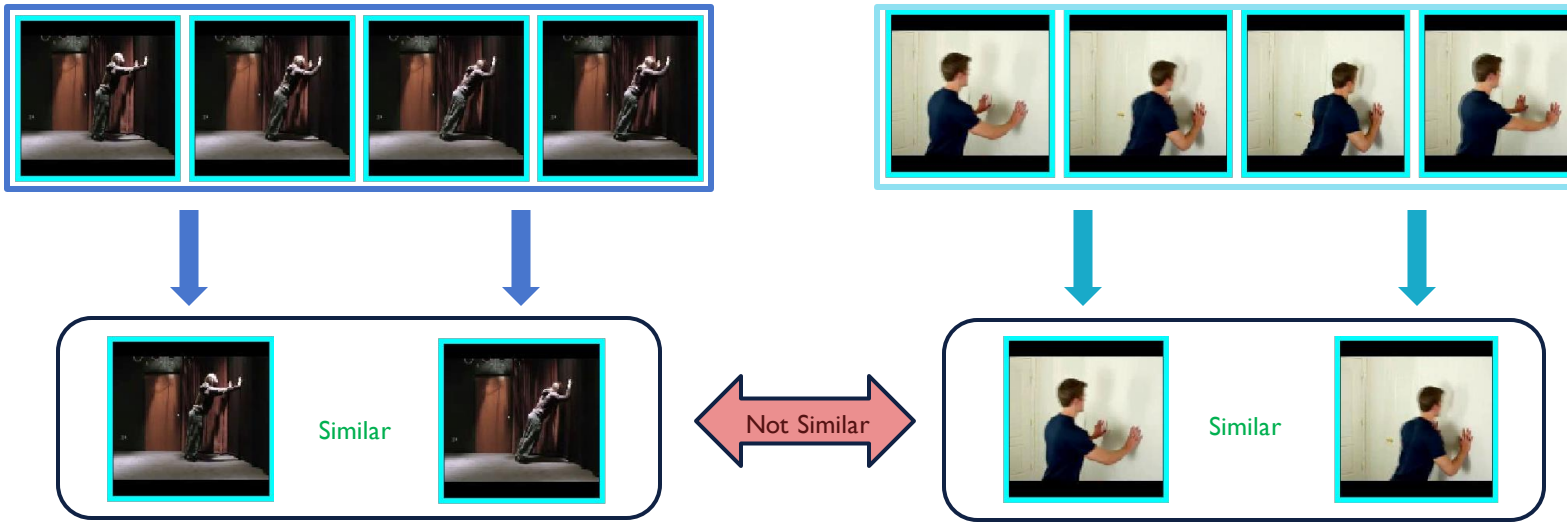
INTRODUCTION

- Self-supervised learning (SSL) - Handle large unlabeled datasets
- Contrastive Learning - Emerged as a promising framework for SSL
- Recent advances show SSL for video representations is key for video analysis

CONTRASTIVE LEARNING

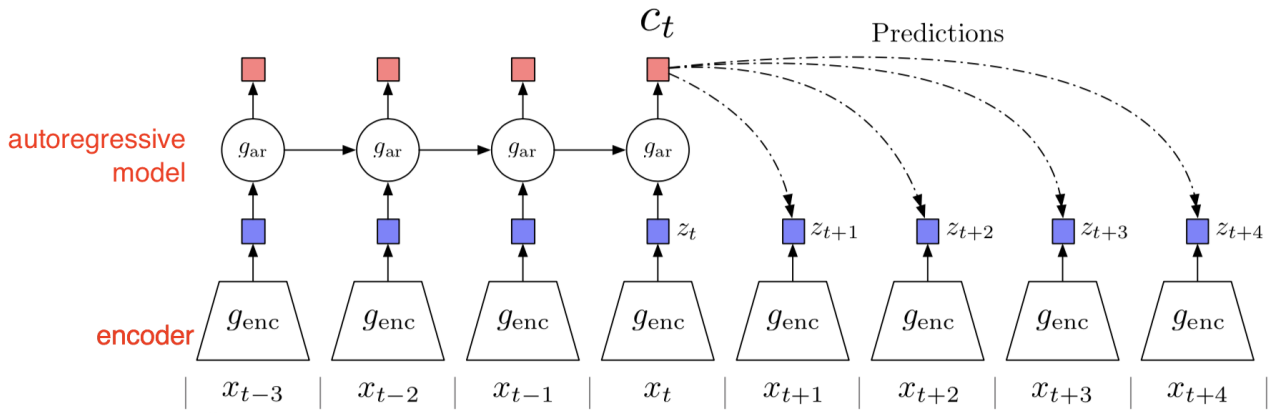
- Main idea
  - Similar instances should be represented similarly
  - Different instances should be represented differently
- Very Intuitive!
- But how do we define similar and different?
  - Simple rigid constraints - samples from same video are similar
  - Everything else is different!

