# Unsupervised Learning of View-invariant Action Representations

Junnan Li, Yongkang Wong, Qi Zhao, Mohan S. Kankanhalli

## Motivation

*We need unsupervised learning!*

Supervised learning achieves good performance for action recognition. However:

- Require significant amount of manually labeled data.
- Human labeling is expensive and time-consuming.

Unsupervised learning:

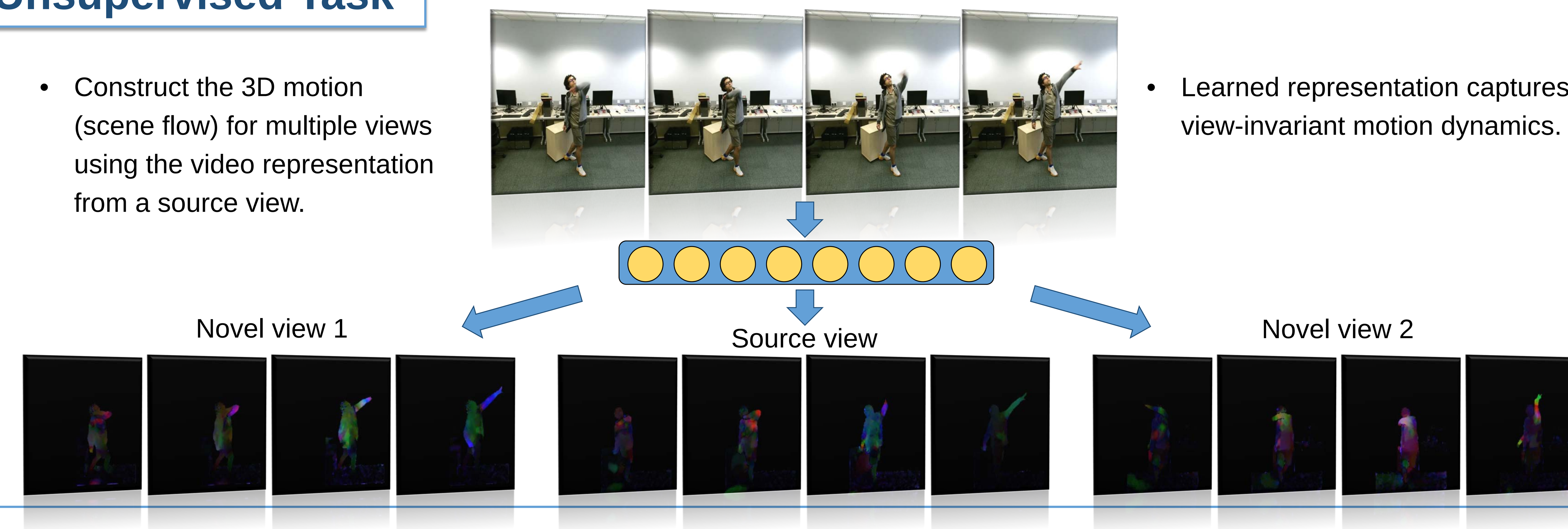
- Leverage *free* unlabeled data.
- A surrogate task that exploits the inherent structure of raw videos.
- Learn representations useful for the supervised task.

