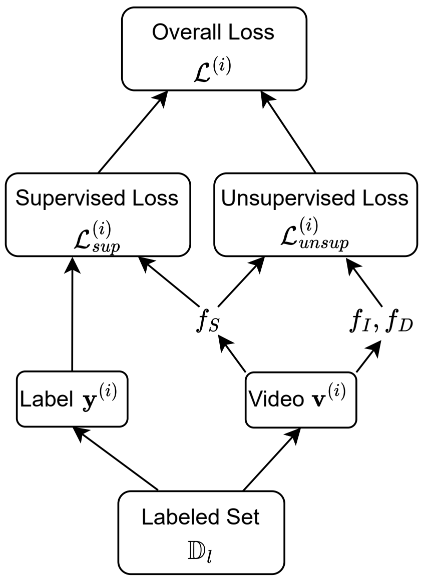
D. Method

D.1. Loss for Labeled and Unlabeled set

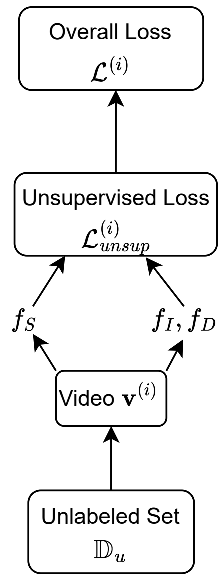
In Fig. 1, we show the handling of labeled and unlabeled data in the semi-supervised training of student f_S. For la-

	Backbone	UCF101		HMDB		Kinetics
		5%	20%	40%	60%	10%
Prior best methods	R3D-50	27.0 [7]	51.7 [7]	32.9 [7]	38.9	58.4 [6]
Ours (scratch student)	R3D-50	53.1(**+26.1**)	83.0 (**+31.3**)	52.1 (**+19.2**)	54.3 (**+15.4**)	60.8 (**+2.4**)
Prior best methods	R3D-18	44.8 [5]	76.1 [5]	46.5 [5]	49.7 [5]	53.7 [6]
Ours (scratch student)	R3D-18	46.7 (**+1.9**)	78.2 (**+2.1**)	49.1 (**+2.6**)	52.9 (**+3.2**)	54.4 (**+0.7**)

Table 2. Experiments with Student trained from Random Initialization. **(+n)** shows absolute improvement over the prior best work



(a) Labeled data



(b) Unlabeled data

Figure 1. Loss computations in labeled and unlabeled data. (a) In case of Labeled data, the student f_S gets supervision from supervised cross-entropy loss from label $\mathbf{y}^{(i)}$ and unsupervised $\mathcal{L}_2$ loss from teachers. (b) For unlabeled set, the student is only trained with the unsupervised loss from teachers. Details are in Sec 3.2 of the main paper.

beled data $\mathbb{D}_l$, the student model has two sources of supervision: (1) Labeled supervision $\mathcal{L}^{(i)}_{sup}$ in the form of standard cross entropy loss which is computed from the student's prediction and given class label $\mathbf{y}^{(i)}$ (2) Unlabeled supervision $\mathcal{L}^{(i)}_{unsup}$ in the form of $\mathcal{L}_2$ distillation loss computed from the weighted average of predictions of teachers (f_D and f_I). For the unlabeled set $\mathbb{D}_u$, the student model gets supervision only in the form of $\mathcal{L}_2$ distillation loss.

D.2. Temporally-Distinctive pretraining using unpooled features

Since $\mathcal{L}_{D1}$ deals with temporally-pooled(averaged) features, it promotes temporal-distinctiveness for the *pooled* features. Similar to that, [1] designs a contrastive objective that promotes temporally-distinctive representation on the *unpooled* features. We call it unpooled temporal-distinctive objective $\mathcal{L}_{D2}$, which is illustrated in Fig. 2.

References

[1] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, page 103406, 2022. 2

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 1

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

[4] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[5] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. In *Pro-*

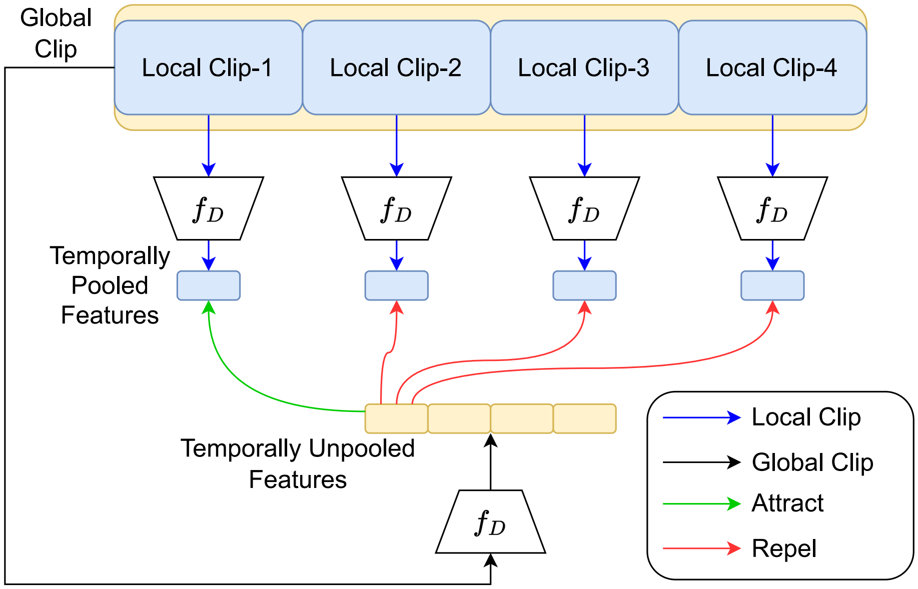


Figure 2. **Temporally-Distinctive Contrastive Objective for Temporally-unpooled features** $\mathcal{L}_{D2}$: A time-duration of the video can be represented in two different ways: (1) Pooled features of the short(local) clip (2) Unpooled feature slice of the long(global) clip. In this contrastive objective, we maximize the agreement between *temporally-aligned* pooled and unpooled features.

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3252–3262, 2022. 2

[6] Yinghao Xu, Fangyun Wei, Xiao Sun, Ceyuan Yang, Yujun Shen, Bo Dai, Bolei Zhou, and Stephen Lin. Cross-model pseudo-labeling for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2959–2968, 2022. 2

[7] Yuliang Zou, Jinwoo Choi, Qitong Wang, and Jia-Bin Huang. Learning representational invariances for data-efficient action recognition. *arXiv preprint arXiv:2103.16565*, 2021. 2