2021 ICCV OCTOBER 11-17 VIRTUAL

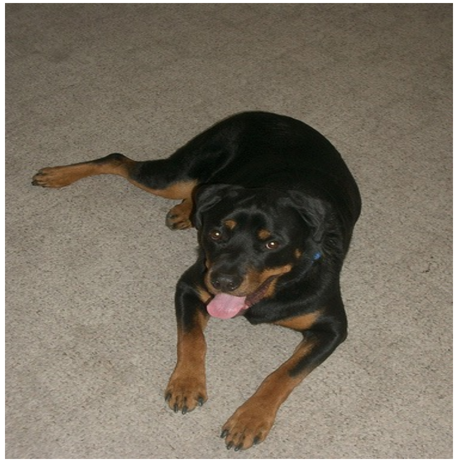
Contrastive Learning of Image Representations with Cross-Video Cycle-Consistency

Haiping Wu Xiaolong Wang

McGill University, Mila UC San Diego

Motivation

- Most contrast learning methods build on *intra-image* invariance learning or *intra-video* invariance learning.



Augmentation $\longrightarrow$ Encoder q

Augmentation’ $\longrightarrow$ Encoder’ k

losses push q and k close
i.e. *intra-image* invariance learning

MoCo, SimCLR, BYOL, etc.

MoCo: He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." *CVPR* 2020.
SimCLR: Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." PMLR 2020.
BYOL: Grill, Jean-Bastien, et al. "Bootstrap your own latent: A new approach to self-supervised learning." NeurIPS 2020
Image credit: ImageNet

Motivation

- Most contrast learning methods build on *intra-image* invariance learning or *intra-video* invariance learning.

image or video clip A

Augmentation

Encoder

q

From the same video

losses push q and k close

i.e. *intra-video* invariance learning

image or video clip B

Augmentation'

Encoder'

k

VINCE, CVRL, ρBYOL, etc.

VINCE: Gordon, Daniel, et al. "Watching the world go by: Representation learning from unlabeled videos." *Arxiv* 2020

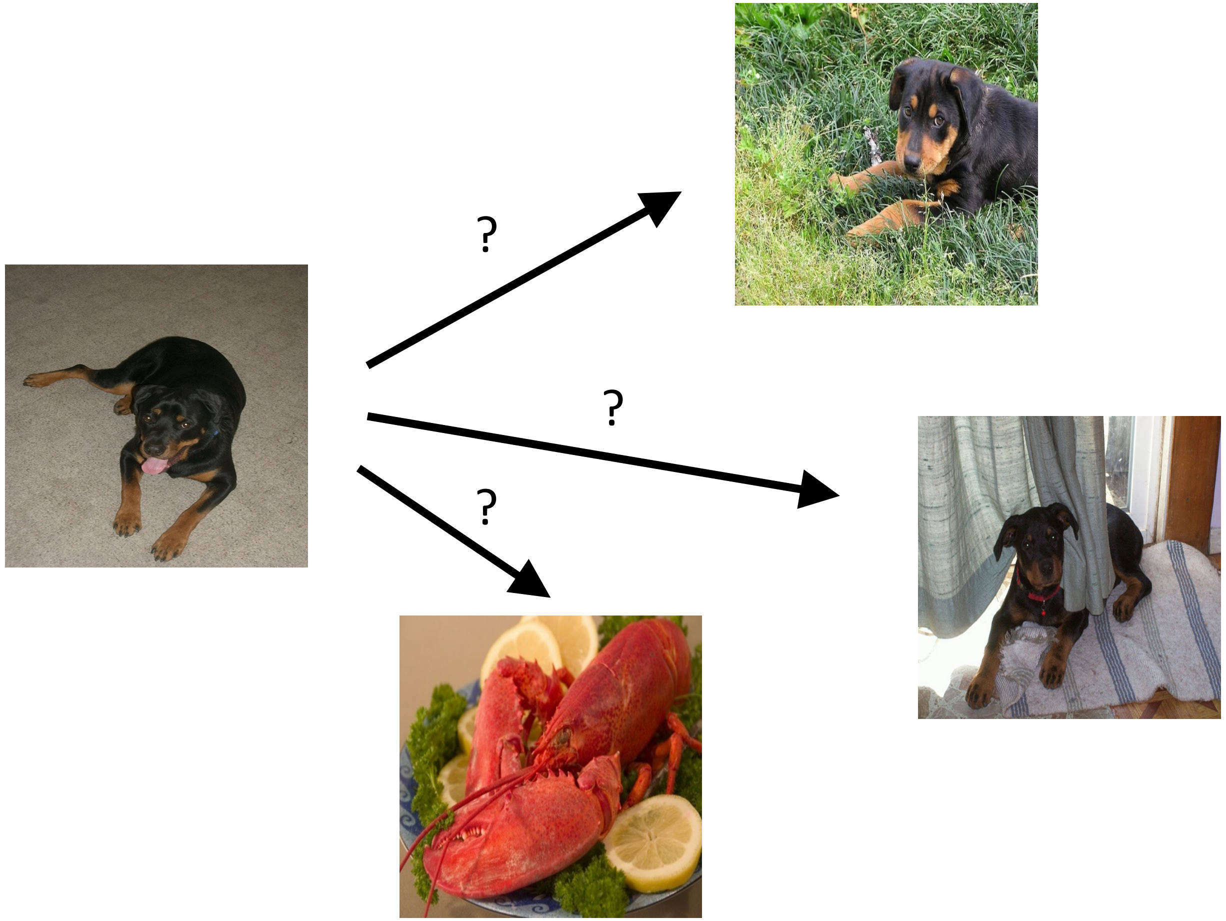
CVRL: Qian, Rui, et al. "Spatiotemporal contrastive video representation learning." CVPR 2021

ρBYOL: Feichtenhofer, Christoph, et al. "A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning." *CVPR* 2021

Image credit: UCF101

Motivation

- What about inter-image relationships?