*We need view-invariant representations!*

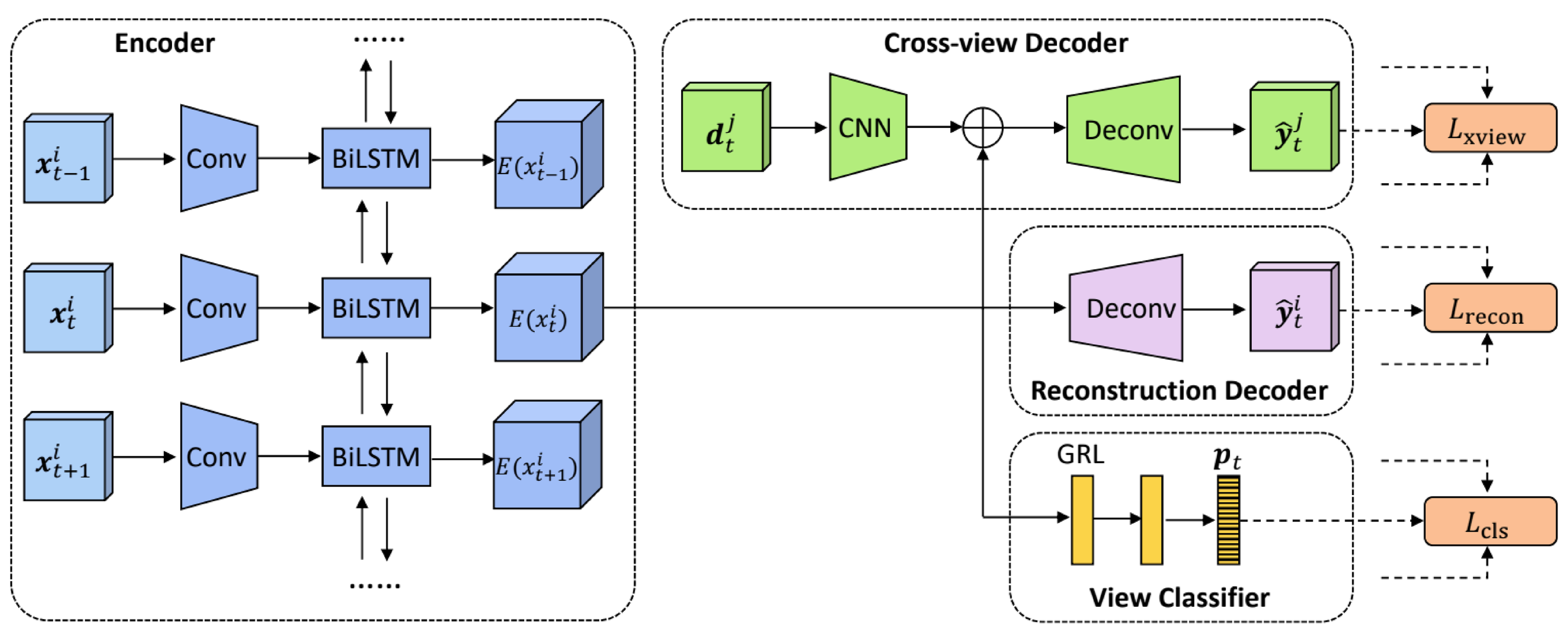
- The same action appears quite different from different views.
- Action recognition from unseen view is difficult.
- Human brains can build view-invariant action representations.

## Unsupervised Task

- Construct the 3D motion (scene flow) for multiple views using the video representation from a source view.
- Learned representation captures view-invariant motion dynamics.




