ICLR 2021

Self-Supervised Learning of Compressed Video Representations

Youngjae Yu *

Sangho Lee *

Gunhee Kim

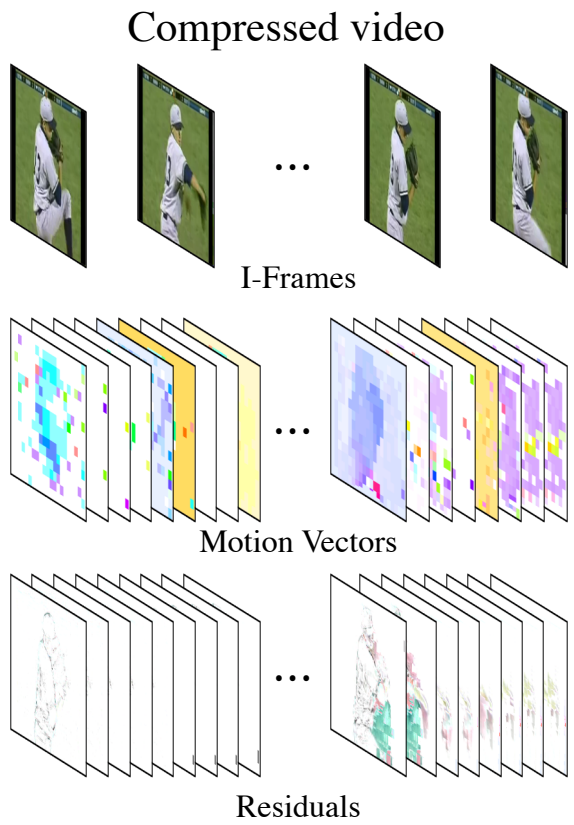
Yale Song

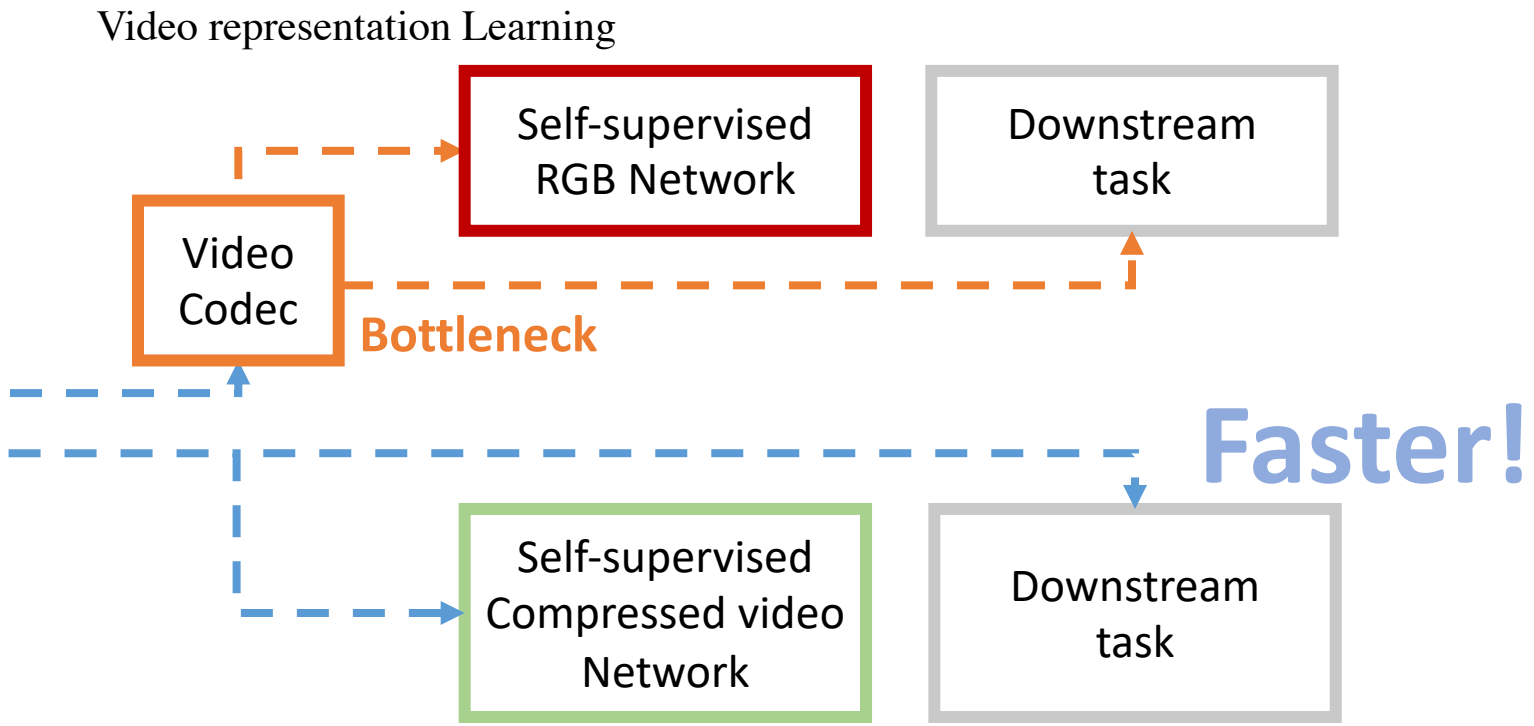
Microsoft

SEOUL NATIONAL UNIV.
VISION & LEARNING

