Multi-view invariance and grouping for self-supervised learning

Ishan Misra

Facebook AI Research

Multi-view

Same data sample
Different ways of looking at it

Grouping

Associating related
data samples

Multi-view invariance

Grouping

Let us fill this table throughout the talk

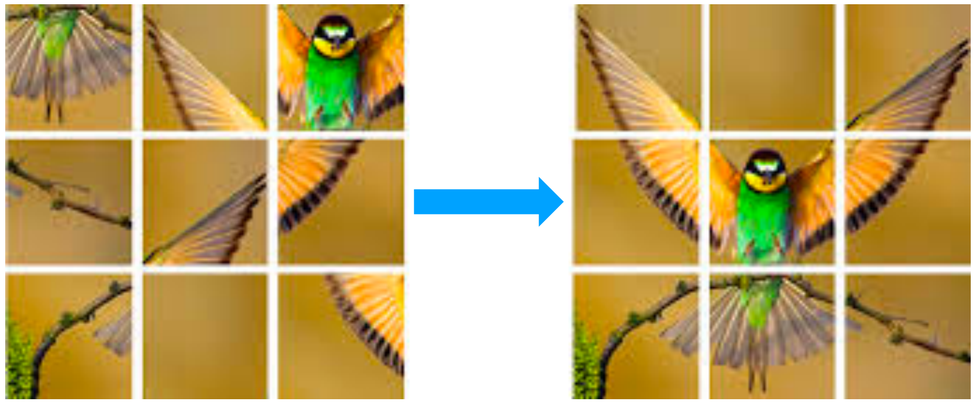
Method	Multi-view Invariance	Grouping
Method Name	?	?

Pre-2019: "Pretext" tasks

- Create proxy tasks

Rotation

(Gidaris et al., 2018)

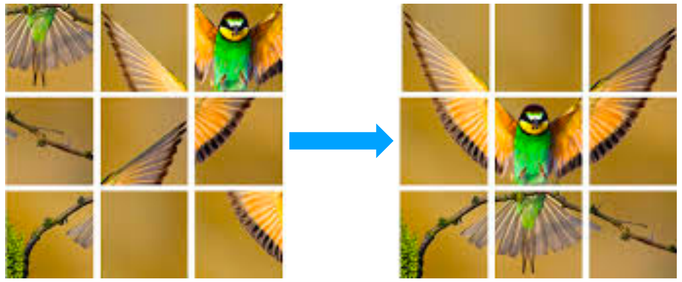


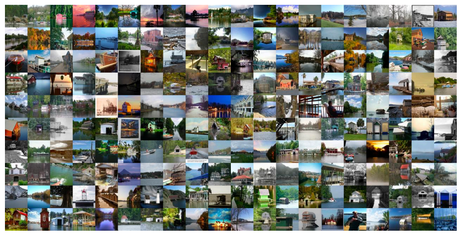
Jigsaw puzzles

(Noroozi et al., 2016)

The hope of generalization

- We **hope** that the pre-training task and the transfer task are "aligned"





Pre-training

Transfer Tasks

Pretext task

Predict property of transform t

ConvNet

$\mathbf{I}^t$

Contain information about transform t

$\mathbf{I}^t$

Less Semantic Features

$\mathbf{I}$

Image Transform t

$\mathbf{I}^t$

Method	Multi-view Invariance	Grouping
Pretext Task	No	?

Pretext task

Predict property of transform t

ConvNet

$\mathbf{I}^t$

Prediction only about $\mathbf{I}$ and t

$\mathbf{I}^t$

No relation to other images

$\mathbf{I}$

Image Transform t

$\mathbf{I}^t$

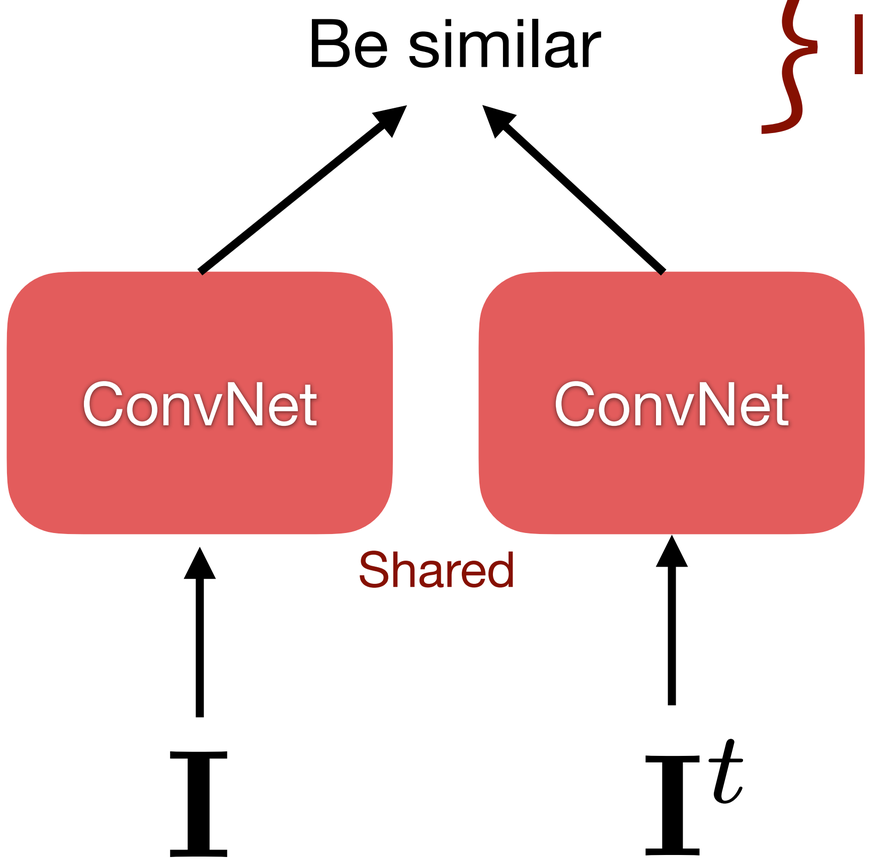
Method	Multi-view Invariance	Grouping	Performance
Pretext Task	No	No	Weak

"Self-supervised learning is a naive/optimistic hope for feature learning"

Pretext-Invariant Representation Learning (PIRL)