## Framework

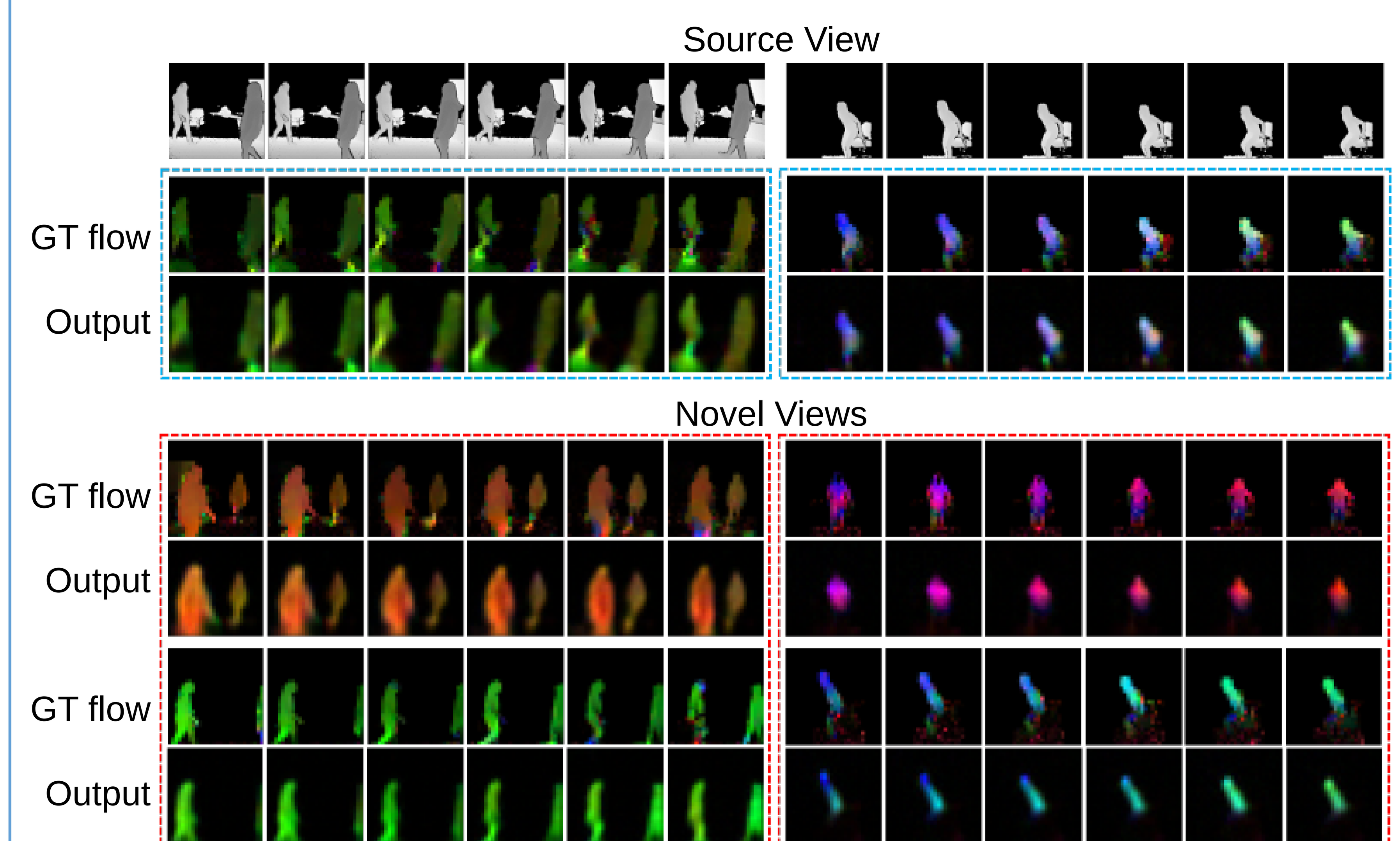



- Simultaneously optimize multiple loss terms: $L = L_{\text{xview}} + \alpha L_{\text{recon}} + \beta L_{\text{cls}}$
- **Encoder** encodes a sequence of frames into a sequence of features. "Conv" is a down-sampling CNN. "BiLSTM" is a bi-directional convolutional LSTM.
- **Cross-view decoder** predicts the 3D flow $y_t^j$ for view $j$ given the encoding $E(x_t^i)$ for view $i$, at timestep $t$. $d_t^j$ is a depth map for view $j$ that provides view-specific information. "Deconv" is an upsampling CNN. Let $\hat{y}_t^j$ denotes the output, we want to minimize the mean squared error between $\hat{y}_t^j$ and $y_t^j$ for all timesteps: $L_{\text{xview}}^j = \sum_{t=1}^T \|y_t^j - \hat{y}_t^j\|_2^2$.
- **Reconstruction decoder** reconstructs the 3D flow $y_t^i$ given the encoding from the same view: $L_{\text{recon}} = \sum_{t=1}^T \|y_t^i - \hat{y}_t^i\|_2^2$.
- **View Adversarial Training**
  - View classifier tries to predict which view the encoded representation belongs to.
  - Encoder tries to confuse the view classifier by generating **view-invariant representations**.
  - "GRL" is a gradient reversal layer that reverses the sign of the gradient.
  - The view classifier is trained to minimize the cross-entropy loss $L_{\text{cls}}$, while the encoder is trained to maximize $L_{\text{cls}}$.