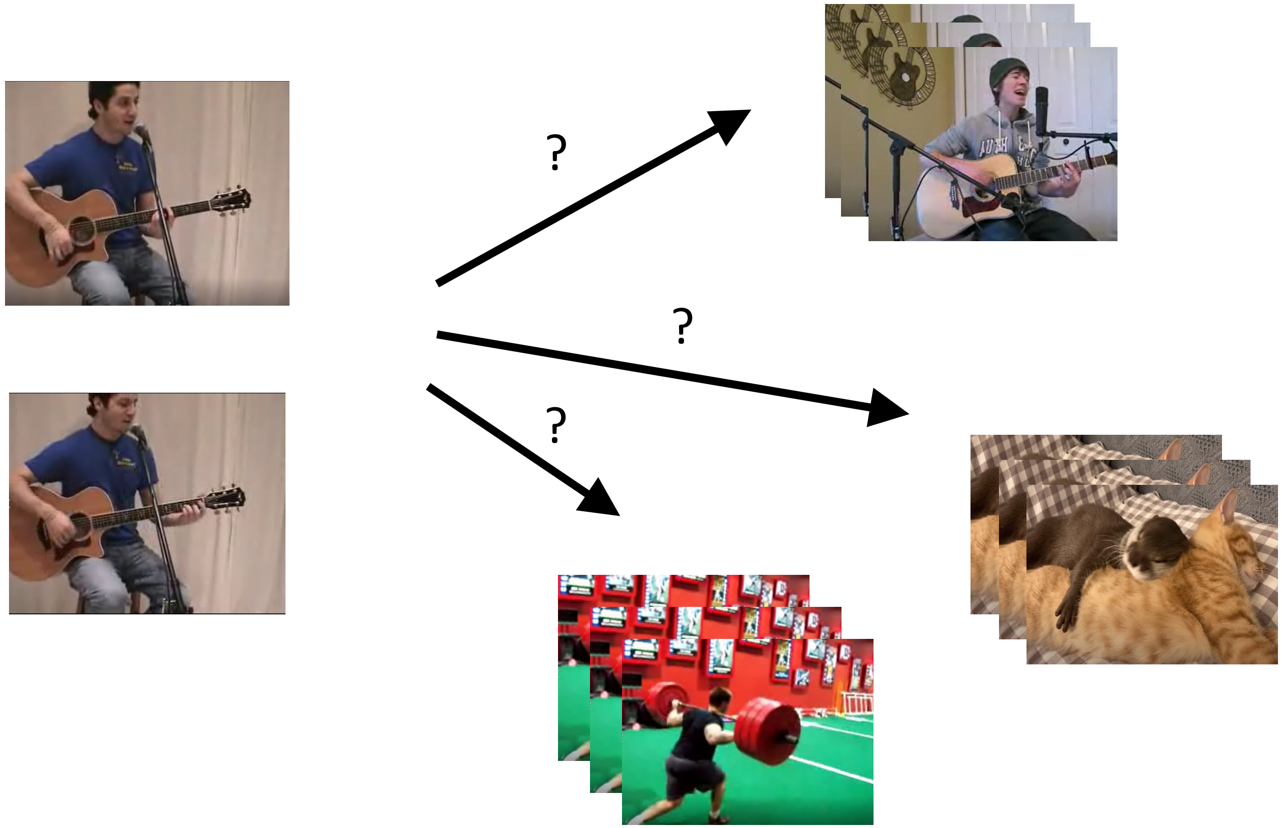


Ideally, **intra-class** invariance is desired, however, no class labels are available

Image credit: ImageNet

Motivation

- What about inter-video relationships?



Ideally, **intra-class** invariance is desired, however, no class labels are available

Image credit: ImageNet

Motivation

- Most contrast learning methods build on *intra-image* invariance learning or *intra-video* invariance learning.
- Fail to explicitly regularize feature presentations belong to the same class.
- How to solve?

-> use clustering to generate pseudo labels (e.g. DeepClustering, Local Aggregation)

DeepClustering: Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." *ECCV 2018*.
Local Aggregation: Zhuang, Chengxu et al. "Local aggregation for unsupervised learning of visual embeddings." *CVPR* 2019.

Motivation

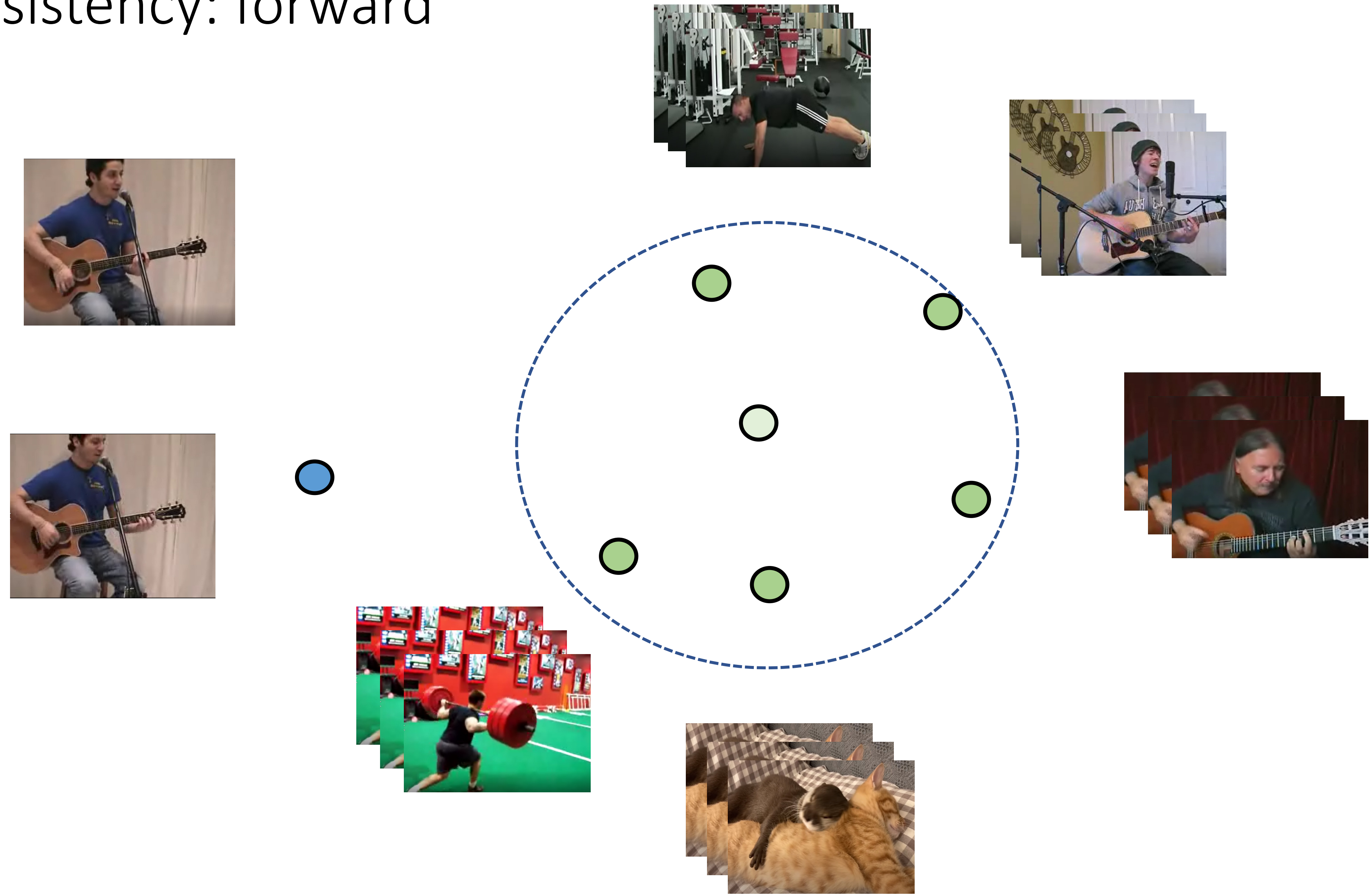
- Most contrast learning methods build on *intra-image* invariance learning or *intra-video* invariance learning.
- Fail to explicitly regularize feature presentations belong to the same class.
- How to solve?