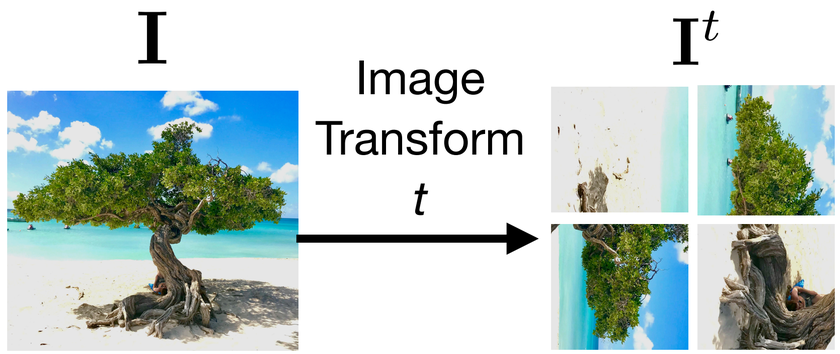
Ishan Misra, Laurens van der Maaten

PIRL

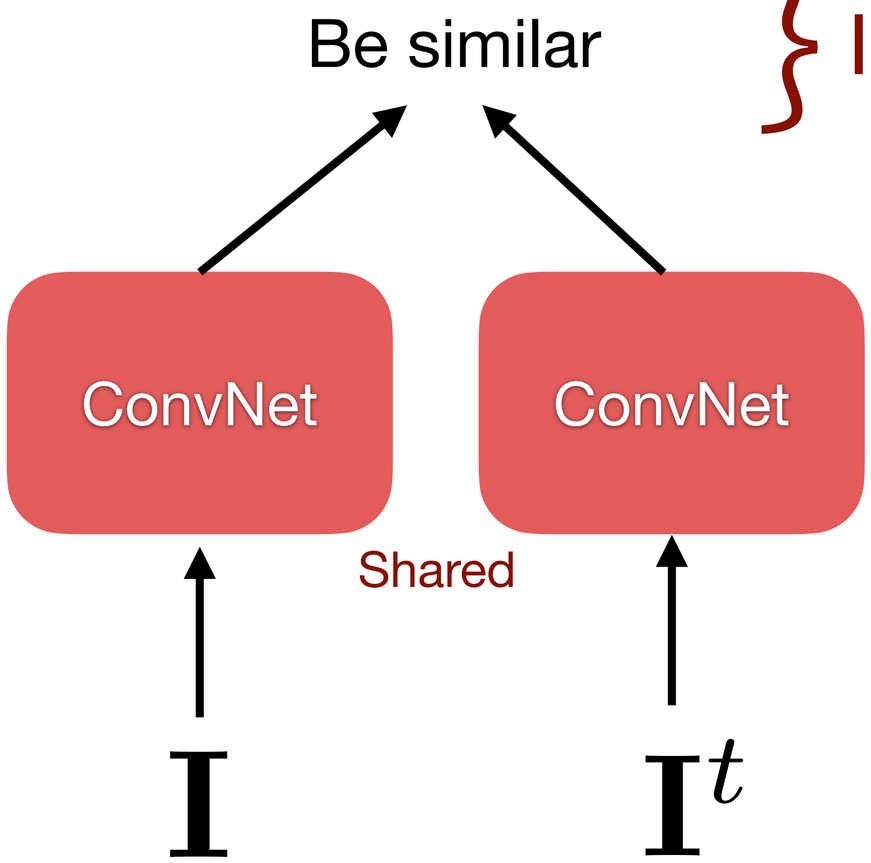


} Invariant to Pretext transform

$$L_{\text{contrastive}}(\mathbf{v_I}, \mathbf{v_I}^t)$$

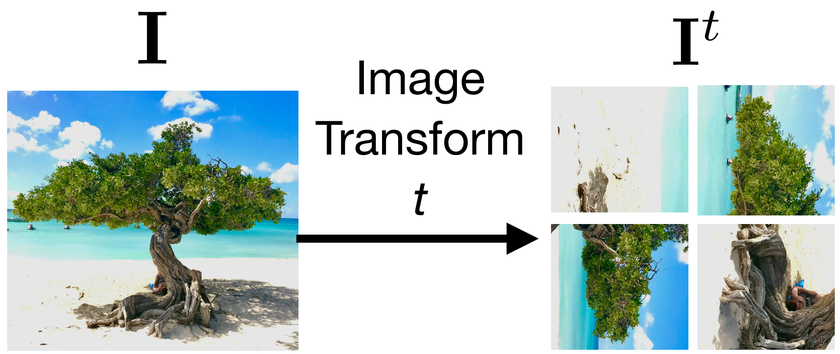


PIRL



Invariant to Pretext transform

$$L_{\text{contrastive}}(\mathbf{v_I}, \mathbf{v_I}^t)$$



Invariance to

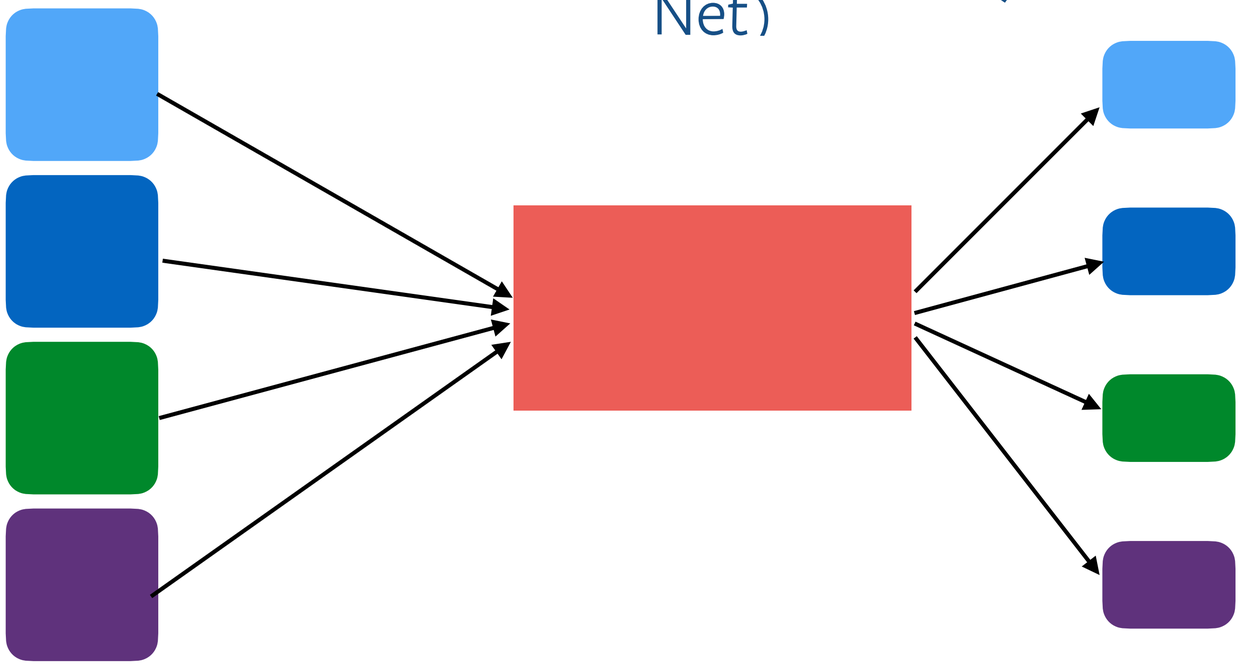
- Data Augmentations
- Multiple views created by pretext task (Jigsaw/Rotation)

Contrastive Learning

Related and Unrelated Images

Shared network (Siamese Net)

Image Features (Embeddings)



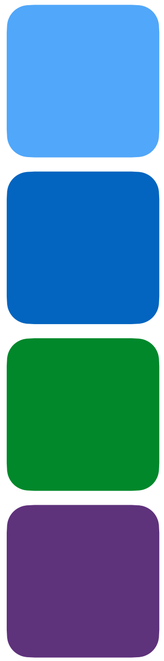
Loss Function

Embeddings from related images should be closer than embeddings from unrelated images

Hadsell et al., 2005, DrLim

Contrastive Learning in PIRL

Dataset

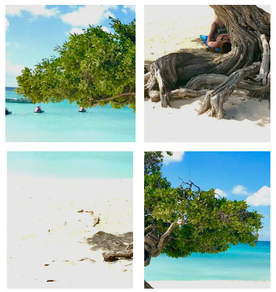


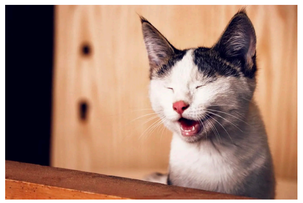
Loss Function

Image Feature & Patch Features

Random Images

$\mathbf{I}$ $\mathbf{I}^t$





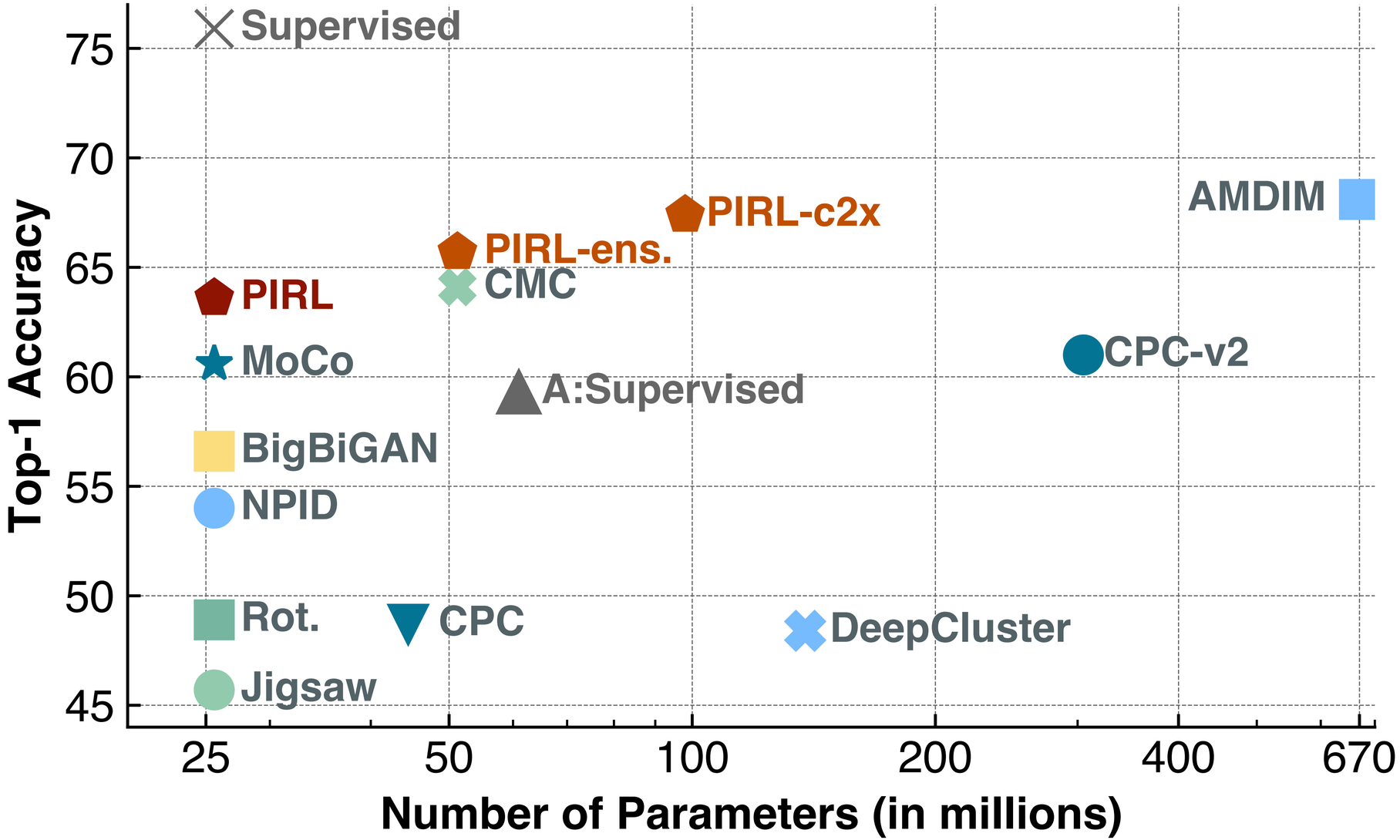
Object Detection

- **Outperforms** ImageNet supervised pre-trained networks
- Full fine-tuning, no bells & whistles
- No extra data, changes in model architecture, fine-tuning schedule

Initialization	VOC07+12		
	AP^{all}	AP^{50}	AP^{75}
ImageNet Supervised	52.6	**81.1**	57.4
	+1.4		**+2.3**
PIRL	**54.0**	80.7	**59.7**

ImageNet Classification

- Linear classifiers on fixed features.



