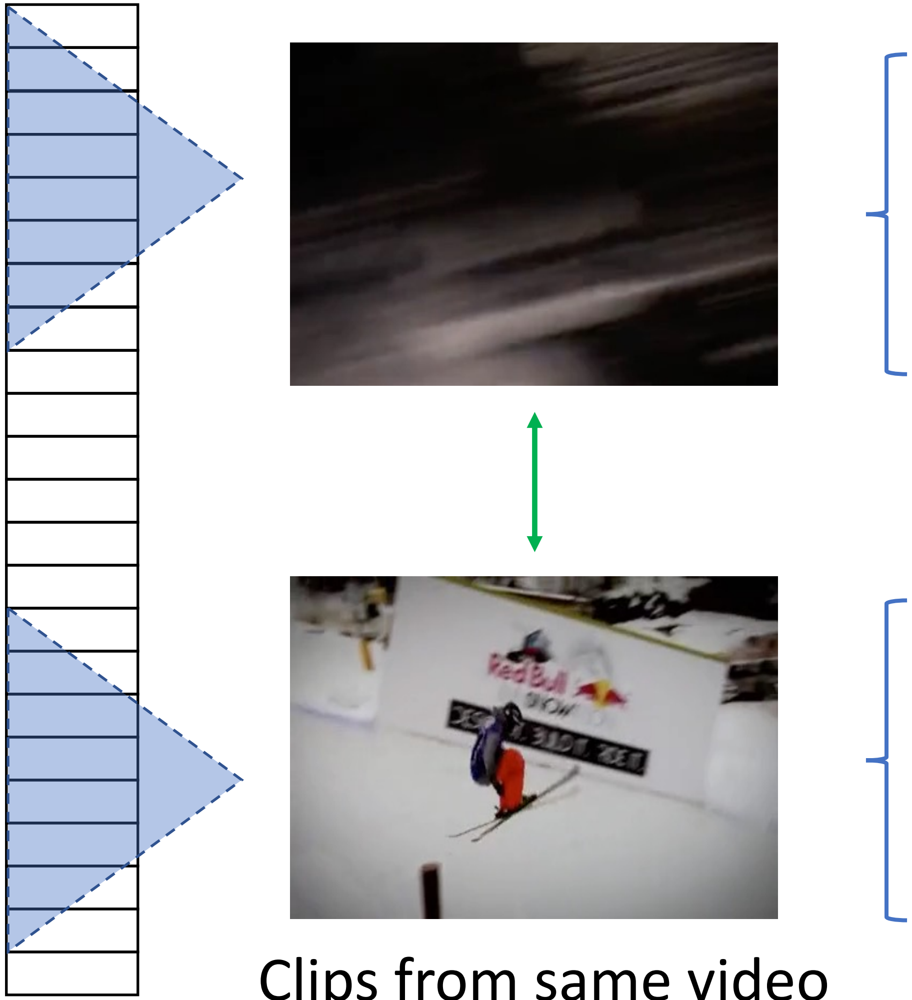
Look Globally and Locally: Inter-Intra Contrastive Learning from Unlabeled Videos

ICLR 2023 ME-FoMo Workshop

David Fan, Deyu Yang, Xinyu Li, Vimal Bhat, Rohith MV
Amazon Prime Video

Dominant self-supervised contrastive learning methods for video focus on learning relationships of clips within the **same** video.

