A Brief Introduction to Unsupervised Representation Learning

王宁

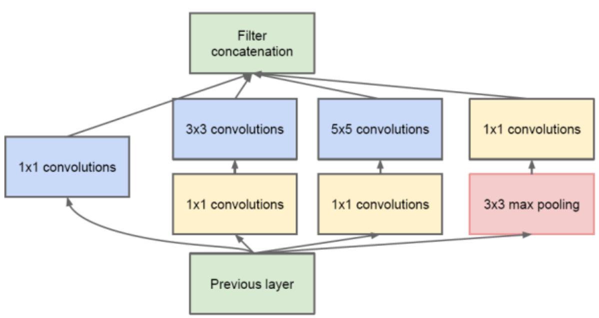
2017.10.27

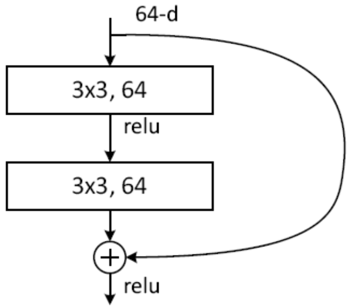
Introduction

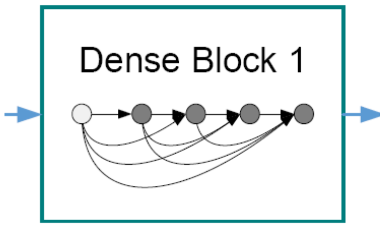
1. What is representation learning?

e.g., **ICLR:** International Conference of Learning Representation

Established by *Lecun*, *Hinton* and *Bengio* in 2013.

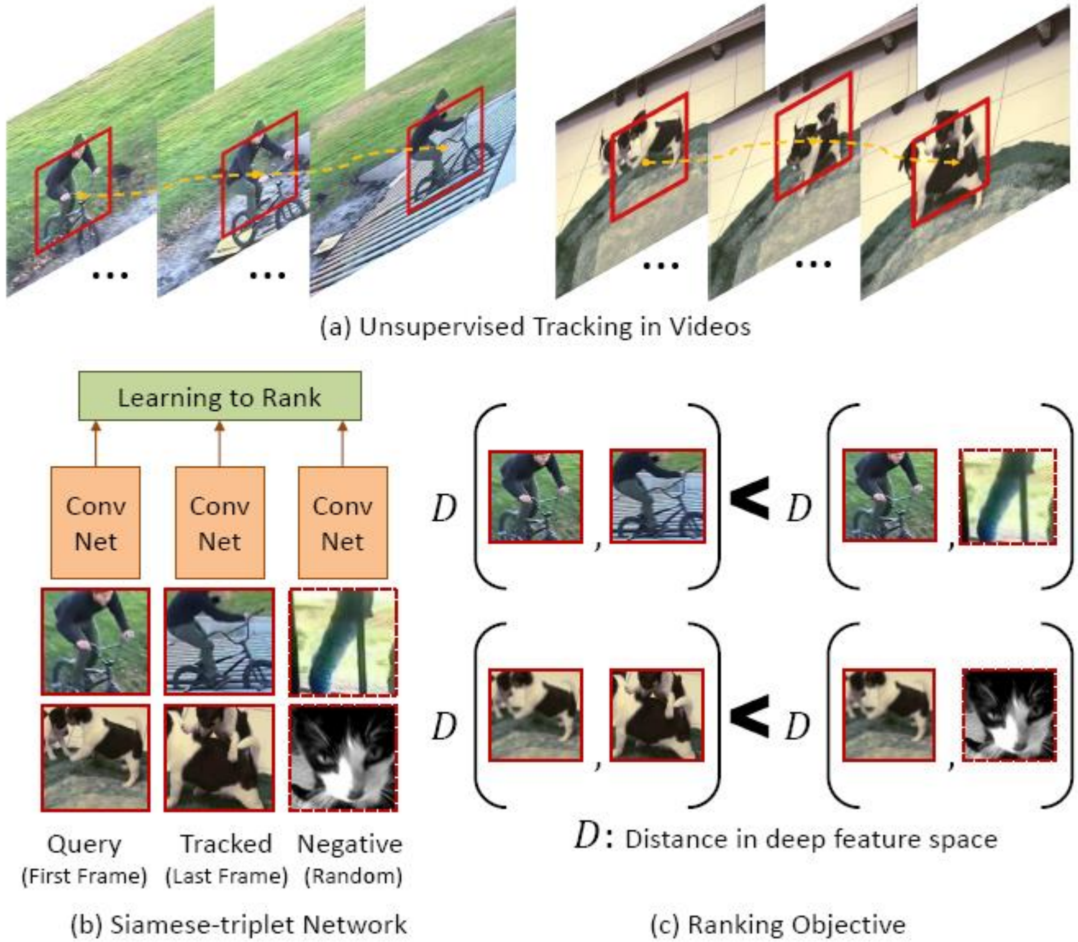






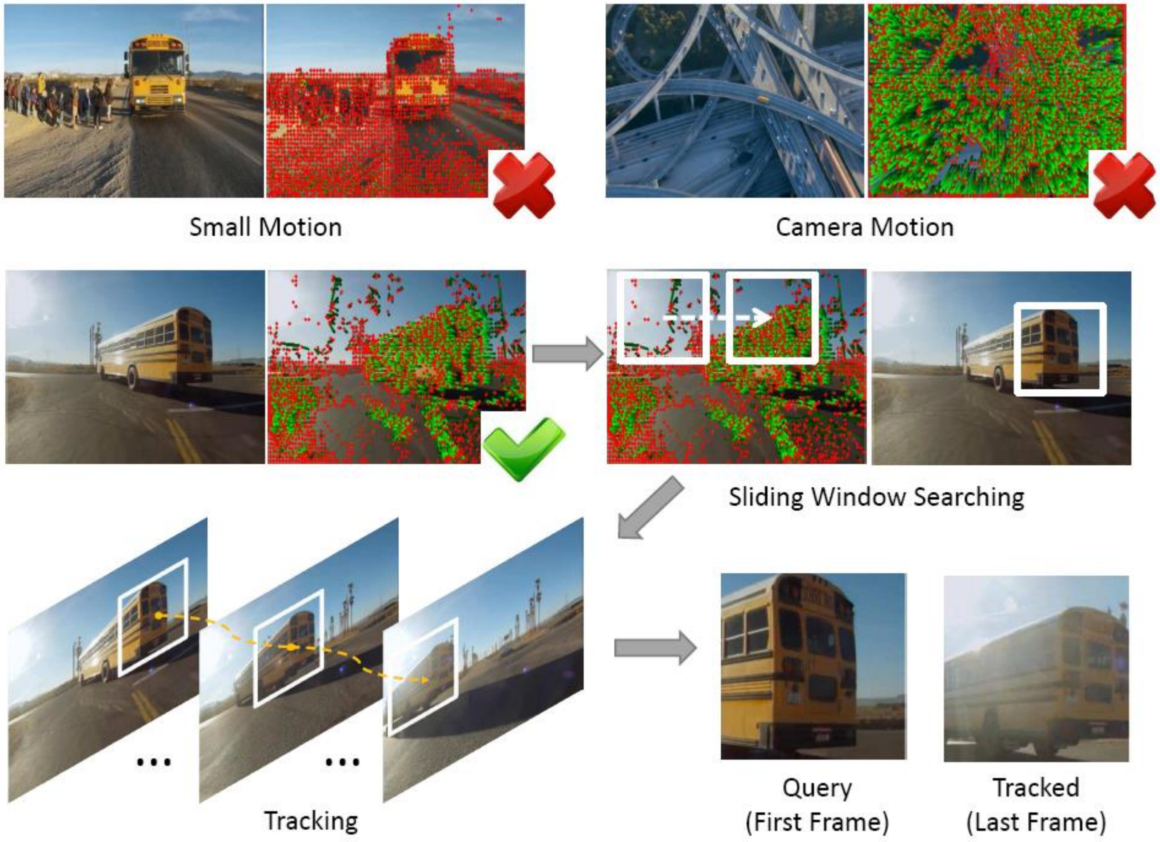
Unsupervised Learning of Visual Representations using Videos