CONTRASTIVE LEARNING



CONTRASTIVE LEARNING

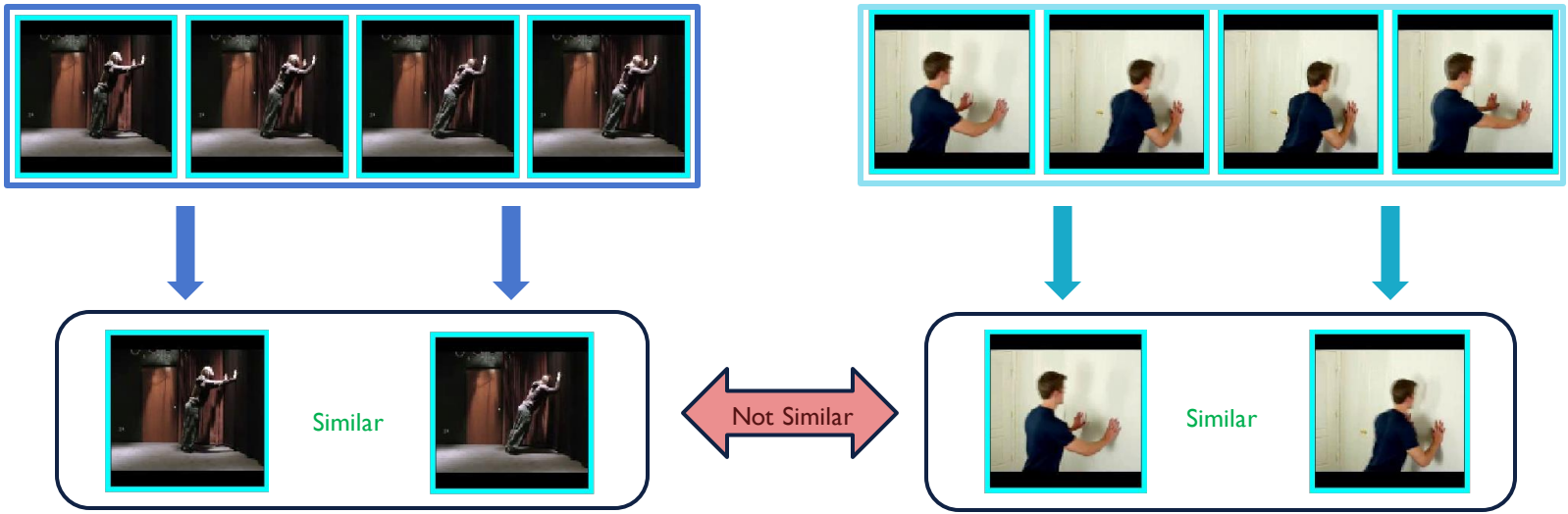
- Utilize Contrastive Predictive Coding [1] to learn video representations
- Model videos as a sequence of frames i.e. $x_1, \ldots, x_t$