-> use clustering to generate pseudo labels (e.g. DeepClustering, Local Aggregation)

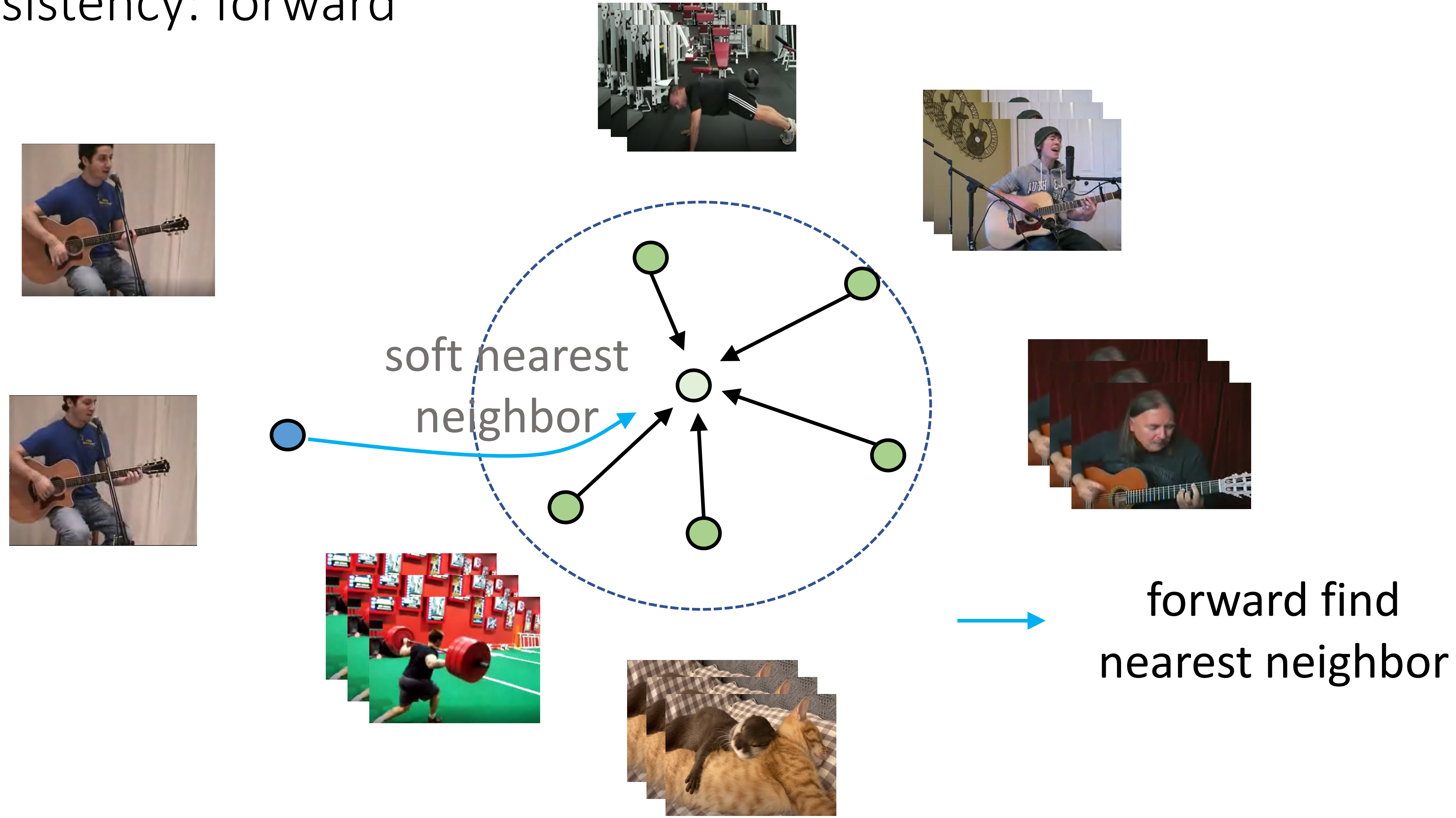
-> this paper: use *cycle-consistency without* generating pseudo labels

DeepClustering: Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." *ECCV 2018*.
Local Aggregation: Zhuang, Chengxu et al. "Local aggregation for unsupervised learning of visual embeddings." *CVPR* 2019.

