2021 ICCV OCTOBER 11-17 VIRTUAL

Motion-Augmented Self-Training for Video Recognition at Smaller Scale

Kirill Gavrilyuk, Mihir Jain, Ilia Karmanov, Cees G.M. Snoek

UNIVERSITY OF AMSTERDAM

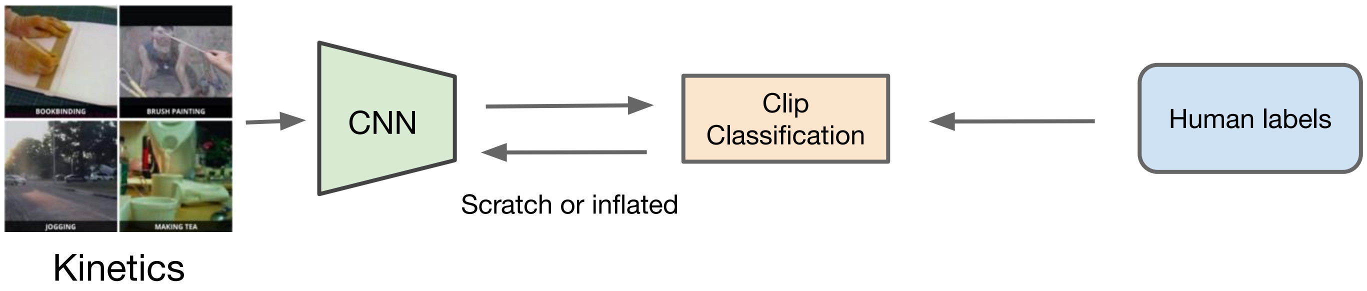
Qualcomm

Goal

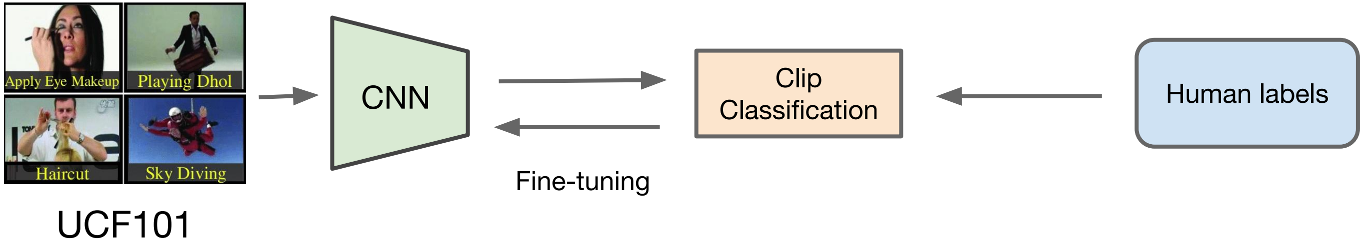
- Self-train a model that can be effectively fine-tuned on **small-scale** datasets with around 10k or even less videos
- Small-scale video datasets benefit more from motion than appearance, but the **flow computation** affects efficiency
- **We strive to train a network using optical flow but avoid its computation during inference on small-scale video datasets**

