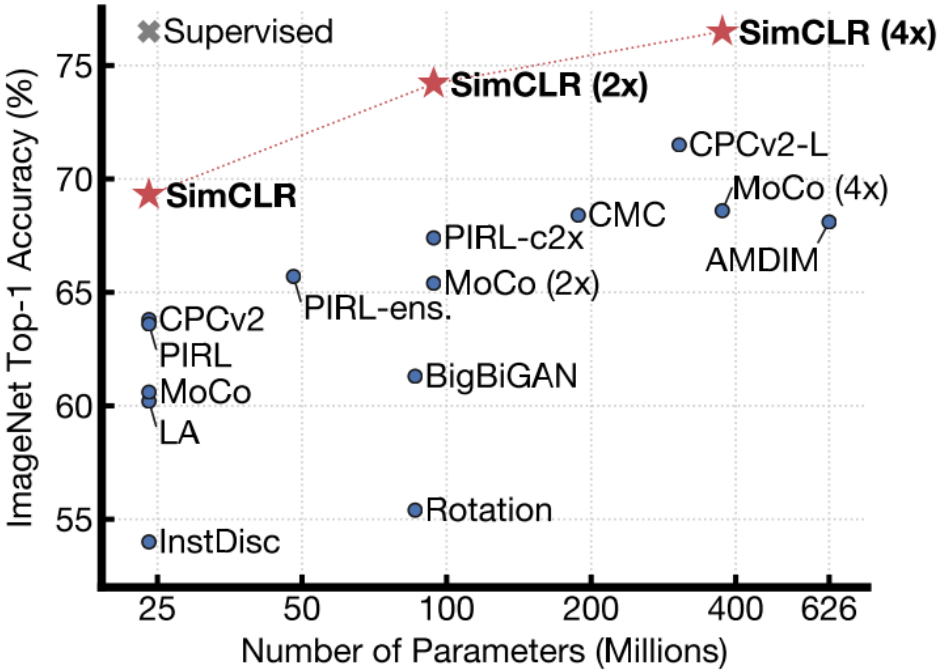
Temporally Coherent Embeddings for Self-Supervised Video Representation Learning

Joshua Knights, Ben Harwood, Daniel Ward, Anthony Vanderkop, Olivia Mackenzie-Ross, Peyman Moghadam