Motivation

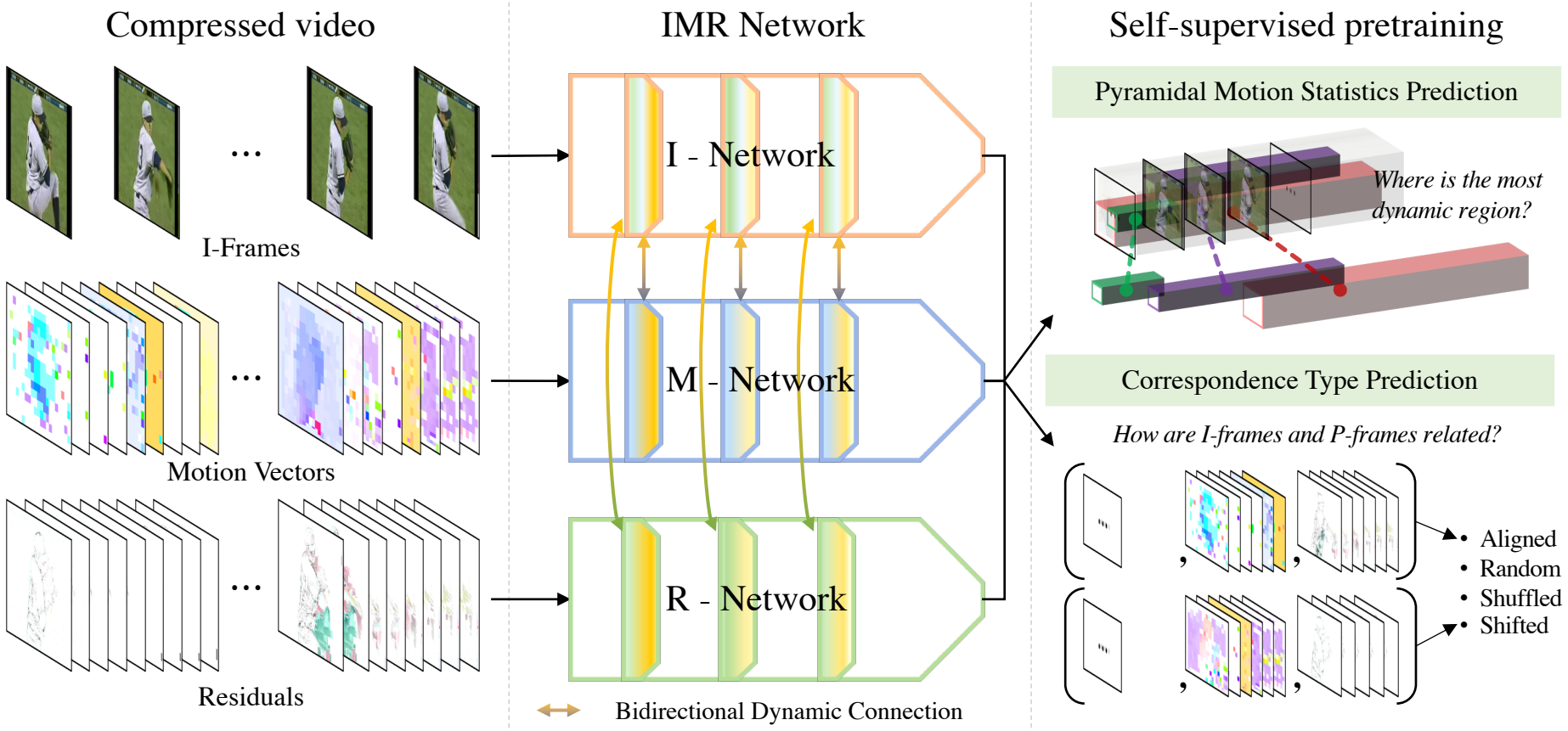
- Self-supervised learning of video representations by eliminating the expensive RGB video frame decoding step
- A novel three-stream video architecture that encodes I-frames and P-frames (Motion, Residuals)





