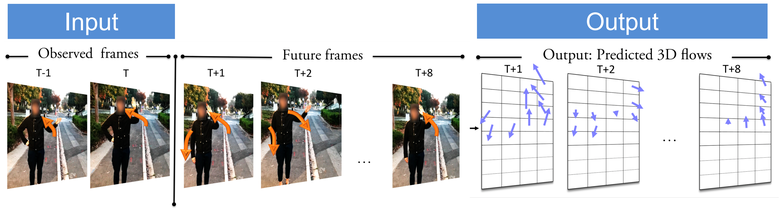
Unsupervised Learning of Long-Term Motion Dynamics for Videos

Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, Li Fei-Fei
Computer Science Department, Stanford University

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

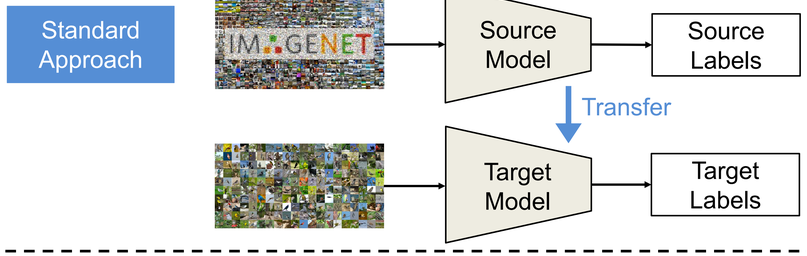
Introduction

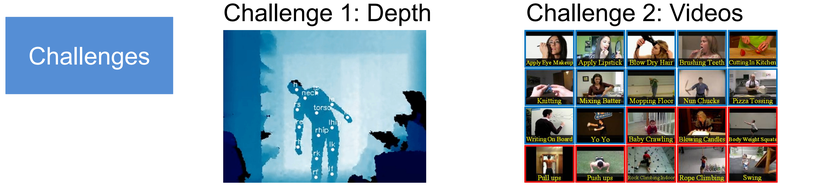
- We propose a self-supervised learning approach that learns a video representation by predicting long-term 3D motions



Motivation

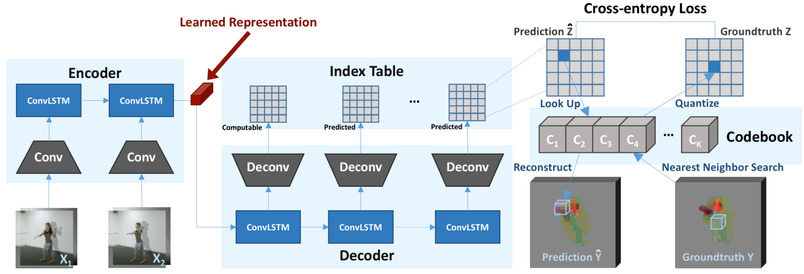
- A good video representation for activity recognition needs to capture:
  - Image-level semantics
  - Long-term motion dependencies
- However,
  - No large-scale depth datasets to pre-train representations
  - Existing unsupervised methods fail to learn long-term motion



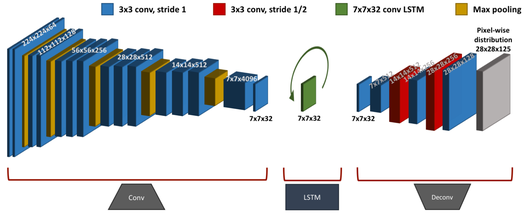


Representation Learning

- Sequence-to-sequence learning problem
- Encoder-decoder framework with ConvLSTM



- **Encoder**: A pair of frames ⇒ Video representation
  - Conv + Encoder ConvLSTM
  - Modalities: RGB, depth, or RGB-D
- **Decoder**: Video representation ⇒ Atomic 3D flows
  - Decoder ConvLSTM + Deconv
  - Atomic 3D flows: A sequence of quantized 3D scene flows



- **Quantization**: Regression ⇒ Classification
  - Predict each non-overlapping flow patch as a distribution over the codebook
  - Codebook: uniform, k-means, or learnable codebook
- **Objective function**: Cross-entropy loss with rebalancing

$$\mathbf{w} \propto \left((1-\lambda)\widetilde{\mathbf{p}}+\frac{\lambda}{K}\right)^{-1}$$

$$\sum_{k=1}^{K}\widetilde{\mathbf{p}}_k\mathbf{w}_k = 1$$

$\widetilde{\mathbf{p}}$: Empirical distribution of the codewords in the codebook
K : Number of quantized classes
λ : Smoothing weight

Experiments and Results

- The learned encoder is used to initialize the video feature extractor
- Fine-tune on new datasets

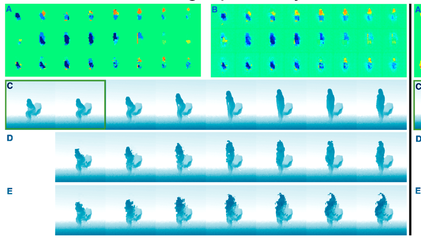
NTU-RGB+D ⇒ NTU-RGB+D

Methods	mAP
HON4D [D]	30.56
Lie Group [S]	50.08
FTP Dynamic Skeletons [S]	60.23
Shuffle and Learn [U]	47.5
Our method [U]	66.2

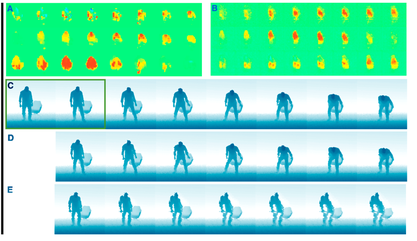
NTU-RGB+D ⇒ MSRDailyActivity3D

Methods	mAP
Dynamic Temporal Warping	54.0
Actionlet Ensemble	85.8
HON4D	85
3D Trajectories	72
Our method [U]	86.9

Standing Up Activity



Sitting Down Activity



Ablation Study

Methods	Depth	RGB
Our architecture only	37.5	34.1
Our method with 2D motion	58.8	-
Our method with 3-step prediction	62.1	54
Our method with 8-step prediction	66.2	56

RGB videos (UCF-101)

Methods	mAP
Dynamic Temporal Warping	54.0
Actionlet Ensemble	85.8
HON4D	85
3D Trajectories	72
Our method [U]	86.9