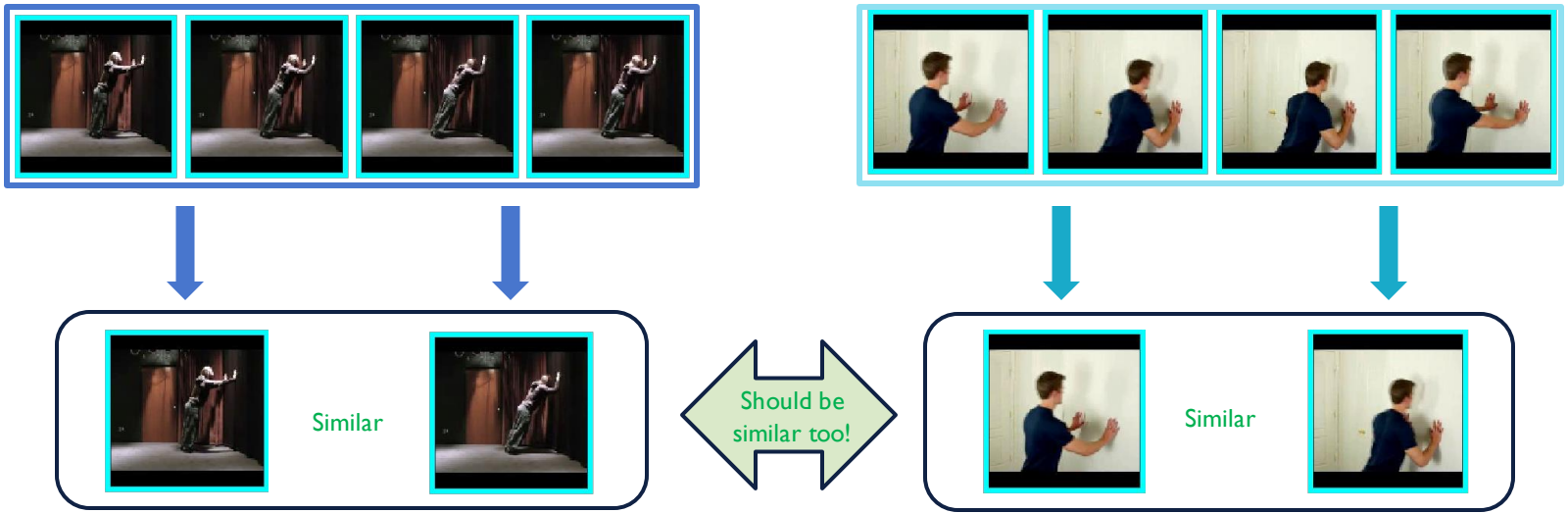
MOTIVATION

- Assumption - Samples from same video are similar, everything else is different
- Is this correct?
- What if two different videos represent the same action?
  - We still discriminate between them – have different representations for them
- Contrastive learning with such an assumption encourages performing low-level discrimination even between semantically similar videos