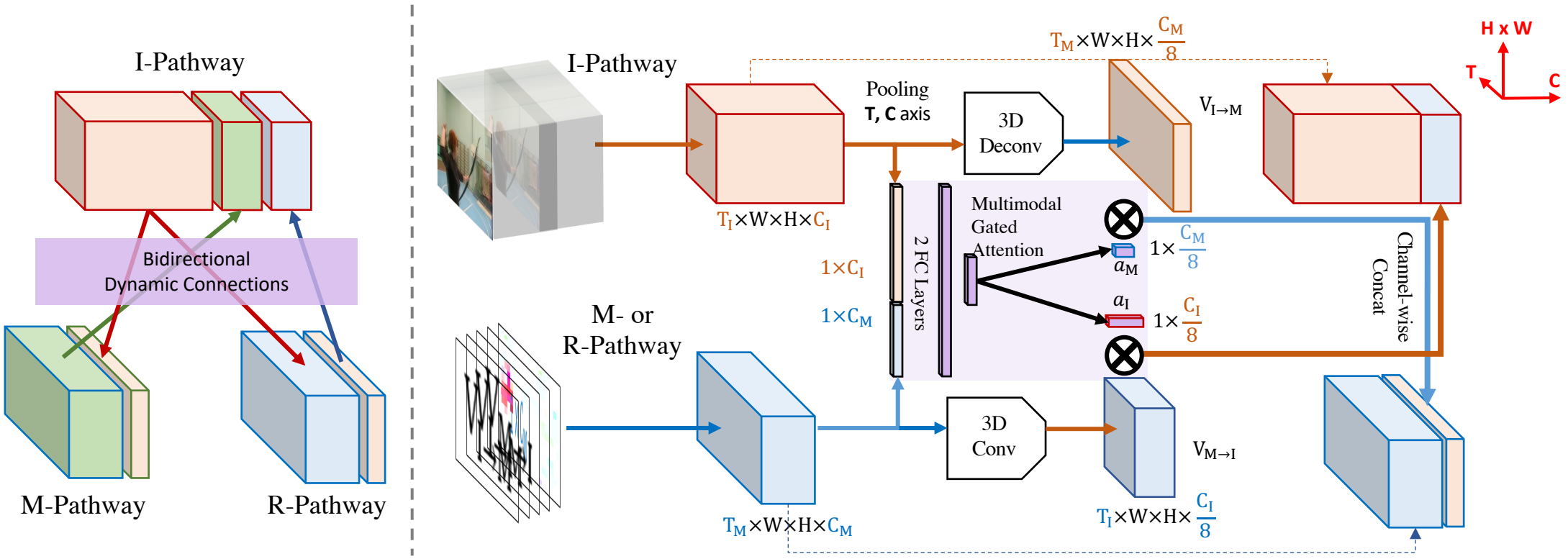
Objective – Representation learning for compressed video streams