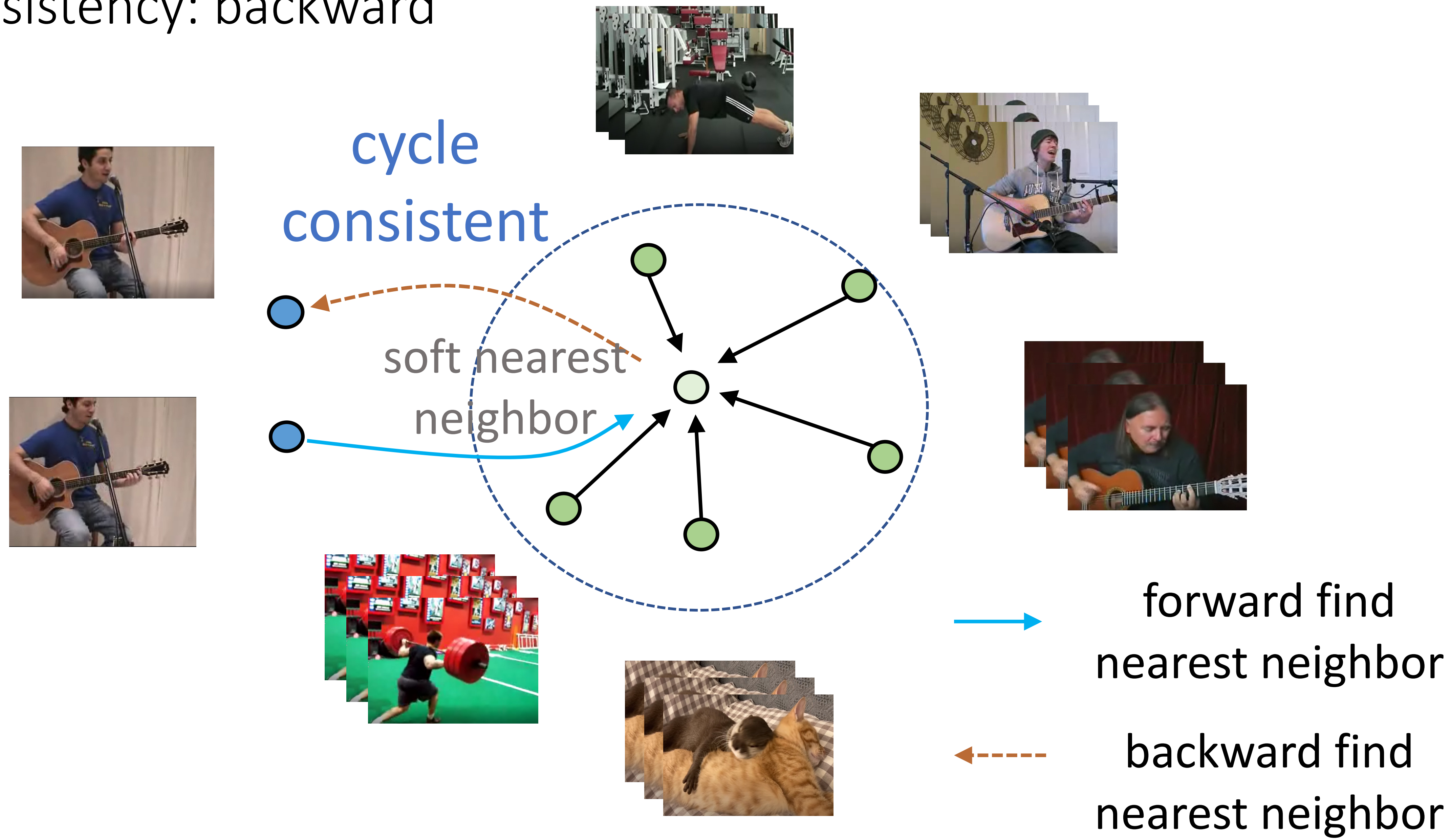
Cycle consistency: forward



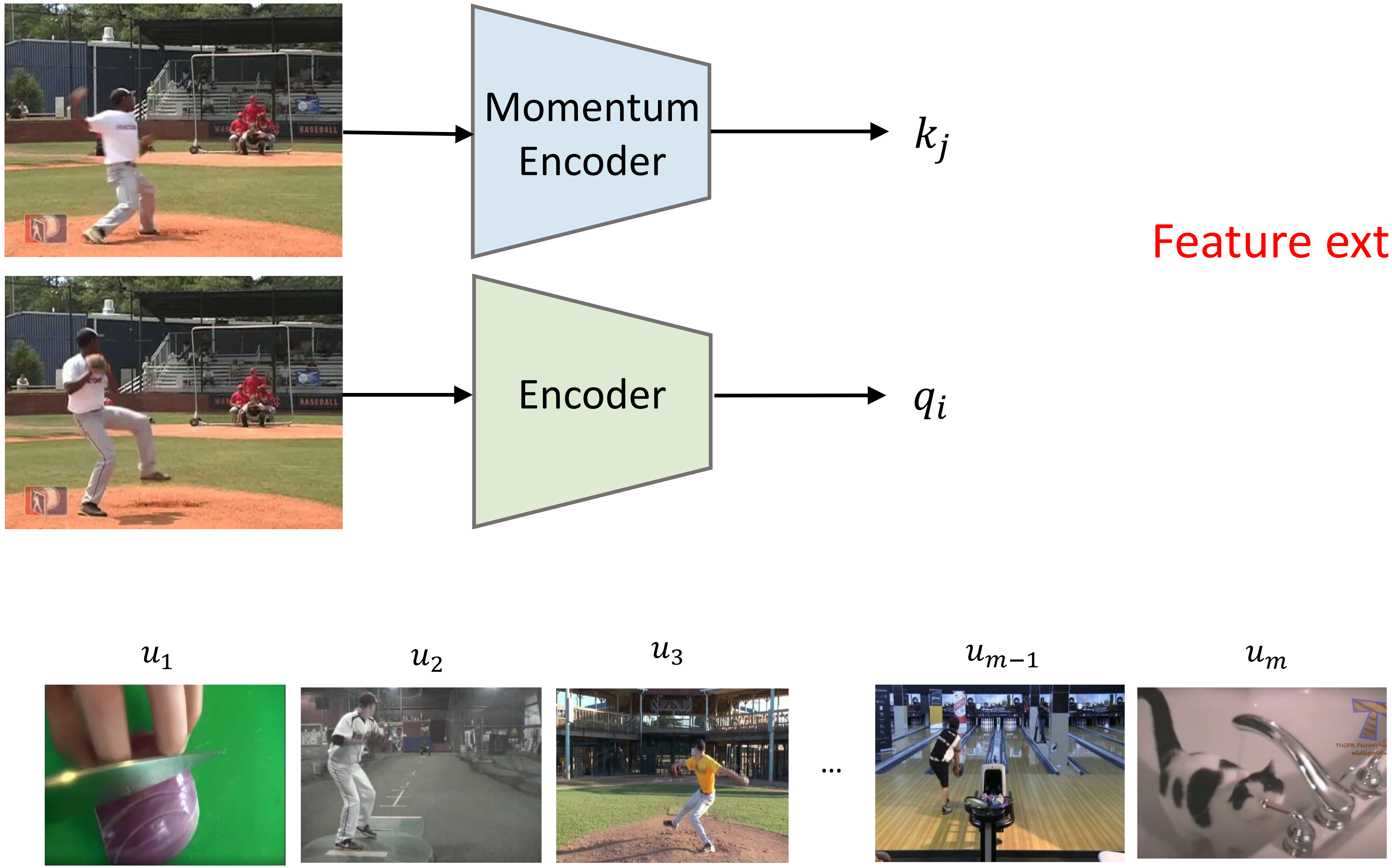