Method	Multi-view Invariance	Grouping	Performance
Pretext Task	No	No	Weak
PIRL	Yes	Weak	Moderate

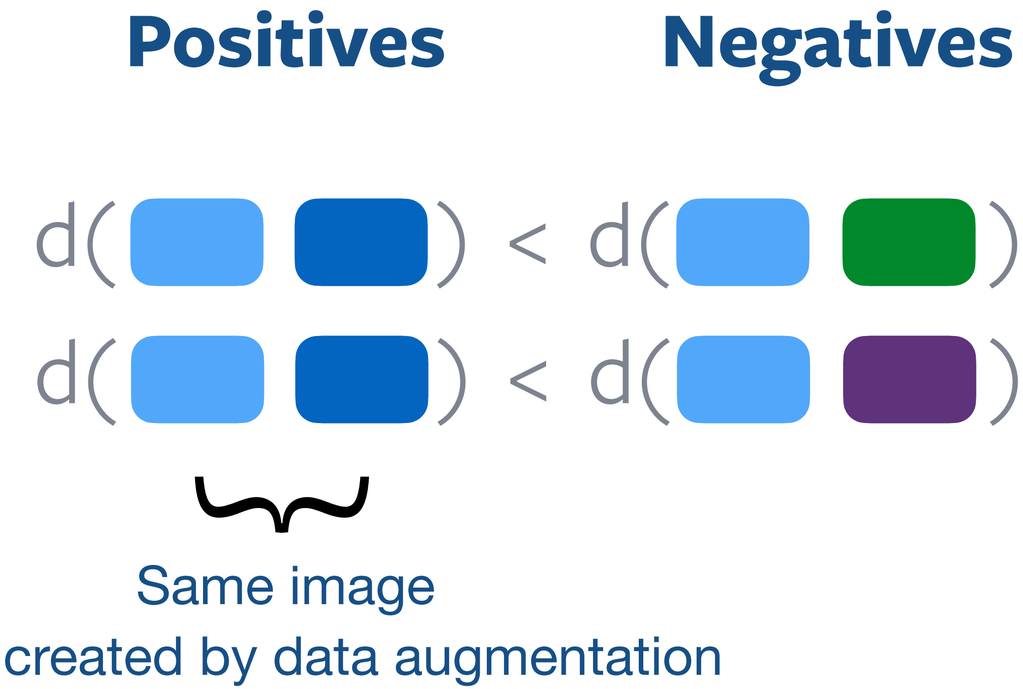
Image & Patch as views

Relate other images using negatives

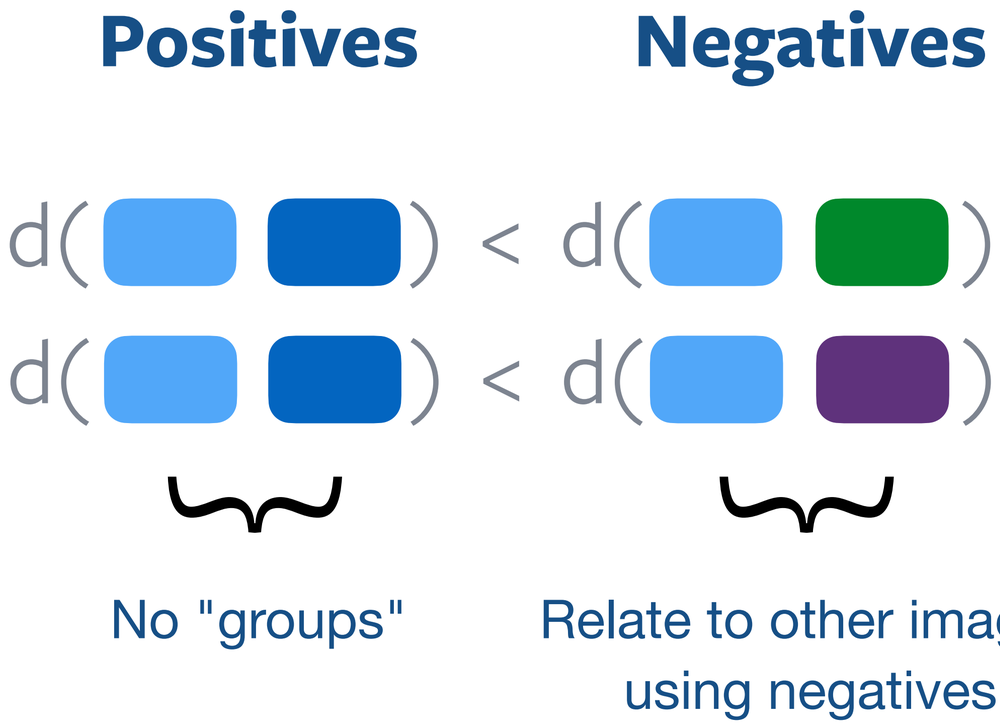
Outperform supervised features on some tasks

See also - CPCv2, MoCo, BoWNet, SimCLR

Lack of Grouping in "Instance" based Contrastive Learning



Lack of Grouping in "Instance" based Contrastive Learning



How to add Grouping?

Audio-Video
(AVID CMA)

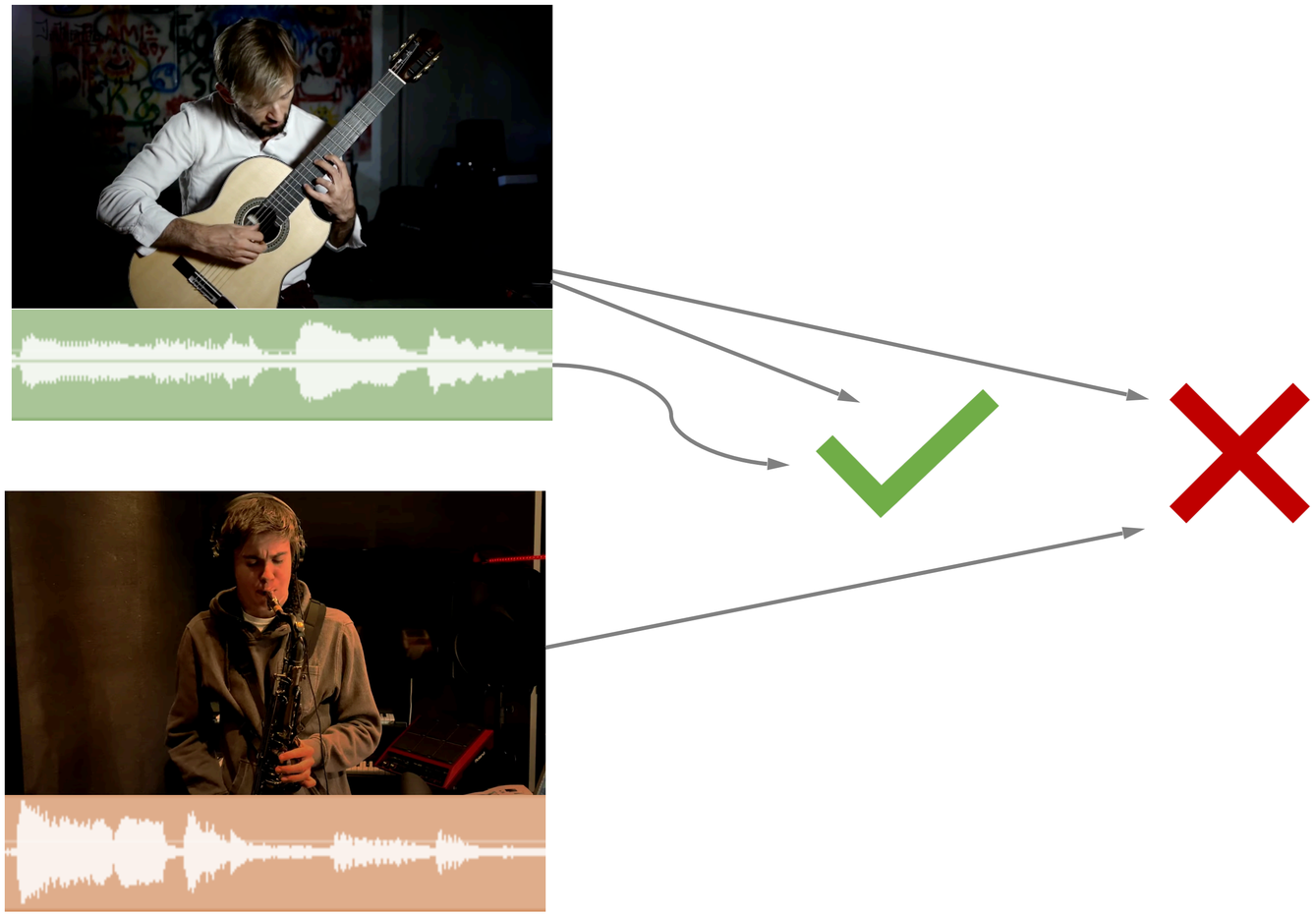
Images
(SwAV)

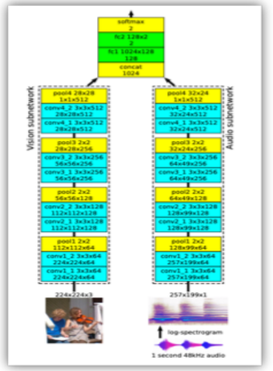
Audio Visual Instance Discrimination with Cross Modal Agreement (AVID + CMA)

Pedro Morgado, Nuno Vasconcelos, Ishan Misra



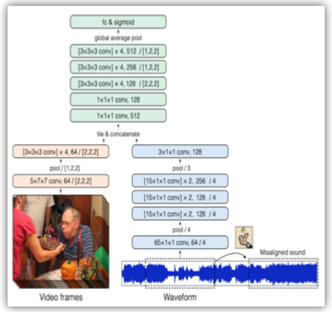
Audio-visual correspondence



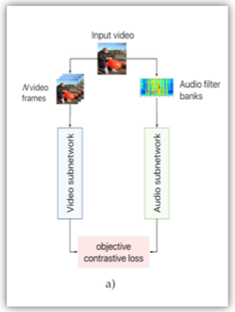


Look, listen & learn
Arandjelovic et al., 2017.