WHAT HAPPENS



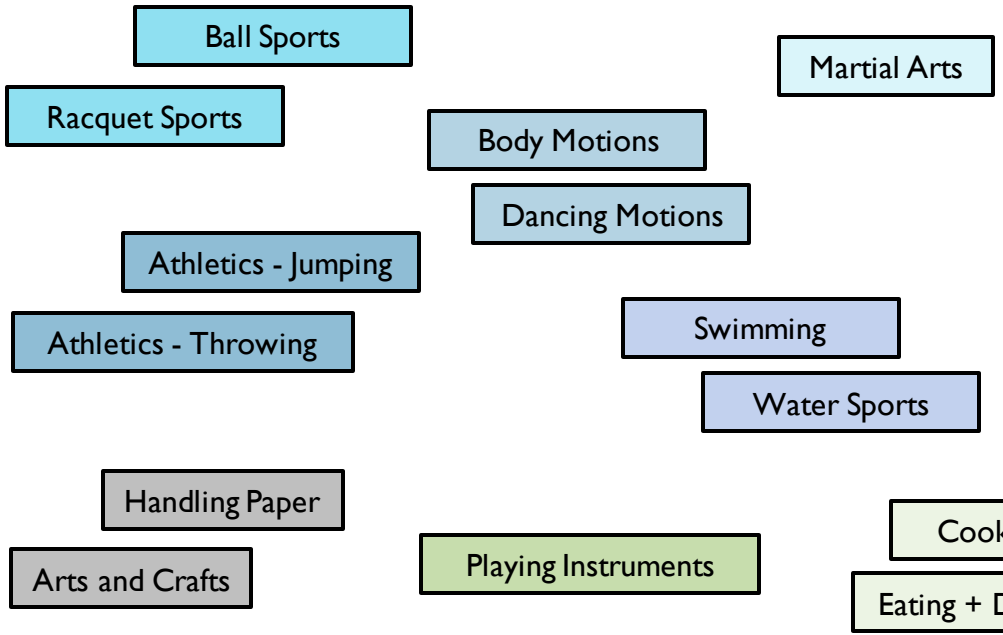
WHAT WE WANT



MOTIVATION (CONT.)

- Utilizing semantic relationships for contrastive learning will allow us to learn high-level features actually representing the involved semantics
- Hold on. Encountering pairs belonging to the exact same class is uncommon
- For Kinetics400 with 400 classes - 0.25% chance of getting such a pair
- Is there still any benefit?

KINETICS-400



- 400 different classes!
- But are they completely different?
- Only **~25** main types of classes
- Notice the subtle relations between types
- **~10** broad categories of types
- Can we use these relationships?

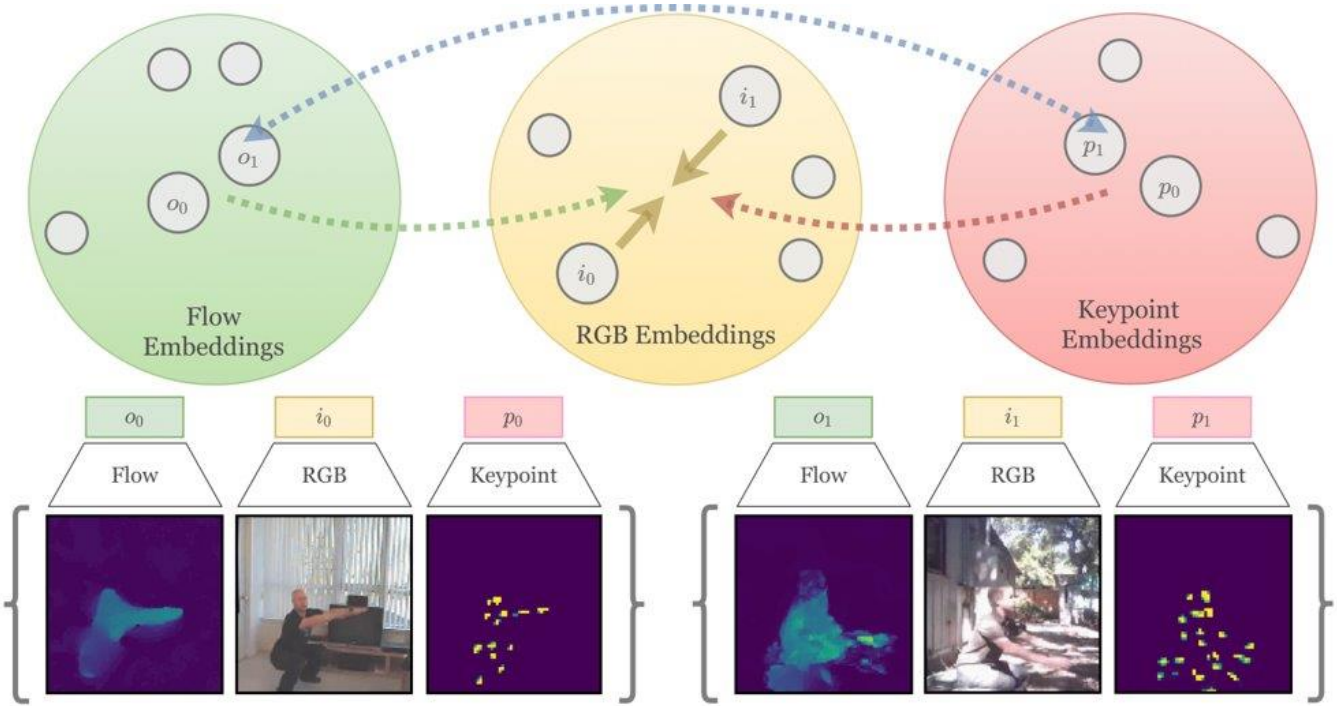
Note:
- We only show a subset of types
- Each type contains 10-30 classes
- Types taken from Kinetics400 [2]