Cycle consistency: forward



Cycle consistency: backward



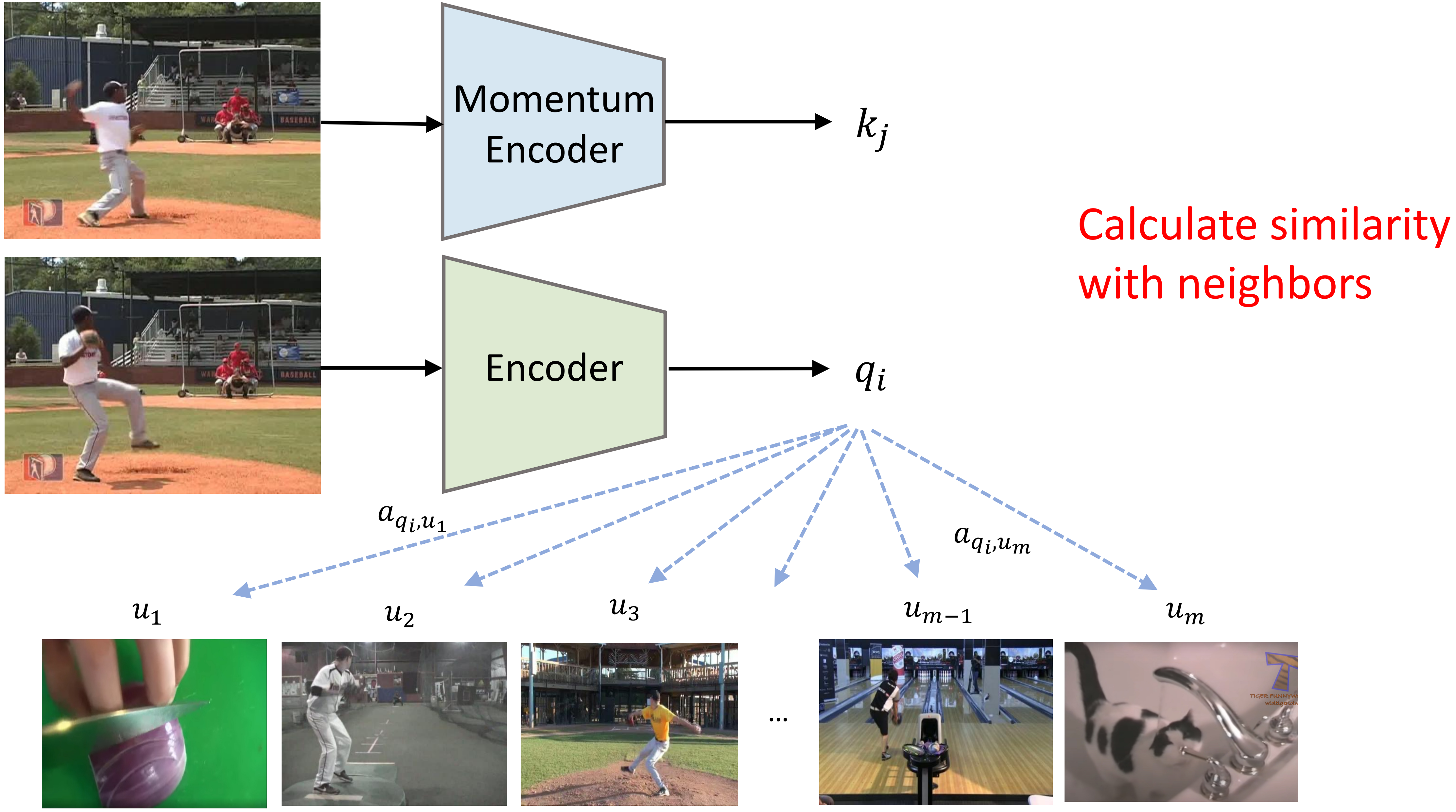
Overall framework



Feature extraction