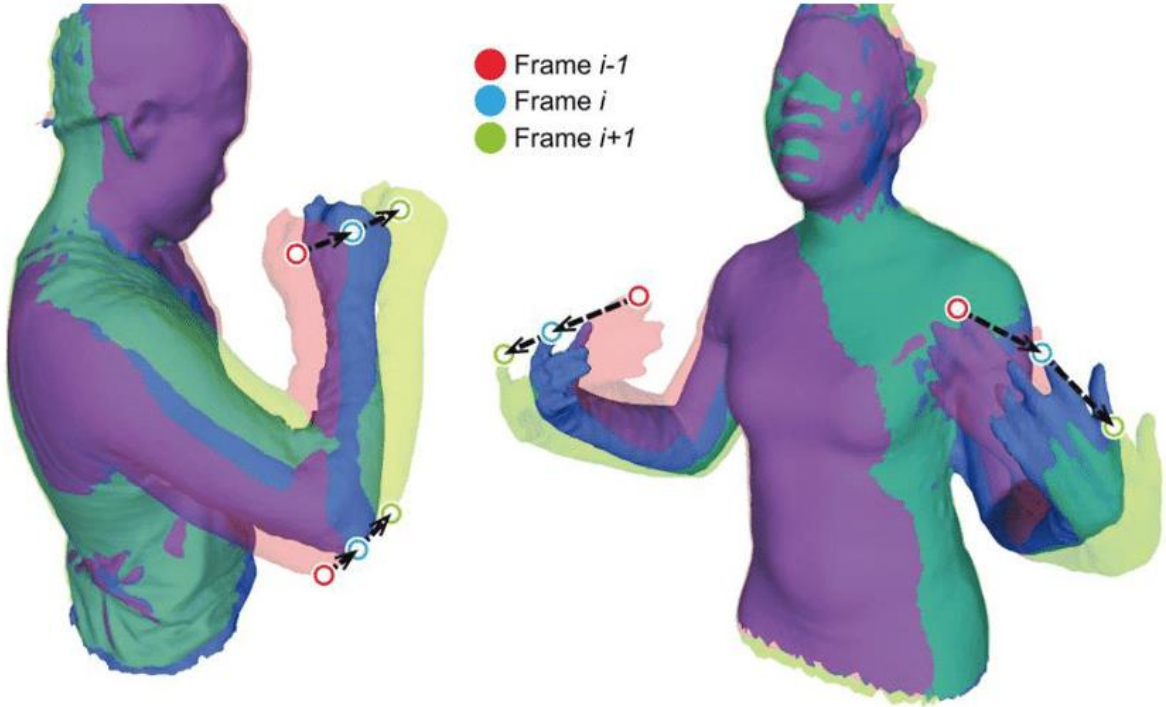
ICPR 2020
Paper ID: 2462

Background



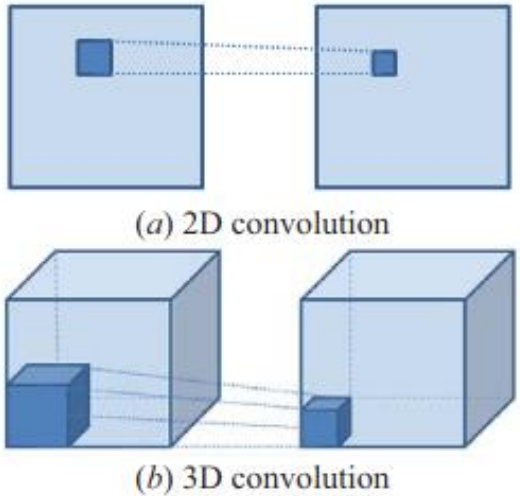
Chen, Ting, et al. "A Simple Framework for Contrastive Learning of Visual Representations."

Motivation



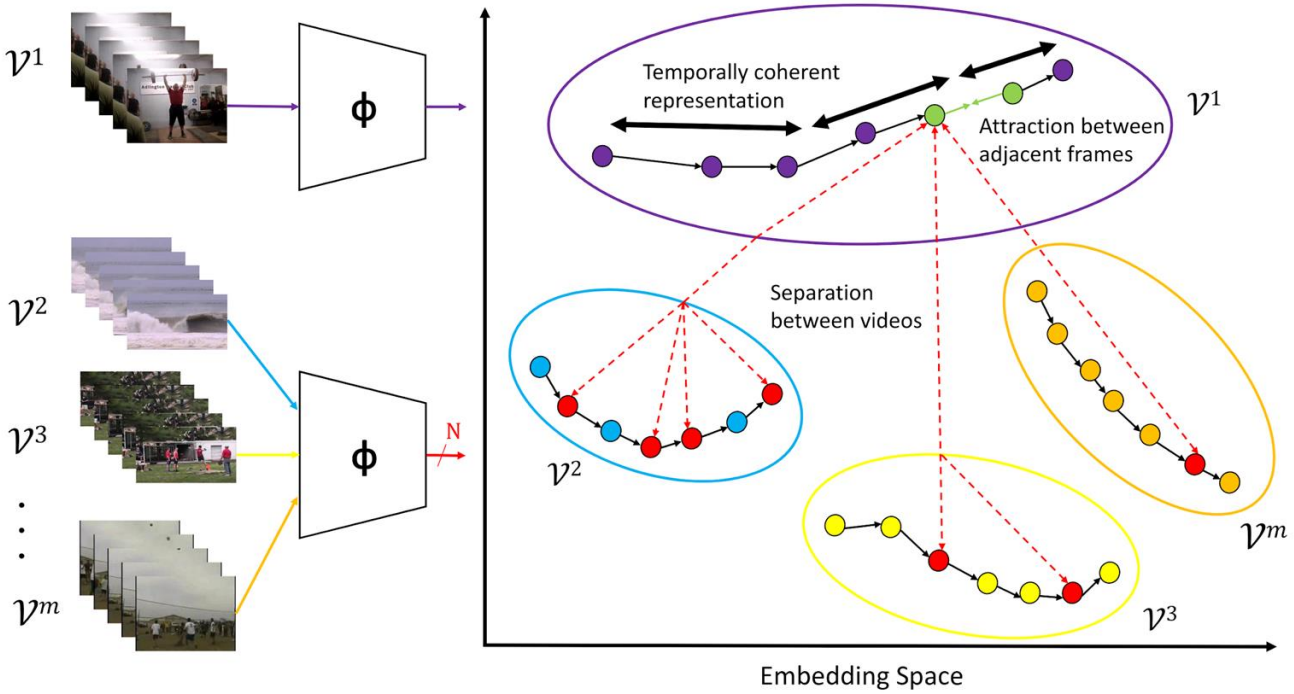
Li, Zhong, et al. "Robust 3D human motion reconstruction via dynamic template construction." *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017.