Xiaolong Wang, Abhinav Gupta
Robotics Institute, Carnegie Mellon University

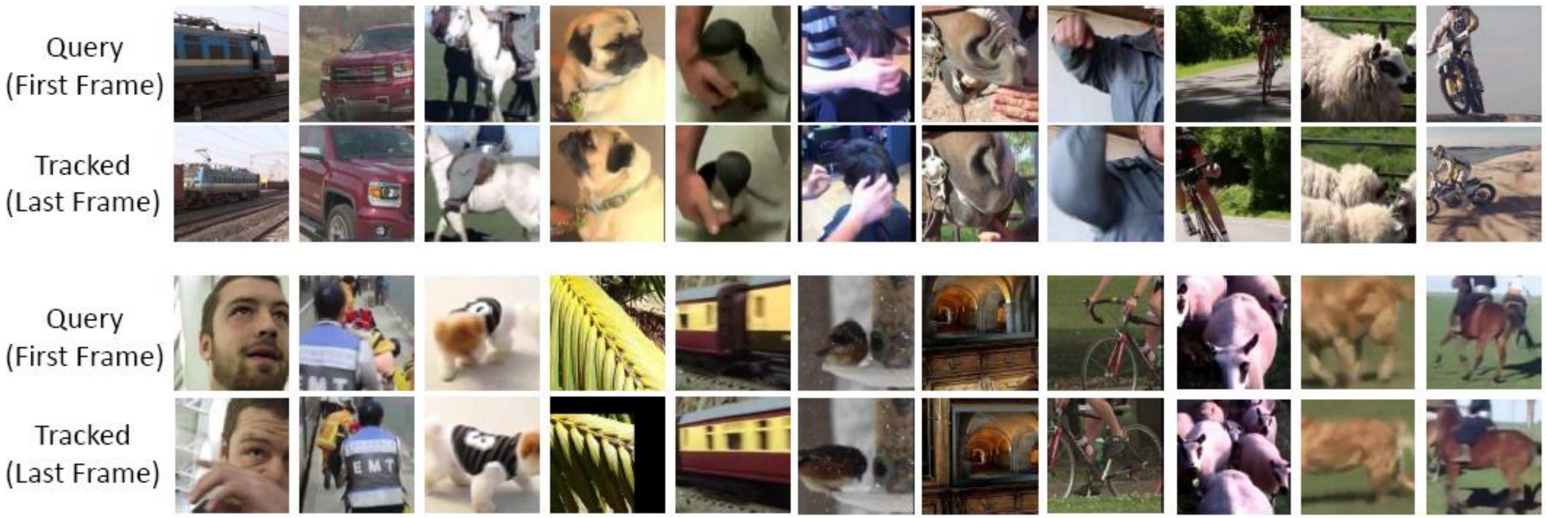


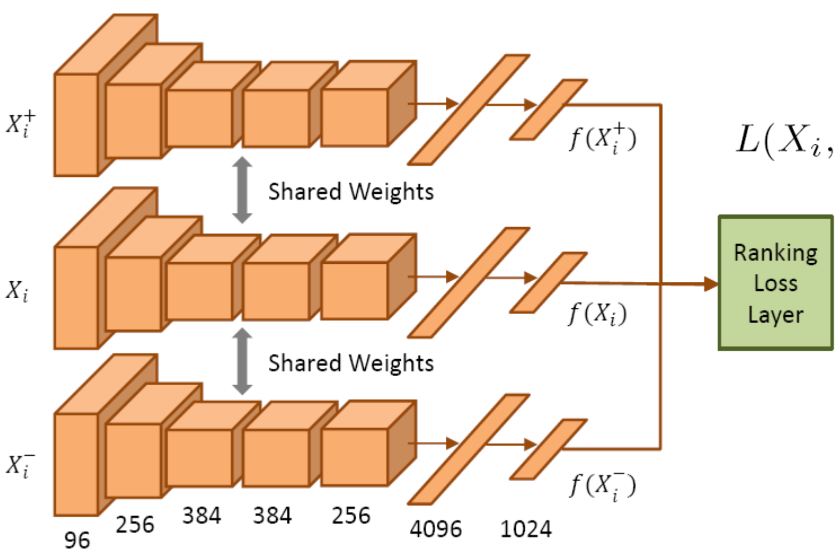
Unsupervised Learning of Visual Representations using Videos

1. Obtain SURF interest points and use Improved Dense Trajectories (IDT) for point motion estimation.
2. Utilize KCF tracker to track the object.



