SLIC: Self-Supervised Learning with Iterative Clustering for Human Action Videos

UNIVERSITY OF TORONTO Robotics Institute

Salar Hosseini Khorasgani* Yuxuan Chen* Florian Shkurti

*Equal Contribution University of Toronto Vector Institute

Robot Vision & Learning

Introduction

Motivation: Self-supervised contrastive learning methods for videos typically underperform compared to fully supervised methods as a result of conservative positive and negative selection.

Contributions:

- Using iterative clustering to provide pseudo labels for **self-supervised learning** of video representations.
- Integrating **iterative clustering** with **multi-view encoding** and a **temporal discrimination loss** to sample harder positives and negatives during pretraining.

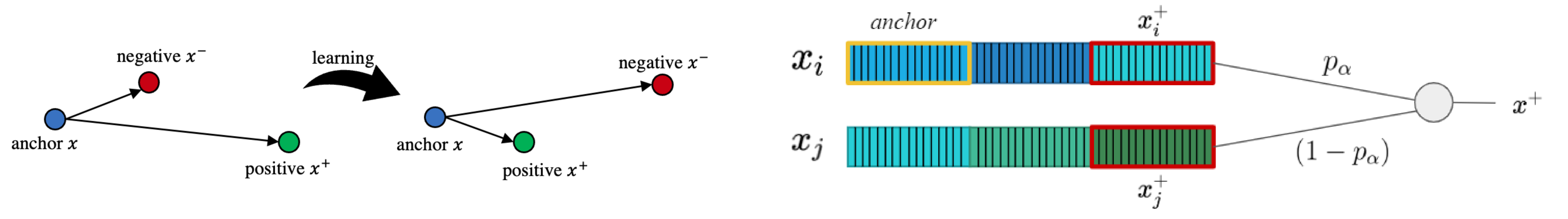
Iterative Clustering

- Extract features using a deep 3D CNN and perform **FINCH clustering** every k epochs in the feature space to obtain cluster assignments.
- The cluster assignments are used as **pseudo labels** to sample positives and negatives for triplet learning.

Instance-based Triplet Loss

Triplet margin loss: $\mathcal{L}_{triplet}(x, x^+, x^-; \theta, m_1) = \max(0, d(f_\theta(x), f_\theta(x^+)) - d(f_\theta(x), f_\theta(x^-)) + m_1)$

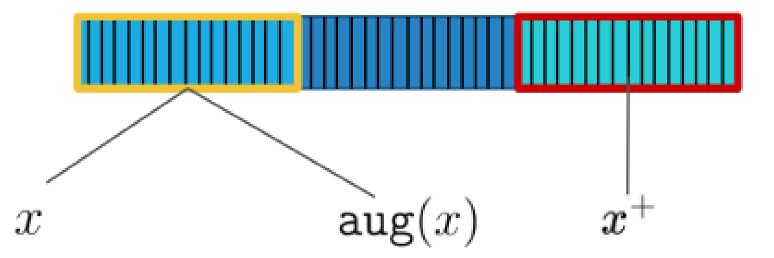
- **Anchor** (x): Clip from a random video instance x_i.
- **Positive** (x^+): Either (i) a clip from the same instance, x_i^+, with probability p_α or ii) a clip from anther instance in the same cluster, x_j^+, with probability $(1 - p_\alpha)$.
- **Negative** (x^-): Clip from a different cluster than x and x^+ that satisfies $d(f_\theta(x), f_\theta(x^-)) \leq d(f_\theta(x), f_\theta(x^+)) + m_1$.