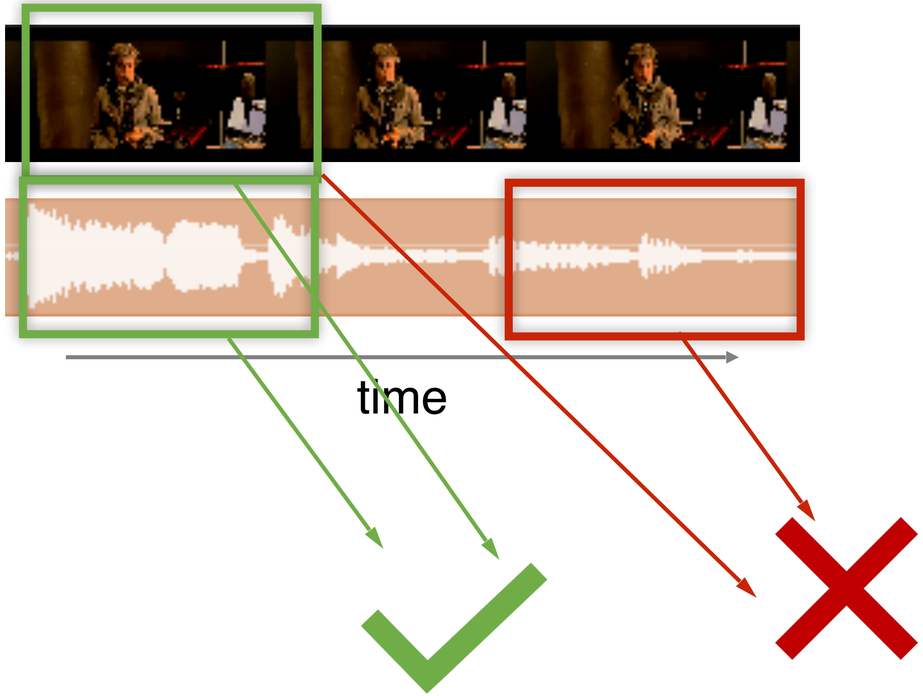
Audio-visual synchronization



Multisensory synchronization
Owens et al 2018.

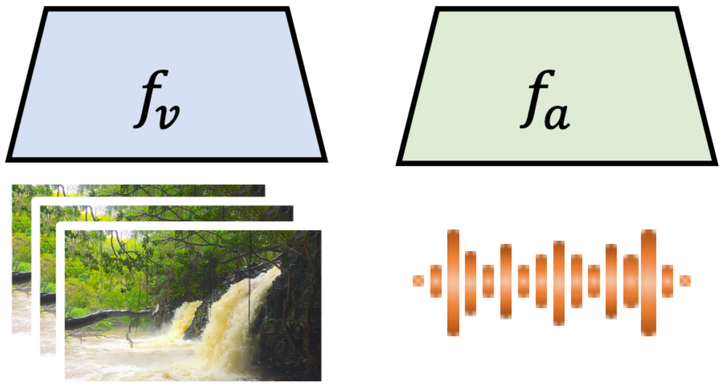


Synchronization w/ curriculum
Korbar et. al 2018.



Method	Multi-view Invariance	Grouping	Performance
AV Sync	Yes	No	Weak

Contrastive (Audio Video Instance Discrimination)



Positives **Negatives**

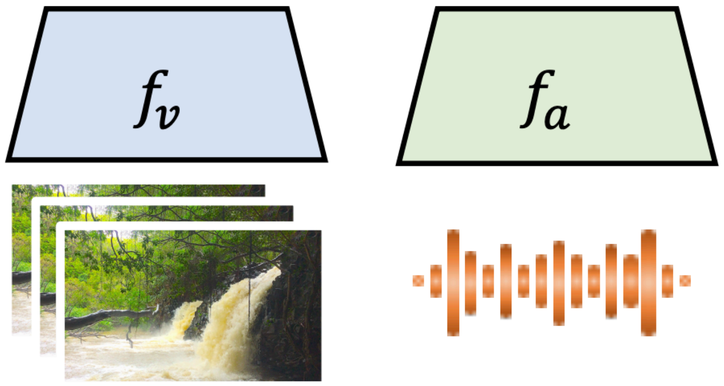
d( ) < d( )

d( ) < d( )

Audio & Video
(same sample)

Relate to other video/audio
using negatives

Contrastive (Audio Video Instance Discrimination)



Positives **Negatives**

d( ) < d( )

d( ) < d( )

Audio & Video (same sample)

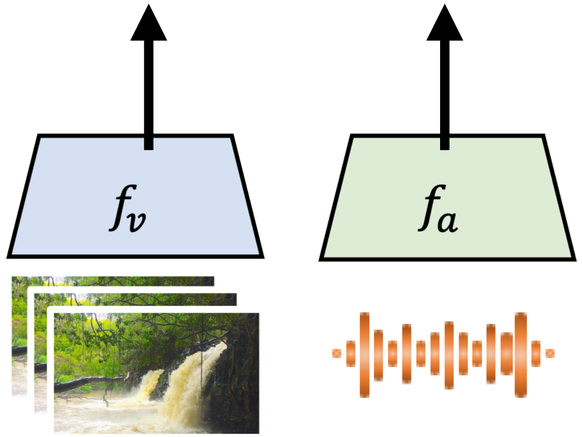
Relate to other video/audio using negatives