## Experiments

NTU RGB+D dataset

- 57K videos, 60 action classes, 40 subjects
- 5 views: front view, left side view, right side view, left side 45 degrees view and right side 45 degrees view
- Cross-subject: half the subjects for training, half for test.
- Cross-view: cameras 2 and 3 for training, camera 1 for test.




(a) walking towards each other. (b) sitting down.

### Action Recognition

Append a one-layer action classifier to the video encoder.

1. scratch: Randomly initialize the weights of encoder and train the model from scratch.
2. fine-tune: Initialize the encoder with unsupervised learned weights and fine-tune it.
3. fix: Keep the pre-trained encoder fixed and only train the action classifier.

Table 1: Accuracy (%) on NTU RGB+D dataset

| Method | Cross-subject | | | Cross-view | | |
|---|---|---|---|---|---|---|
| | RGB | Depth | Flow | RGB | Depth | Flow |
| scratch | 36.6 | 42.3 | 70.2 | 29.2 | 37.7 | 72.6 |
| fix | 48.9 | 60.8 | 77.0 | 40.7 | 53.9 | 78.8 |
| fine-tune w/o view-adversarial | 53.4 | 66.0 | 80.3 | 46.2 | 60.1 | 81.9 |
| fine-tune | 55.5 | 68.1 | **80.9** | 49.3 | 63.9 | **83.4** |

Table 2: Comparison with state-of-the-art methods on NTU RGB+D Dataset

| Method | Modality | Cross-subject | Cross-view |
|---|---|---|---|
| HOG [35] | Depth | 32.24 | 22.27 |
| Super Normal Vector [60] | Depth | 31.82 | 13.61 |
| HON4D [37] | Depth | 30.56 | 7.26 |
| *Shuffle and Learn* [32] | Depth | 46.2 | 40.9 |
| *Luo et al.* [31] | Depth | 61.4 | 53.2 |
| Ours | Depth | **68.1** | **63.9** |
| Lie Group [52] | Skeleton | 50.08 | 52.76 |
| FTP Dynamic Skeletons [16] | Skeleton | 60.23 | 65.22 |
| HBRNN-L [7] | Skeleton | 59.07 | 63.97 |
| 2 Layer P-LSTM [47] | Skeleton | 62.93 | 70.27 |
| ST-LSTM [28] | Skeleton | 69.2 | 77.7 |
| GCA-LSTM [29] | Skeleton | 74.4 | 82.8 |
| Ensemble TS-LSTM [24] | Skeleton | 74.60 | 81.25 |
| Depth+Skeleton [40] | Skeleton | 75.2 | 83.1 |
| VA-LSTM [61] | Skeleton | 79.4 | **87.6** |
| Ours | Flow | **80.9** | **83.4** |

### Representation Transfer

Table 3: Cross-subject action recognition accuracy on MSRDailyActivity3D dataset

| Method | Accuracy |
|---|---|
| Actionlet Ensemble [56] (S) | 85.8 |
| HON4D [37] (D) | 80.0 |
| MST-AOG [57] (D) | 53.8 |
| SNV [60] (D) | 86.3 |
| HOPC [41] (D) | 88.8 |
| *Luo et al.* [31] (D) | 75.2 |
| Ours (scratch) | 42.5 |
| Ours (fine-tune) | 82.3 |

Table 4: Cross-view action recognition accuracy on Northwestern-UCLA dataset

| Method | Accuracy |
|---|---|
| Actionlet Ensemble [56] (S) | 69.9 |
| Hankelets [25] | 45.2 |
| MST-AOG [57] (D) | 53.6 |
| HOPC [41] (D) | 71.9 |
| R-NKTM [43] (S) | 78.1 |
| *Luo et al.* [31] (D) | 50.7 |
| Ours (scratch) | 35.8 |
| Ours (fine-tune) | 62.5 |