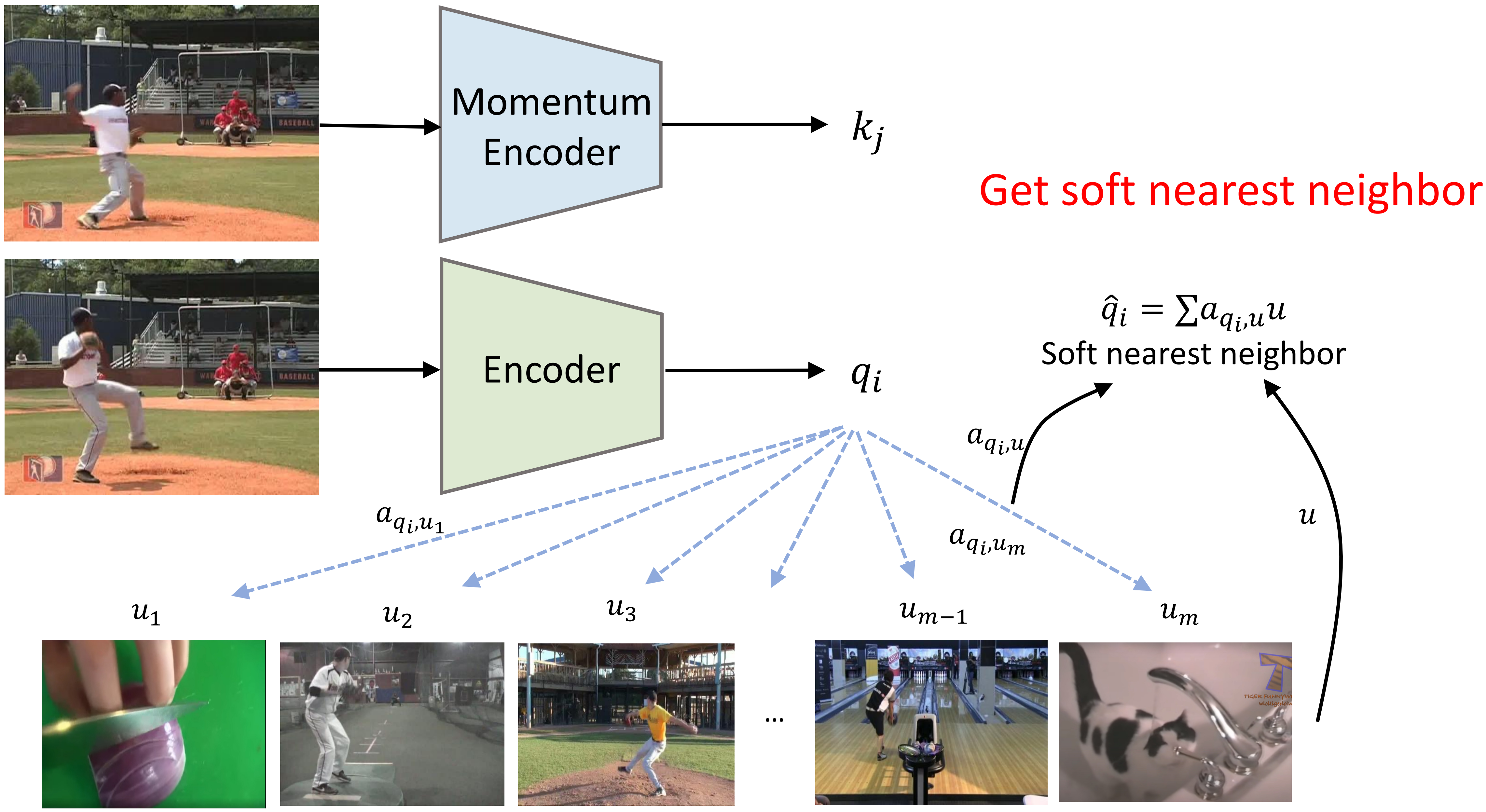
Neighbor representation set $U = \{u_1, u_2, \ldots, u_m\}$ from random videos