Limited Grouping

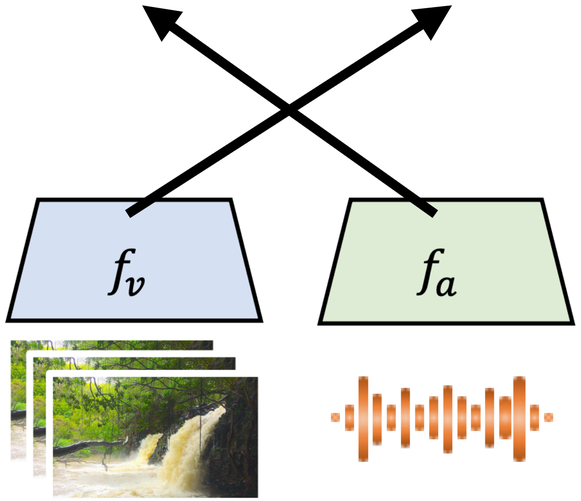
Importance of Multi-view

Video only

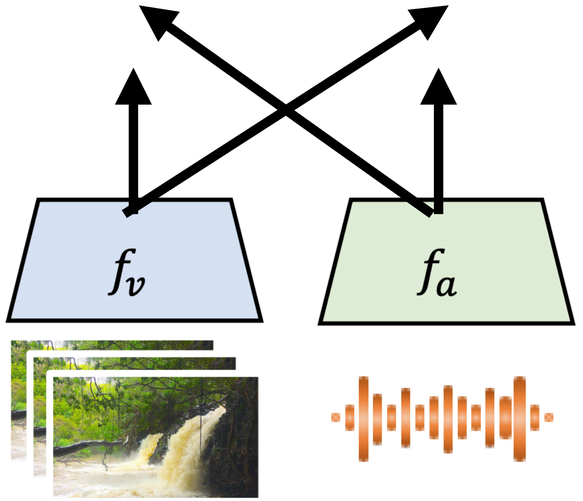
Audio only



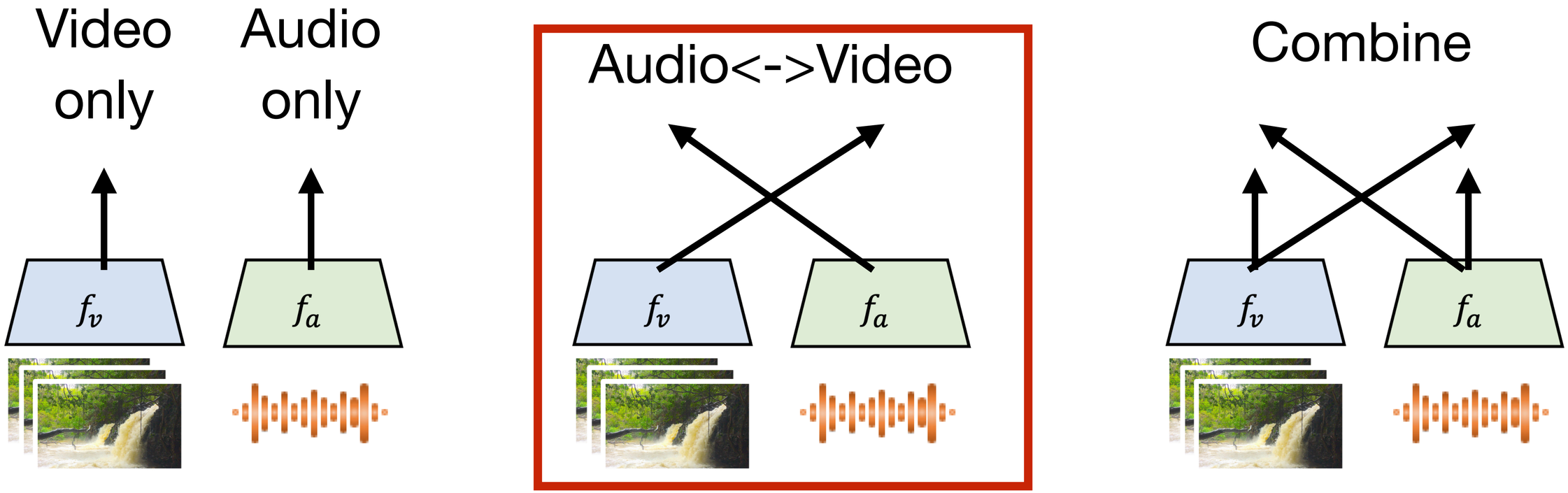
Audio<->Video



Combine

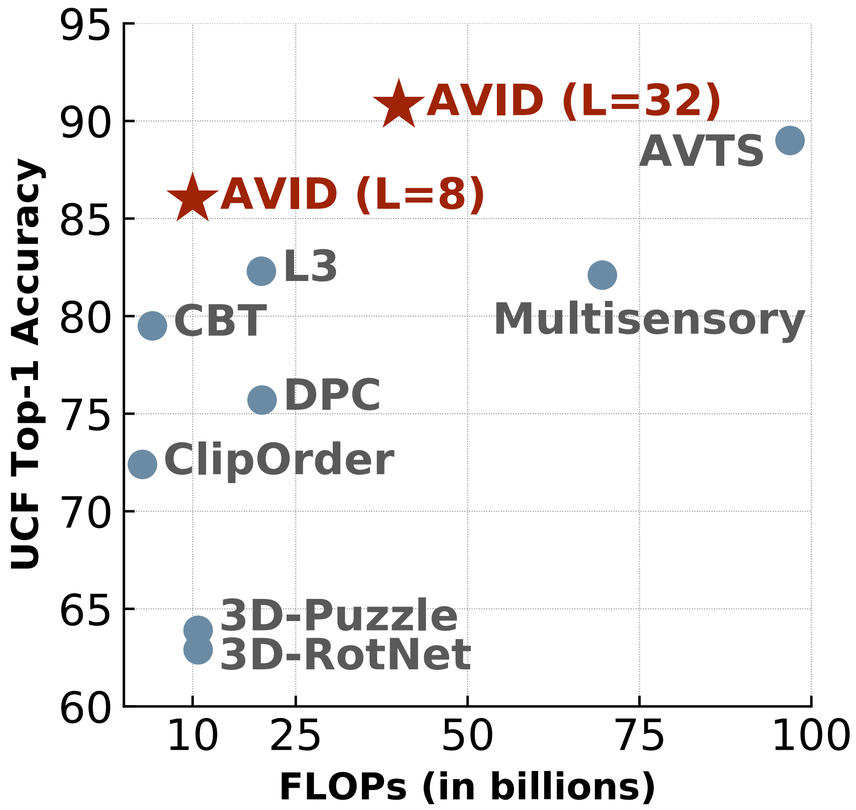


Importance of Multi-view

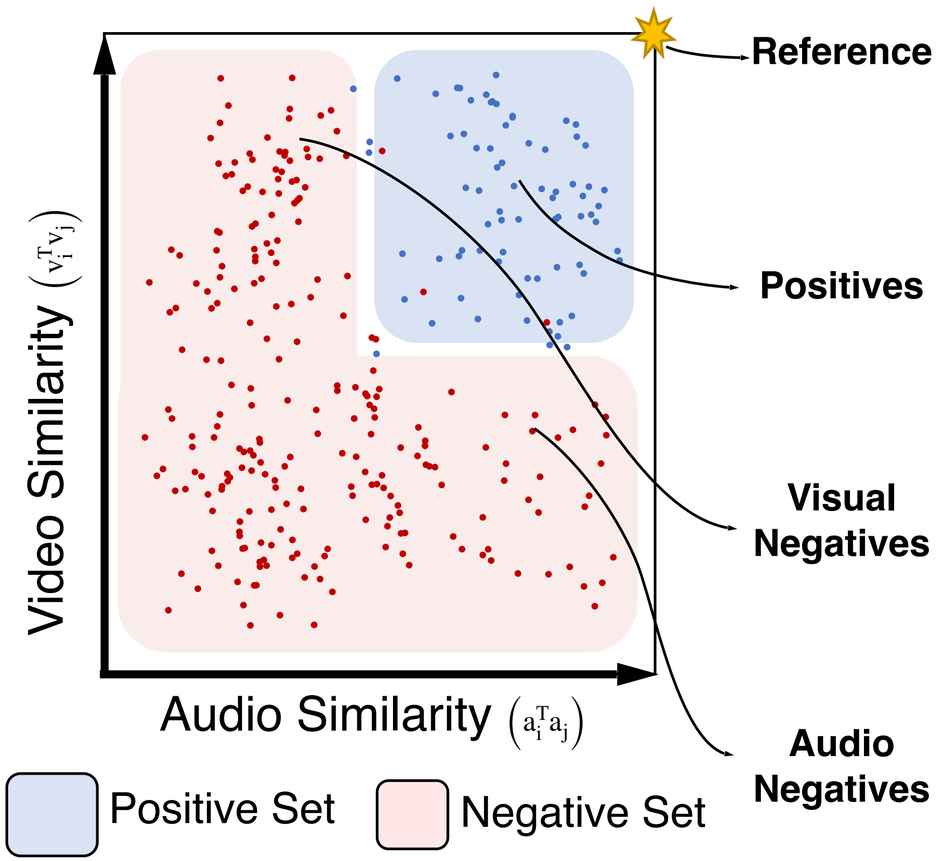


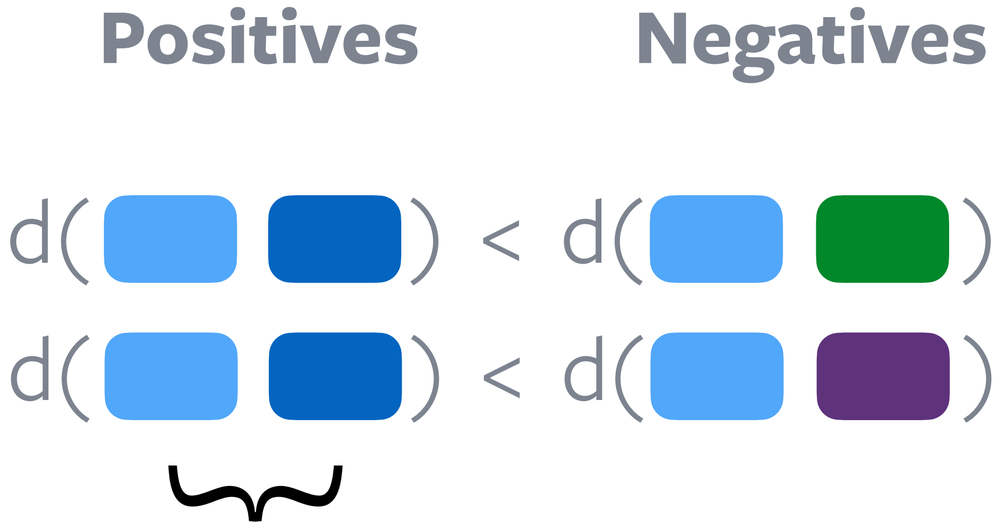
Outperforms all variants

Method	Multi-view Invariance	Grouping	Performance
AV Sync	Yes	No	Weak
AVID	Yes	Weak	Good



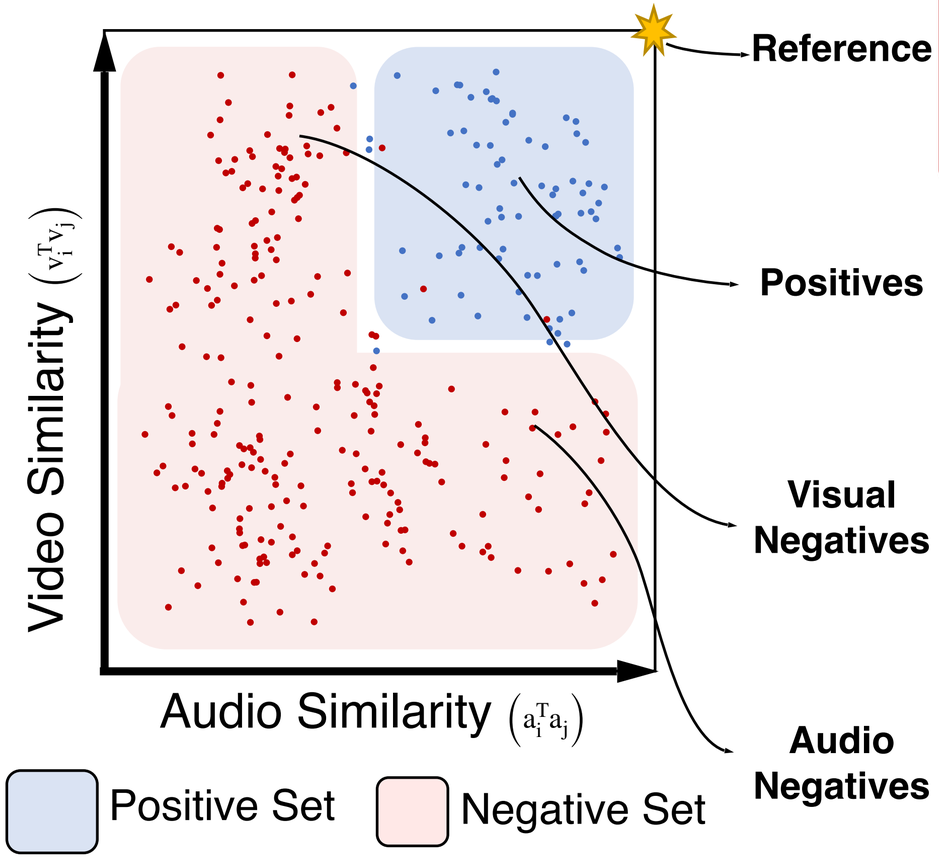
Grouping using Audio-visual Agreements (CMA)