Video

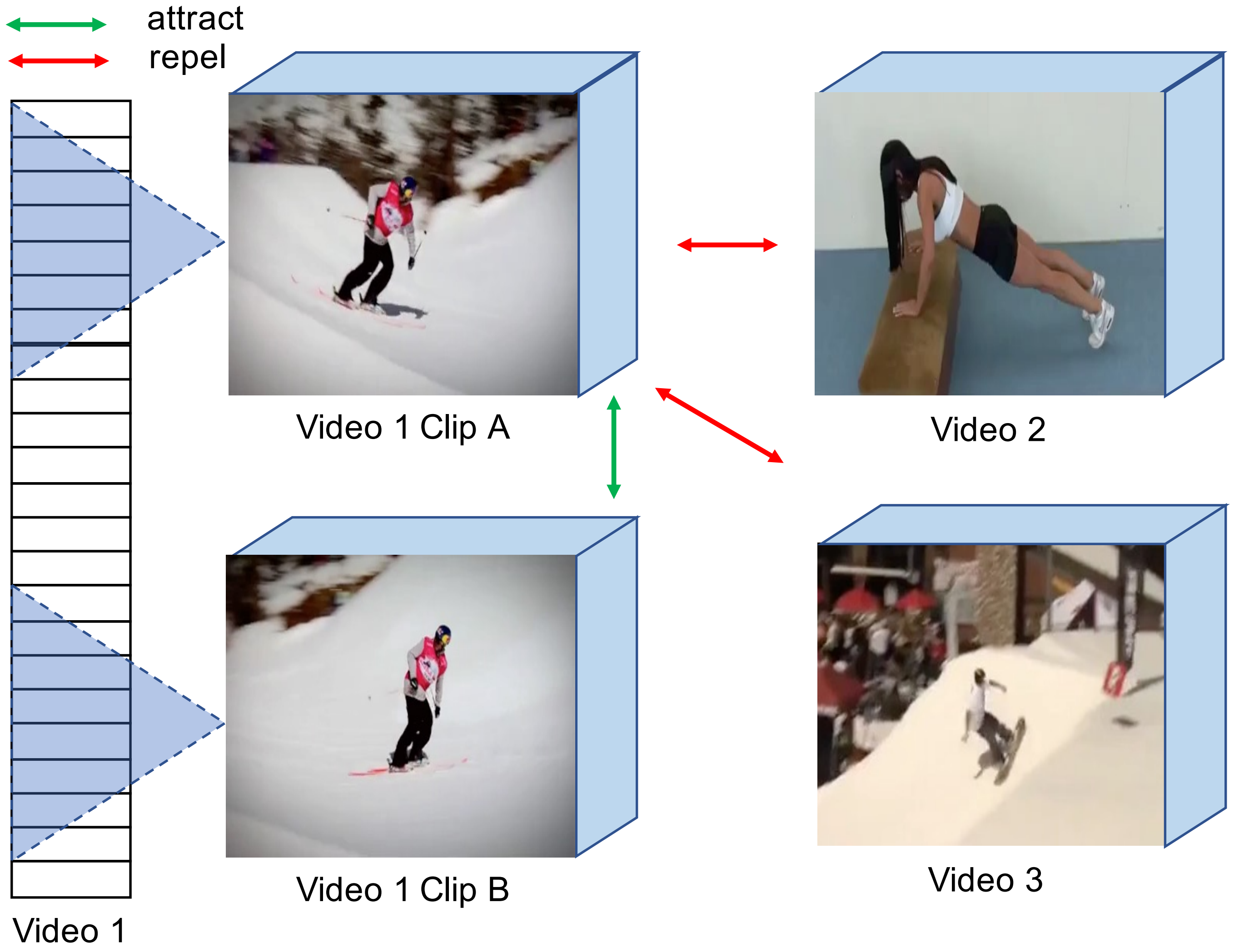


Clips from same video

Same video!

- Spatiotemporal augmentation
- Clip shuffling / frame shuffling
- Speed-up
- Flip directionality
- Frame inpainting

What happens to other videos?



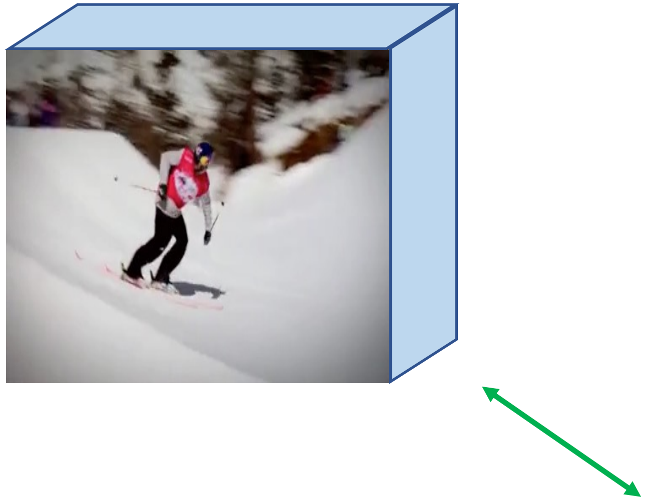
- Pushup and *snowboarding* are both negative to skiing anchor.
  - But *snowboarding*-skiing is more similar than pushup-skiing.
- When positives are only sampled from the <u>same video</u> (”intra-video”), other similar videos will never be leveraged as positives.