Overall framework



Neighbor representation set $U = \{u_1, u_2, \ldots, u_m\}$ from random videos

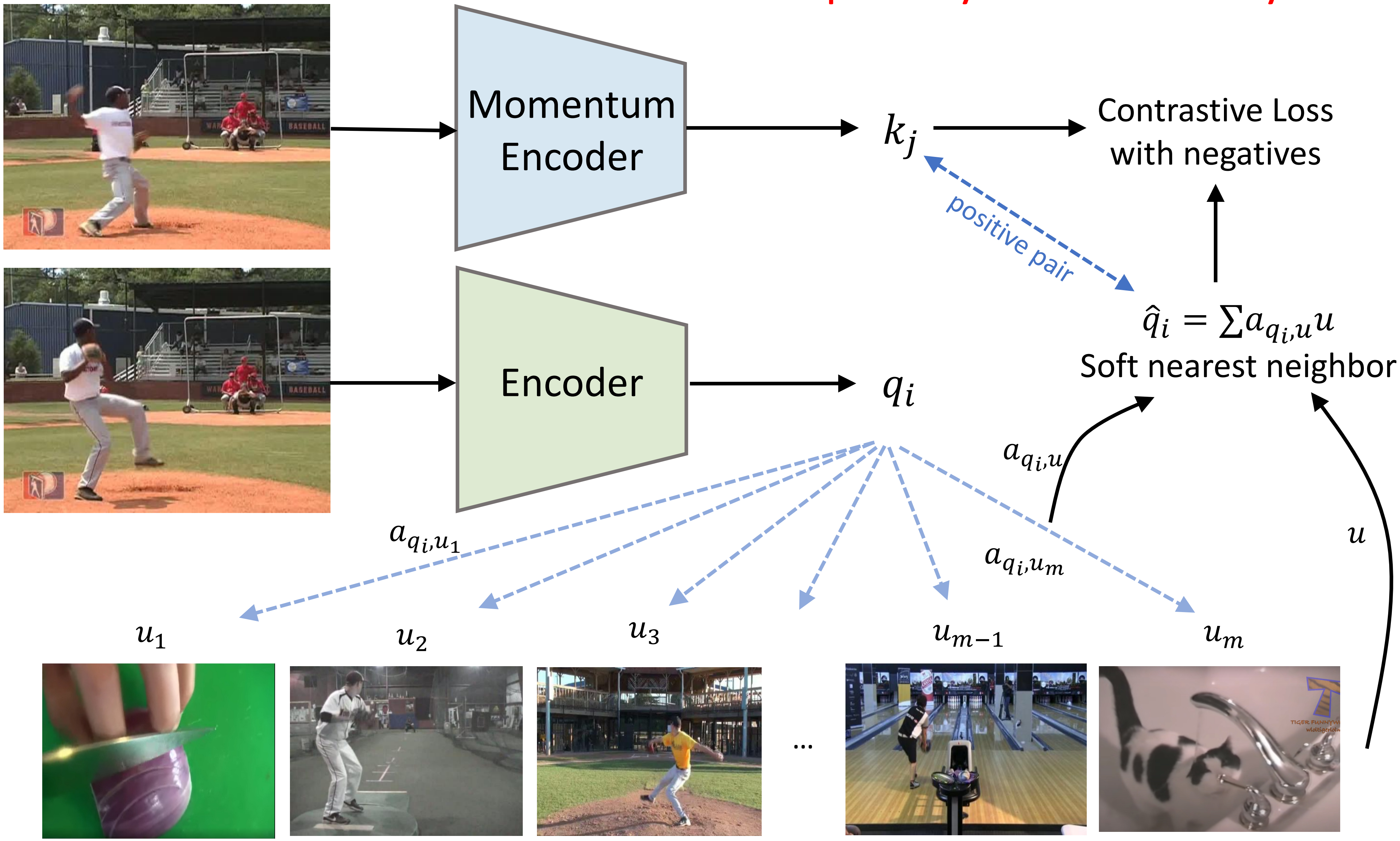
Overall framework



Neighbor representation set $U = \{u_1, u_2, \dots, u_m\}$ from random videos

Overall framework

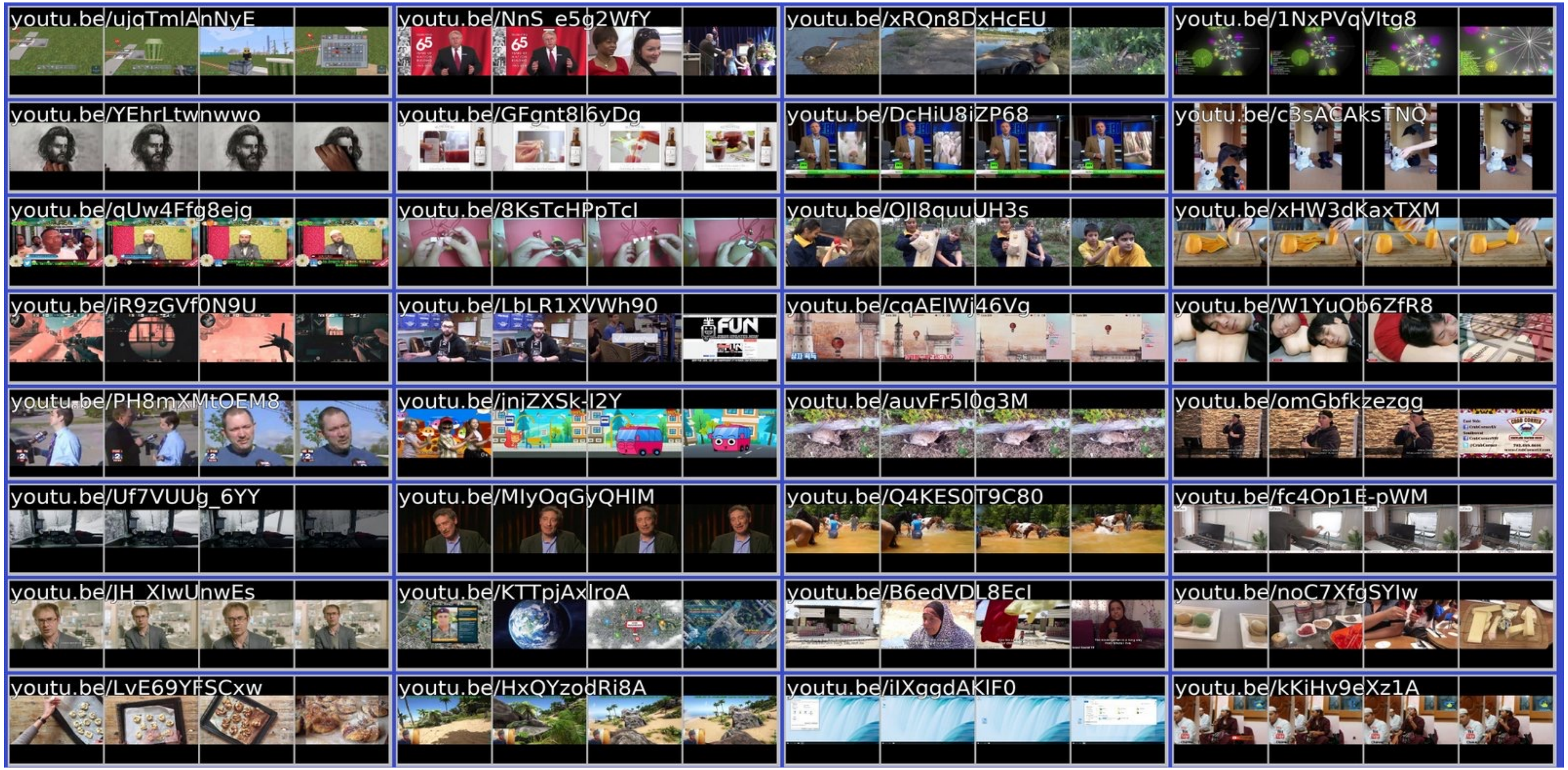
Impose cycle consistency with loss



Neighbor representation set $U = \{u_1, u_2, \ldots, u_m\}$ from random videos

Pretrain Dataset

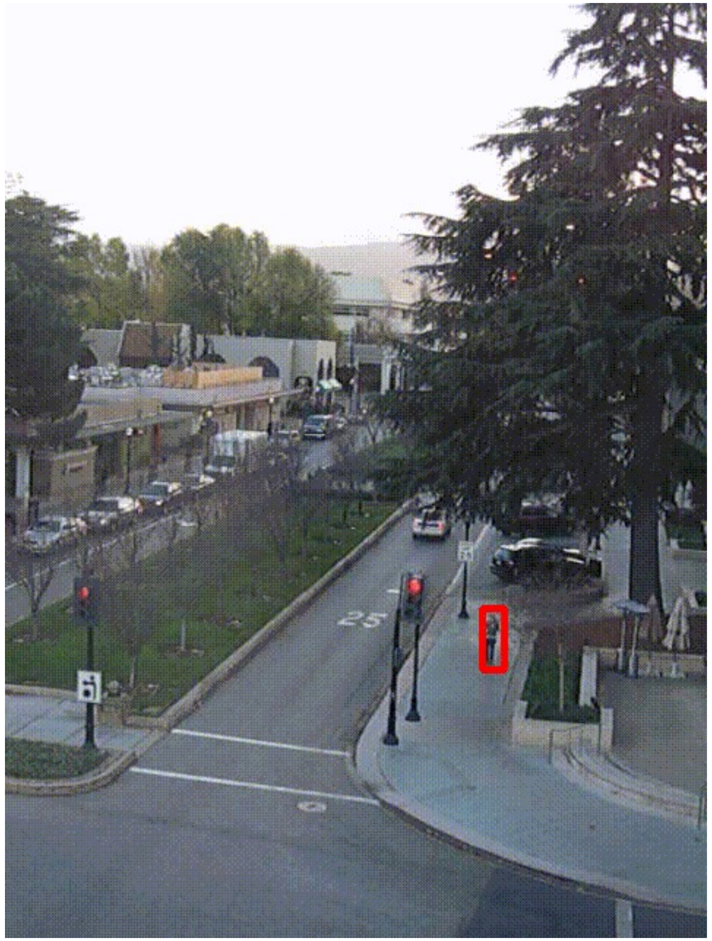
Random Related Video Views (R2V2)



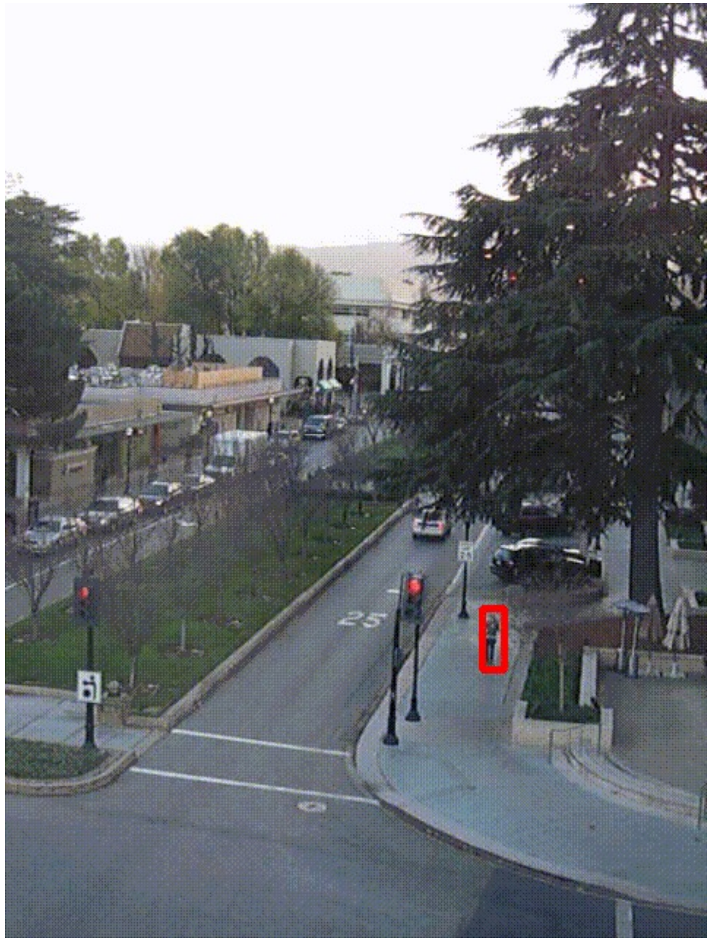
VINCE: Gordon, Daniel, et al. "Watching the world go by: Representation learning from unlabeled videos." *Arxiv* 2020

Results – Visual Object Tracking

Tracking results compared to baseline, on OTB-2015



MoCo



Ours

Results – Visual Object Tracking

Visual object tracking on OTB2015 compared to State-of-the-art

Methods	Precision	Success
SimSiam	61.0	43.2
MoCo	63.7	46.5
VINCE	40.2	30.0
Ours	**72.7**	**53.3**

SimSiam: Chen, Xinlei, and Kaiming He. "Exploring simple siamese representation learning." *CVPR* 2021.
MoCo: He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." *CVPR* 2020
VINCE: Gordon, Daniel, et al. "Watching the world go by: Representation learning from unlabeled videos." *Arxiv* 2020

Results – Linear Image Classification

Linear Image classification compared to state-of-the-art, on ImageNet

Methods	Pretrain dataset	ImageNet Acc Top-1
MoCo	R2V2	53.6
VINCE	R2V2	54.4
Ours	R2V2	**55.6**

MoCo: He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." *CVPR* 2020
VINCE: Gordon, Daniel, et al. "Watching the world go by: Representation learning from unlabeled videos." *Arxiv* 2020

Results – Video Action Recognition

Video action recognition on UCF101

Methods	Backbone	Model Params	Pretrain dataset	Accuracy
3D-RotNet	3D-ResNet18-full	33.6M	Kinetics-400	62.9
SpeedNet	I3D	12.1M	Kinetics-400	66.7
Dense Predictive Coding	3D-ResNet18	14.2M	Kinetics-400	68.2
MemDPC	R-2D3D	32.4M	Kinetics-400	78.1
Ours	ResNet18	11.69M	R2V2	76.8
Ours	Resnet50	25.56M	Kinetics-400	81.6
Ours	ResNet50	25.56M	R2V2	**82.1**

3D-RotNet: Jing et al. "Self-supervised spatiotemporal feature learning by video geometric transformations." *arXiv 2018* .
SpeedNet: Benaim, Sagie, et al. "Speednet: Learning the speediness in videos." CVPR 2020
Dense Predictive Coding: Han et al. "Video representation learning by dense predictive coding." *CVPR Workshops*. 2019.
MemDPC: Han et al. "Memory-augmented dense predictive coding for video representation learning." *ECCV* 2020.

Takeaways and conclusions

- Explore **cross-image/cross-video relation** helps general image representation learning
- Use cross-video ***cycle consistency*** to regularize cross-video representations without pseudo labels