Standard pipeline

1. Pretraining on large labeled dataset



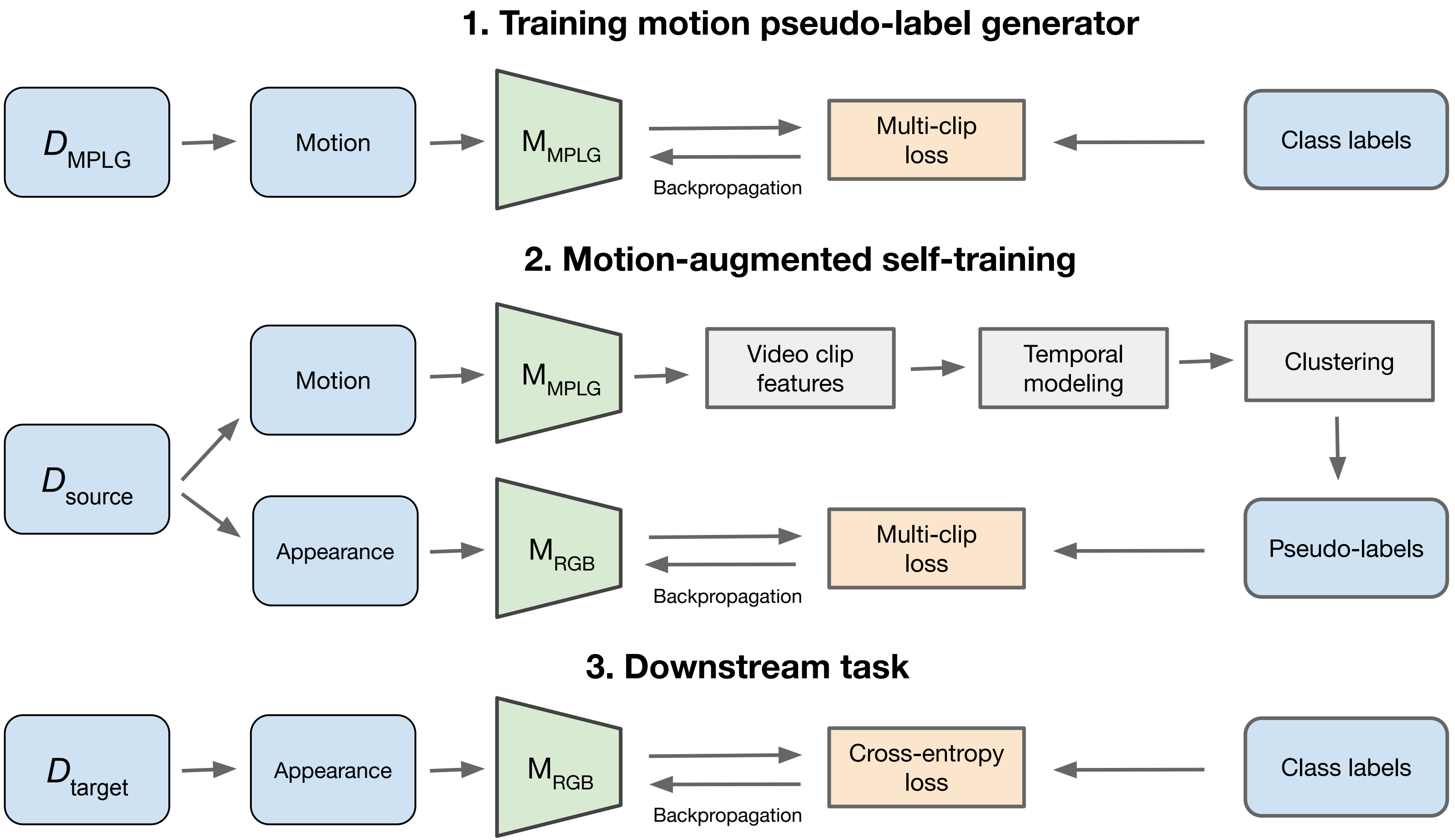
2. Fine-tuning on small labeled dataset



Shortcomings

- Requires large amount of **human annotated** videos to pre-train the model
- Treats different video representations (motion, appearance) mostly **equally**

Model

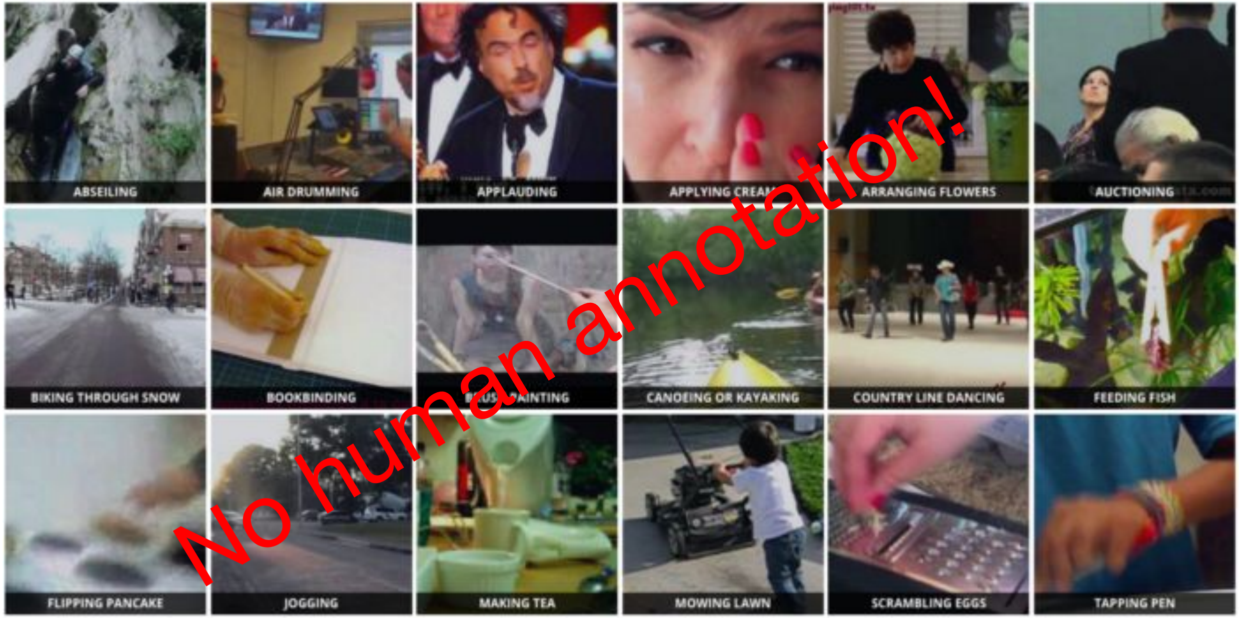


1. Training motion pseudo-label generator
D_{MPLG}
Motion
M_{MPLG}
Backpropagation
Multi-clip loss
Class labels
2. Motion-augmented self-training
D_{source}
Motion
M_{MPLG}
Video clip features
Temporal modeling
Clustering
Appearance
M_{RGB}
Backpropagation
Multi-clip loss
Pseudo-labels
3. Downstream task
D_{target}
Appearance
M_{RGB}
Backpropagation
Cross-entropy loss
Class labels

Contributions

- **MotionFit**: A motion-augmented self-training procedure to transfer motion knowledge to the appearance model
- **Empirical study** to discover form of video pseudo-labels at smaller scale
- **Boosting** the performance on small-scale video datasets in comparison to the state-of-the-art methods

Source dataset



Kinetics-400: ~246k train videos, 50k validation videos

Target datasets

UCF101: 13k videos

HMDB51: 7k videos

Motion pseudo-label generator

		Clip length		Multi-clips (R)			
Dataset	**Representation**	32	64	1	2	3	4
UCF101	Appearance	59.4	60.3	58.9	57.0	59.1	58.4
	Motion	80.8	81.1	78.2	82.2	82.6	82.8
HMDB51	Appearance	20.2	20.4	20.1	20.9	19.1	18.9
	Motion	35.9	35.0	29.7	35.1	35.9	37.9

Training on motion representation is more effective.

Multi-clips helps even more than using larger temporal extent

MotionFit: temporal modeling

Temporal granularity			
Video	ActionBytes	TSN	Clip
76.5	79.0	77.3	80.3

Simple clip-level beats *semantic* partitions and video-level

MotionFit: knowledge transfer comparison

	Backbone	Frames	Resolution	Additional labels	UCF101	HMDB51
Random initialisation	R(2+1)D-18	16	112	–	58.9	22.0
MERS	R(2+1)D-18	16	112	–	78.3	42.1
MARS	R(2+1)D-18	16	112	–	82.2	48.7
STC	STC-ResNext	16	112	ImageNet	84.7	-
DistInit	R(2+1)D-18	32	112	ImageNet	85.7	54.9
Supervised	R(2+1)D-18	16	112	Kinetics-400	95.0	70.4
MotionFit (*ours*)	R(2+1)D-18	16	112	–	87.4	56.4

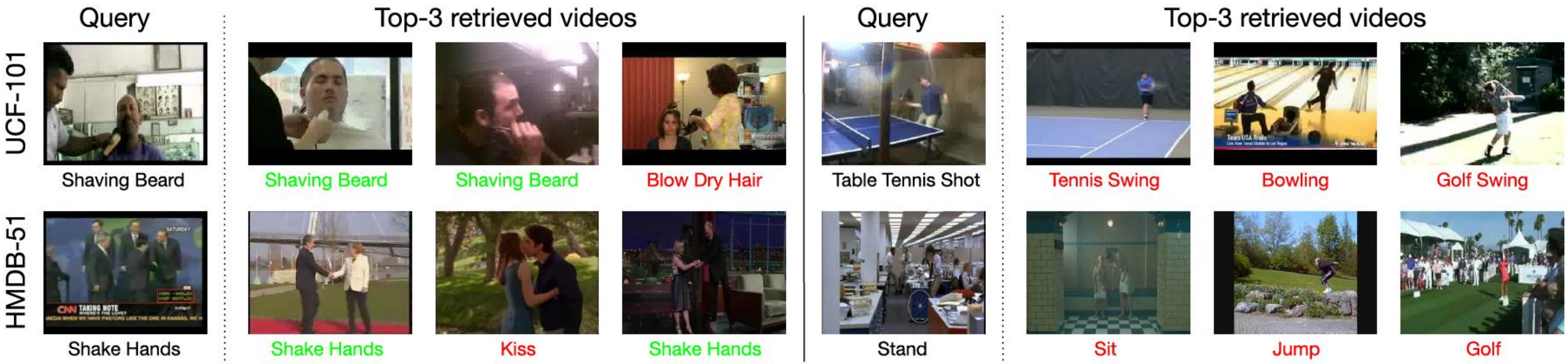
Our approach outperforms knowledge transfer methods even when they rely on additional labels

MotionFit: self-supervised comparison

		Backbone	Frames	Resolution	Modality	UCF101	HMDB51
Multi-modal	Sun *et al.*	S3D	16	112	V + T	79.5	44.6
	Asano *et al.*	R(2+1)D-18	30	112	V + A	83.1	47.1
	Alwassel *et al.*	R(2+1)D-18	32	224	V + A	86.8	52.6
	Xiao *et al.*	SlowFast	64	224	V + A	87.0	54.6
	Morgado *et al.*	R(2+1)D-18	32	224	V + A	87.5	60.8
	Patrick *et al.*	R(2+1)D-18	32	224	V + A	89.3	60.0
Vision-only	Kim *et al.*	R3D-18	16	112	V	65.8	33.7
	Kong *et al.*	R3D-18	8	112	V	69.4	37.8
	Han *et al.*	R-2D3D-34	25	224	V	75.7	35.7
	Jing *et al.*	R3D-18	64	112	V	76.6	47.0
	Zhuang *et al.*	SlowFast	16	112	V	77.0	46.5
	Han *et al.*	R-2D3D-18	25	224	V	78.1	41.2
	Benaim *et al.*	S3D-G	64	224	V	81.1	48.8
	Han *et al.*	S3D	32	128	V	87.9	54.6
	MotionFit *(ours)*	R(2+1)D-18	32	112	V	88.9	61.4
	MotionFit *(ours)*	S3D-G	64	224	V	90.1	50.6

Our approach outperforms most video-only methods being on par with multi-modal self-supervised methods

Retrieval examples



While our model may retrieve videos from different action classes, it still captures distinctive motion patterns like hand motion and human poses

Conclusion

- Motion representation can be successfully transfered to the appearance model via pseudo-labeling self-training on large unlabeled dataset

- Does not require costly optical flow computation during inference

- Well suited for deployment on small-scale video with compute budgets