Should we?

prime video

Intra-Video Similarity

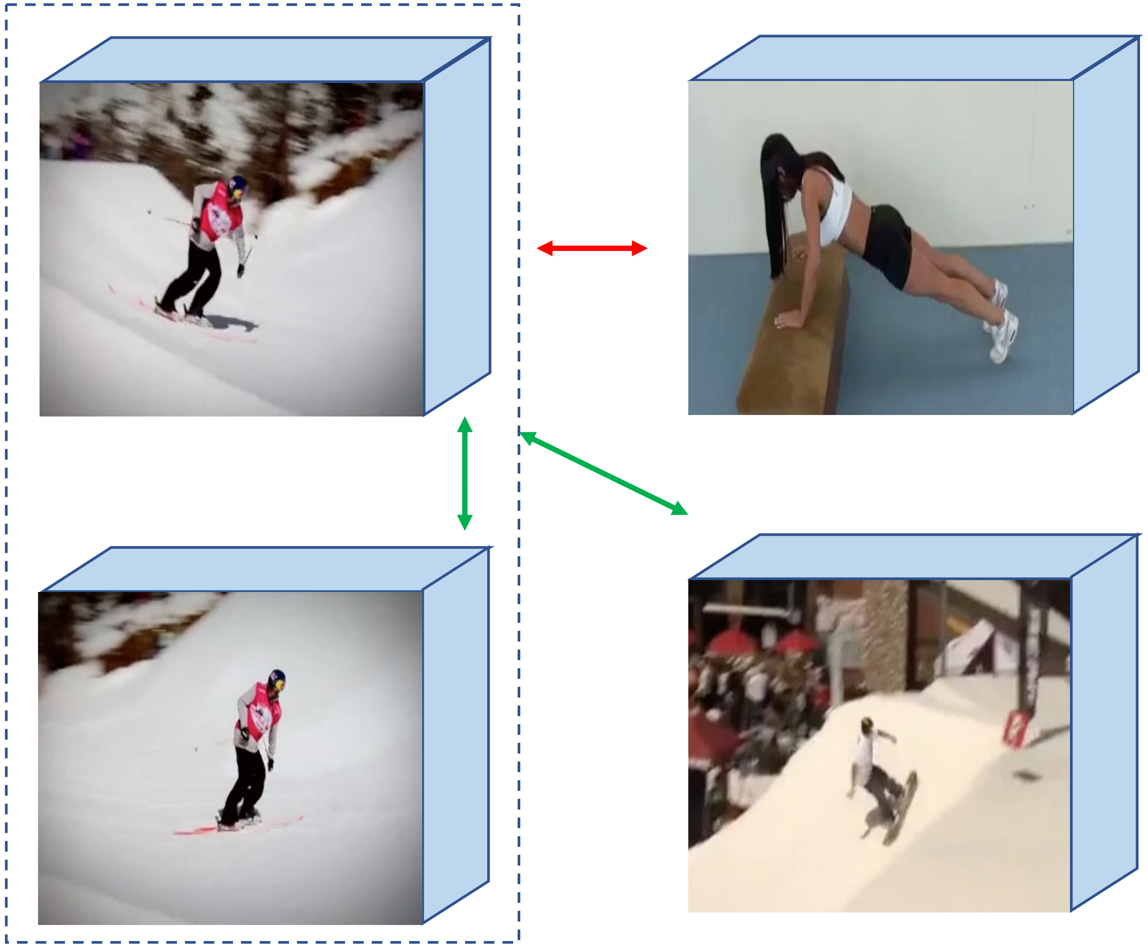
Inter-Video Similarity



- <u>Intra</u>-video similarity -> fine-grained details.
- <u>Inter</u>-video similarity -> hierarchies of visual concepts.
  - E.g. winter sports tend to be more similar to each other than to random sports.
- Is local or global information more important for video representations?

Inter-Intra Contrastive Learning (IICVL)