2D vs 3D CNN

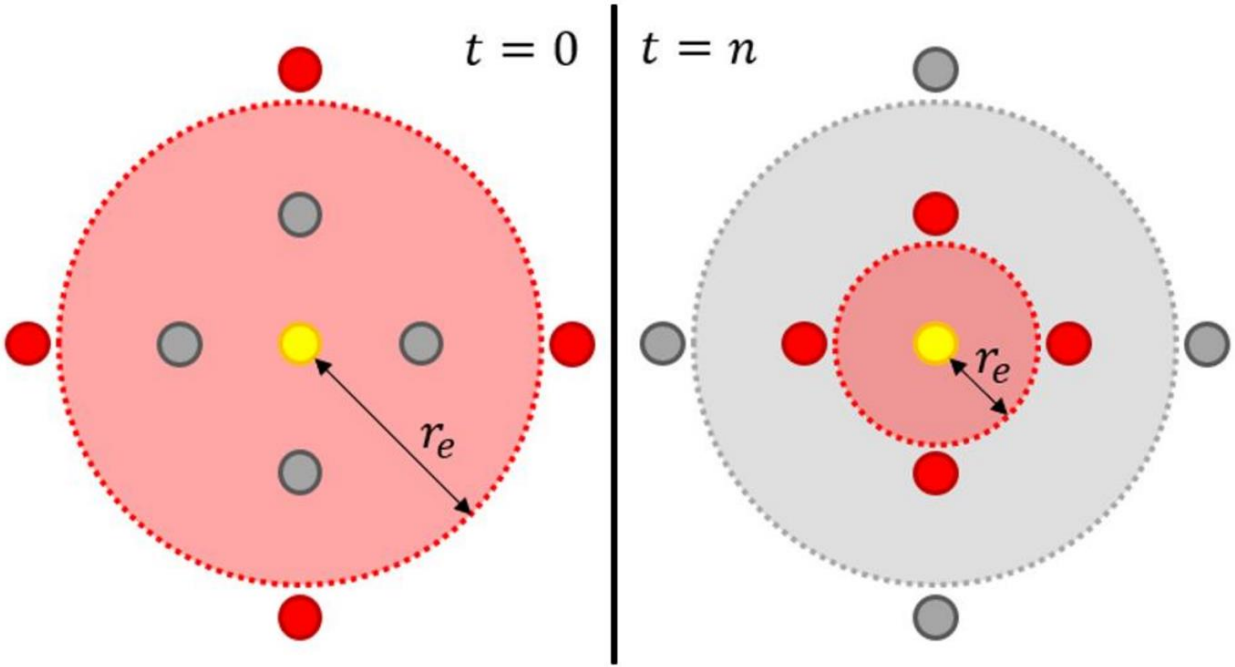


Fan, Lei, et al. "Lung nodule detection based on 3D convolutional neural networks." *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*. IEEE, 2017.