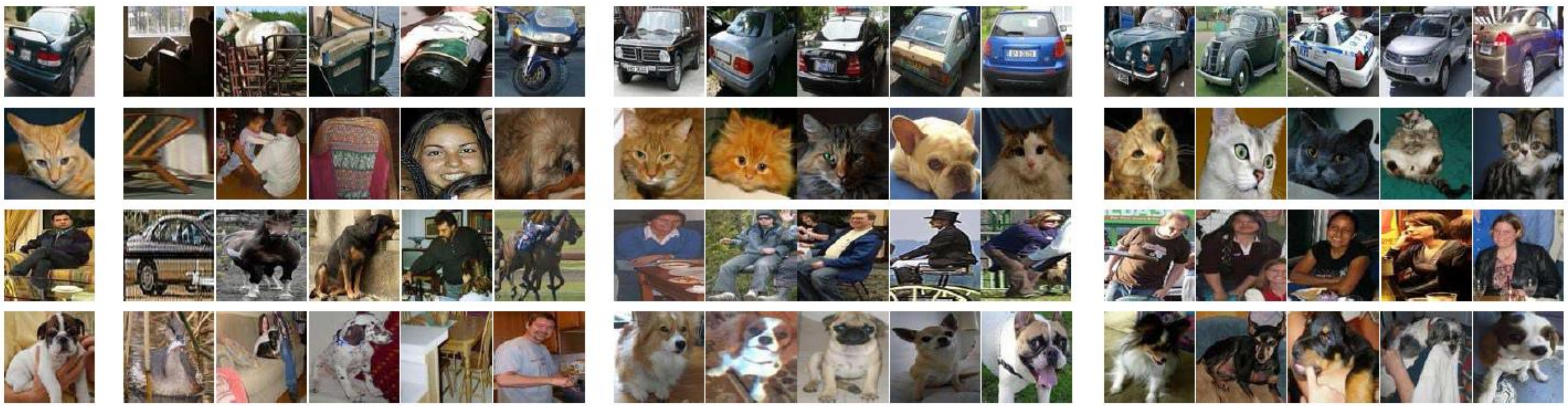
Unsupervised Learning of Visual Representations using Videos





$$L(X_i, X_i^+, X_i^-) = \max\{0, D(X_i, X_i^+) - D(X_i, X_i^-) + M\}$$

Unsupervised Learning of Visual Representations using Videos



Query (a) Random AlexNet (b) Imagenet AlexNet (c) Unsupervised AlexNet

Table 1. mean Average Precision (mAP) on VOC 2012. "external" column shows the number of patches used to pre-train unsupervised-CNN.

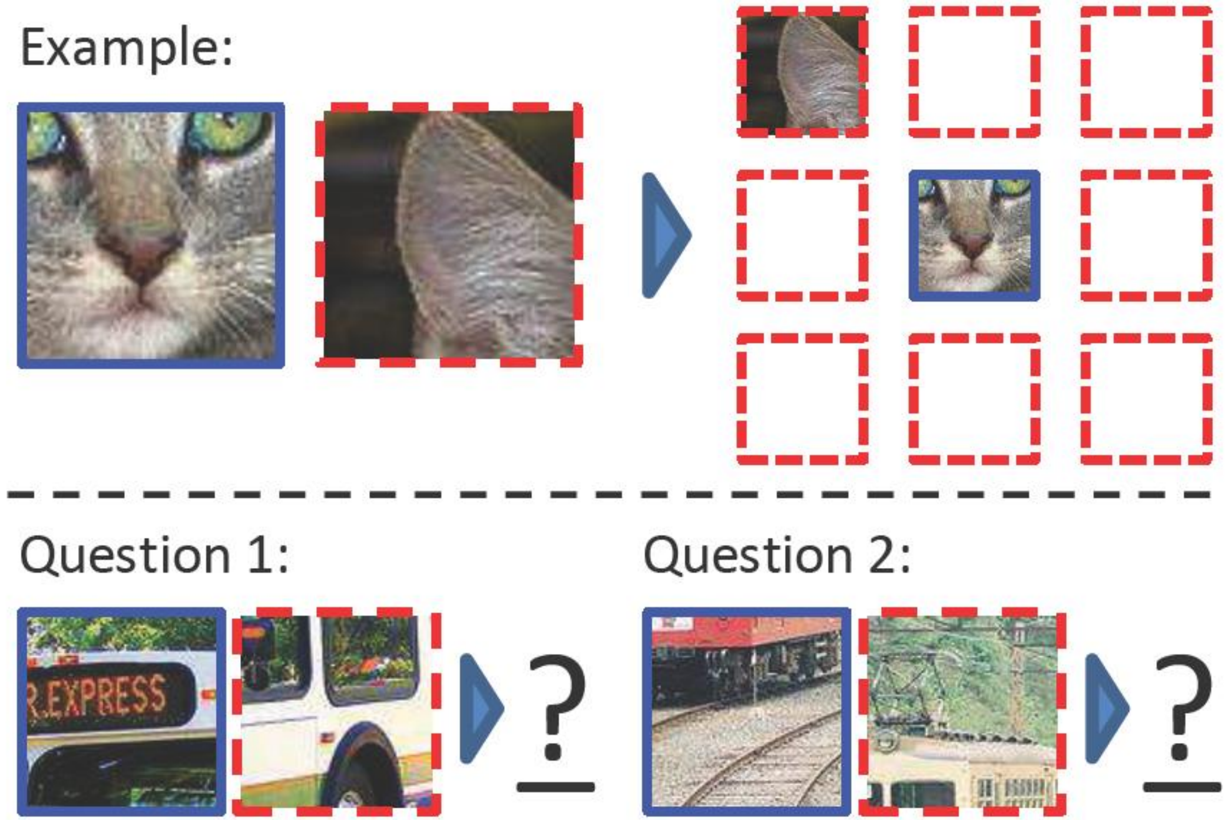
VOC 2012 test	external	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
scratch	0	66.1	58.1	32.7	23.0	21.8	54.5	56.4	50.8	21.6	42.2	31.8	49.2	49.8	61.6	52.1	25.1	52.6	31.3	50.0	49.1	44.0
scratch (3 ensemble)	0	68.7	61.2	36.1	25.7	24.3	58.9	58.8	55.3	24.4	43.5	36.7	53.0	53.8	65.6	54.3	27.3	53.5	38.3	54.6	51.8	47.3
unsup + ft	1.5M	68.8	62.1	34.7	25.3	26.6	57.7	59.6	56.3	22.0	42.6	33.8	52.3	50.3	65.6	53.9	25.8	51.5	32.3	51.7	51.8	46.2
unsup + ft	5M	69.0	64.0	37.1	23.6	24.6	58.7	58.9	59.6	22.3	46.0	35.1	53.3	53.7	66.9	54.1	25.4	52.9	31.2	51.9	51.8	47.0
unsup + ft	8M	67.6	63.4	37.3	27.6	24.0	58.7	59.9	59.5	23.7	46.3	37.6	54.8	54.7	66.4	54.8	25.8	52.5	31.2	52.6	52.6	47.5
unsup + ft (2 ensemble)	6.5M	72.4	66.2	41.3	26.4	26.8	61.0	61.9	63.1	25.3	51.0	38.7	58.1	58.3	70.0	56.2	**28.6**	56.1	38.5	55.9	54.3	50.5
unsup + ft (3 ensemble)	8M	73.4	67.3	44.1	30.4	27.8	**63.3**	**62.6**	64.2	27.7	51.1	40.6	60.8	59.2	71.2	**58.5**	28.2	55.6	39.4	**58.0**	56.1	52.0
unsup + iterative ft	5M	67.7	64.0	41.3	25.3	27.3	58.8	60.3	60.2	24.3	46.7	34.4	53.6	53.8	68.2	55.7	26.4	51.1	34.3	53.4	52.3	48.0
RCNN 70K		72.7	62.9	49.3	31.1	25.9	56.2	53.0	70.0	23.3	49.0	38.0	69.5	60.1	68.2	46.4	17.5	57.2	46.2	50.8	54.1	50.1
RCNN 70K (2 ensemble)		**75.3**	68.3	53.1	35.2	27.7	59.6	54.7	73.4	26.5	53.0	42.2	73.1	**66.1**	71.0	48.5	21.7	59.2	50.8	55.2	**58.0**	53.6
RCNN 70K (3 ensemble)		74.6	**68.7**	**54.9**	**35.7**	29.4	61.0	54.4	**74.0**	**28.4**	**53.6**	**43.0**	**74.0**	**66.1**	**72.8**	50.3	20.5	**60.0**	**51.2**	57.9	**58.0**	**54.4**
RCNN 200K (big stepsize)		73.3	67.1	46.3	31.7	**30.6**	59.4	61.0	67.9	27.3	53.1	39.1	64.1	60.5	70.9	57.2	26.1	59.0	40.1	56.2	54.9	52.3

