Self-Supervised Learning by Cross-Modal Audio-Video Clustering

Humam Alwassel[1], Dhruv Mahajan[2], Bruno Korbar[2], Lorenzo Torresani[2], Bernard Ghanem[1], Du Tran[2]

[1] KAUST
[2] FACEBOOK AI

Motivation

- Fully supervised pre-training, followed by fine-tuning paradigm
  - Pros: work well with large enough data/annotations
  - Cons: **NOT** ***scalable*** and ***taxonomy dependent***.
- Audio-Visual correlation nature of videos
- Is it possible for self-supervised pre-training outperform fully-supervised ones?

Single-Modality vs. Multi-Modality Deep Clustering

Dataset	SDC	MDC	CDC	XDC
UCF101	61.8	68.4	72.9	**74.2**
HMDB51	31.4	37.1	37.5	**39.0**
ESC50	66.5	70.3	74.8	**78.0**

	same-modality-XDC	
Dataset	2 visual encoders	2 audio encoders
UCF101	61.3	N/A
HMDB51	30.5	N/A
ESC50	N/A	66.0

Pretraining Data Type and Size

	Pretraining		Downstream Dataset		
Method	Dataset	Size	UCF101	HMDB51	ESC50
Scratch	None	0	54.5	24.1	54.3
Superv	ImageNet	1.2M	79.9	44.5	NA
Superv	Kinetics	240K	90.9	58.0	82.3
Superv	AudioSet-240K	240K	76.6	40.8	78.3
Superv	AudioSet	2M	84.0	53.5	**90.3**
XDC	Kinetics	240K	74.2	39.0	78.0
XDC	AudioSet-240K	240K	77.4	42.6	78.5
XDC	AudioSet	2M	84.9	48.8	85.8
XDC	IG-Random	65M	88.8	61.2	86.3
XDC	IG-Kinetics	65M	**91.5**	**63.1**	84.8

Curated vs. Uncurated Pretraining Data

UCF101

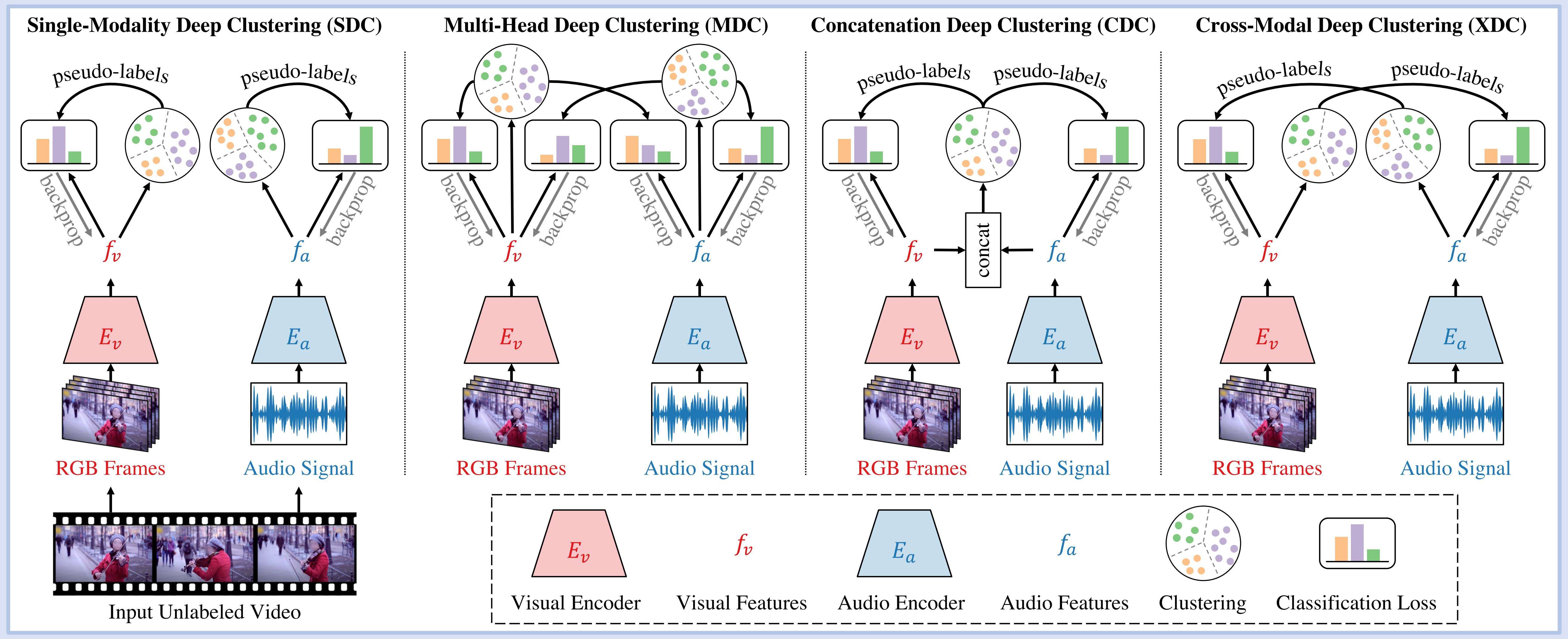
Pretraining Size	1M	16M	65M
IG-Random	79.6	84.1	88.8
IG-Kinetics	**84.2**	**87.6**	**91.5**
Δ	-4.6	-3.5	-2.7