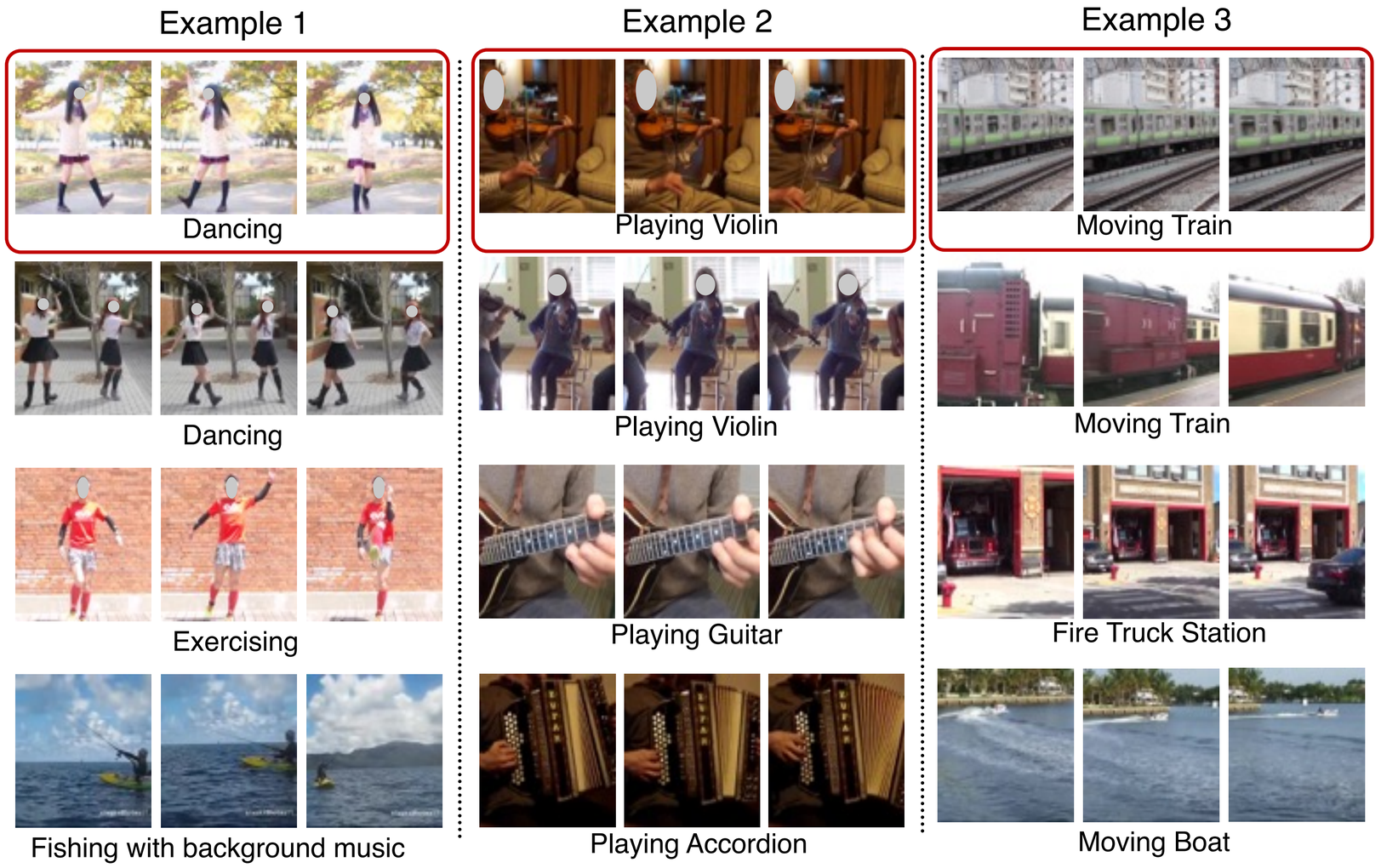


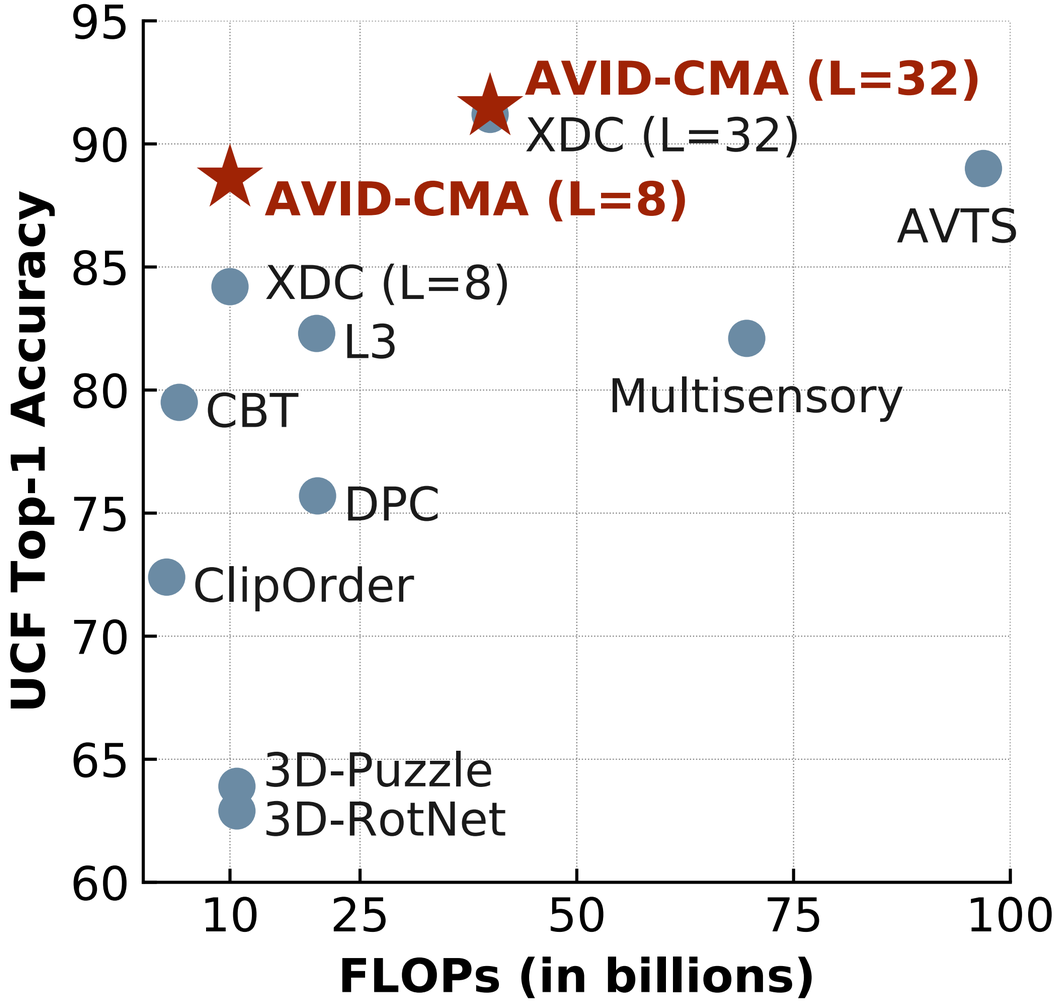


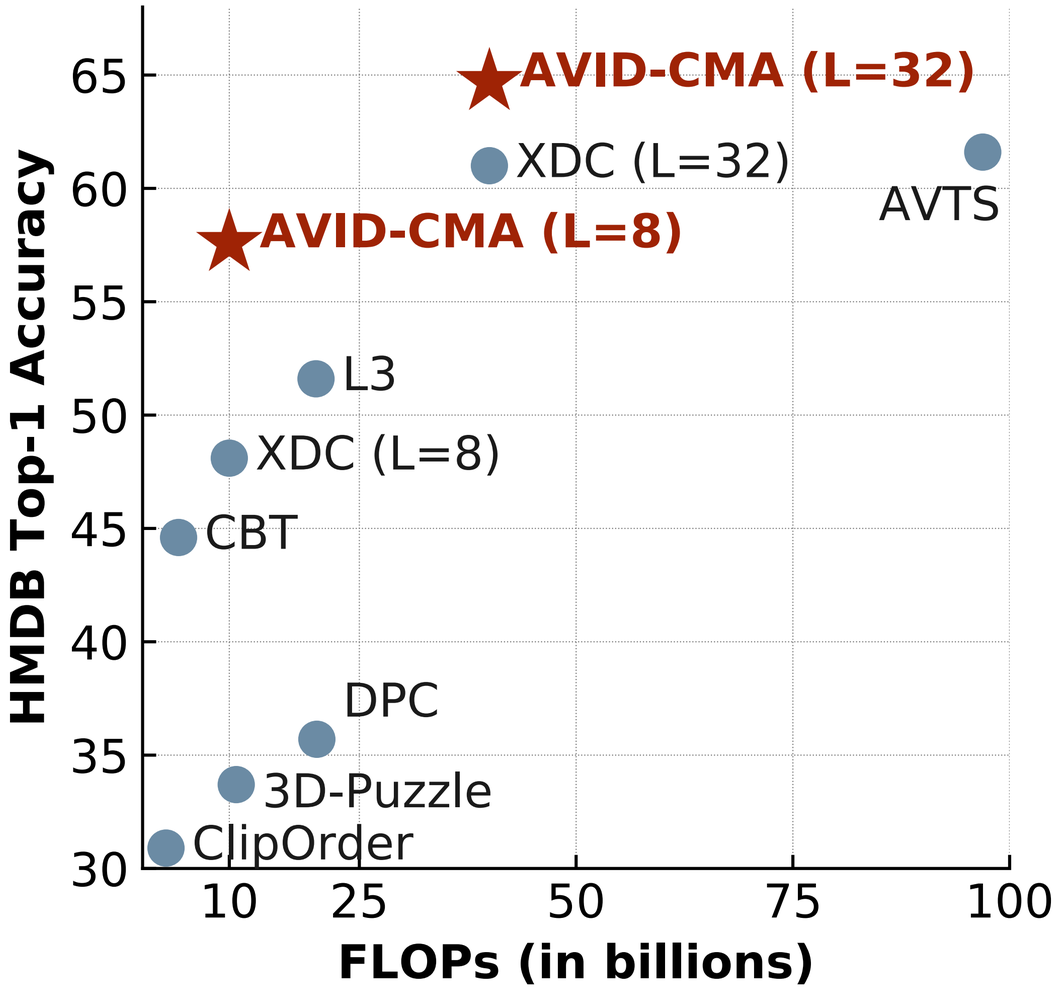
Videos that are similar in audio & video features