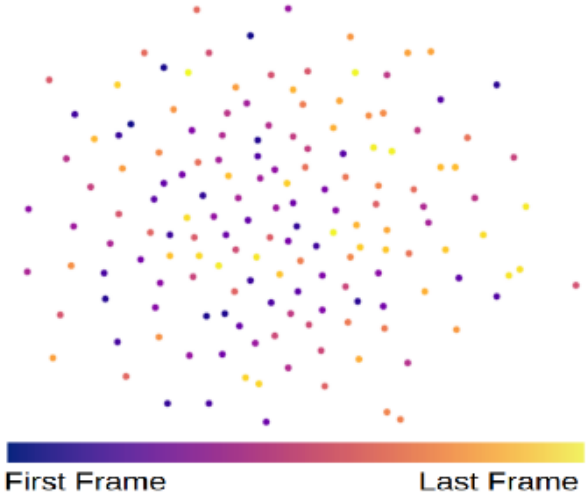
Method: TCE


$\mathcal{V}^1$
$\mathcal{V}^2$
$\mathcal{V}^3$
$\mathcal{V}^m$
ϕ
ϕ
N
Temporally coherent representation
Attraction between adjacent frames
Separation between videos
Embedding Space

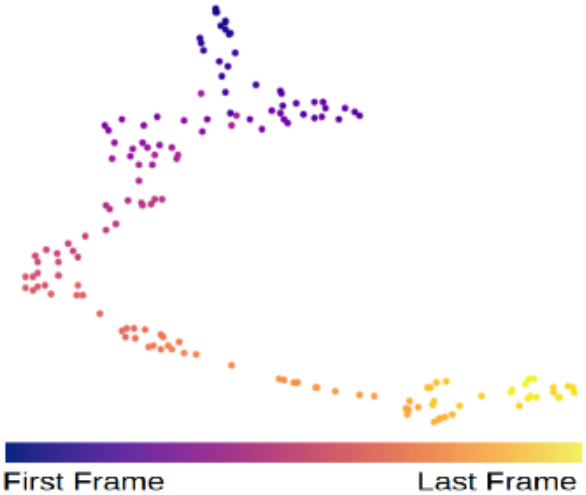
Method: Hard Negative Mining



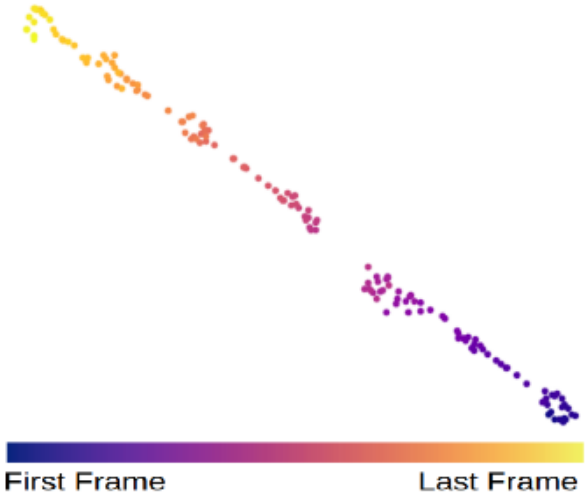
Results: t-SNE Visualisations



(a) Random Initialisation

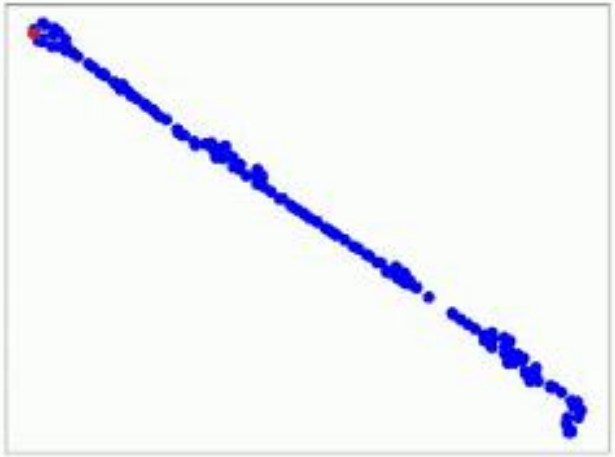


(b) TCE Epoch 25

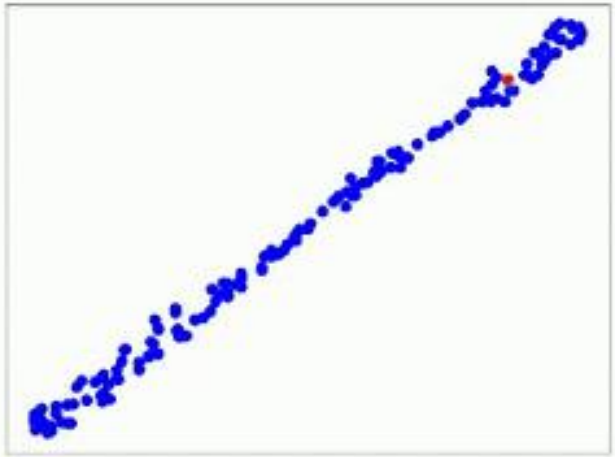


(c) TCE Epoch 50

Results: t-SNE Visualisations



Results: t-SNE Visualisations



Results: State-of-the-art Comparison

Method	Backbone	Params	2D-CNN	Pre-Training	UCF101(%)	HMDB51(%)
3DRotNet [34]	3D ResNet-18	34×10^6	✗	Kinetics400	62.9	33.7
3DCubicPuzzles [10]	3D ResNet-18	34×10^6	✗	Kinetics400	65.8	33.7
DPC [5]	3D ResNet-18†	14×10^6	✗	Kinetics400	68.2	34.5
TCE (Ours)	2D ResNet-18	11×10^6	✓	Kinetics400	68.8^+	34.2
TCE (Ours)	2D ResNet-50	23×10^6	✓	Kinetics400	71.2	**36.6**
DPC [5]	3D ResNet-34†	33×10^6	✗	Kinetics400	**75.7**	35.7
Motion & Appearance [35]	C3D	11×10^6	✗	UCF101	48.6	20.3
Shuffle and Learn [12]	AlexNet	61×10^6	✓	UCF101	50.9^+	19.8
VideoGAN [4]	C3D	11×10^6	✗	UCF101	52.1	-