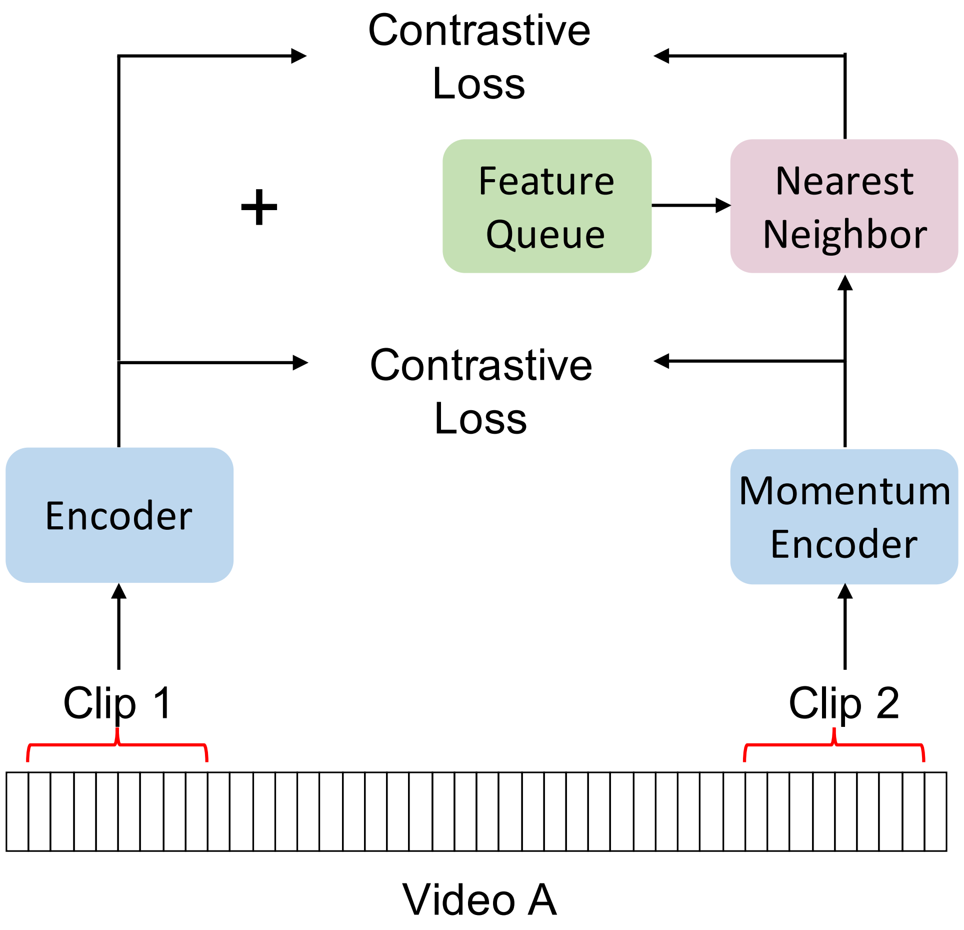
- We propose balancing local and global information through leveraging **nearest-neighboring videos** sampled from a global space, as additional positives for a second contrastive loss term.



Existing Works that use Intra-Video Sampling

Inter-Intra Contrastive Learning (IICVL)

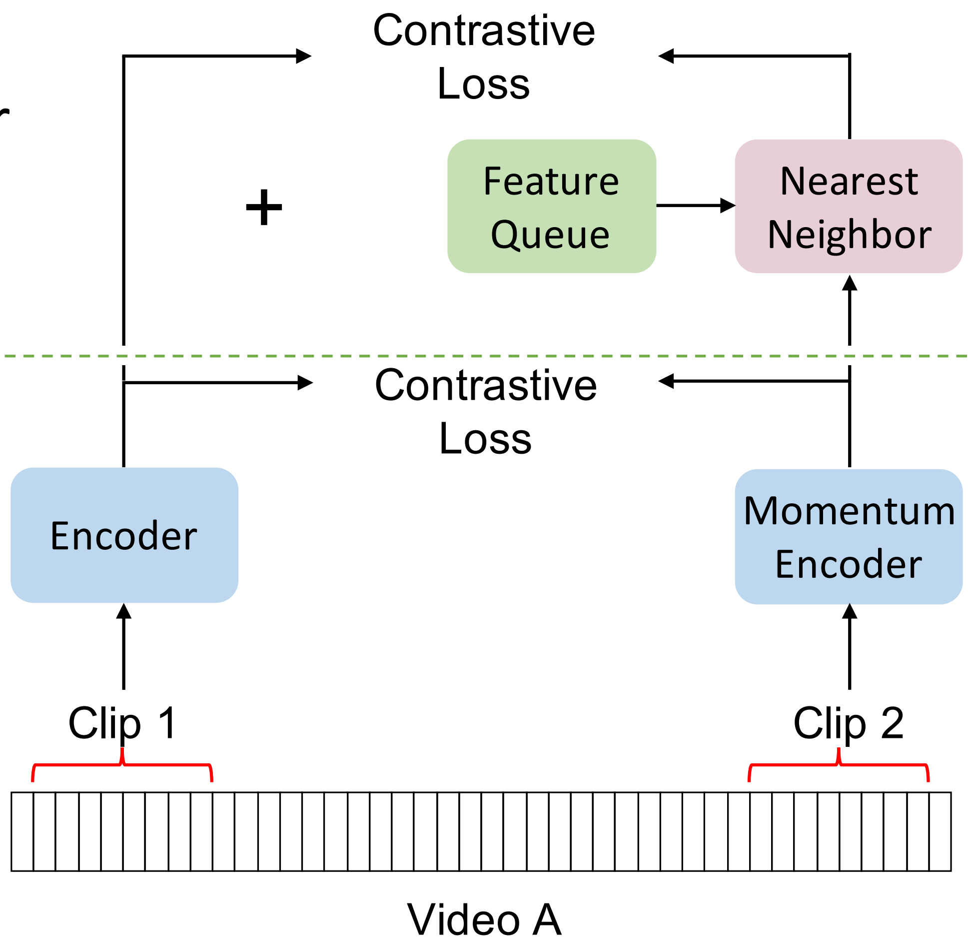
- We leverage a dynamically evolving feature queue to find video NNs.
- Unlike other works, we do not use clustering nor multiple modalities.



prime video

Inter-Intra Contrastive Learning (IICVL)

- Additional contrastive loss term using nearest-neighbor positives improves performance over CVRL and p-MoCo baseline.
- Simple, light-weight, yet effective.