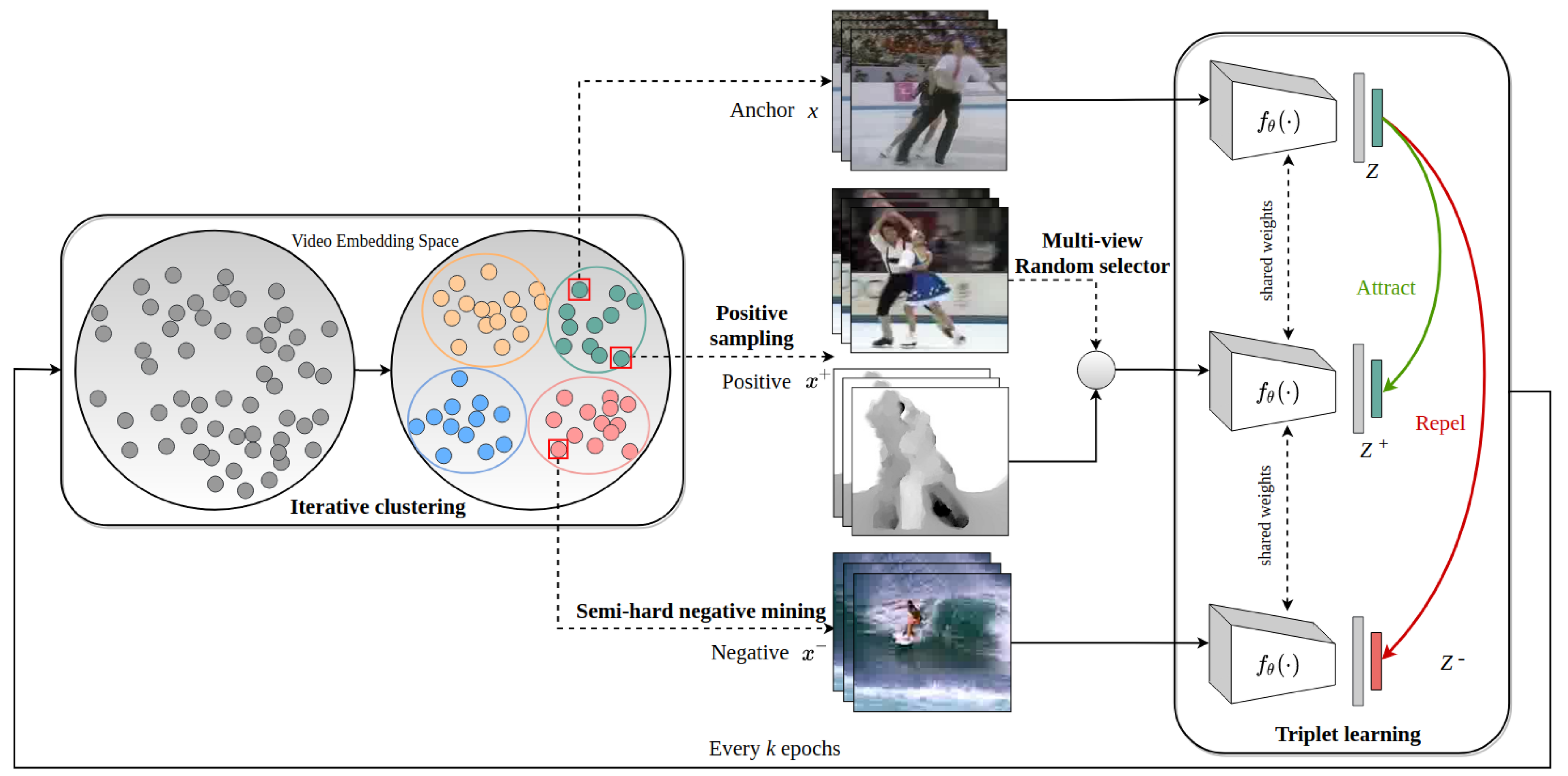
Temporal Discrimination Loss

Temporal discrimination loss: $\mathcal{L}_{temporal} = \mathcal{L}_{triplet}(x, \text{aug}(x), x^+; \theta, m_2)$

- **Anchor:** Same as instance-based triplet loss.
- **Positive:** Spatial augmentation of the anchor clip, $\text{aug}(x)$.
- **Negative:** The positive from the instance-based triplet loss, x^+ (some temporally non-overlapping clip from the same video instance or a different instance in the same cluster).
- **Margin** (m_2): Chosen such that $m_2 < m_1$ so x^+ is not pushed too far from x.