- We train the model using two novel pretext tasks designed by exploiting the underlying structure of compressed videos



Solution – IMR Network

- Encode different information streams provided in compressed videos
- Use bidirectional dynamic connections to facilitate information sharing across stream

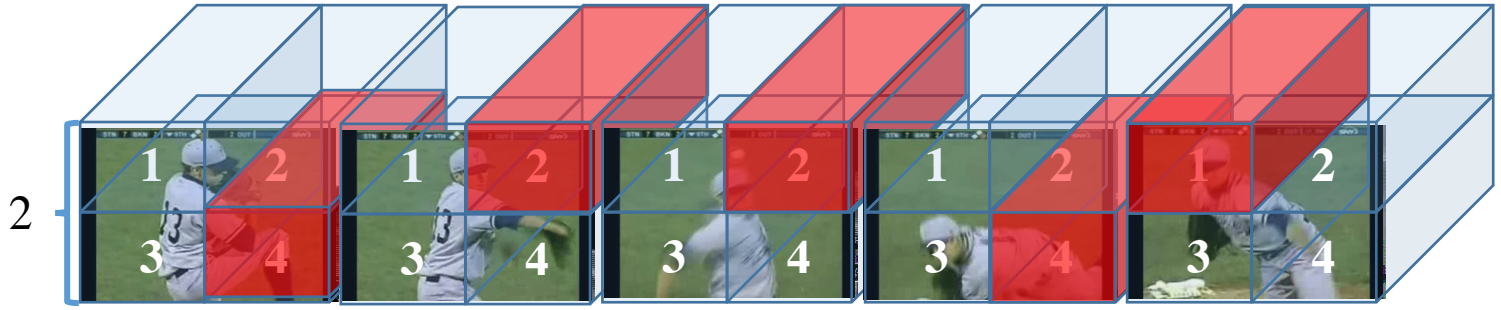


Solution – Self-supervised Pretext task (PMSP)

Pyramidal motion statistics prediction task

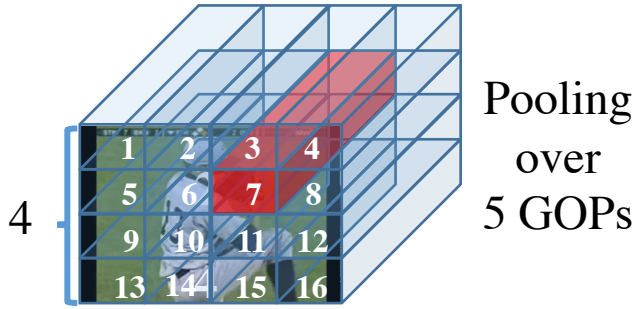
- Make our network find a region with the highest energy of motion
- 2-layer MLP with a *softmax* classifier predict the most vibrant region in the given 3D grid

(a) r = ([2×2], 5), Ans : [4, 2, 2, 4, 1]



4 way Classification × [t=5]

(b) r = ([4×4], 1), Ans : 7



16 way Classification × [t=1]

Solution – Self-supervised Pretext task (PMSP)