Arrow of time [36]	AlexNet	61×10^6	✓	UCF101	55.3	-
CMC [17]	CaffeNet $\times 2^*$	$58 \times 10^6 \times 2$	✓	UCF101	55.3	-
OPN [11]	VGG-M-2048	8.6×10^6	✓	UCF101	59.8	23.8
DPC [5]	3D ResNet-18†	14×10^6	✗	UCF101	60.6^+	-
Skip-Clip [8]	3D ResNet-18	34×10^6	✗	UCF101	64.4^+	-
Video Clip Ordering [13]	R3D	14×10^6	✗	UCF101	64.9^+	29.5
***TCE* (Ours)**	2D ResNet-18	11×10^6	✓	UCF101	**68.2^+**	**31.7**

Results: Higher-Order Coherence Ablation

Method	UCF101
First Order Only	68.2
First + Second Order	66.88

- In addition to enforcing similarity between frames, we also investigate minimising curvature of a video's path through the embedding space
- Results suggest that while temporal coherency is a valuable property to train for, enforcing higher order coherencies may result in the loss of other salient properties in the embedding space

Code Release

Github: https://github.com/csiro-robotics/TCE

Webpage: https://csiro-robotics.github.io/TCE-Webpage/

Thank you

Joshua Knights

Embodied AI Team
Commonwealth Scientific and Industrial Research Organisation
joshua.knights@csiro.au

School of Electrical Engineering and Computer Science
Queensland University of Technology
joshua.knights@hdr.qut.edu.au