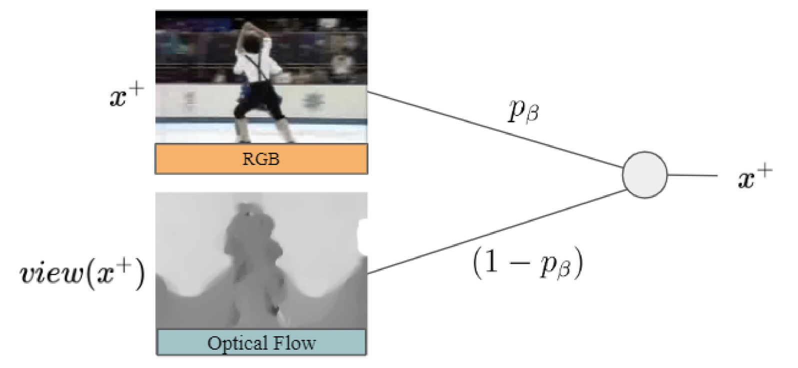


Method Overview



Multi-view Positives

- We sample the RGB clip as the positive with probability p_β, or replace it with optical flow with probability $(1 - p_\beta)$.
- We use a shared encoder for RGB and optical flow.



Evaluation on Video Retrieval and Action Classification

Video retrieval: Given a query video from test set, retrieve k-nearest neighbours from training set.

Action classification: Attach linear classifier on backbone, and evaluate i) linear probing, ii) end-to-end finetuning.

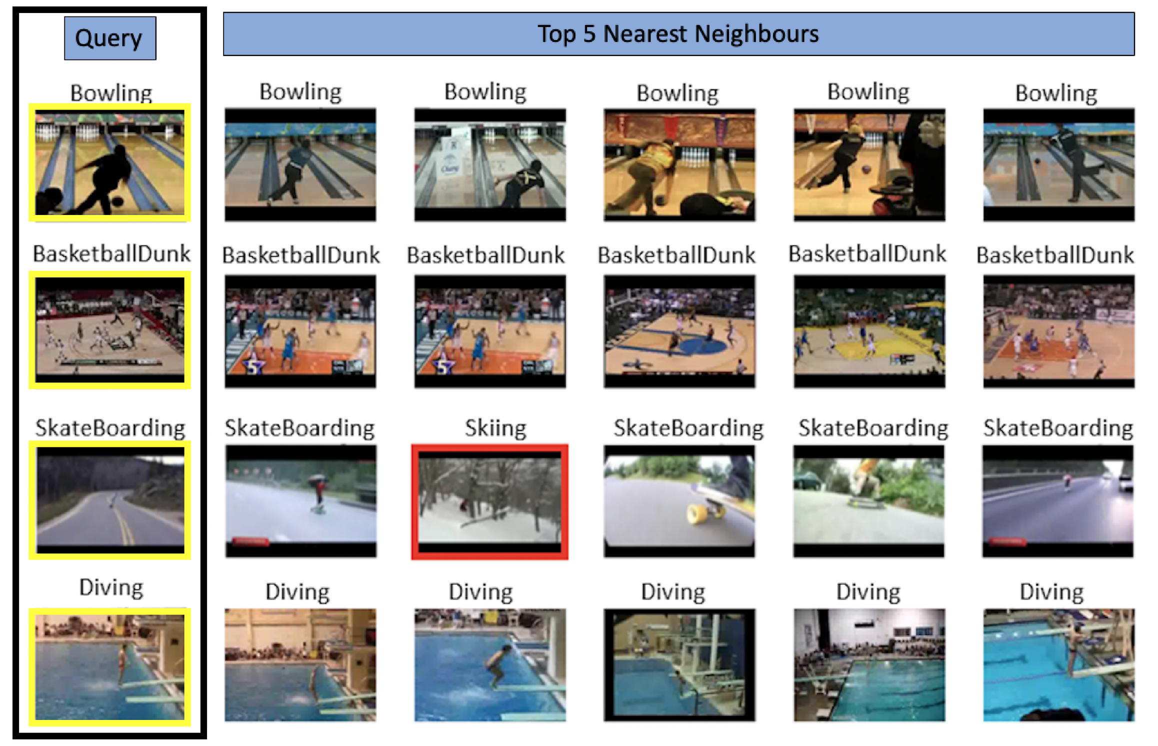
Method	Arch.	UCF R@1	UCF R@5	HMDB R@1	HMDB R@5
CoCLR-RGB	S3D-23	53.3	69.4	23.2	43.2
TCLR	R3D-18	56.2	72.2	22.8	45.4
SLIC	S3D-23	69.8	79.2	26.8	**52.9**
SLIC	R3D-18	**71.6**	**82.4**	**28.9**	52.8
Supervised SLIC	S3D-23	72.5	79.1	19.1	45.1
Supervised SLIC	R3D-18	81.0	84.9	33.0	57.4

Table 1. **Nearest neighbour retrieval results** on UCF101 and HMDB51. Temporal window: 32 frames.

