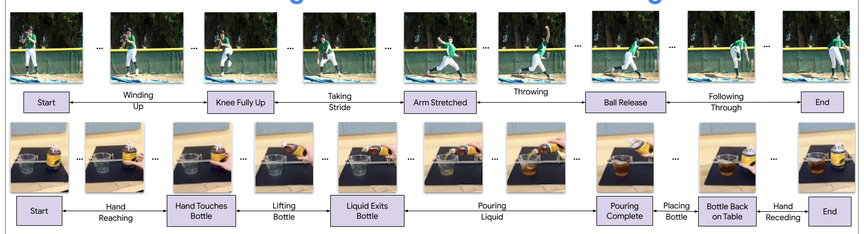
Temporal Cycle-Consistency Learning

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet and Andrew Zisserman

Problem Setting

Fine-grained Video Understanding

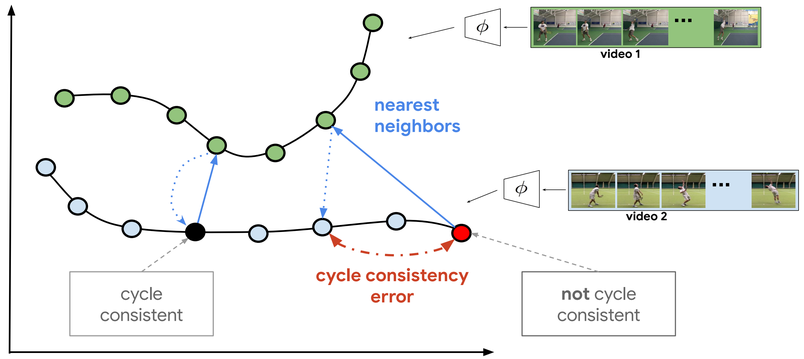


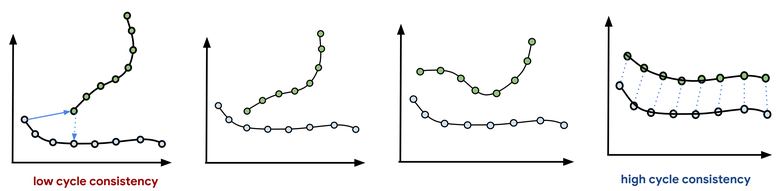
Research Questions

1. How do we train models that understand each frame of a video?
2. Can we avoid annotating each frame for training such models?

Temporal Cycle-Consistency (TCC) Learning

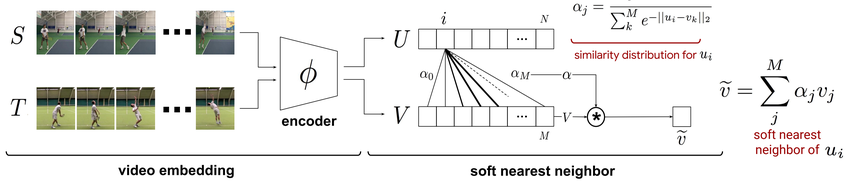
1. Learn per-frame representations by video alignment.
2. Principle of mutual nearest neighbors is used as a training signal.



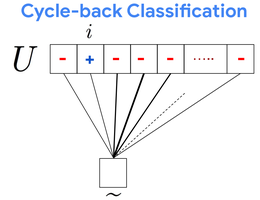


Embedding Spaces with varying levels of Cycle-Consistency

Differentiable Cycle-Consistency Losses



Cycle-back Classification



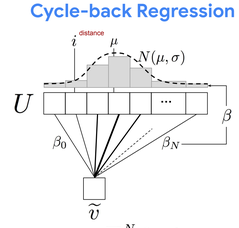
logits $x_k = -||\widetilde{v} - u_k||_2$

predictions: $\hat{y} = softmax(x)$

label: $y = i$

$$L_{cbc} = -\sum_{j}^{N} y_j \log(\hat{y}_j)$$

Cycle-back Regression



$$\mu = \sum_k^N \beta_k * k$$

$$\sigma^2 = \sum_k^N \beta_k * (k - \mu)^2$$

$$L_{cbr} = \frac{|i - \mu|^2}{\sigma^2} + \lambda \log(\sigma)$$

Metrics for Fine-grained Understanding

1. **Phase Classification Accuracy:** For each frame, predict action phase.
2. **Progress Prediction:** For each frame predict, how far ahead/behind the frame is from key events (like beginning of an action phase)
3. **Kendall's Tau:** For every pair of videos, measures alignment between sequences.