MOTIVATION (CONT.)

- Around 10% chance of getting pairs belonging to the same broad type!
- Even higher chance when considering inter-type relationships
- Focusing on what makes these videos similar is key to learning meaningful video representations
- Sounds good. But how to get such pairs if we don't have labels!?

PROPOSED APPROACH

- CoCon: Cooperative Contrastive Learning
- Utilize multiple views of data to gain insights
- Utilize these insights in a cooperative manner to improve representations for all the views involved
- Intuition: Instances undiscernible in one view, can be easily discerned in another



Two instances of people doing squats. Note these instances are similar through Optical Flow and Keypoint views, but not through RGB space. CoCon leverages this inconsistency and encourages distances in all views to be similar. CoCon pushes i_0, i_1 towards each other in the RGB space

PROPOSED APPROACH

- Utilize Contrastive Predictive Loss to learn view-specific encoders

$$\mathcal{L}_{cpc} = -\sum_{i,k}\left(\log\frac{\exp(\tilde{z}_{i,k}\cdot z_{i,k}\,/\,\tau)}{\sum_{j,m}\exp(\tilde{z}_{i,k}\cdot z_{j,m}\,/\,\tau)}\right)$$

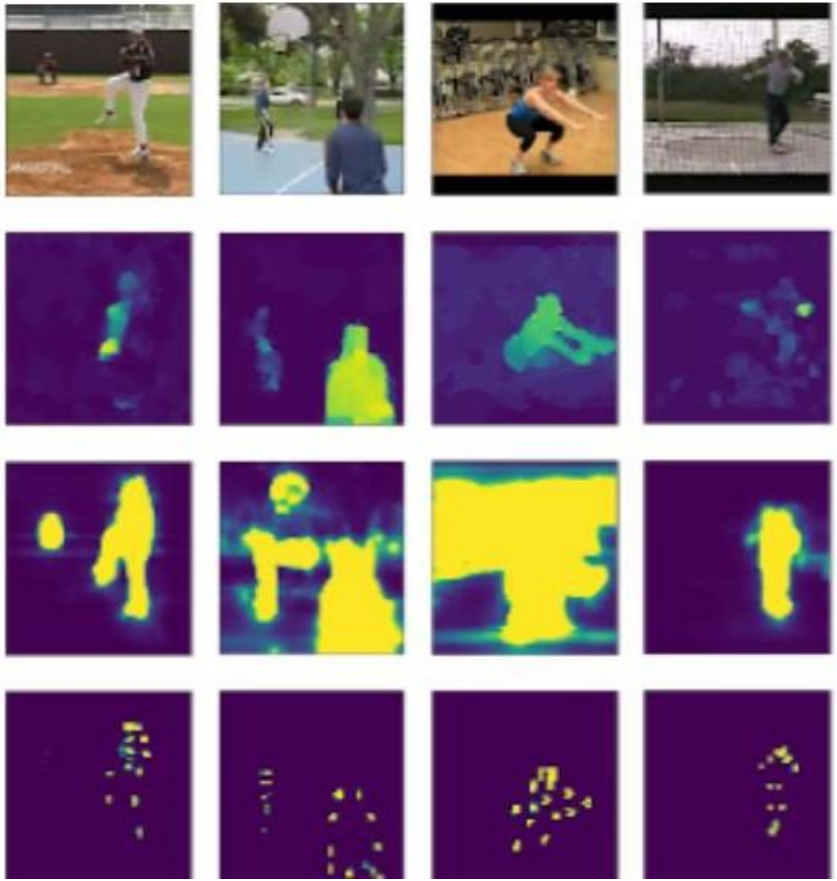
- Losses to promote diversity and consistency of similarities between views

$$\mathcal{L}_{sync} = \sum_{v_0,v_1}\sum_{a,b}\left(\mathcal{D}(h^a_{v_0}, c^b_{h_0}) - \mathcal{D}(h^a_{v_1}, h^b_{v_1})\right)^2$$

$$\mathcal{L}_{sim} = \sum_{v_0,v_1}\sum_{a}\mathcal{D}(h^a_{v_0}, h^a_{v_1}) + \mu\sum_{a\neq b} max\left(0, 1 - \mathcal{D}(h^a_{v_0}, h^b_{v_1})\right)$$

VIEWS

- To learn features representative of actions, we use views such as,
- RGB Images
- Optical Flow
- Human Segmentation Masks
- Human Pose Keypoints
- Note the prevalent noise



RESULTS

- Use UCF101, HMDB51, Kinetics400 for evaluation
- We utilize downstream performance on action classification as a proxy for representation quality
- Pre-train model using CoCon, proceed to fine-tune with supervision
- Perform both quantitative and qualitative evaluations for detailed analysis

ABLATION STUDY

- We study the role of each component in CoCon
- Best results achieved when utilizing both *sim* and *sync* loss
- We argue CoCon performs better as it utilizes cross view relations

View	Random	$\mathcal{L}_{cpc}$	$\mathcal{L}_{sim}^{cpc}$	$\mathcal{L}_{sync}^{cpc}$	$\mathcal{L}_{cocon}$
RGB	46.7	63.7	66.0	62.7	**67.8**
Flow	65.3	69.8	71.4	69.2	**72.5**

IMPACT OF VIEWS

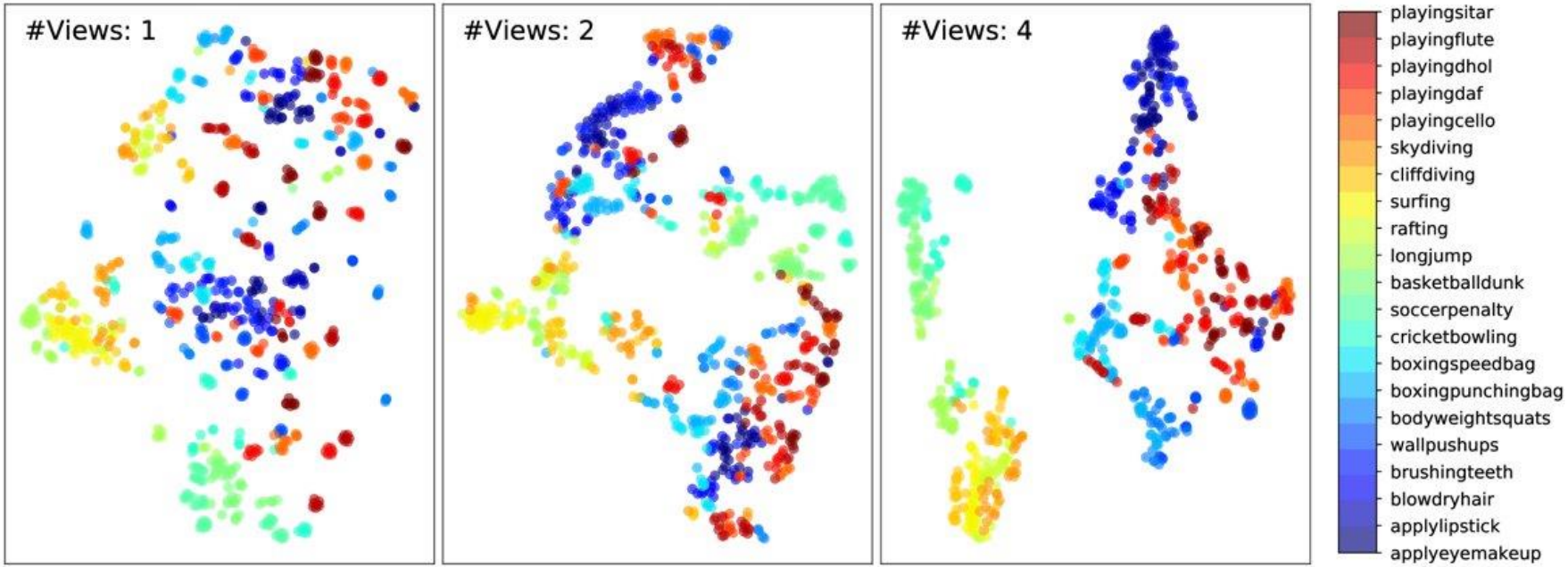
Method	RGB		Flow		PoseHM		SegMask	
	UCF	HMDB	UCF	HMDB	UCF	HMDB	UCF	HMDB
Random	46.7	20.6	65.3	31.2	51.7	33.0	42.7	26.3
CPC	63.7	33.1	71.2	44.6	56.4	42.0	53.7	32.8
CoCon	**71.0**	**39.0**	**74.5**	**45.4**	**58.7**	**42.6**	**55.8**	**34.0**

Adding additional views improves performance across all views. Even low dimensional modalities such as PoseHM and SegMask are able to improve performance through cooperative training.

Method	Resolution	Backbone	# Views	Pre-train	UCF	HMDB
Random Initialization	128 × 128	ResNet18	1		46.7	20.6
ImageNet [30]	224 × 224	VGG-M-2048	1	ImageNet	73.0	40.5
Shuffle and Learn [23]	227 × 227	CaffeNet	1	UCF-HMDB	50.2	18.1
OPN [18]	80 × 80	VGG-M-2048	1	UCF-HMDB	59.8	23.8
DPC [8]	128 × 128	ResNet18	1	UCF101	60.6	-
VGAN [36]	N/A	C3D	2	Flickr [36]	52.1	-
LT-Motion [21]	N/A	RNN [21]	2	NTU	53.0	-
Cross and Learn [29]	224 × 224	CaffeNet	2	UCF101	58.7	27.2
Geometry [6]	N/A	CaffeNet	2	UCF101	55.1	23.3
CMC [34]	128 × 128	CaffeNet	3	UCF101	59.7	26.1
CoCon - RGB	128 × 128	ResNet18	4	UCF101	**70.5**	**38.4**
CoCon - Ensemble	128 × 128	ResNet18	4	UCF101	**82.4**	**52.0**
3D-RotNet [13]	112 × 112	ResNet18	1	Kinetics	62.9	33.7
DPC [8]	128 × 128	ResNet18	1	Kinetics	68.2	34.5
CoCon - RGB	128 × 128	ResNet18	2	Kinetics	**71.6**	**46.0**
CoCon - Ensemble	128 × 128	ResNet18	2	Kinetics	**78.1**	**52.0**
ST-Puzzle [15]	224 × 224	ResNet18	1	Kinetics	65.8	33.7
DPC [8]	224 × 224	ResNet34	1	Kinetics	75.7	35.7
CoCon - RGB	224 × 224	ResNet34	2	Kinetics	**79.1**	**48.5**
CoCon - Ensemble	224 × 224	ResNet34	2	Kinetics	**82.0**	**53.1**

Table 6: Comparison of classification accuracies on UCF101 and HMDB51, averaged over all splits.

TSNE VISUALIZATION



We visualize CoCon's video features. We get more meaningful clusters as #views increases.

Soft Alignment between Videos. Similar phases of actions are framed with boxes of similar color. Heatmaps represent similarities between different timesteps of the videos.