(CVRL)

Spatiotemporal Contrastive Video Representation Learning

Rui Qian* [1,2,3] Tianjian Meng* [1] Boqing Gong[1] Ming-Hsuan Yang[1]
Huisheng Wang[1] Serge Belongie[1,2,3] Yin Cui[1]

[1]Google Research [2]Cornell University [3]Cornell Tech

(p-MoCo)

A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning

Christoph Feichtenhofer Haoqi Fan Bo Xiong Ross Girshick Kaiming He
Facebook AI Research (FAIR)

Inter-Intra Contrastive Learning (IICVL)

$$\mathcal{L}(z_1^{\text{Intra}}, z_2^{\text{Intra}}, z_1^{\text{NN}}, z_2^{\text{NN}}) = \boxed{\lambda_{Intra}} \cdot \mathcal{L}_{Intra}(z_1^{\text{Intra}}, z_2^{\text{Intra}}, Q_{\text{Intra}}) + \boxed{\lambda_{NN}} \cdot \mathcal{L}_{NN}(z_1^{\text{NN}}, z_2^{\text{NN}}, Q_{\text{NN}}) \quad (5)$$

- We use weights to control the contribution of intra and inter-video similarity
- When λ_{Intra} is 1.0 and λ_{NN} is 0.0, our method is equivalent to p-MoCo and CVRL.
- But code is not available for neither works. We thus reimplemented p-MoCo.
- For fair comparison, we compare to our reimplementation as the baseline.

Spatiotemporal Contrastive Video Representation Learning

Rui Qian* [1,2,3] Tianjian Meng* [1] Boqing Gong[1] Ming-Hsuan Yang[1]
Huisheng Wang[1] Serge Belongie[1,2,3] Yin Cui[1]

[1]Google Research [2]Cornell University [3]Cornell Tech

A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning

Christoph Feichtenhofer Haoqi Fan Bo Xiong Ross Girshick Kaiming He

Facebook AI Research (FAIR)

Results

- We pretrained on unlabeled K400 (~240K videos) for 200 epochs.
- Then we transfer our model weights to downstream tasks to evaluate our model's generalizability.
  - Action recognition:
    - UCF101 (finetune)
    - HMDB51 (finetune)
    - K400 (linear eval)
    - SSv2 (finetune)
  - Action Detection
    - AVA (finetune)
  - Few-shot action recognition
  - Video retrieval

Action Recognition on UCF, HDMB, K400

Method	Date	Backbone	Pretrain Data (duration)	Pretrain Epochs	Pretrain Input Size	UCF	HMDB	K400
Supervised		R3D-50	scratch		8×224^2	68.8	22.7	74.7
DPC [26]	2019	R2D-3D34	K400 (28d)	110	40×224^2	75.7	35.7	-
CBT [56]	2019	S3D	K600+(273d)	130	112^2	79.5	44.6	-
DynamoNet [16]	2019	STCNet	YT8M-1 (58d)	-	32×112^2	88.1	59.5	-
SpeedNet [2]	2020	S3D-G	K400 (28d)	-	16×224^2	81.1	48.8	-
MemDPC [27]	2020	R2D-3D34	K400 (28d)	-	40×224^2	86.1	54.5	-
VideoMoCo [50]	2021	R(2+1)D18	K400 (28d)	200	32×224^2	78.7	49.2	-
TCLR [13]	2021	R(2+1)D18	K400 (28d)	400	16×112^2	84.1	53.6	-
VCLR [38]	2021	R2D-50	K400 (28d)	400	32×224^2	85.6	54.1	64.1
LSFD [1]	2021	R3D-18	K400 (28d)	500	16×224^2	77.2	53.7	-
TECVRL [33]	2021	R3D-18	K400 (28d)	200	16×128^2	87.1	63.6	-
IIVCL		R3D-18	K400 (28d)	200	8×128^2	89.4	60.2	59.2
VTHCL [68]	2020	R3D-50	K400 (28d)	200	8×224^2	82.1	49.2	37.8
CVRL [52]	2020	R3D-50	K400 (28d)	1000	16×224^2	92.2	66.7	66.1
ρ-MoCo† [20]	2021	R3D-50	K400 (28d)	200	8×224^2	91.1	65.3	65.4
ρ-MoCo† [20]	2021	R3D-50	K400 (28d)	400	8×224^2	92.5	-	67.4
IIVCL		R3D-50	K400 (28d)	200	8×224^2	92.6	65.8	65.7
IIVCL		R3D-50	K400 (28d)	400	8×224^2	93.3	68.1	67.1

Table 1: **Comparison with state-of-the-art self-supervised approaches**. Reported results are top-1 accuracy under finetune protocol (UCF, HMDB) and linear protocol (K400). We do not compare against two-stream methods.
† refers to our reimplementation (see Sec. 4.1).

prime video

Action Recognition on Something-Something

Method	Backbone	Pretrain Data	Top-1 Acc
Supervised [19]	R3D-50	K400	52.8
ρ-MoCo [20]	R3D-50	K400	53.6
IIVCL	R3D-50	K400	53.8

Table 2: **Action recognition on Something-Something.** We finetune on SSv2 using a clip size of 8×8 and report top-1 accuracy.

- SSv2 is very different from K400. Videos in SSv2 are highly fine-grained
  - E.g. “putting something **into** something” vs. “putting something **next to** something”
- Our model still outperforms baseline for a downstream dataset that is very different from the unlabeled pretraining data source.

Action Detection on AVA

- Besides action recognition … what if we evaluate on a new task?

Method	Pretrain Data	Top-1 Acc
Supervised [19]	K400	21.9
CVRL [52]	K400	16.3
ρ-MoCo [20]	K400	18.6
IIVCL	K400	19.0

Table 3: **Action detection on AVA.** We finetune on AVA using a clip size of 8×8 and report mAP@0.5 IOU.

- Our method outperforms the baseline on five different downstream datasets spread across two tasks. Thus more generalizable.

Few-Shot Learning

- What if the amount of downstream training data is limited to 1-10% of the original dataset size?

	UCF Finetune			K400 Linear	
Method	1%	5%	20%	1%	10%
Supervised (Scratch)				3.2	39.6
ρ-MoCo [20]	41.8	68.0	84.7	34.3	53.3
IIVCL	44.3	68.9	85.0	34.9	54.2
Δ	**+2.5**	+0.9	+0.3	+0.6	+0.9

Table 4: **Few-shot learning on UCF101 and K400.** Rows indicate different pretrained models on K400. Columns vary the % of UCF training data used for finetuning and % of K400 training data used for linear eval.

- Delta between our method and baseline is highest for **1%** subset. Thus our method is highly performant in low-data settings.

prime video

Video Retrieval

- Embeddings are extracted using model pretrained on K400. No finetuning is done on downstream dataset in retrieval setting.

			UCF				HMDB			
Method	Network	Pretrain	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
SpeedNet [2]	S3D-G	K400	13.0	28.1	37.5	49.5				
GDT [51]	R(2+1)D	K400	57.4	73.4	80.8	88.1	25.4	51.4	63.9	75.0
VCLR [38]	R2D-50	K400	70.6	80.1	86.3	90.7	35.2	58.4	68.8	79.8
ρ-MoCo [20]	R3D-50	K400	73.2	87.0	91.8	95.5	36.3	61.9	72.0	82.5
IIVCL	R3D-50	K400	74.2	87.6	92.1	95.1	37.6	62.2	72.9	82.5

Table 7: **Zero-shot video retrieval on UCF101 and HMDB.** We do not compare against two-stream methods. This work only uses RGB.

prime video

Intra vs. Inter Video Similarity

- To address the question from the intro …
- Our results suggest balancing local + global similarity is ideal for video.
- But surprisingly, pure NN is nearly on-par with the pure Intra baseline.
- How to best balance this local and global information? Future work.

	Model		Action Recognition				Action Detection	
			Finetune			Linear	Finetune	**Avg.**
	λ_{Intra}	λ_{NN}	UCF	HMDB	SSv2	K400	AVA	**Rank**
p-MoCo / CVRL baseline	1.0	0.0	91.1 (#3)	65.3 (#3)	53.6 (#2)	65.4 (#2)	18.6 (#2)	2.4
	1.0	1.0	**92.6** (#1)	65.8 (#2)	**53.8** (#1)	**65.7** (#1)	**19.0** (#1)	1.2
Video analog of pure NNCLR	0.0	1.0	91.2 (#2)	**66.2** (#1)	53.2 (#3)	63.7 (#3)	18.4 (#3)	2.4

Table 6: **Do NNs lead to better generalization?** The first row corresponds to the *ρ-MoCo* baseline and second row corresponds to IIVCL. All models are pretrained on full K400 for 200 epochs. Downstream eval uses clip size of 8×8. λ_{Intra}=0.0 means no intra-video positives are used (NNCLR for video). We denote rank in blue parenthesis (where 1st = best) on each task to show the generalization of each model.

prime video

Does Longer Pretraining and More NNs Help?

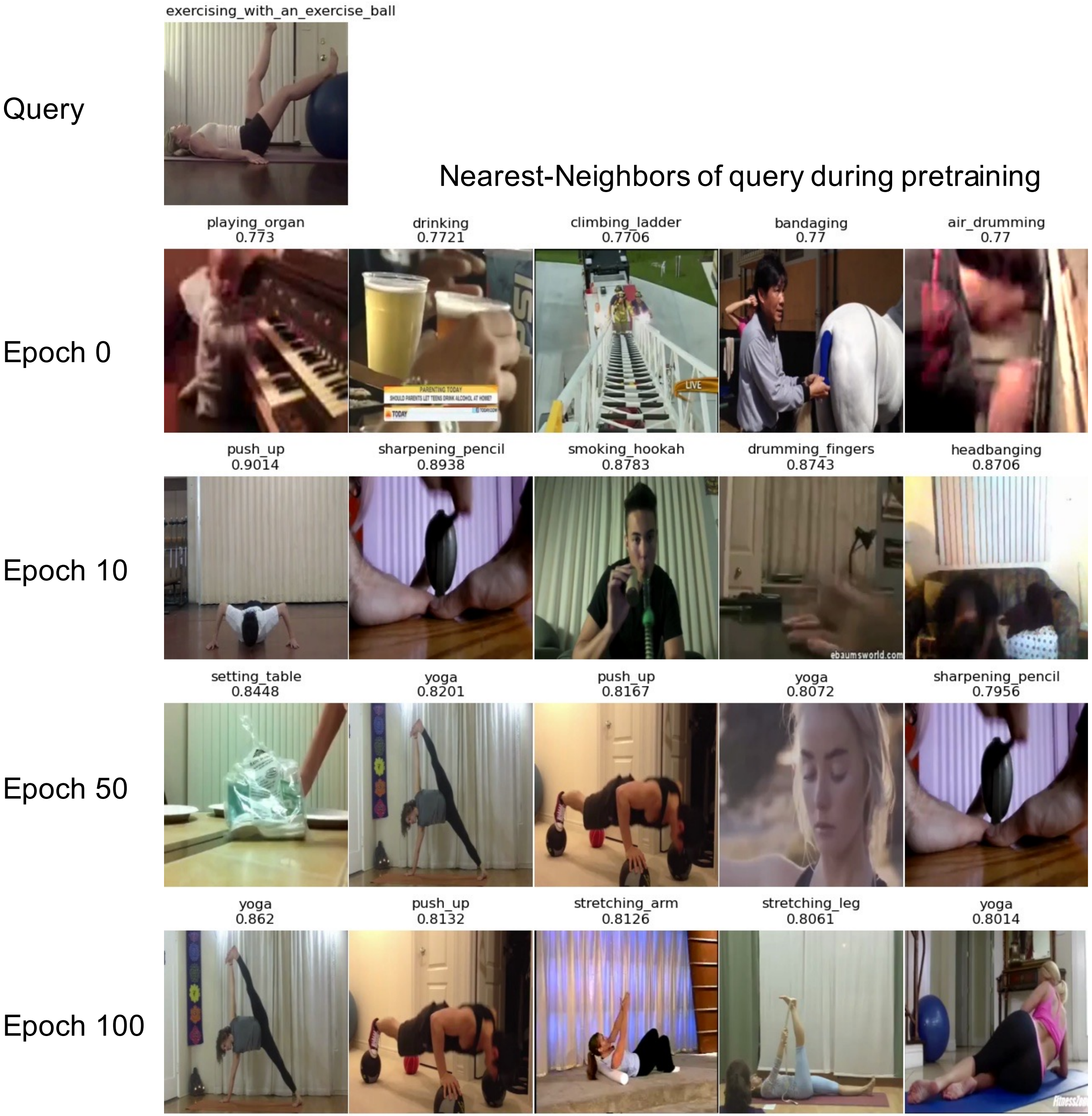
- Yes! But more expensive.

Epochs	ρ	UCF	HMDB	K400	SSv2
200	2	92.6	65.8	65.7	53.8
200	4	93.3	67.8	66.6	54.6
400	2	93.3	68.1	67.1	54.2

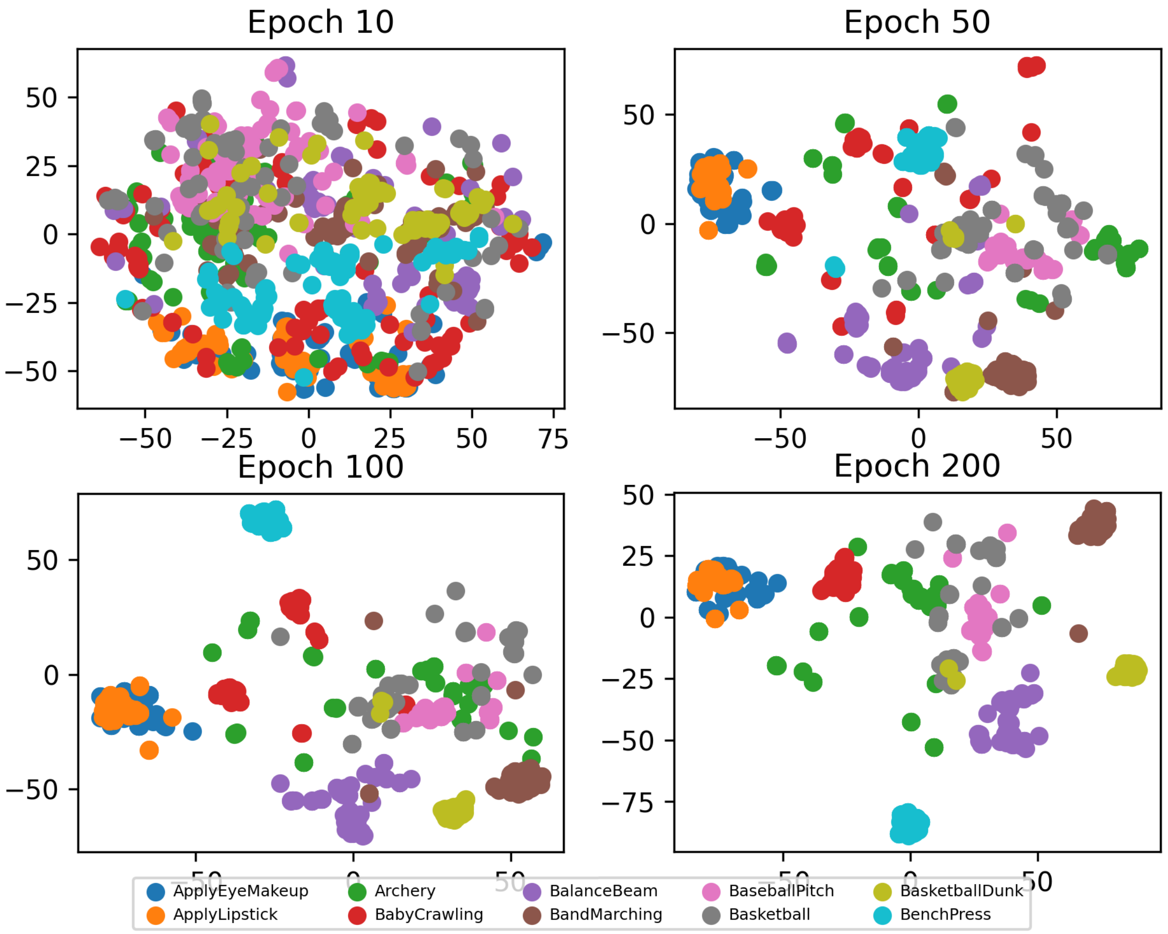
Table 5: **More pretraining epochs and NNs.** Data is unlabeled K400.

What does the model learn?

- From random initialization, our model is able to progressively learn semantic similarity.
- Our model can leverage similarity <u>across class boundaries</u>. Here query is “exercise ball” and the top-NN at epoch 100 is “yoga”.
- More diverse notion of similarity balanced with intra-video sampling leads to improved representation.



T-SNE Visualization during SSL Pretraining



Our method learns a progressively improved notion of semantic similarity without explicit supervision.

Features extracted from UCF101 using model pretrained on unlabeled K400 (no finetuning on UCF)

prime video

Conclusions

- Dominant contrastive learning works are focused on intra-video similarity.
- We are the first to leverage both intra and inter-video similarity for contrastive learning, to learn a balanced view of global and global information.
- Existing video works that go beyond single-video use clustering, but instead we use nearest-neighbors. Our method scales and is simple.
- We outperform baseline on five different datasets and present detailed analysis of model performance.

Thanks for watching!

Reach out to me at fandavi@amazon.com for further questions or interest in collaborations.

Nearest-Neighbor Inter-Intra Contrastive Learning from Unlabeled Videos

David Fan Deyu Yang Xinyu Li Vimal Bhat Rohith MV
Amazon Prime Video
{fandavi, deyu, xxnl, vimalb, kurohith}@amazon.com

prime video