Grouping using Audio-visual Agreements (CMA)






AVID-CMA (L=32)
XDC (L=32)
AVID-CMA (L=8)
AVTS
XDC (L=8)
L3
Multisensory
CBT
DPC
ClipOrder
3D-Puzzle
3D-RotNet
UCF Top-1 Accuracy
FLOPs (in billions)
95
90
85
80
75
70
65
60
10
25
50
75
100

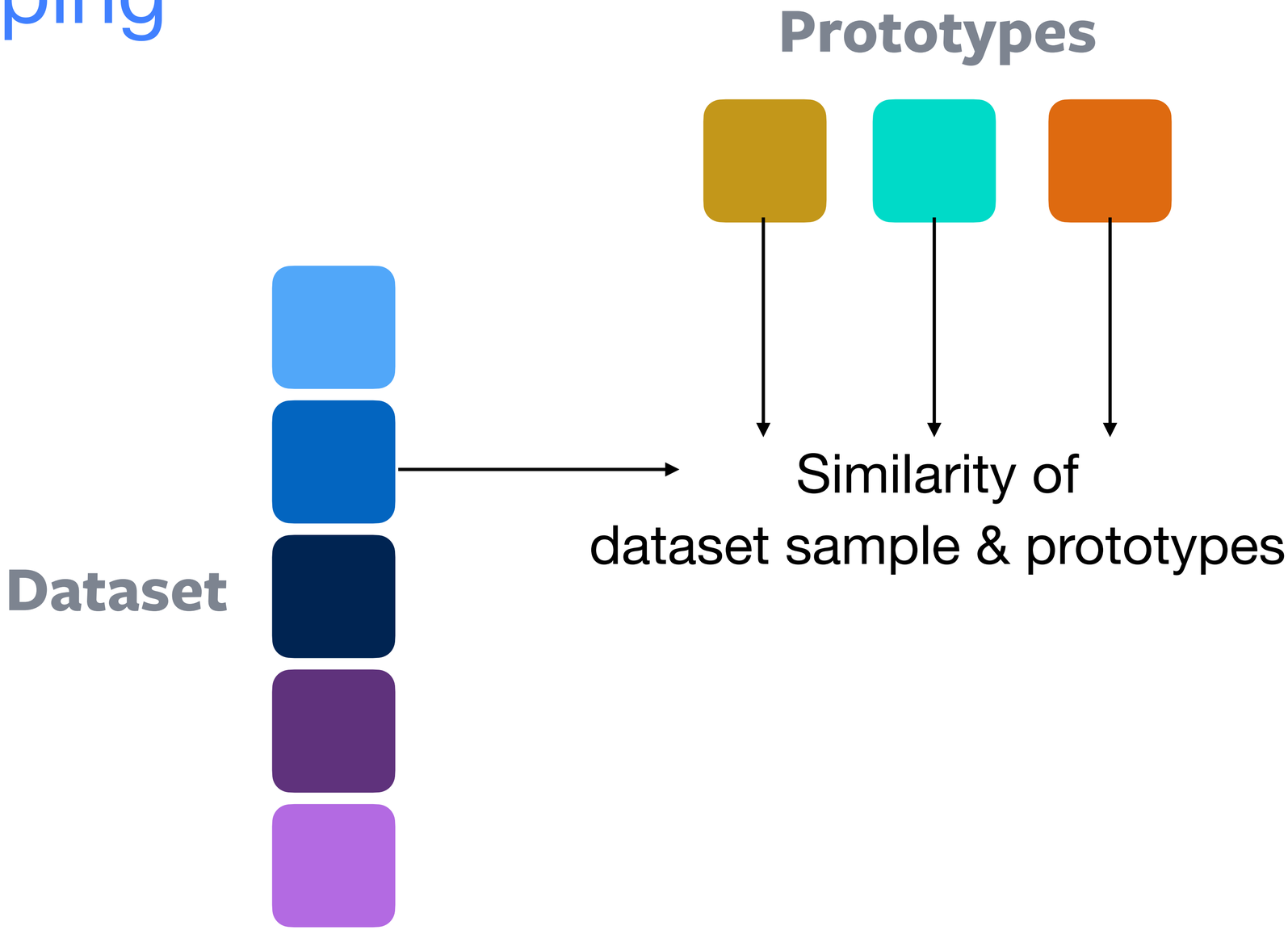

AVID-CMA (L=32)
XDC (L=32)
AVTS
AVID-CMA (L=8)
L3
XDC (L=8)
CBT
DPC
3D-Puzzle
ClipOrder
HMDB Top-1 Accuracy
FLOPs (in billions)
65
60
55
50
45
40
35
30
10
25
50
75
100

Method	Multi-view Invariance	Grouping	Performance
AV Sync	Yes	No	Weak
Contrastive	Yes	Weak	Good
CMA	Yes	Yes	Better

Swapping Assignments between Views (SwAV)

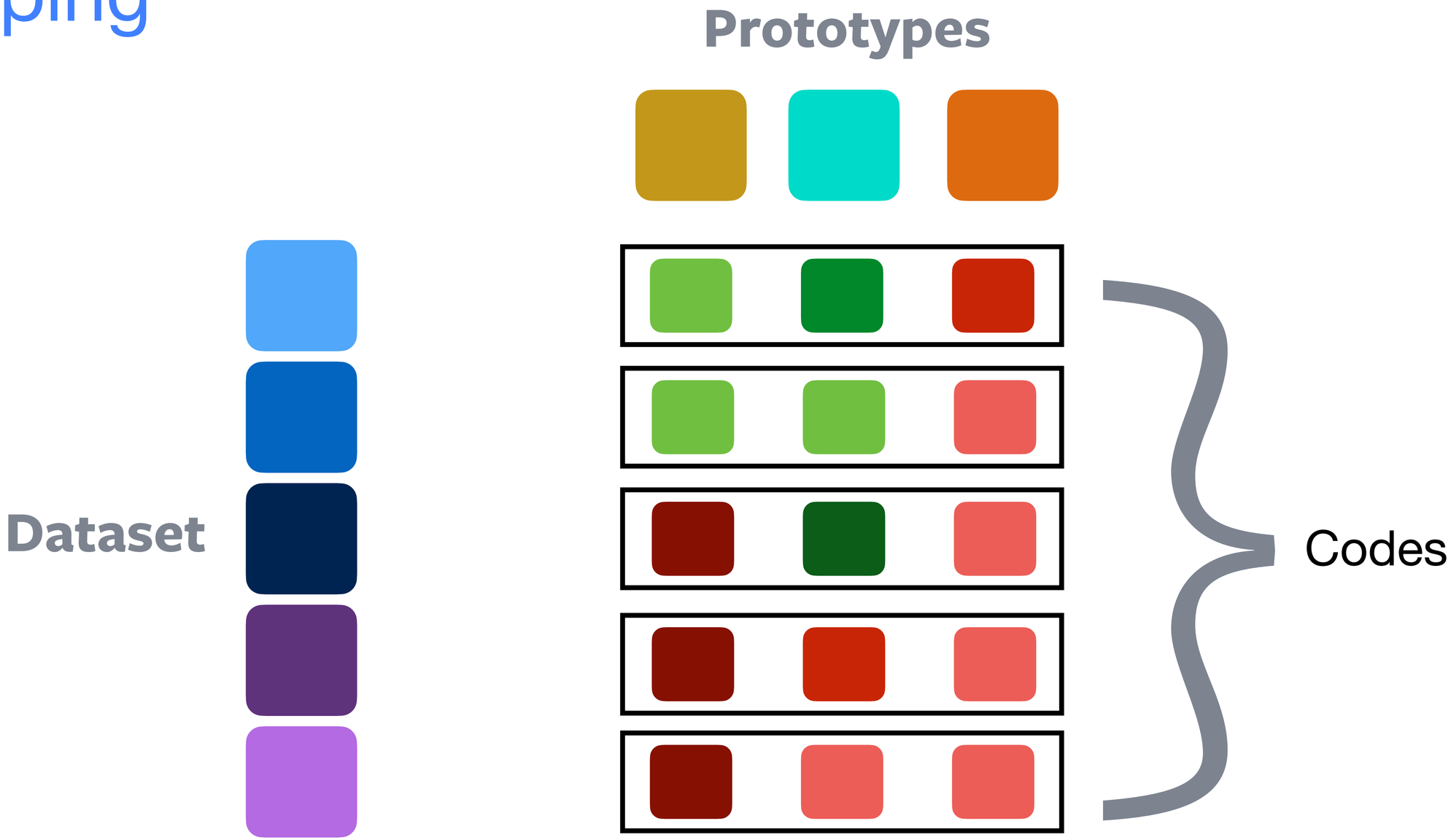
Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin

Grouping

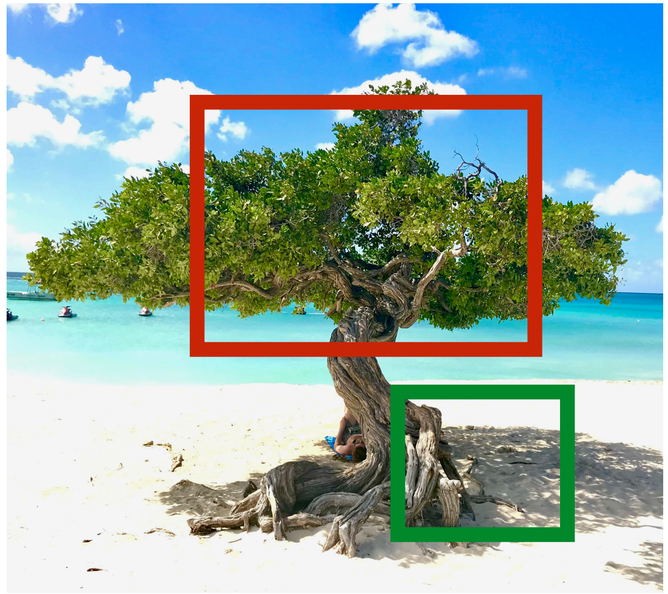


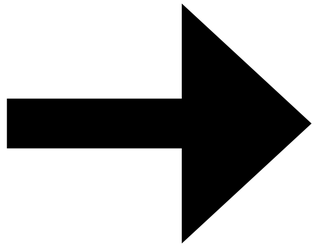
See also - *SeLa* by Asano et al., 2019

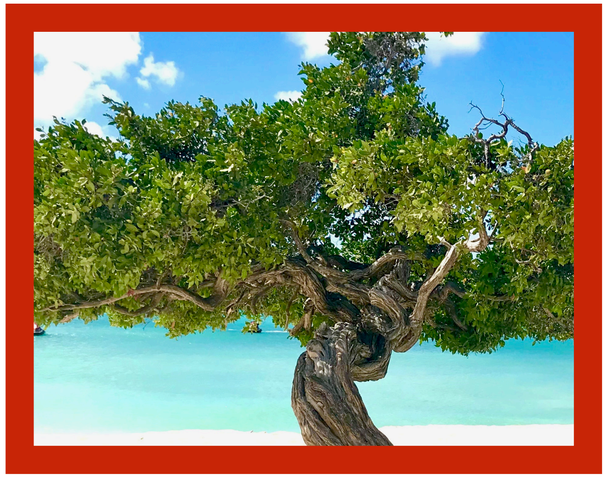
Grouping



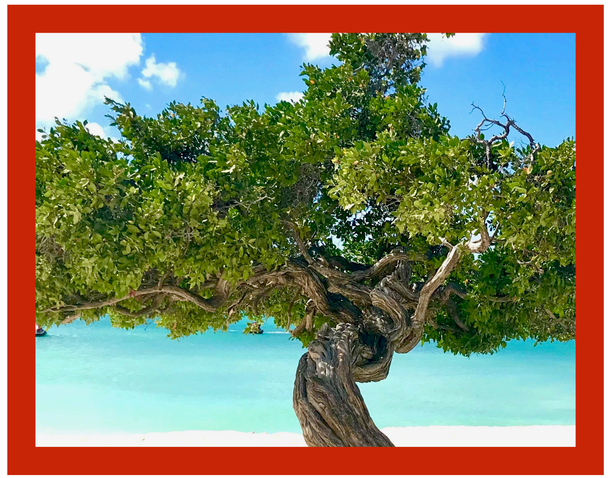
Multi-crop (Multi-view)

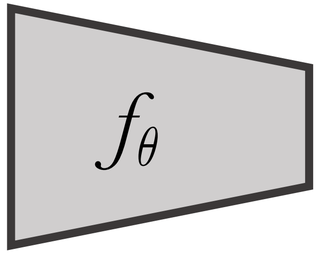


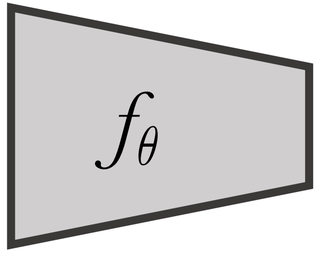




Compare crops of
different sizes & resolutions

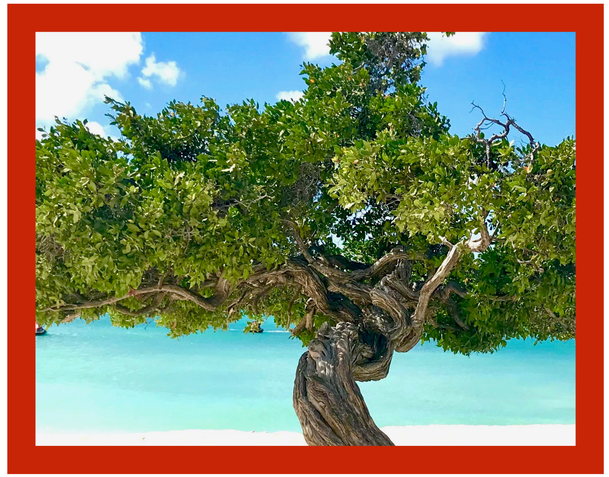






Code 1

Code 2



f_θ

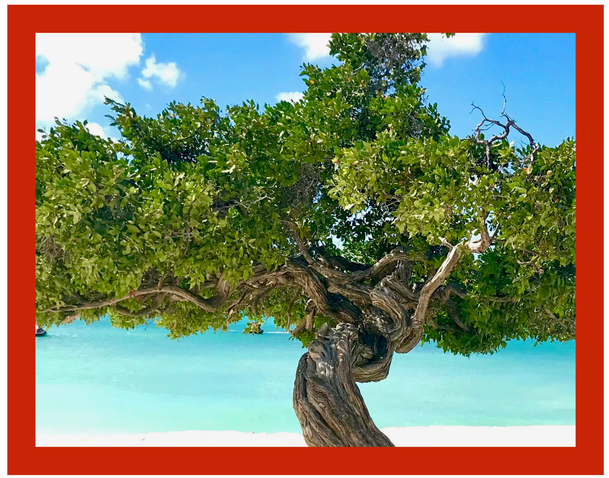
f_θ

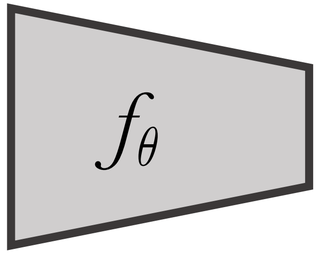
Prototypes

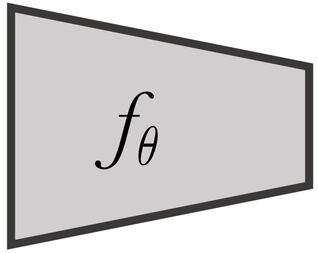
Code 1

Predict

Code 2







Backprop

Backprop

Prototypes

Code 1

Code 2

Not contrastive!

Multi-crop (Multi-view)

Method	ImageNet Top-1		Δ
		w/ Multi-crop	Multi-crop
SimCLR	68.2	70.6	+2.4
SeLa-v2	67.2	71.8	+4.6
DeepCluster-v2	70.2	74.3	+4.1
SwAV	70.1	74.1	+4.0

Multi-view improves ALL methods

Key Results

	Linear Classifier (Fixed Features)			Detection	
	ImageNet	**Places**	**iNaturalist**	**VOC07+12**	**COCO**
Supervised	76.5	53.2	46.7	81.3	40.8
Prior self-supervised	71.1 (-5.4)	52.1	38.9	82.5	42.0
SwAV	75.3 (-1.2)	**56.7**	**48.6**	**82.6**	**42.1**

Practical advantages of SwAV

- Trains on 4-8 GPUs
- **Faster convergence** than prior work (SimCLR, MoCov2)
  - Smaller compute requirements.
  - **2x faster** than MoCo-v2 on 8 GPUs
    - 72% after 100h vs. 71% after 200h

- Better results

Code & Models - https://github.com/facebookresearch/swav

PyTorch Lightning implementation on the way

Method	Multi-view Invariance	Grouping	Performance
Pretext Task	No	No	Weak
PIRL	Yes	Weak	Moderate
SwAV	Yes	Yes	Good

Outperforms ImageNet supervised pretraining on **all transfer tasks**

What invariances matter?

- Our set of invariances are overfit to ImageNet
- Need to evaluate on **different image distributions (uncurated data)**

Horizontal flipping may not be a good idea ...

Scalable objectives

- Contrastive learning converges very slowly & scales poorly to large data
- Notion of grouping is important

The Future ...

Method	Multi-view Invariance	Grouping	???	Performance
Pretext Task	No	No	No	Weak
PIRL	Yes	Weak	No	Moderate
SwAV	Yes	Yes	No	Good
Your method	Yes	Yes	Yes	**BEST**

Thanks!