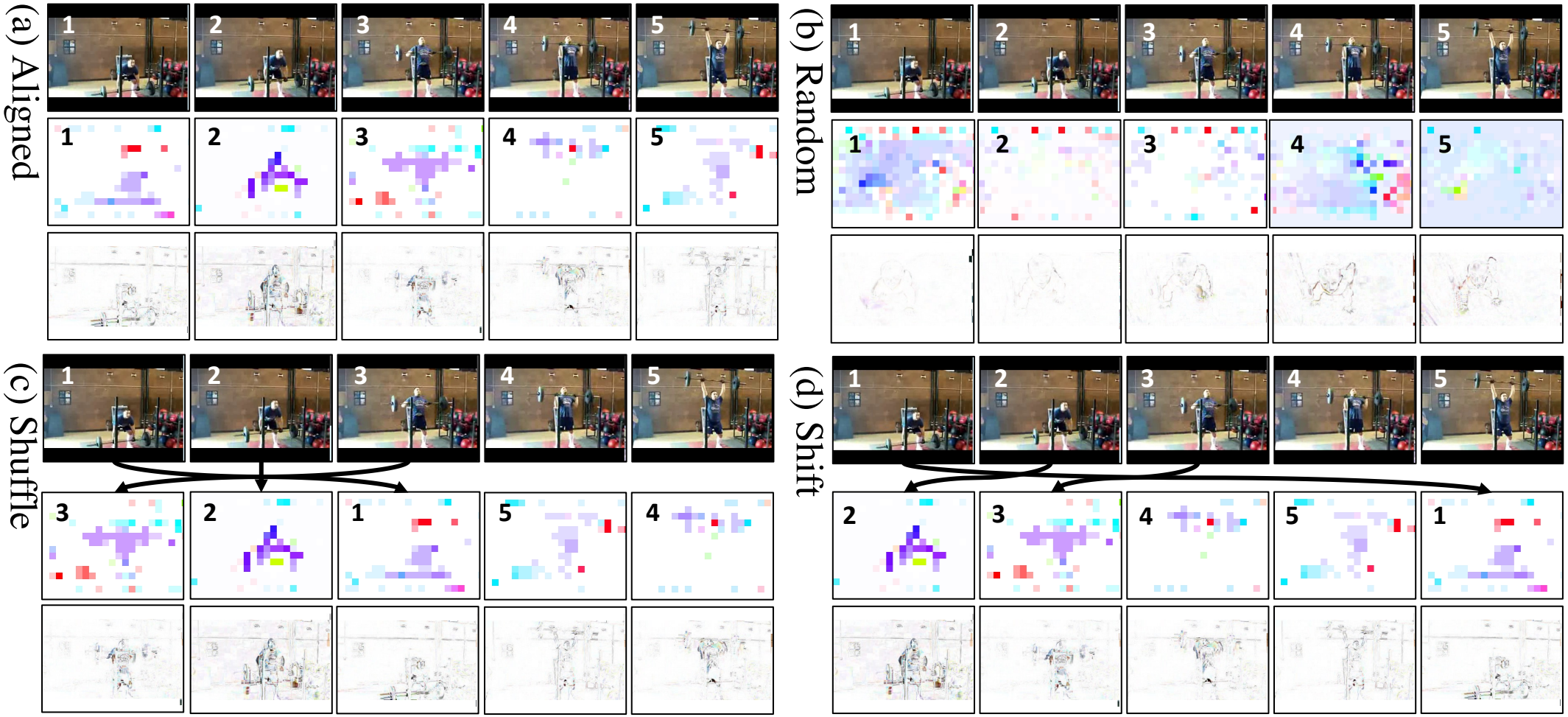
Pyramidal motion statistics prediction task

- Implicit videographer bias captured in videos that naturally reflect visual saliency

Solution – Self-supervised Pretext task (CTP)

Correspondence type prediction task

- Make our network categorize different types of transformations applied on P-frame



Experiments – Compressed Video Classification

Achieve state-of-the-art performance in both self-supervised/supervised regimes

- While maintaining a similar computational efficiency as existing compressed video recognition approaches