HMDB51

Pretraining Size	1M	16M	65M
IG-Random	45.1	55.2	61.2
IG-Kinetics	**50.3**	**57.3**	**63.1**
Δ	-5.2	-2.1	-1.9

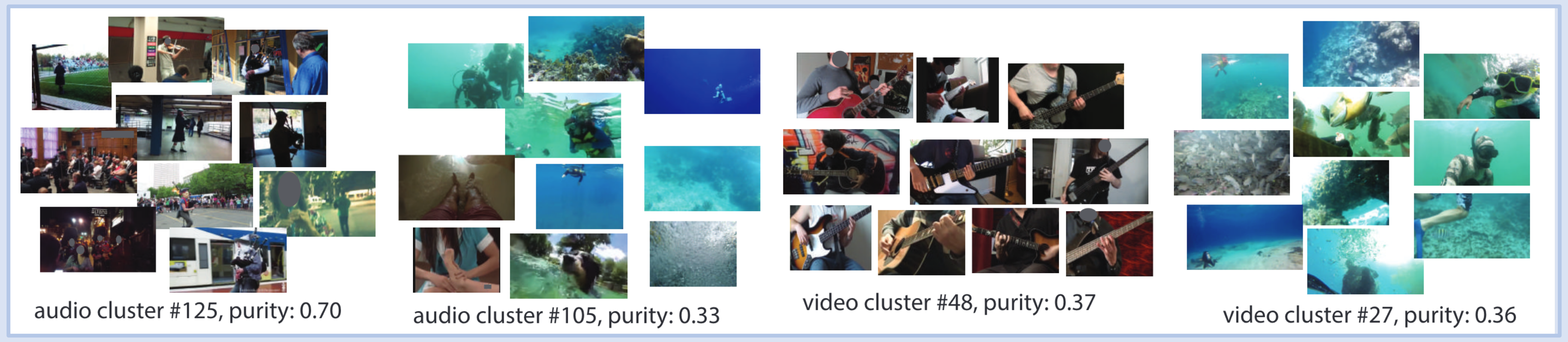
ESC50

Pretraining Size	1M	16M	65M
IG-Random	77.8	**84.3**	**86.3**
IG-Kinetics	**79.5**	82.5	84.8
Δ	-1.7	+1.8	+1.5



XDC is the first to show *self-supervision* **outperforming** large-scale *full-supervision* pretraining for action recognition when pretrained on the same architecture and a larger number of uncurated videos.

XDC Clusters Visualization



audio cluster #125, purity: 0.70
"playing bagpipes"

audio cluster #105, purity: 0.33
"scuba diving", "snorkeling"

video cluster #48, purity: 0.37
"play bass guitar", "play guitar", "tap guitar"

video cluster #27, purity: 0.36
"scuba diving", "feeding fish"

State-of-the-art Comparison

	Pretraining		Evaluation	
Method	Architecture	Dataset	UCF101	HMDB51
ClipOrder [79]	R(2+1)D-18	UCF101	72.4	30.9
MotionPred [72]	C3D	Kinetics	61.2	33.4
ST-Puzzle [28]	3D-ResNet18	Kinetics	65.8	33.7
DPC [18]	3D-ResNet34	Kinetics	75.7	35.7
CBT [64]	S3D	Kinetics	79.5	44.6
SpeedNet [4]	S3D	Kinetics	81.1	48.8
AVTS [29]*	MC3-18	Kinetics	84.1	52.5
AVTS [29]†	R(2+1)D-18	Kinetics	86.2	52.3
XDC (ours)	R(2+1)D-18	Kinetics	86.8	52.6
AVTS [29]*	MC3-18	AudioSet	87.7	57.3
AVTS [29]†	R(2+1)D-18	AudioSet	89.1	58.1
XDC (ours)	R(2+1)D-18	AudioSet	93.0	63.7
MIL-NCE [38]	S3D	HowTo100M	91.3	61.0
ELo [50]	R(2+1)D-50	YouTube-8M	93.8	67.4
XDC (ours)	R(2+1)D-18	IG-Random	94.6	66.5
XDC (ours)	R(2+1)D-18	IG-Kinetics	**95.5**	**68.9**
Fully supervised	R(2+1)D-18	ImageNet	84.0	48.1
Fully supervised	R(2+1)D-18	Kinetics	94.2	65.1

Method	ESC50
Piczak ConvNet [47]	64.5
SoundNet [2]	74.2
L^3-Net [1]	79.3
AVTS [28]	82.3
ConvRBM [52]	**86.5**
XDC (AudioSet)	84.8
XDC (IG-Random)	85.4

Method	DCASE
RNH [50]	77
Ensemble [56]	78
SoundNet [2]	88
L^3-Net [1]	93
AVTS [28]	94
XDC (AudioSet)	**95**
XDC (IG-Random)	**95**

XDC for Temporal Action Localization on THUMOS14

	mAP @ tIoU				
Features Type	0.3	0.4	0.5	0.6	0.7
Superv (Kinetics)	50.9	44.4	36.6	28.4	19.8
XDC (IG-Random)	**51.5**	44.8	36.9	28.6	**20.0**
XDC (IG-Kinetics)	**51.5**	**44.9**	**37.2**	**28.7**	**20.0**

Conclusion

- **Cross-modal correlation** helps self-supervised learning
- XDC is **simple**, **scalable**, **taxonomy-** and **downstream task-independent**
- XDC **outperforms** Kinetics and ImageNet ***fully-supervised*** pretraining