Unsupervised Visual Representation Learning by Context Prediction

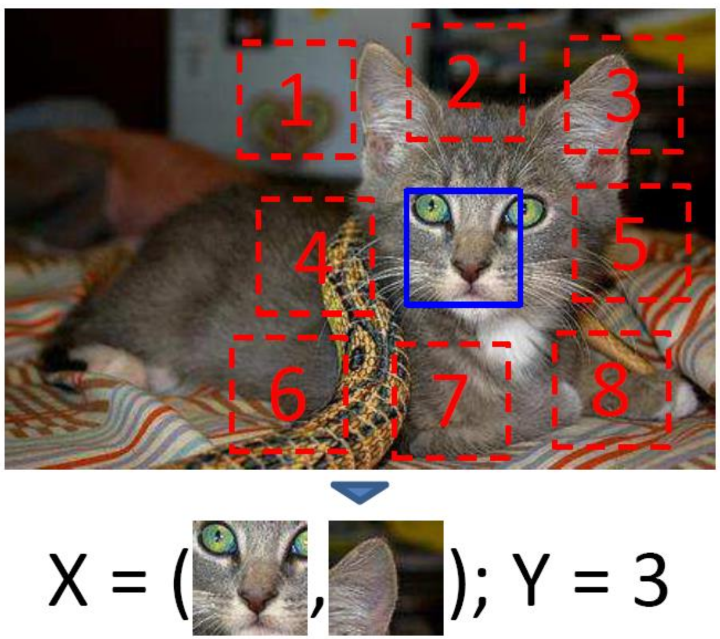
Carl Doersch[1,2] Abhinav Gupta[1] Alexei A. Efros[2]

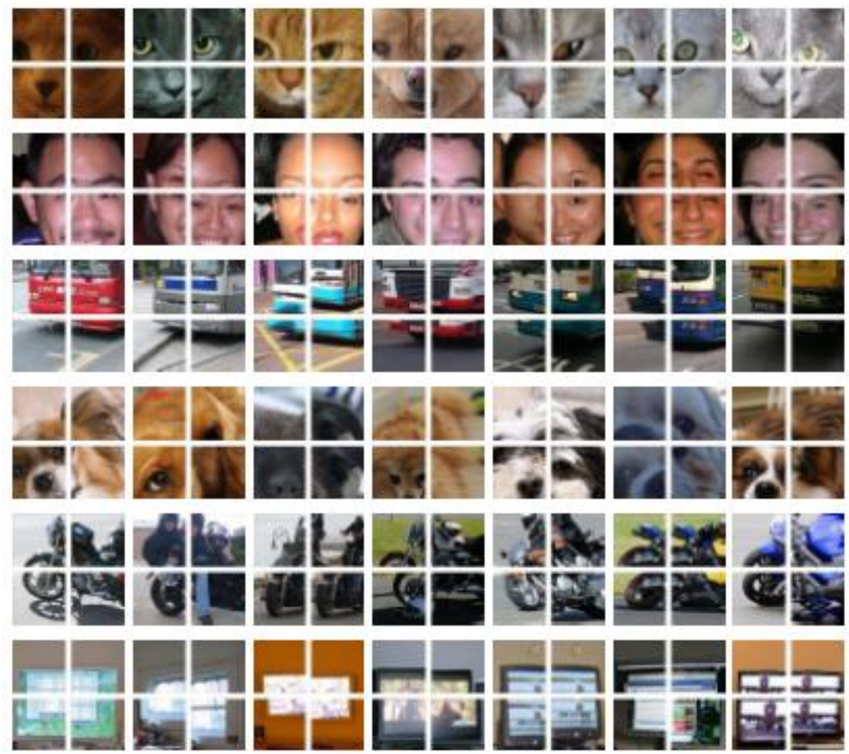
[1] School of Computer Science
Carnegie Mellon University

[2] Dept. of Electrical Engineering and Computer Science
University of California, Berkeley

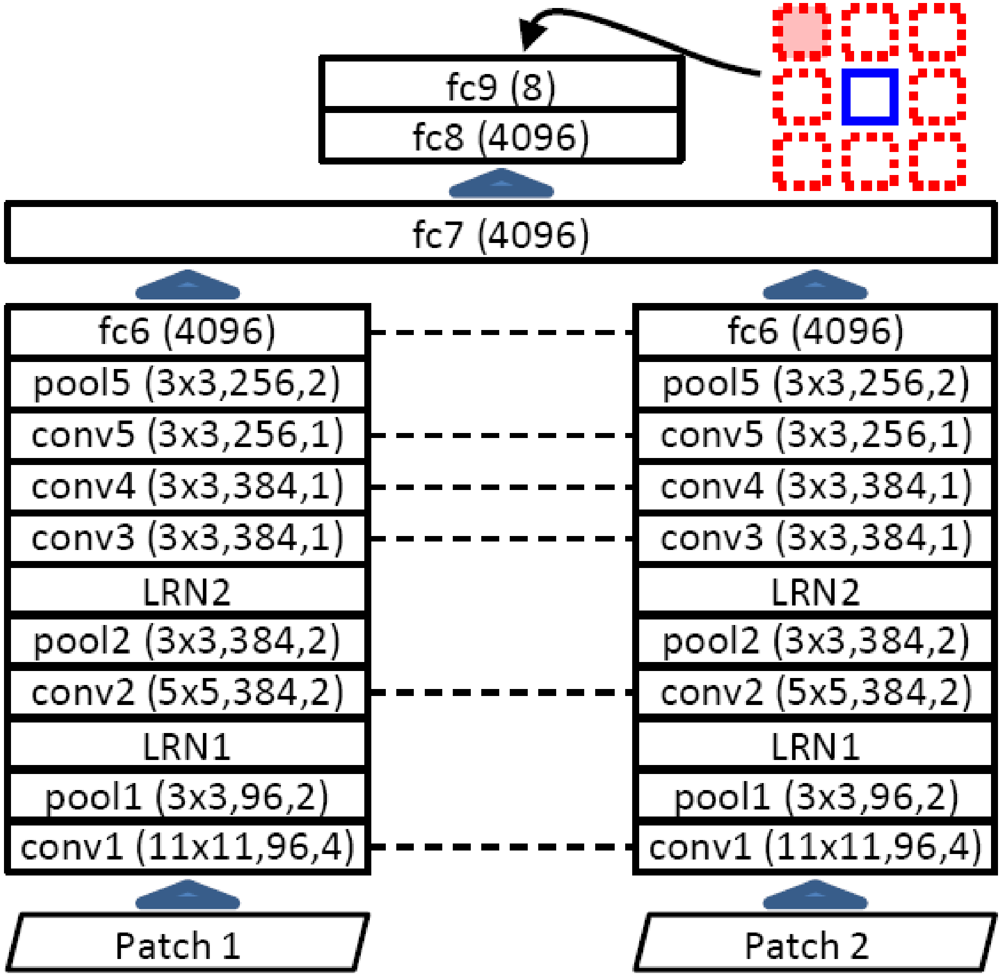


Unsupervised Visual Representation Learning by Context Prediction



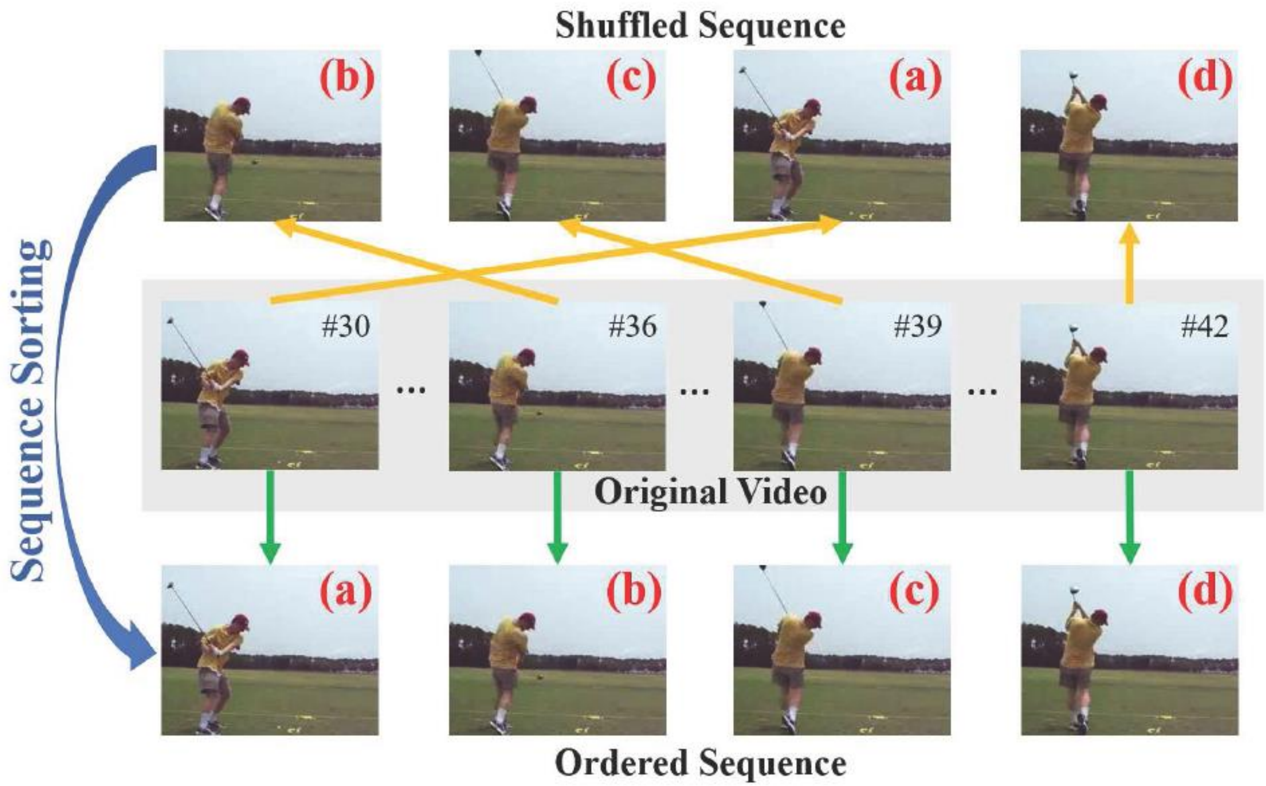


Unsupervised Visual Representation Learning by Context Prediction

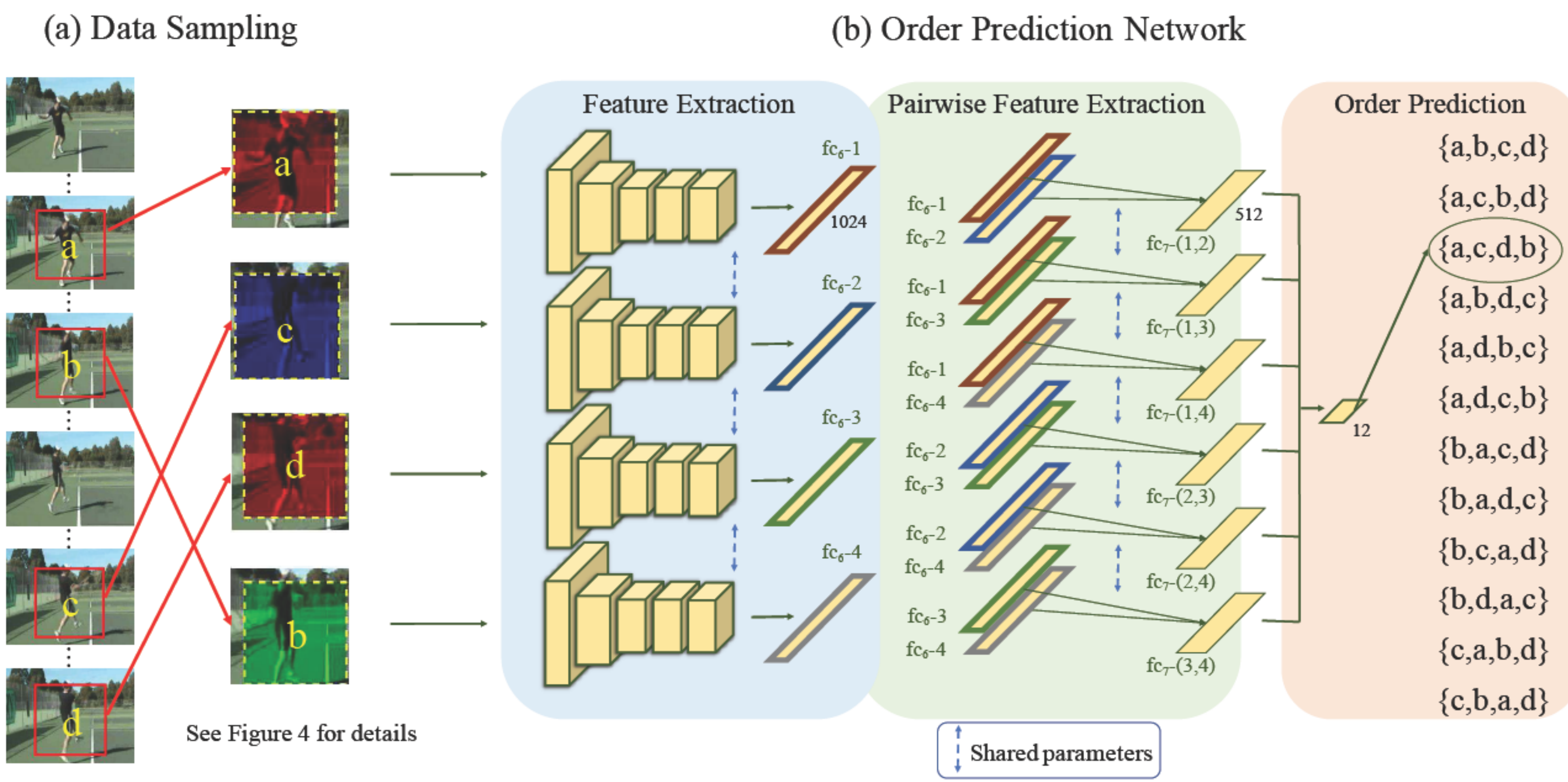


Unsupervised Representation Learning by Sorting Sequences

Hsin-Ying Lee[1] Jia-Bin Huang[2] Maneesh Singh[3] Ming-Hsuan Yang[1]
[1]University of California, Merced [2]Virginia Tech [3]Verisk Analytics
http://vllab1.ucmerced.edu/~hylee/OPN/



Unsupervised Representation Learning by Sorting Sequences



Unsupervised Representation Learning by Sorting Sequences

Initialization	CaffeNet	VGG-M-2048
random	47.8	51.1
ImageNet	67.7	70.8
Misra et al. [25]	50.2	-
Purushwalkam et al. [31]*	-	55.4
Vondrick et al. [40][†]	52.1	-
binary	51.6	56.8
3-tuple Concat	52.8	57.0
3-tuple OPN	53.2	58.3
4-tuple Concat	55.2	59.0
4-tuple OPN	**56.3**	**59.8**

Initialization	CaffeNet	VGG-M-2048
random	16.3	18.3
Imagenet	28.0	35.3
Misra et al. [25]	18.1	-
Purushwalkam et al. [31]*	-	23.6
binary	20.9	21.0
3-tuple OPN	21.3	21.5
4-tuple OPN	21.6	21.9
Misra et al. [25] (UCF)	15.2	-
4-tuple OPN (UCF)	**22.1**	**23.8**

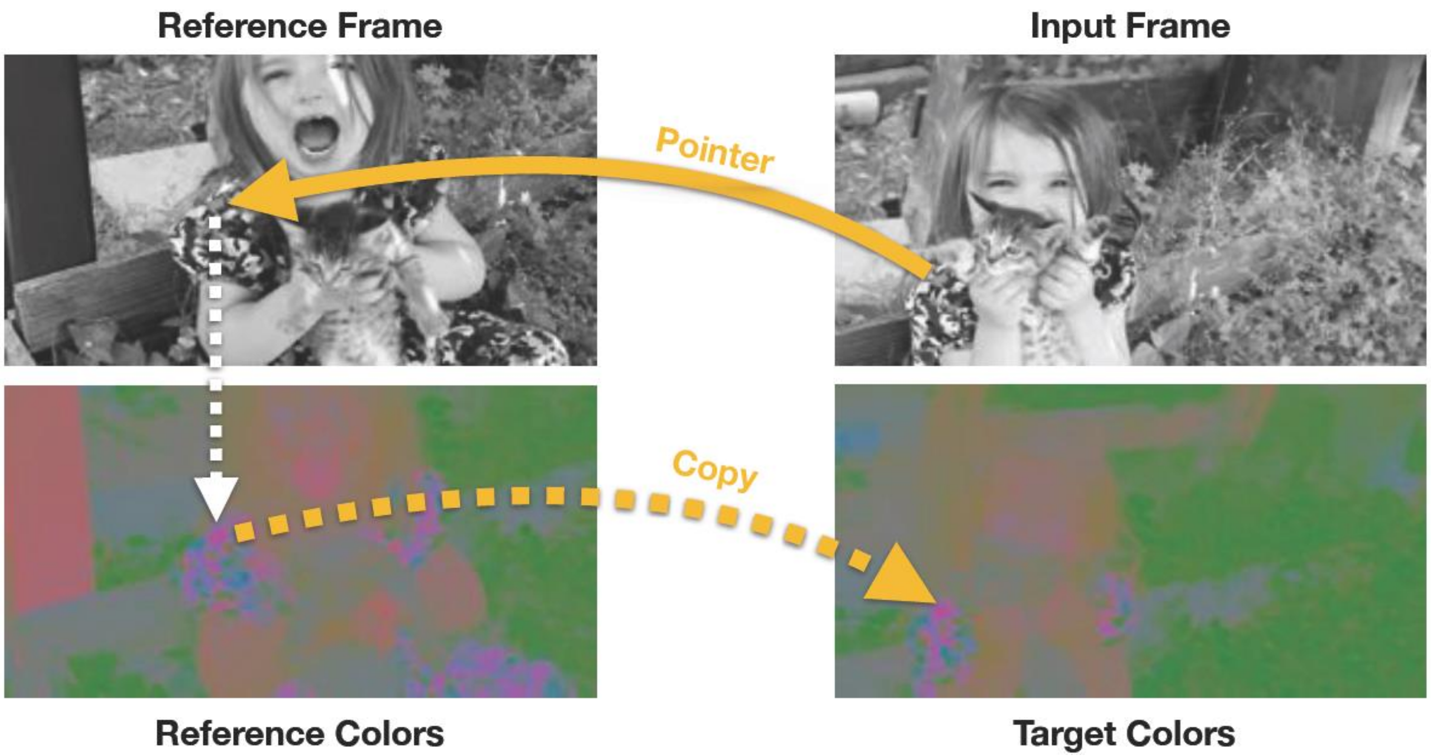
Table 4: **Results of the Pascal VOC2007 classification and detection datasets.**

Method	Pretraining time	Source	Supervision	Classification	Detection
Krizhevsky et al. [18]	3 days	ImageNet	labeled classes	78.2	56.8
Doerch et al. [7]	4 weeks	ImageNet	context	55.3	46.6
Pathak et al. [30]	14 hours	ImagetNet+StreetView	context	56.5	44.5
Norrozi et al. [27]	2.5 days	ImageNet	context	**68.6**	**51.8**
Zhang et al. [44]	-	ImageNet	reconstruction	67.1	46.7
Wang and Gupta (color) [42]	1 weeks	100k videos, VOC2012	motion	58.4	44.0
Wang and Gupta (grayscale) [42]	1 weeks	100k videos, VOC2012	motion	62.8	**47.4**
Agrawal et al. [2]	-	KITTI, SF	motion	52.9	41.8
Misra et al. [25]	-	$<$ 10k videos	motion	54.3	39.9
Ours (OPN)	$<$ 3 days	$<$ 30k videos	motion	**63.8**	46.9

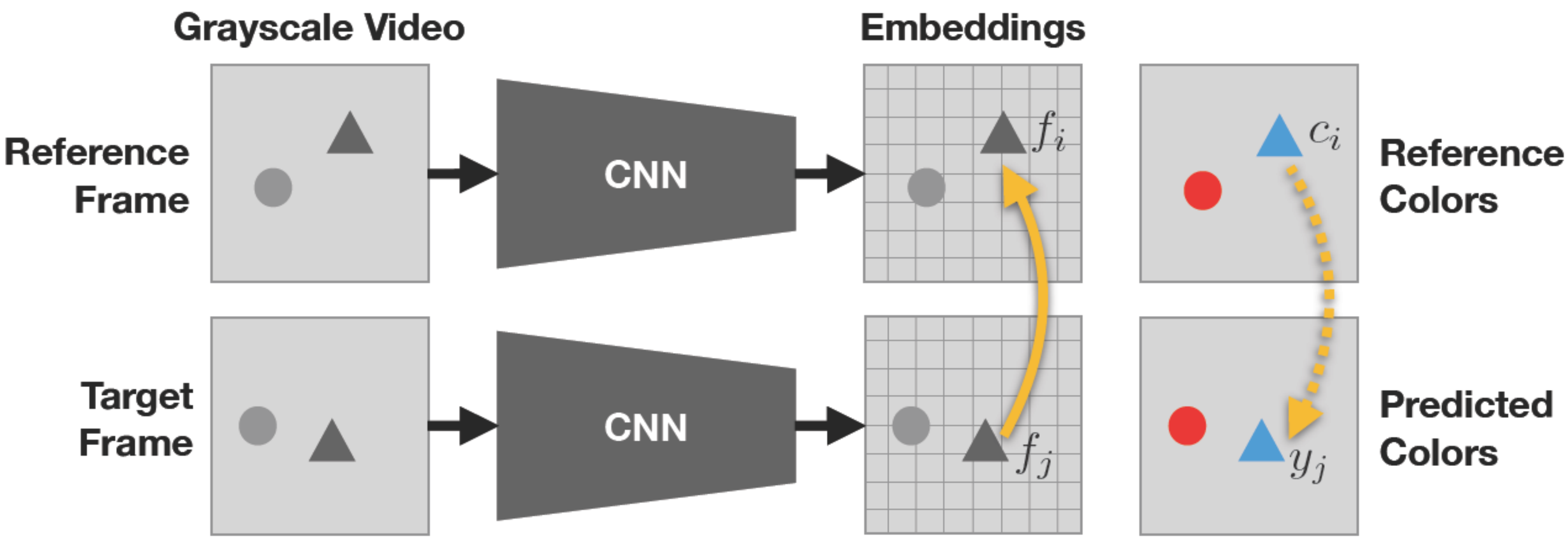
Tracking Emerges by Colorizing Videos

Carl Vondrick, Abhinav Shrivastava, Alireza Fathi,
Sergio Guadarrama, Kevin Murphy

Google Research



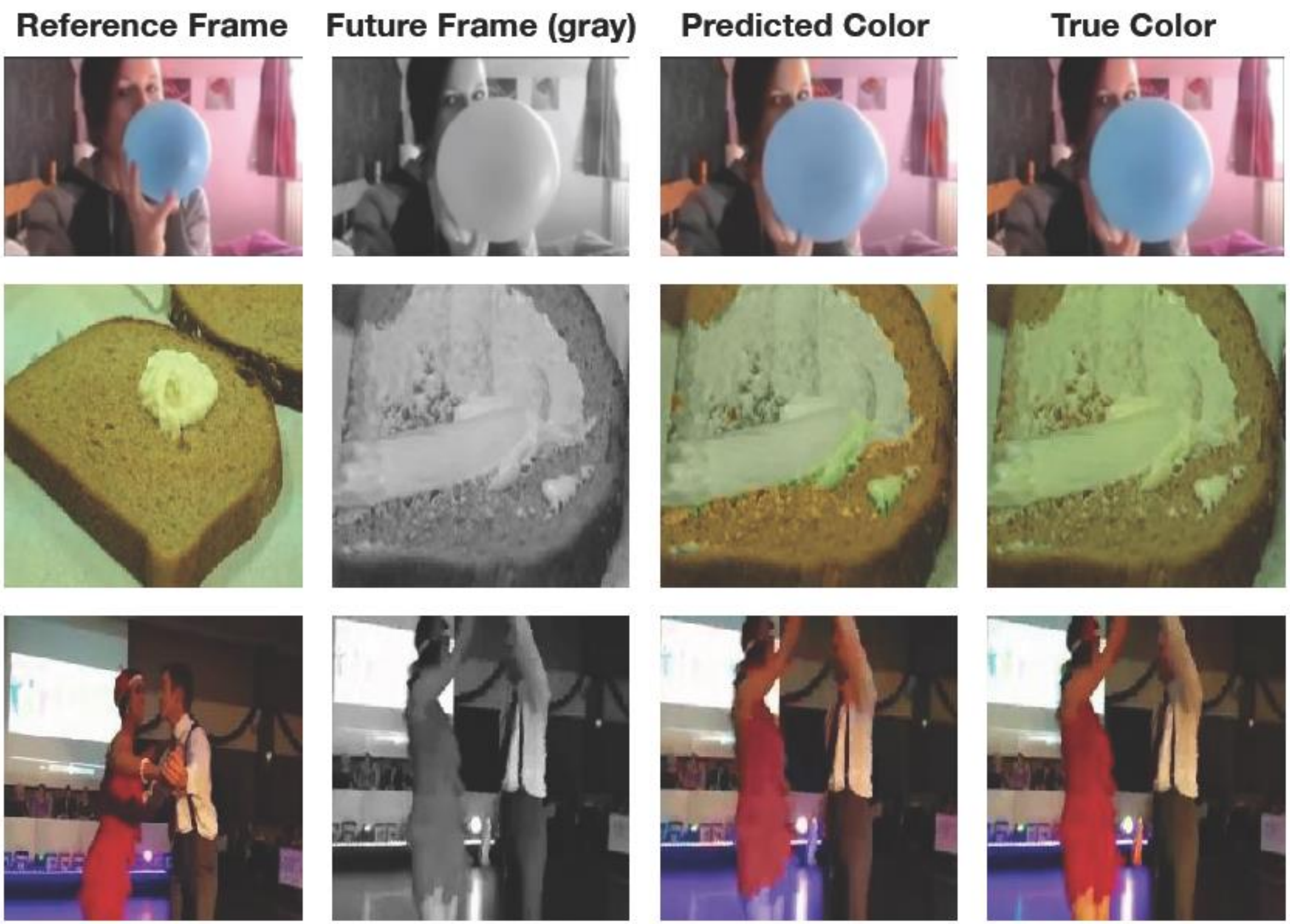
Tracking Emerges by Colorizing Videos



$$A_{ij} = \frac{\exp\left(f_i^T f_j\right)}{\sum_k \exp\left(f_k^T f_j\right)}$$

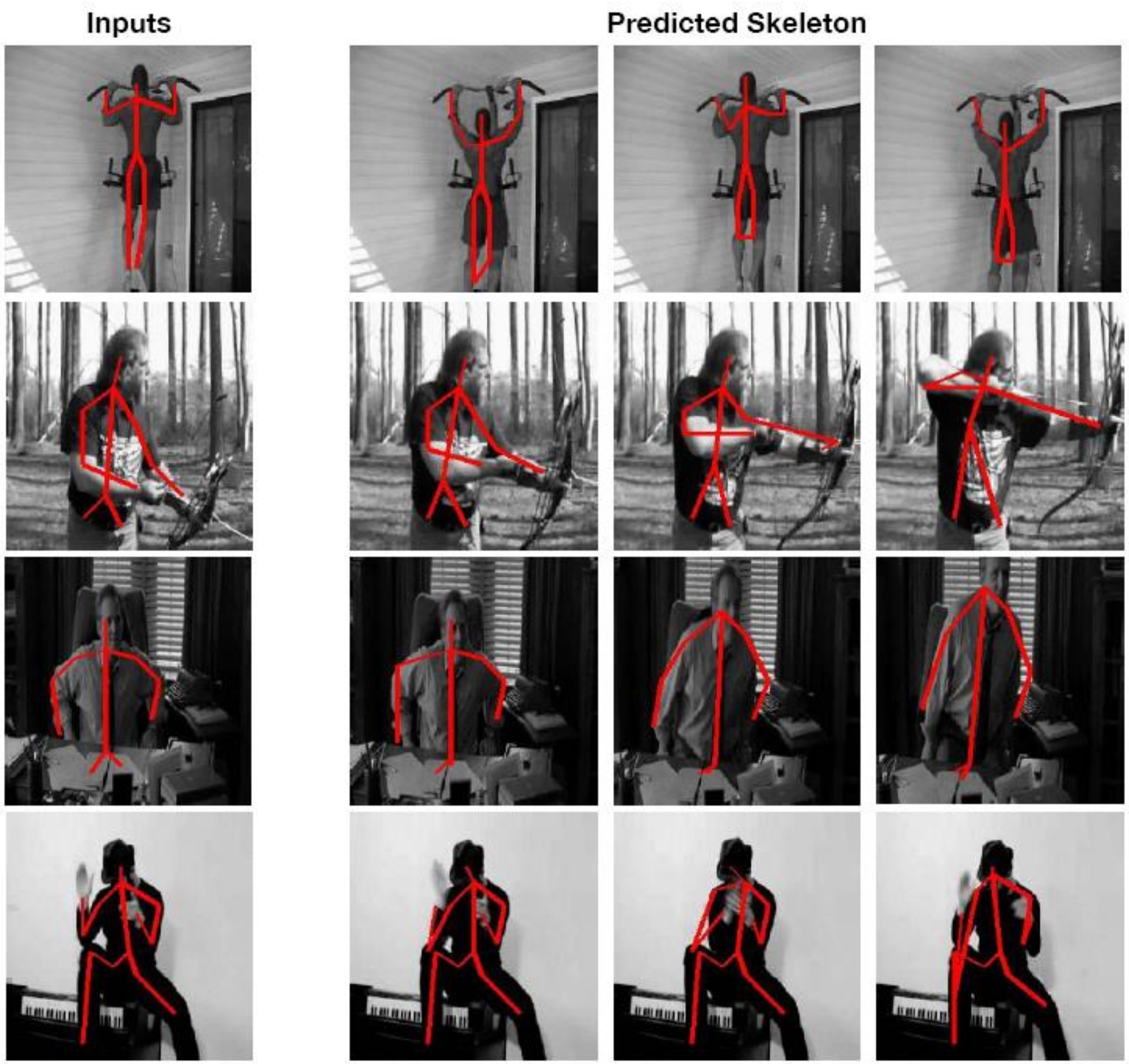
$$y_j = \sum_i A_{ij} c_i \quad \Longrightarrow \quad \min_\theta \sum_j \mathcal{L}\left(y_j, c_j\right)$$

Tracking Emerges by Colorizing Videos



Tracking Emerges by Colorizing Videos

Tracking Emerges by Colorizing Videos



Summary

1. Encoder-Decoder
2. Context
3. Motion
4. Color

....

Typically, erase some known information and further recover it for self-supervised (or unsupervised) learning.

Thank you ！