Models	Compressed	Modality	Pretext	Pretrain	Backbone	UCF101	HMDB51
C3D	✗	V	MotPred	Kinetics400	C3D	61.2	33.4
3D-ResNet18	✗	V	RotNet3D	Kinetics600	3D-ResNet18	62.9	33.7
3D-ResNet18	✗	V	ST-Puzzle	Kinetics400	3D-ResNet18	65.8	33.7
R(2+1)D-18	✗	V	ClipOrder	UCF101	R(2+1)D-18	72.4	30.9
3D-ResNet34	✗	V	DPC	Kinetics400	3D-ResNet34	75.7	35.7
Multisensory	✗	A+V	Multisensory	Kinetics400	Audio-VisualNet	82.1	–
AVTS	✗	A+V	AVTS	Audioset	MC3	89.0	61.6
ELo	✗	A+V	ELo	Kinetics400	(2+1)D ResNet-50	93.8	67.4
CoViAR‡	✓	V	Scratch	None	ResNet152	43.8	27.3
IMRNet	✓	V	Scratch	None	3D-ResNet18	74.1	43.7
CoViAR‡	✓	V	AOT	Kinetics400	ResNet152	53.6	29.3
CoViAR‡	✓	V	Rotation	Kinetics400	ResNet152	56.7	31.4
IMRNet	✓	V	InfoNCE	Kinetics400	3D-ResNet18	73.9	43.7
IMRNet	✓	V	AOT	Kinetics400	3D-ResNet18	74.6	44.0
IMRNet	✓	V	Rotation	Kinetics400	3D-ResNet18	75.1	44.3
CoViAR‡	✓	V	PMSP	Kinetics400	ResNet152	63.5	35.9
CoViAR‡	✓	V	CTP	Kinetics400	ResNet152	64.4	37.4
CoViAR‡	✓	V	CTP (Binary)	Kinetics400	ResNet152	63.7	37.1
IMRNet	✓	V	PMSP	Kinetics400	3D-ResNet18	76.0	44.9
IMRNet	✓	V	CTP	Kinetics400	3D-ResNet18	76.7	44.8
IMRNet	✓	V	CTP (Binary)	Kinetics400	3D-ResNet18	74.6	44.2
IMRNet	✓	V	PMSP+CTP	Kinetics400	3D-ResNet18	**76.8**	**45.0**

Self-supervised setting

Models	OF	Pretrain	Backbone	UCF101	HMDB51
CoViAR‡	✗	Scratch	ResNet152	43.8	27.3
IMR (No connection)	✗	Scratch	3D-ResNet18	52.7	34.6
IMR (Unidirectional)	✗	Scratch	3D-ResNet18	69.7	40.8
IMR (No conv)	✗	Scratch	3D-ResNet18	71.7	42.6
IMR (No attention)	✗	Scratch	3D-ResNet18	73.2	43.5
IMRNet	✗	Scratch	3D-ResNet18	74.1	43.7
IMRNet	✗	Scratch	3D-ResNet50	80.2	55.9
CoViAR†	✗	ImageNet	ResNet152 (I), ResNet18 (P)	90.4	59.1
CoViAR‡	✗	Kinetics400	ResNet152	90.8	59.2
IMRNet (Ours)	✗	Kinetics400	3D-ResNet18	91.4	62.8
IMRNet (Ours)	✗	Kinetics400	3D-ResNet50	**92.6**	**67.8**
CoViAR†	✓	ImageNet	ResNet152 (I), ResNet18 (P, OF)	94.9	70.2
DMC-Net†	✓	ImageNet	ResNet152 (I), ResNet18 (P)	90.9	62.8
DMC-Net†	✓	ImageNet	ResNet152 (I), I3D (P)	92.3	71.8
IMRNet (Ours)	✓	Kinetics400	3D-ResNet50 (I, P), I3D (OF)	**95.1**	**72.2**

Supervised setting

Models	ResNet152*	R(2+1)D†	CoViAR‡	DMC‡	IMR‡ (R18)	IMR‡ (R50)
Preprocess (ms)	75.00	75.00	2.87	2.87	2.87	2.87
Inference (ms)	7.50	1.74	1.30	1.91	1.36	1.44
Total (ms)	82.50	76.74	4.17	4.78	4.23	4.31
GFLOPs	11.3	0.96	4.2	4.4	0.66	1.04

Runtime Analysis

Self-Supervised Learning of Compressed Video Representations

- Efficient self-supervised approach to learn video representations
- IMR Network - three-stream video architecture that encodes a compressed video
- For details, please refer to our paper
  - Project page : http://vision.snu.ac.kr/projects/compressedvideo/