Method	Pretrain	Arch.	Frozen	UCF	HMDB
TCLR	UCF	R3D-18	✓	69.9	-
CoCLR-RGB	UCF	S3D-23	✓	70.2	39.1
SLIC	UCF	R3D-18	✓	**77.7**	**48.3**
CoCLR-RGB	UCF	S3D-23	✗	81.4	52.1
TCLR	UCF	R3D-18	✗	82.4	52.9
SLIC	UCF	R3D-18	✗	**83.2**	**54.5**
TCLR	K400	R3D-18	✗	84.1	53.6
CoCLR-RGB	K400	S3D-23	✗	**87.9**	**54.6**
SLIC	K400	R3D-18	✗	83.1	52.0

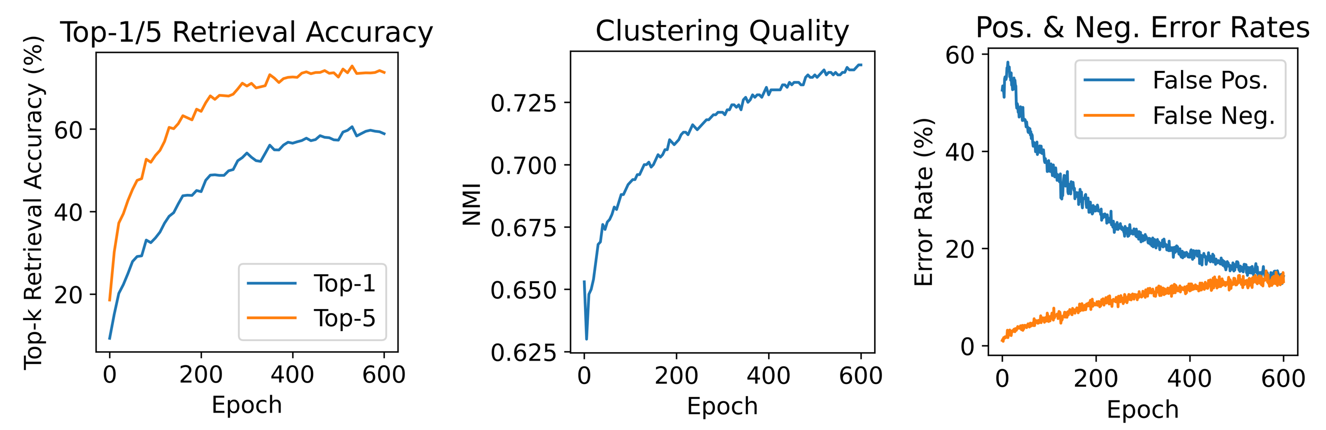
Table 2. **Top-1 accuracy results for action classification**. 'Frozen ✓' indicates classification with a frozen backbone; 'Frozen ✗' indicates end-to-end finetuning. Temporal window: 32 frames.

Qualitative Evaluation



Evolution of Clustering Quality and Retrieval Accuracy

During training, we monitor i) **top-k retrieval accuracy**, ii) **NMI** between cluster assignments and ground-truth labels, iii) **false positive and negative sample rates** from clusters (false according to ground truth labels).



Ablation Study

Clustering	Multi-view	Temporal Loss	UCF101 R@1	UCF101 R@5	HMDB51 R@1	HMDB51 R@5
✗	✓	✓	45.0	62.3	19.5	45.1
✓	✗	✗	54.7	65.6	18.2	41.5
✓	✓	✗	59.9	69.8	19.1	41.1
✓	✗	✓	59.2	69.8	20.1	43.6
✓	✓	✓	**66.7**	**77.3**	**25.3**	**49.8**

Table 3. Ablation study on the impact of different training components, with input size set to 16×128^2.

Conclusion

- Proposed a self-supervised, iterative clustering based contrastive learning framework for video representations.
- Achieved competitive or state-of-the-art performance across various downstream video understanding tasks.

{salar.hosseinikhorasgani, yuxuansherry.chen}@mail.utoronto.ca