Experiments and Results

1. Which cycle-consistency loss is better?

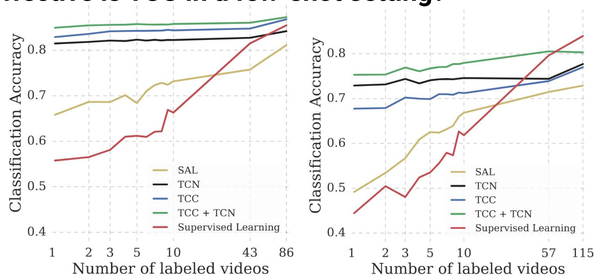
Loss	Phase Classification	Phase Progression	Kendall's Tau
Mean Squared Error	86.2	0.65	0.61
Cycle-back Classification	88.1	0.67	0.67
Cycle-back Regression	**91.8**	**0.80**	**0.85**

Regression uses more temporal information and results in better performance.

2. How effective is TCC for phase classification?

		% of Labels		
Dataset	**Method**	**0.1**	**0.5**	**1.0**
Penn Action [3]	Supervised Learning	50.7	72.9	80.0
	SAL [2]	66.2	71.1	72.5
	TCN [1]	69.7	71.4	72.2
	TCC	**74.7**	**76.4**	**77.3**
Pouring [1]	Supervised Learning	62.0	77.7	88.4
	SAL [2]	74.5	81.0	83.2
	TCN [1]	76.0	83.3	84.6
	TCC	**86.8**	**89.4**	**90.2**

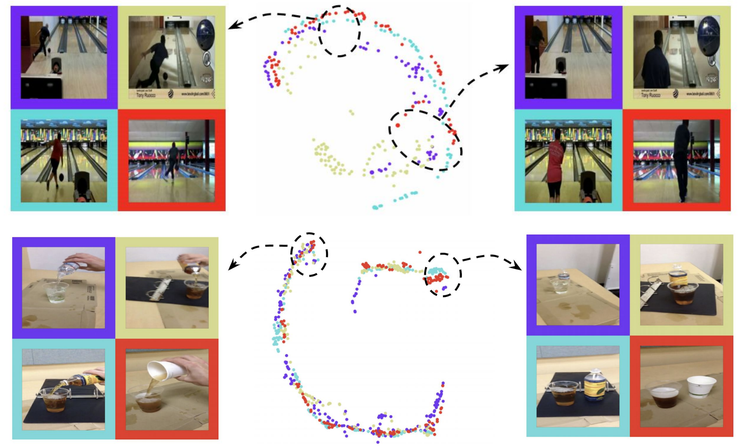
3.How effective is TCC in a few-shot setting?



(a) Golf Swing **(b)** Tennis Serve

TCC + one labeled video is as good as supervised learning on ~50 labeled videos.

t-SNE Visualization of TCC Embeddings



Applications

Unsupervised Video Alignment

Video alignment is as easy as looking up nearest-neighbors in TCC space.

Label/Modality Transfer

Transfer modalities (like sound) and temporal labels from one video to all other videos of the same action

Fine-grained Retrieval

Each frame of a video can be used for retrieval in other videos.

References

[1] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, and S. Levine, "Time-contrastive networks: Self-supervised learning from video"
[2] I. Misra, C. L. Zitnick, and M. Hebert. "Unsupervised Learning using Sequential Verification for Action Recognition"
[3] W. Zhang, M. Zhu, and K. G. Derpanis. " From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding"