# Generative Multi-View Human Action Recognition

Presenter: Andre Von Zuben

Lichen Wang
Zhengming Ding
Zhiqiang Tao
Yunyu Liu
Yun Fu

ICCV 2019

# Outline

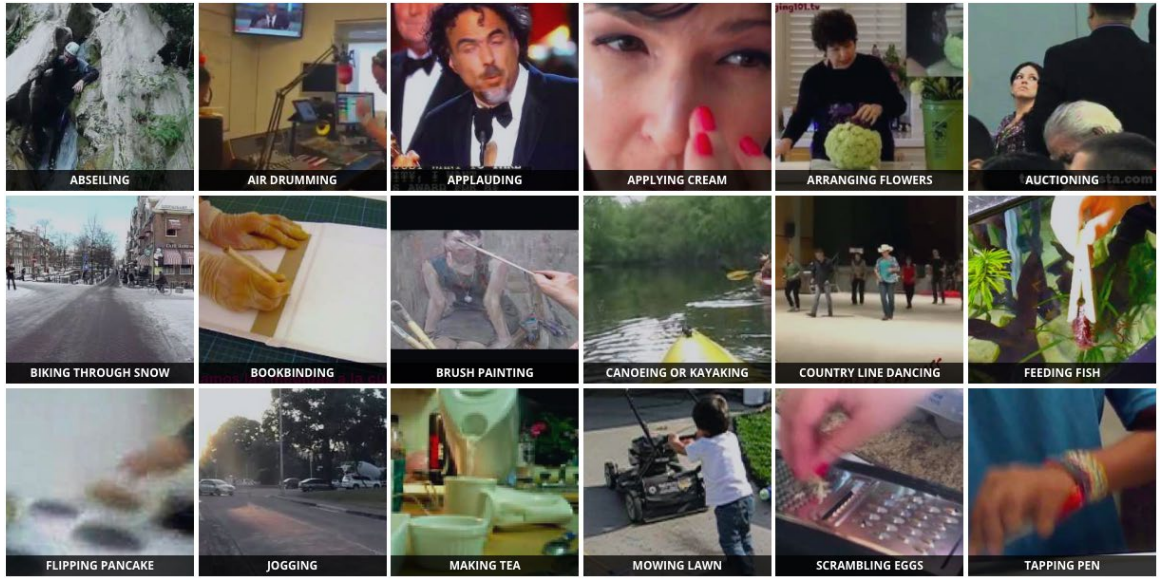
- Introduction
- Related Works
- Proposed Method
- Experiments
- Conclusion

# Introduction

- Action Recognition



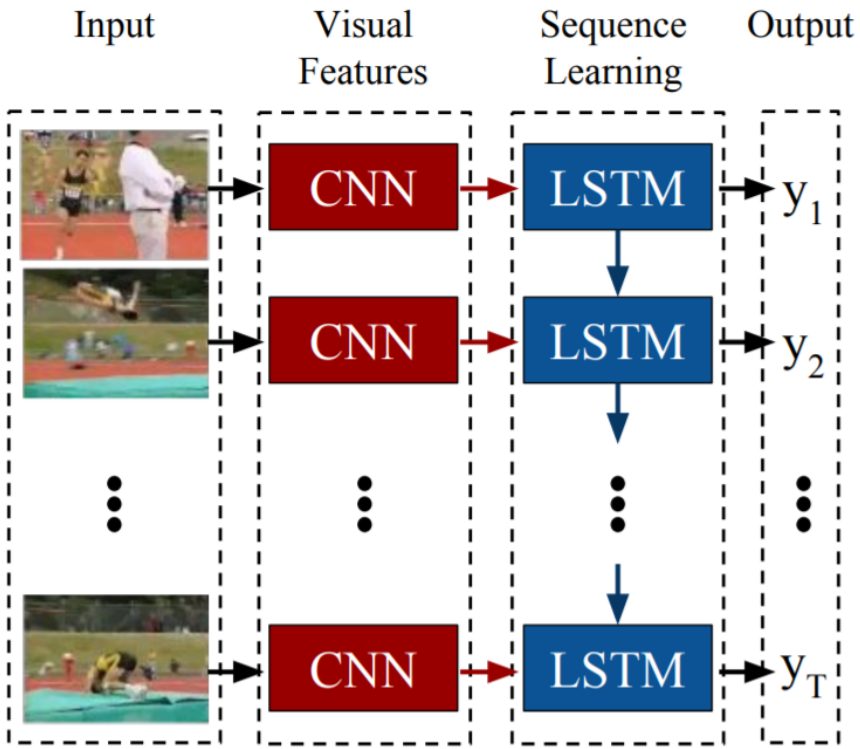

Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, CRCV-TR-12-01, November, 2012



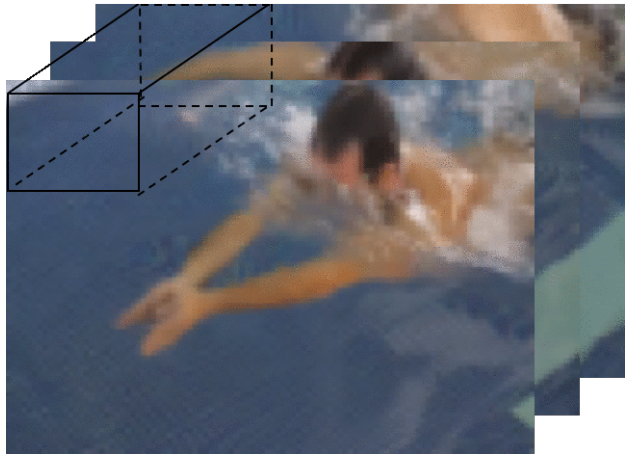

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about Kinetics600. arXiv:1808.01340, 2018

# Introduction

- Action Recognition – Single View



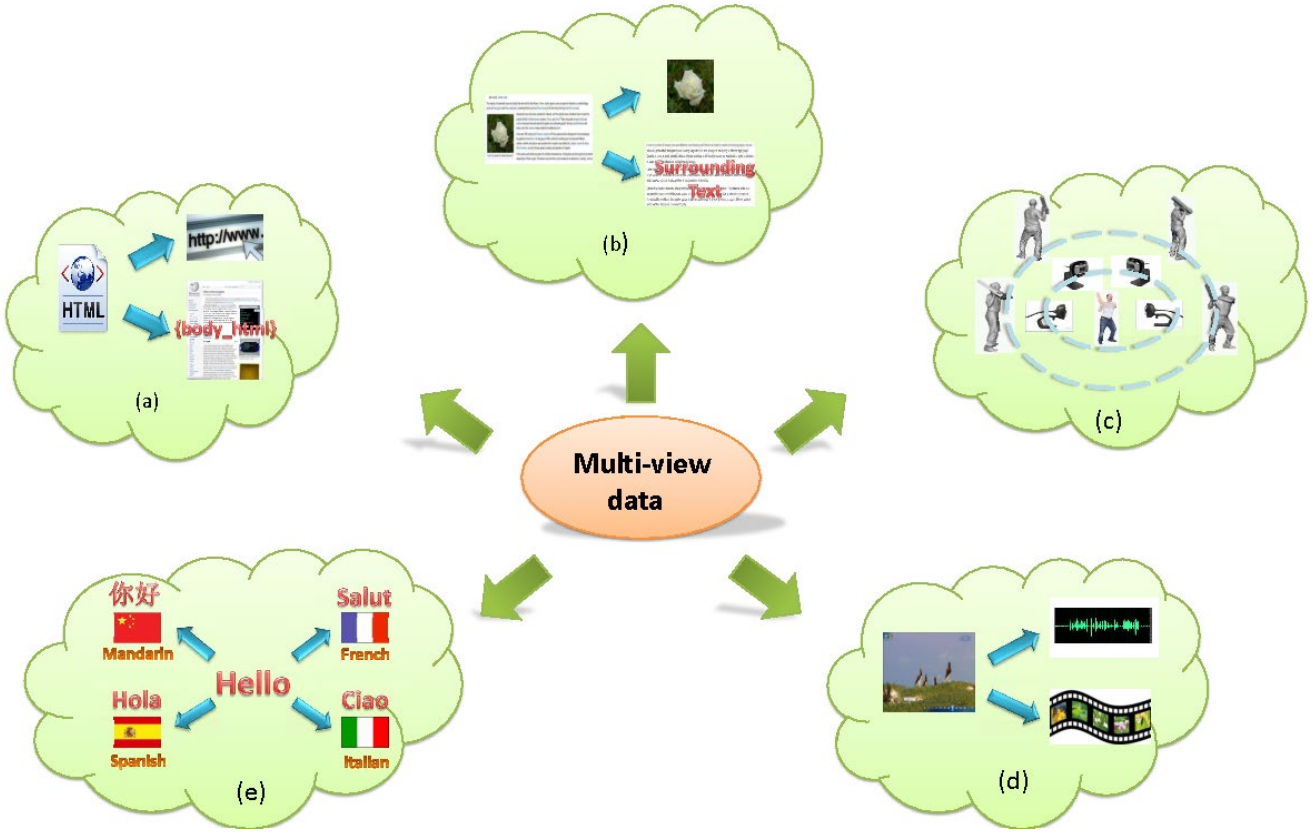

Donahue, Jeff, Hendrikcs, Lisa Anne, Guadarrama, Sergio, Rohrbach, Marcus, Venugopalan, Subhashini, Saenko, Kate, and Darrell, Trevor. Long-term recurrent convolutional networks for visual recognition and description. arXiv:1411.4389v2 [cs.CV], November 2014



http://blog.qure.ai/notes/deep-learning-for-videos-action-recognition-review

# Introduction

- Multi-View
  - Complementary information among different views

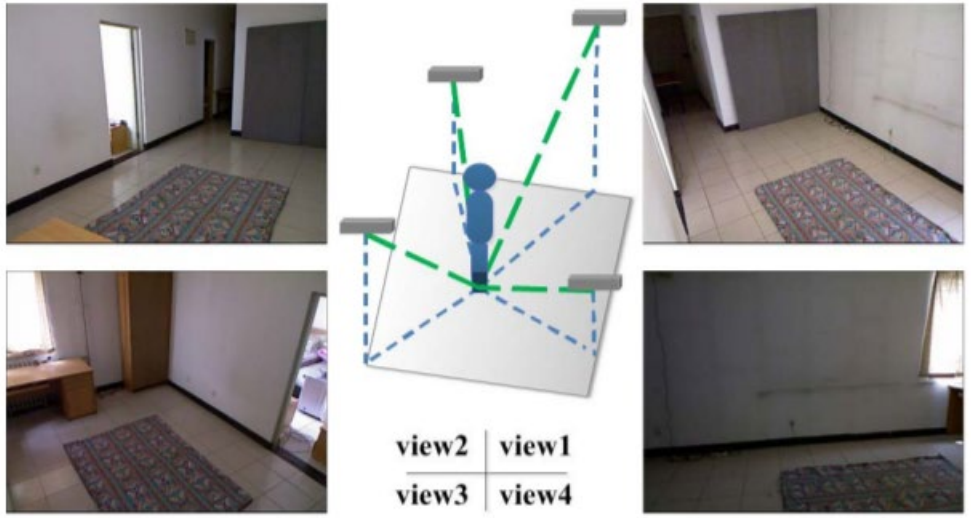



Chang Xu, Dacheng Tao, and Chao Xu. A survey on multiview learning. arXiv preprint arXiv:1304.5634, 2013

# Introduction

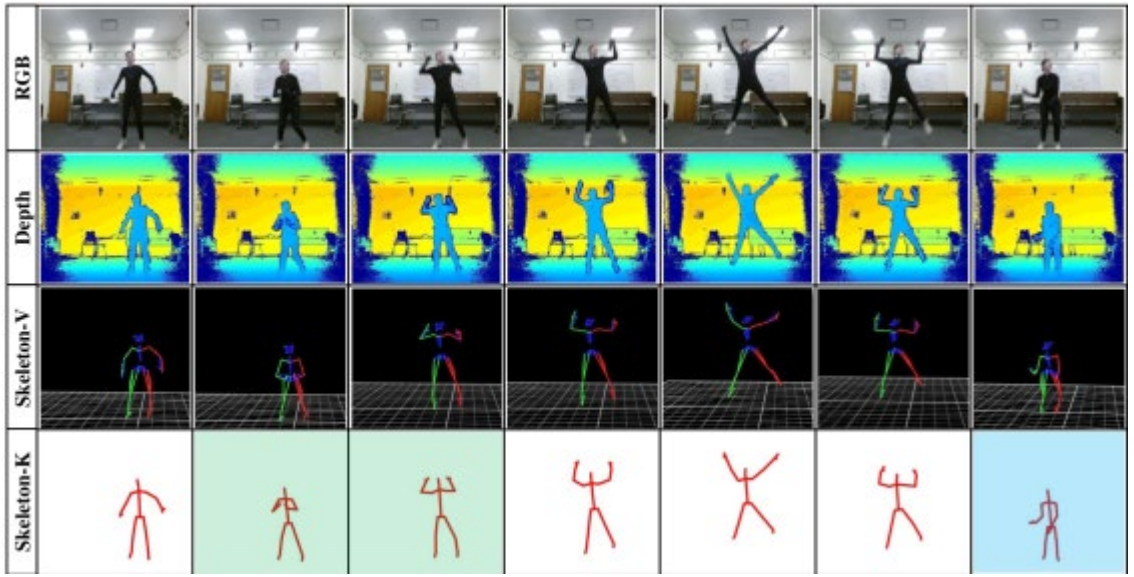
- Multi-View Action Recognition

Multiple sensors from the same visual modality




Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. Human daily action analysis with multi-view and color-depth data. In Proc. ECCV, pages 52–61. Springer, 2012

Different types of sensors




Lichen Wang, Bin Sun, Joseph Robinson, Taotao Jing, and Yun Fu. EV-Action: Electromyography-Vision multi-modal action dataset. arXiv preprint arXiv:1904.12602, 2019.

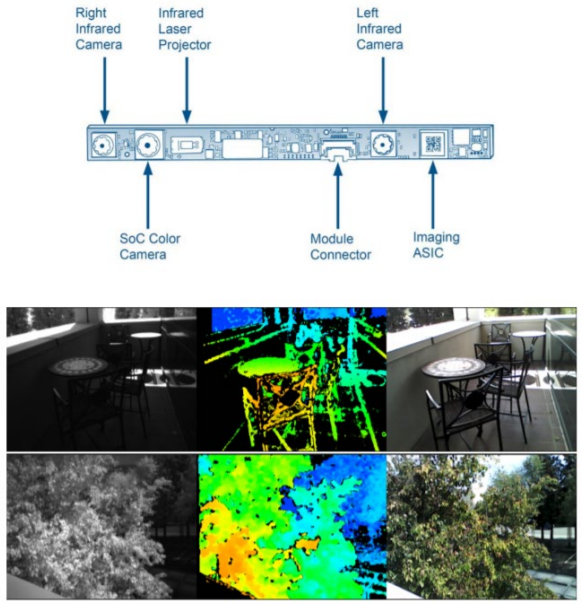
# Introduction

- RGB-Depth (RGB-D) action recognition
  - one of the most important research directions
    - popularity of depth/3D sensors and the corresponding applications




Microsoft Kinect

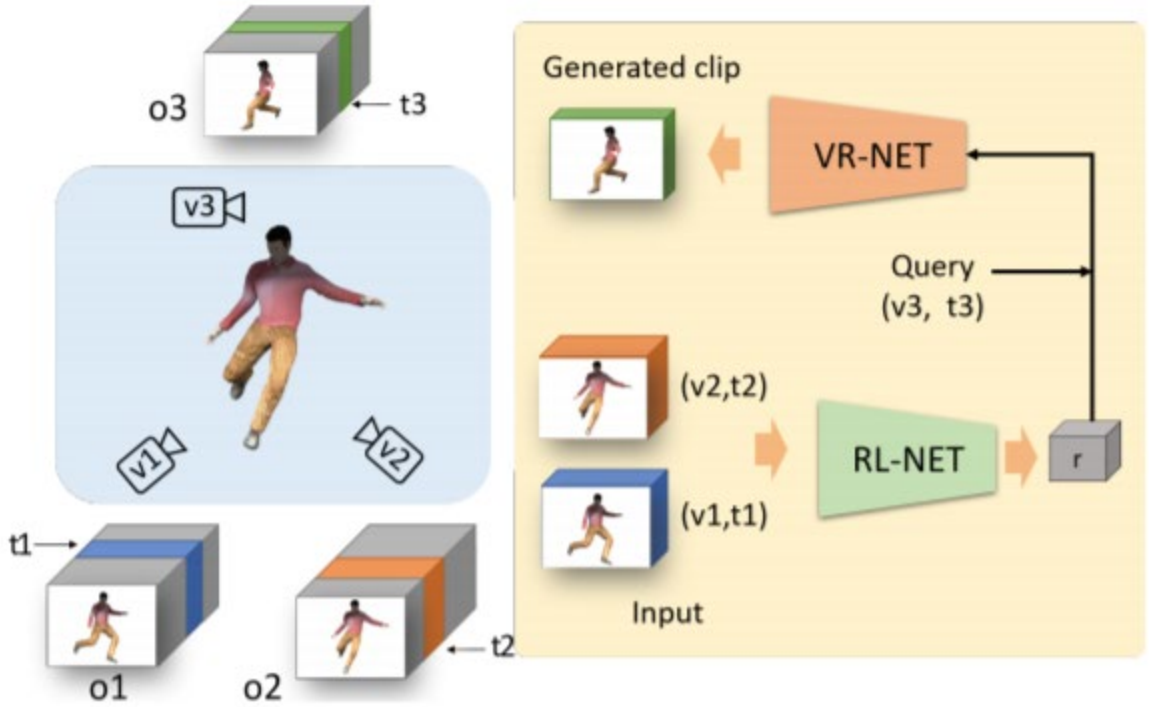
Zhengyou Zhang. Microsoft kinect sensor and its effect. IEEE Multimedia, 19(2):4–10, 2012




Intel RealSense

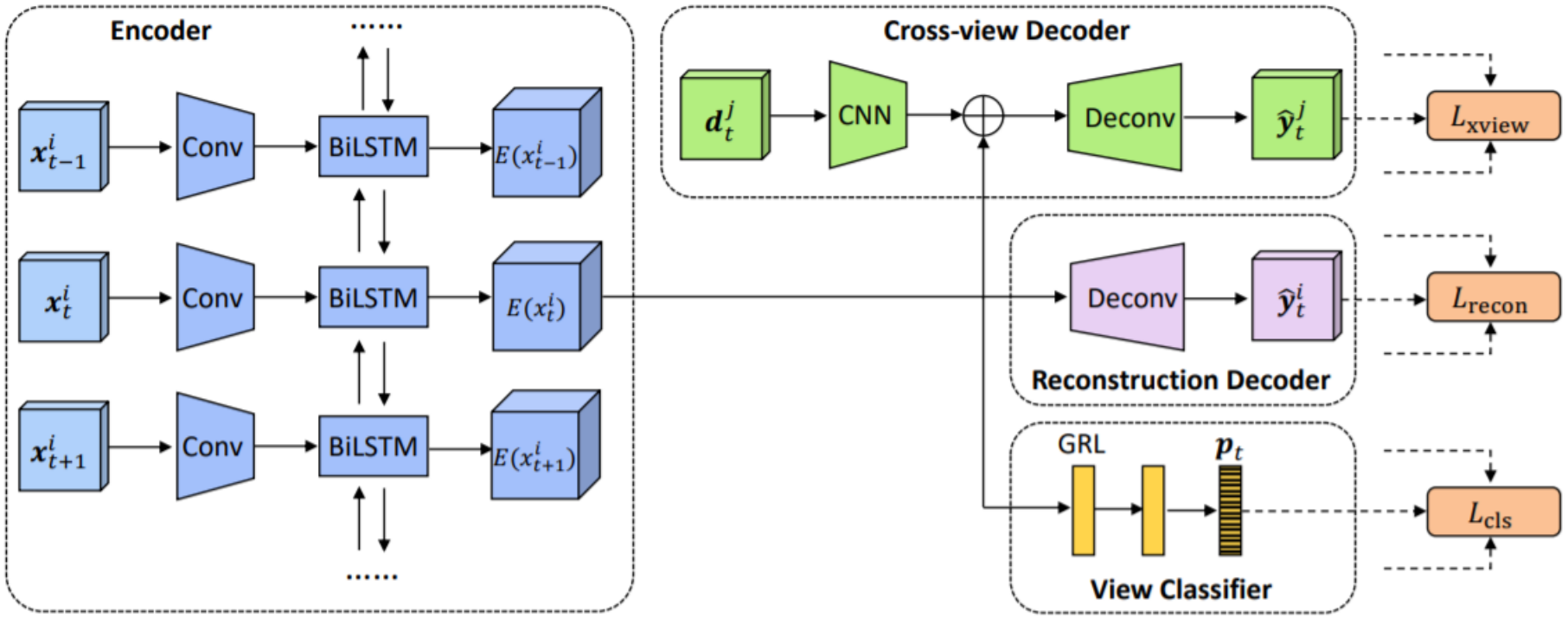
Leonid Keselman, John Iselin Woodfill, Anders GrunnetJepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In Proc. IEEE CVPR workshop, pages 1–10, 2017.

# Time-aware and View-aware Video Rendering for Unsupervised Representation Learning



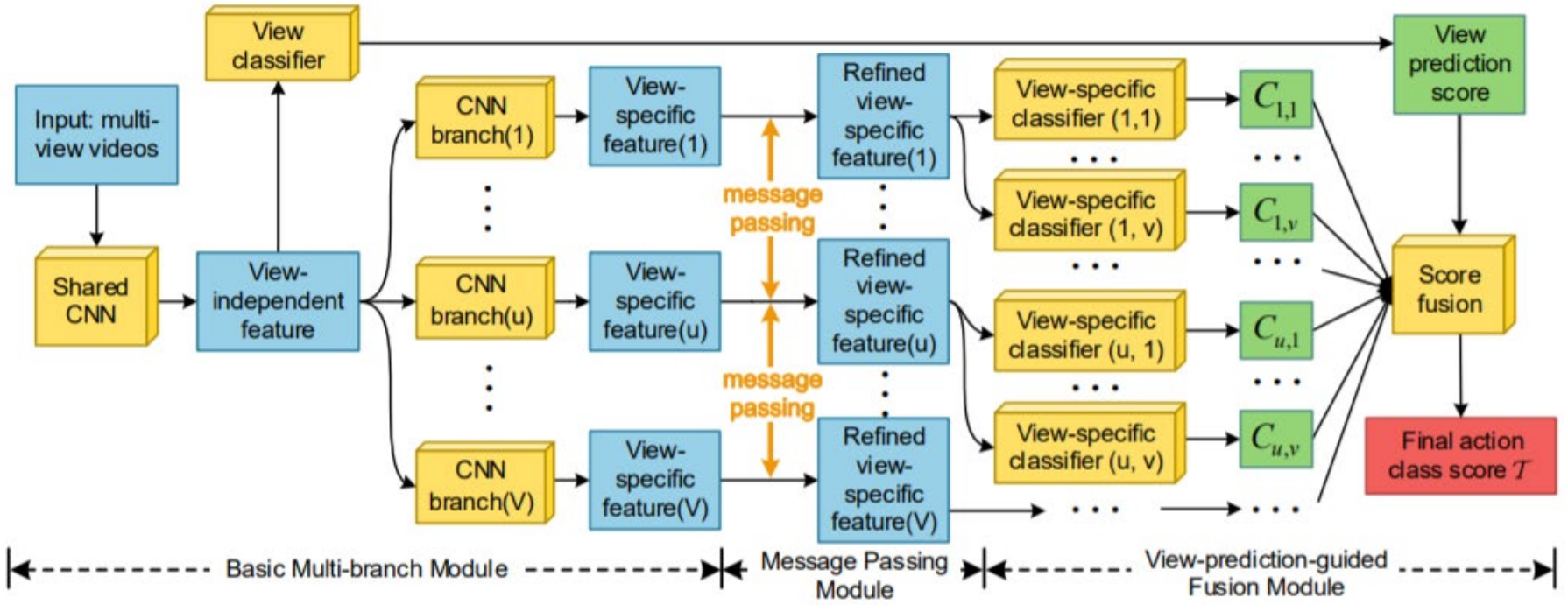

Shruti Vyas, Yogesh Singh Rawat, and Mubarak Shah. Time-aware and view-aware video rendering for unsupervised representation learning. In CoRR, volume abs/1811.10699, 2018.

# Unsupervised Learning of View-invariant Action Representations



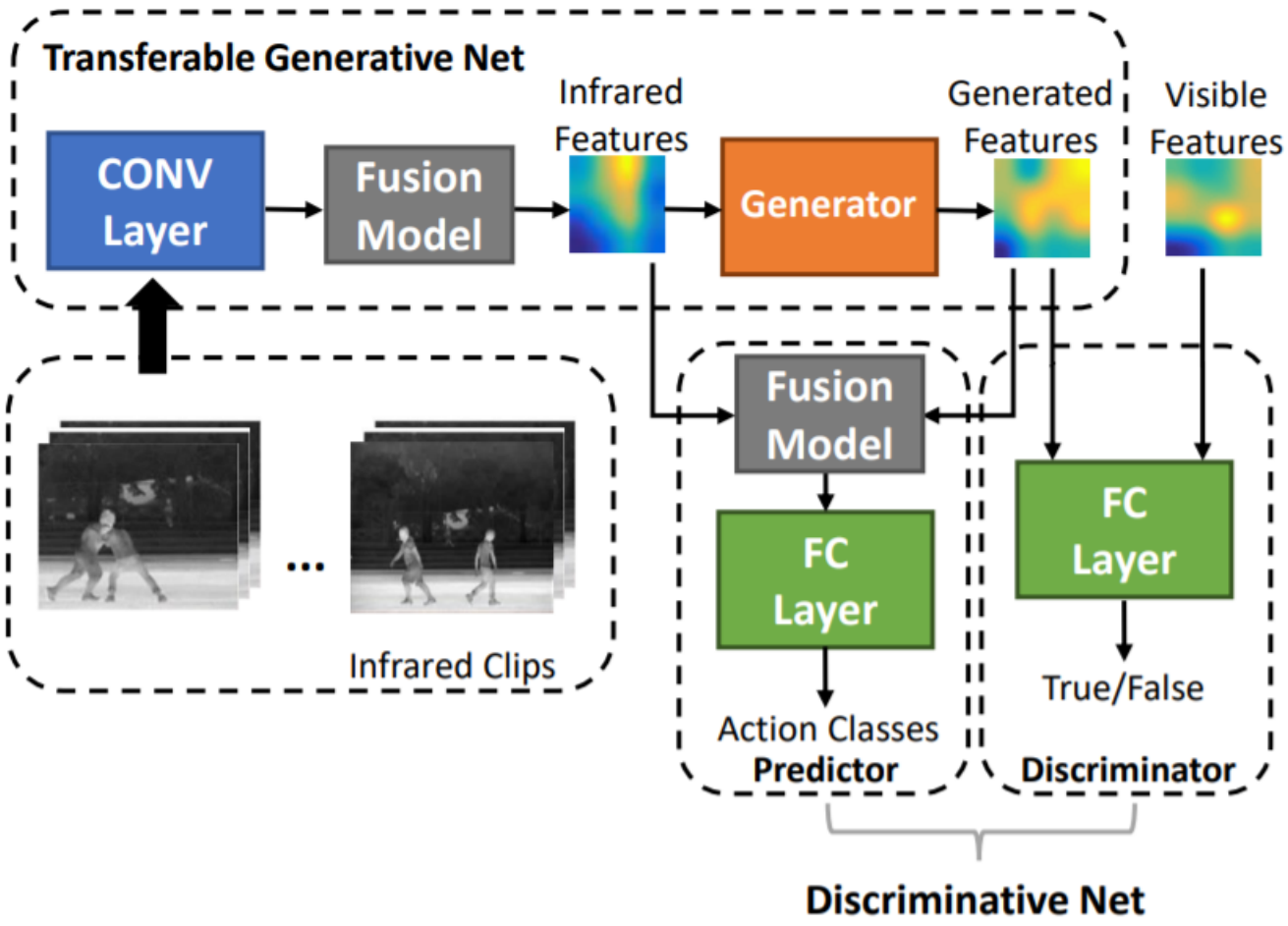

J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli. Unsupervised learning of view-invariant action representations. arXiv preprint arXiv:1809.01844, 2018

# Dividing and Aggregating Network for Multi-view Action Recognition (DA-net)




Dongang Wang, Wanli Ouyang, Wen Li, and Dong Xu. Dividing and aggregating network for multi-view action recognition. In Proc. ECCV, September 2018

# PM-GANs: Discriminative Representation Learning for action Recognition Using Partial Modalities




Lan Wang, Chenqiang Gao, Luyu Yang, Yue Zhao, Wangmeng Zuo, and Deyu Meng. PM-GANs: Discriminative representation learning for action recognition using partial modalities. In Proc. ECCV, pages 384–401, 2018

# Multi-view Existent Approaches

- Cross-view
- View-invariant
- Generative learning
  - Unseen views

- Goal:
  - Extract good features from each modality

# Challenges

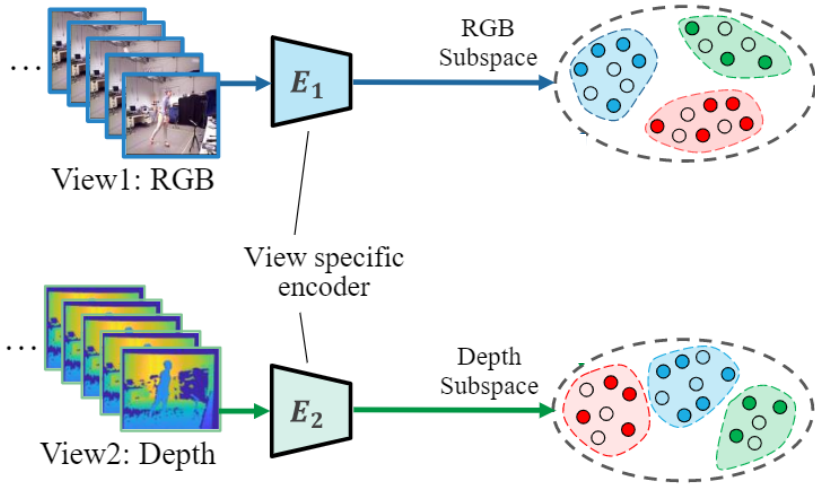
- Distinct properties among heterogeneous modalities
- Incomplete or missing view sequences
- Inconsistent view-specific predictions
- Naively fusing multi-view features could induce a negative effect
  - Concatenation
  - Summation

# Proposed Method

- Three major components

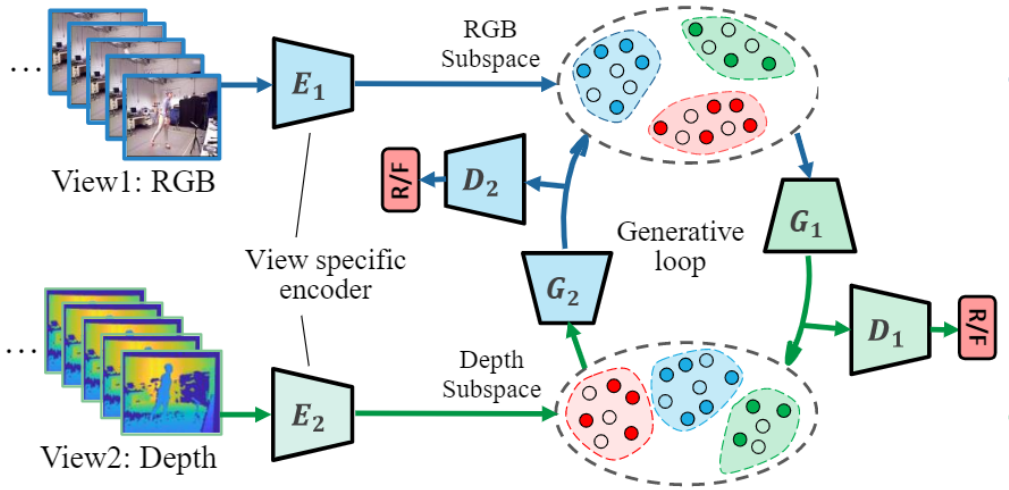
# Proposed Method

- Three major components
  - View-specific Encoders



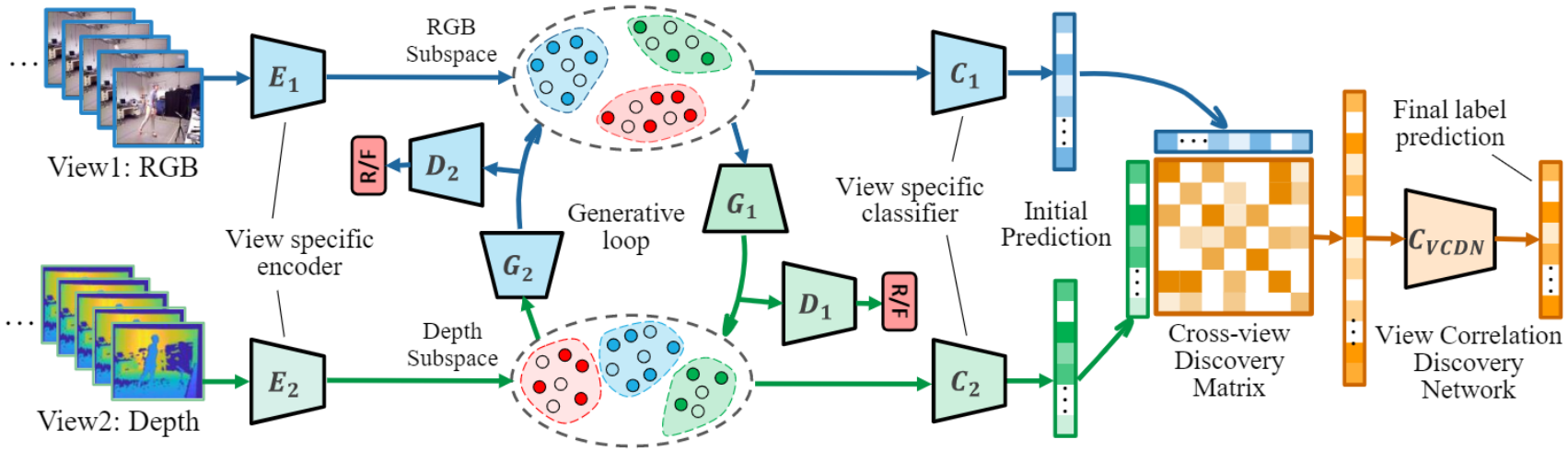
# Proposed Method

- Three major components
  - View-specific Encoders
  - Cross-view Adversarial Generators



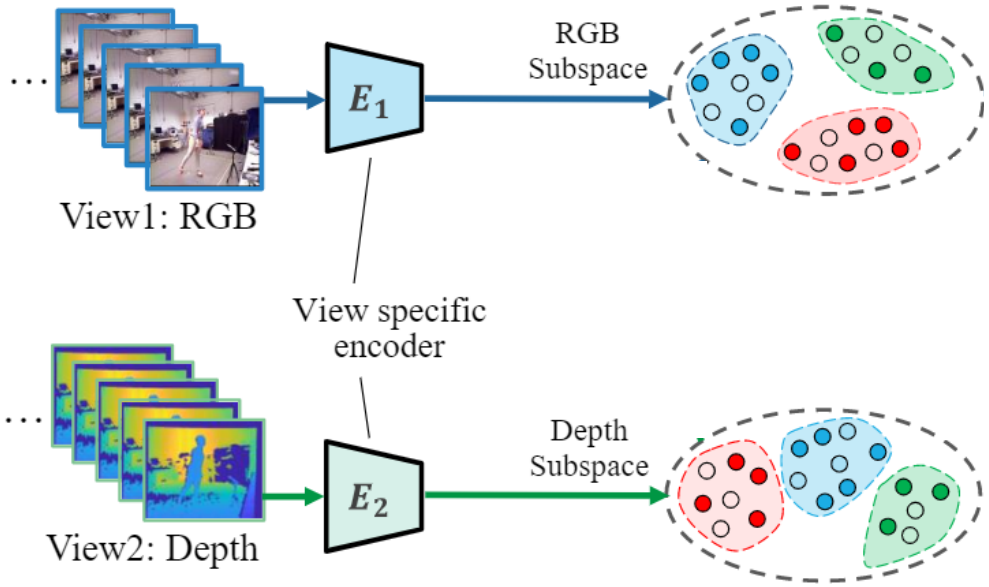
# Proposed Method

- Three major components
  - View-specific Encoders
  - Cross-view Adversarial Generators
  - View Correlation Discovery Network (VCDN)



# View-specific Encoders

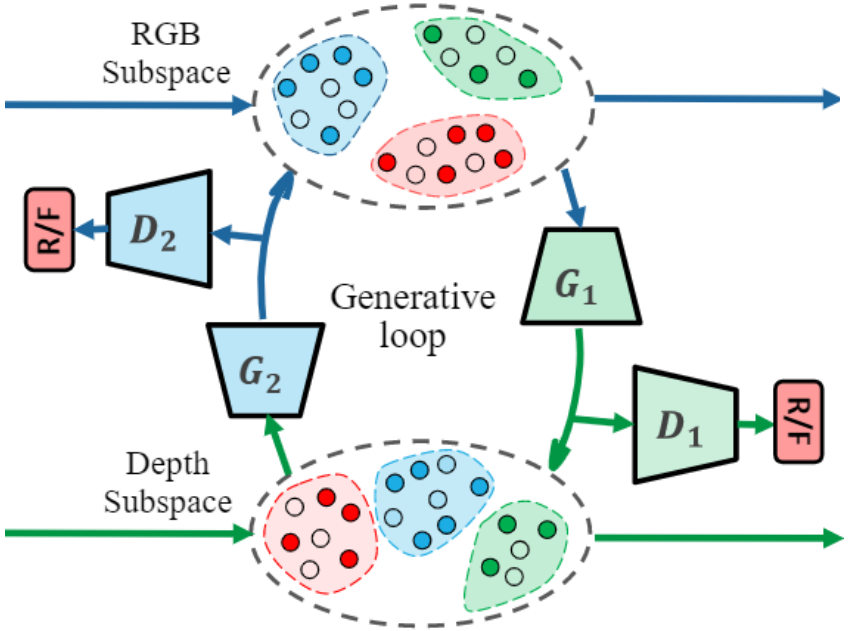
- Seek distinctive action representations in subspaces




$$L_{E_m} = \sum_{i=1}^{M} \max\Big(\big[\|E_m(X^a_{tr_i}) - E_m(X^p_{tr_i})\|_2^2 - \|E_m(X^a_{tr_i}) - E_m(X^n_{tr_i})\|_2^2 + \alpha\big], 0\Big)$$

# Cross-view Adversarial Generators

- Increase cross-view representation diversity
- Enhance model robustness
- Handle missing or incomplete view sequences




$$L_{D_1} = E_{X\sim p_X(X)} \log D_1\left(E_2(X_{tr}^2)\right) + E_{z\sim p_z(z)} \log\left(1 - D_1\left(G_1(z|E_1(X_{tr}^1))\right)\right)$$

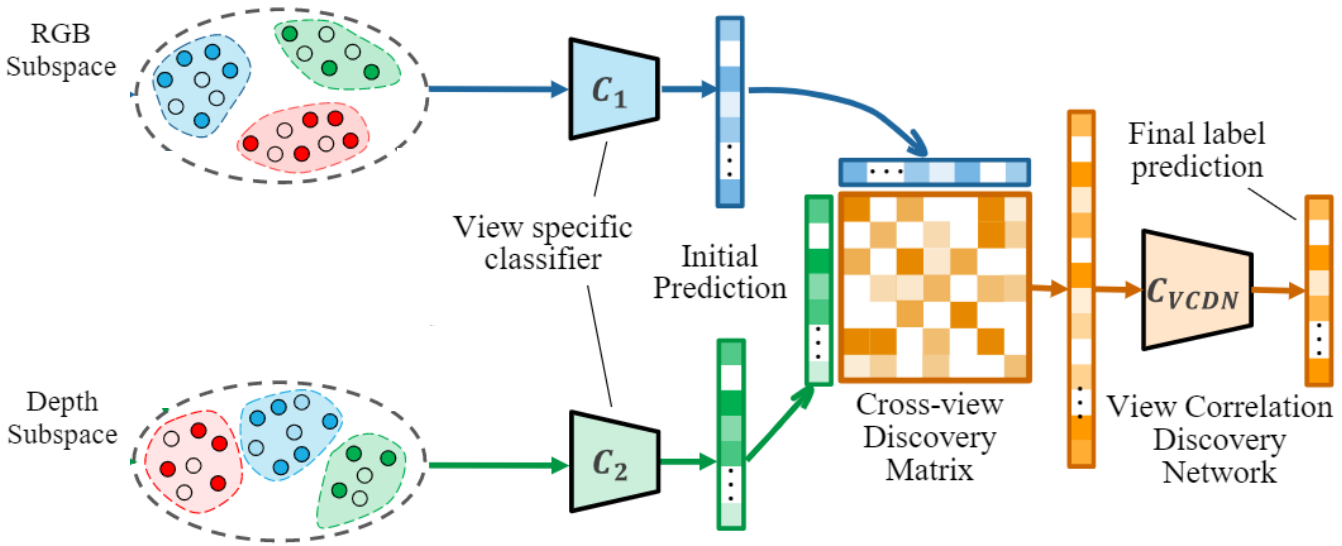
$$L_{G_1 d} = -E_{z\sim p_z(z)} \log\left(1 - D_1\left(G_1(z|E_1(X_{tr}^1))\right)\right)$$

$$L_{G_1 s} = E_{z\sim p_z(z)}\left(\|G_1(z|E_1(X_{tr}^1)) - E_2(X_{tr}^2)\|_{\mathrm{F}}^2\right)$$

$$L_{G_1} = L_{G_1 d} + \lambda L_{G_1 s}$$

# View Correlation Discovery Network (VCDN)

- View-specific classification
- Pair-wise label correlation matrix
- VCDN explore the latent high-level label correlation




$$L_{C_{m_r}} = \|Y_{tr} - C_m(E_m(X_{tr}^m))\|_{\mathrm{F}}^2$$

$$L_{C_{1_g}} = \|Y_{tr} - C_1(G_2(z|E_2(X_{tr}^2))\|_{\mathrm{F}}^2$$

$$L_{C_{2_g}} = \|Y_{tr} - C_2(G_1(z|E_1(X_{tr}^1))\|_{\mathrm{F}}^2$$

$$L_{C_m} = \beta L_{C_{m_r}} + (1-\beta) L_{C_{m_g}}$$

$$L_{C_{VCDN}^{rr}} = \sum_{i=1}^{n_{tr}} \|y_i - C_{VCDN}(y_{tr_i}^2 \cdot y_{tr_i}^{1\top})\|_2^2$$

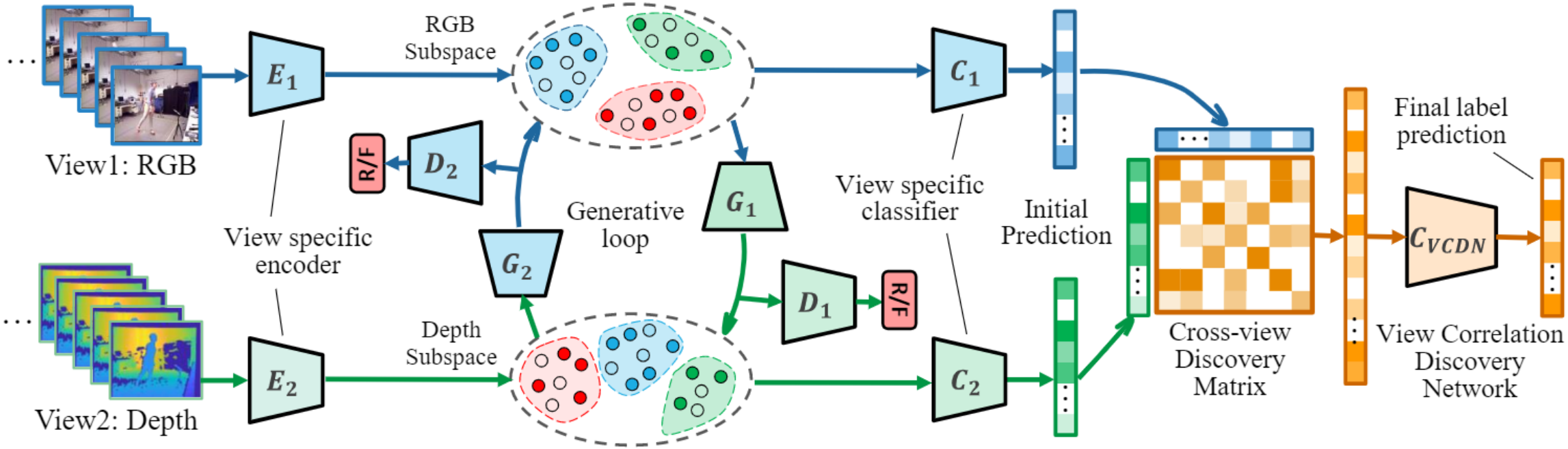
$$L_{C_{VCDN}^{rf}} = \sum_{i=1}^{n_{tr}} \|y_i - C_{VCDN}(y_{f_i}^2 \cdot y_{tr_i}^{1\top})\|_2^2$$

$$L_{C_{VCDN}^{fr}} = \sum_{i=1}^{n_{tr}} \|y_i - C_{VCDN}(y_{tr_i}^2 \cdot y_{f_i}^{1\top})\|_2^2$$

$$L_{C_{VCDN}} = \gamma L_{C_{VCDN}^{rr}} + \frac{1-\gamma}{2}(L_{C_{VCDN}^{rf}} + L_{C_{VCDN}^{fr}})$$

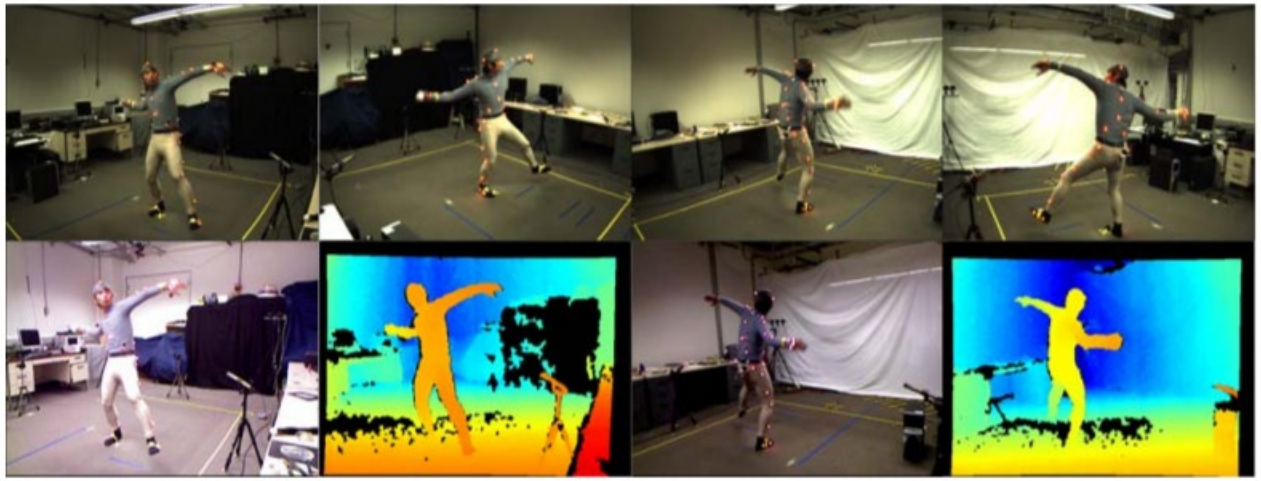
# Generative Multi-View Action Recognition (GMVAR)

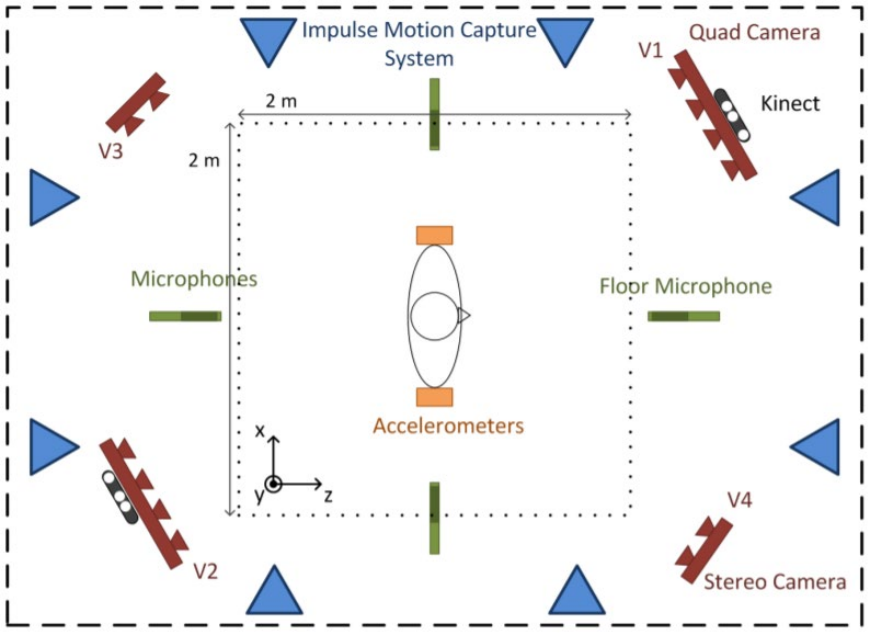
- Complete Framework



# Datasets

- Berkeley Multimodal Human Action Database (MHAD)
  - RGB, depth, skeleton, acceleration, and audio views
  - 660 action sequences
    - 11 actions
    - 12 subjects
    - 5 repetitions of each action

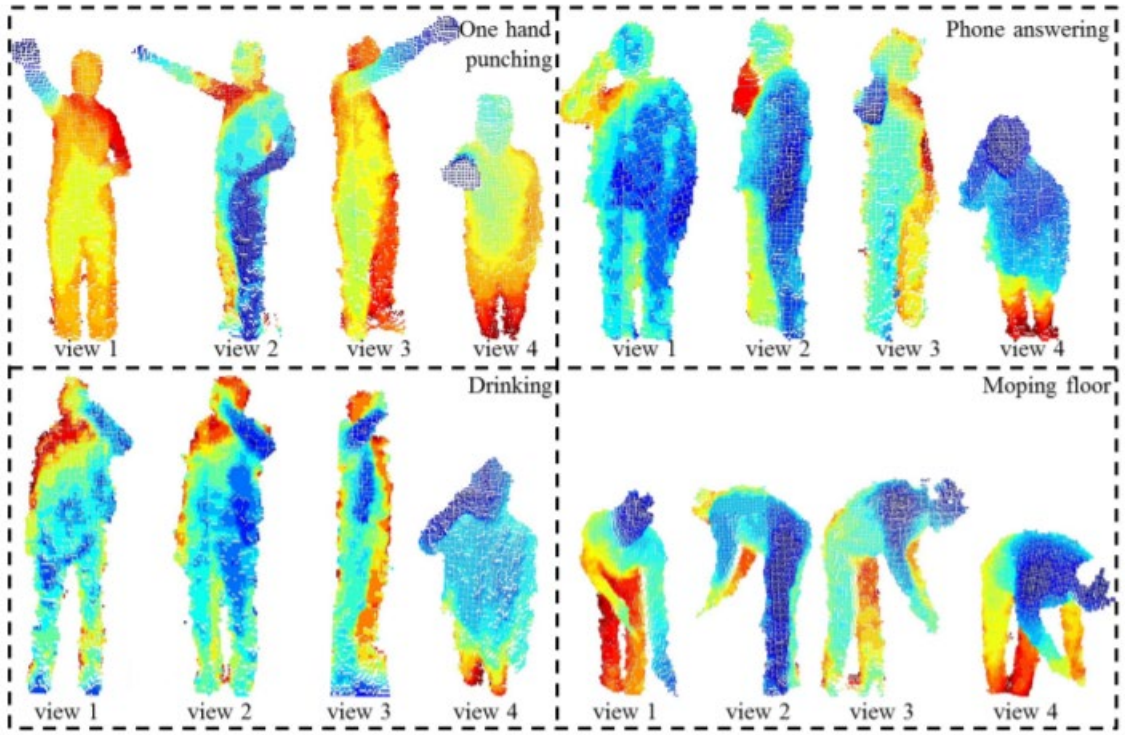





Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, Rene Vidal, and Ruzena Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In Proc. IEEE WACV, pages 53–60, 2013

# Datasets

- UWA3D Multiview Activity (UWA)
  - varying viewpoints, self-occlusion and high similarity among activities
  - 30 actions
  - 10 subjects




Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for crossview action recognition. IEEE Trans. PAMI, 38(12):2430– 2443, 2016

# Datasets

- Depth-included Human Action dataset (DHA)
  - RGB images, human masks and depth data
  - 483 video clips
    - 23 categories
    - 21 subjects

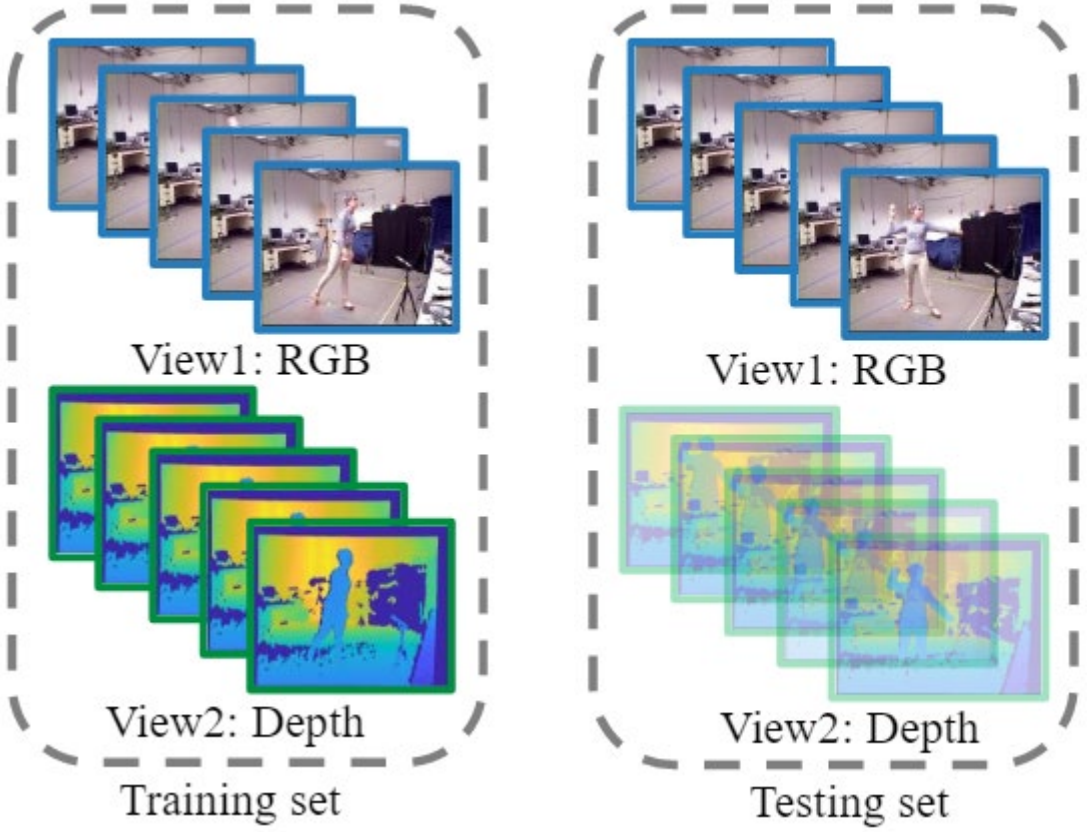


Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, YungHuan Hsieh, and Hong-Ming Chen. Human action recognition and retrieval using sole depth information. In Proc. ACM MM, pages 1053–1056, 2012

# Datasets

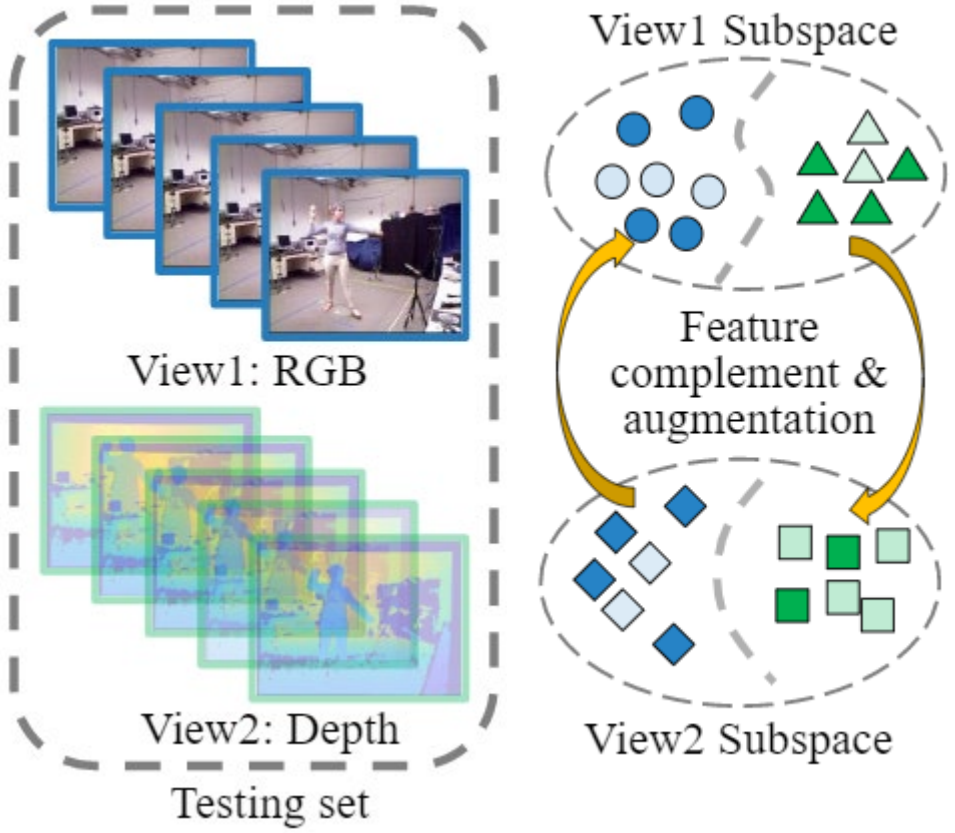
- Half of the available samples for training and another half for test

- Training
  - RGB and depth
- Tests
  - Single-view
    - RGB
    - Depth
  - Multi-view
    - RGB-D



# Experiments

- Single-view
  - RGB → Depth
  - Depth → RGB
- Multi-view
  - RGB-D



# Performance Analysis

MHAD

| Method | RGB | R→D | Depth | D→R | R+D |
|---|---|---|---|---|---|
| LSR | 96.46 | 97.17 | 47.63 | 42.51 | 97.17 |
| SVM [36] | 96.09 | 96.80 | 45.39 | 45.13 | 96.80 |
| VLAD [14] | 97.17 | - | - | - | - |
| TSN [51] | 97.31 | - | - | - | - |
| WDMM [1] | - | - | 66.41 | - | - |
| AMGL [30] | 96.46 | 97.11 | 30.03 | 29.96 | 94.70 |
| MLAN [29] | 96.05 | 96.10 | 41.48 | 41.25 | 96.46 |
| PM-GANs [49] | - | 96.76 | - | 66.84 | - |
| Ours | - | **98.23** | - | **68.32** | **98.94** |

UWA

| Method | RGB | R→D | Depth | D→R | R+D |
|---|---|---|---|---|---|
| LSR | 67.59 | 69.17 | 45.45 | 37.73 | 68.77 |
| SVM [36] | 69.44 | 68.53 | 34.92 | 34.33 | 72.72 |
| VLAD [14] | 71.54 | - | - | - | - |
| TSN [51] | 71.01 | - | - | - | - |
| WDMM [1] | - | - | 46.58 | - | - |
| AMGL [30] | 69.17 | 71.54 | 39.92 | 35.96 | 68.53 |
| MLAN [29] | 67.19 | 67.19 | 33.28 | 33.61 | 66.64 |
| PM-GANs [49] | - | 71.36 | - | 49.01 | - |
| Ours | - | **73.53** | - | **50.35** | **76.28** |

DHA

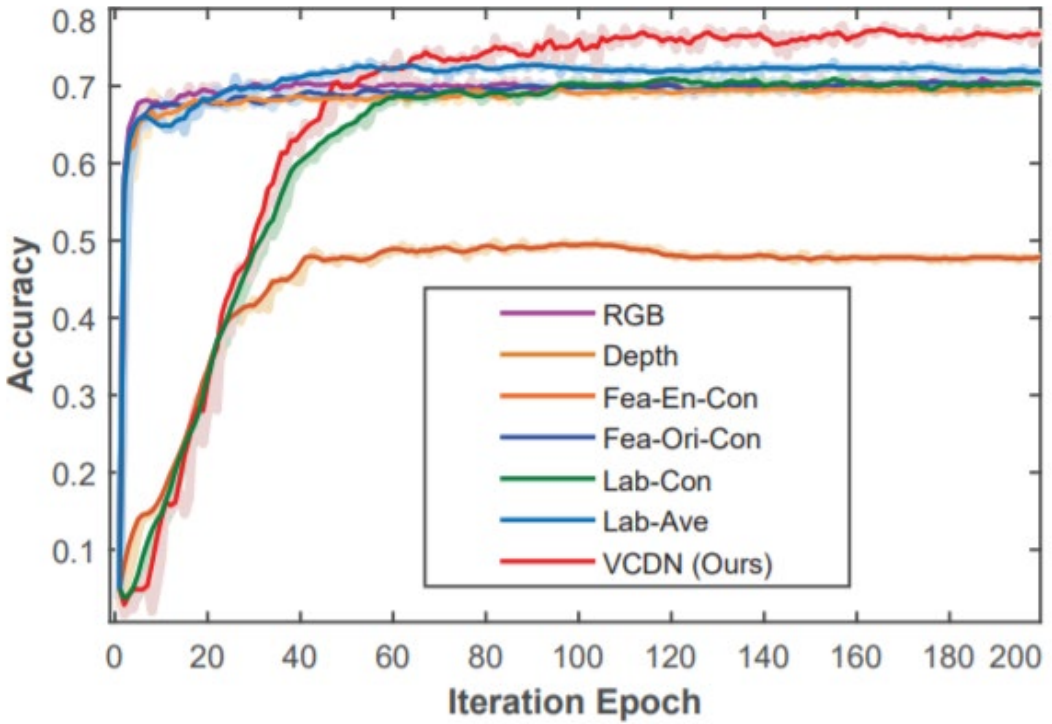
| Method | RGB | R→D | Depth | D→R | R+D |
|---|---|---|---|---|---|
| LSR | 65.02 | 65.43 | 82.30 | 48.56 | 77.36 |
| SVM [36] | 66.11 | **70.24** | 78.92 | 78.18 | 83.47 |
| VLAD [14] | 67.13 | - | - | - | - |
| TSN [51] | 67.85 | - | - | - | - |
| WDMM [1] | - | - | 81.05 | - | - |
| AMGL [30] | 64.61 | 59.05 | 72.84 | 67.33 | 74.89 |
| MLAN [29] | 67.91 | 67.91 | 72.96 | 72.83 | 76.13 |
| PM-GANs [49] | - | 68.72 | - | 76.02 | - |
| Ours | - | 69.72 | - | **83.48** | **88.72** |

# Ablation Studies

- VCDN studies
  - Different label fusion/correlation learning models
    - Feature/label concatenation
    - Label average/weighted fusion

| Setting | UWA | MHAD | DHA |
|---|---|---|---|
| RGB-$C_1$ | 69.18 | 96.42 | 68.15 |
| Depth-$C_2$ | 45.28 | 63.05 | 79.79 |
| RGBD-Fea-En-Con | 68.78 | 96.82 | 70.85 |
| RGBD-Fea-Ori-Con | 69.22 | 97.32 | 70.83 |
| RGBD-Lab-Con | 70.38 | 96.28 | 80.95 |
| RGBD-Lab-Ave | 71.84 | 97.56 | 83.28 |
| RGBD-Lab-Wei | 71.15 | 97.17 | 83.95 |
| RGBD-VCDN (Ours) | **74.07** | **98.06** | **84.32** |

UWA



# Ablation Studies

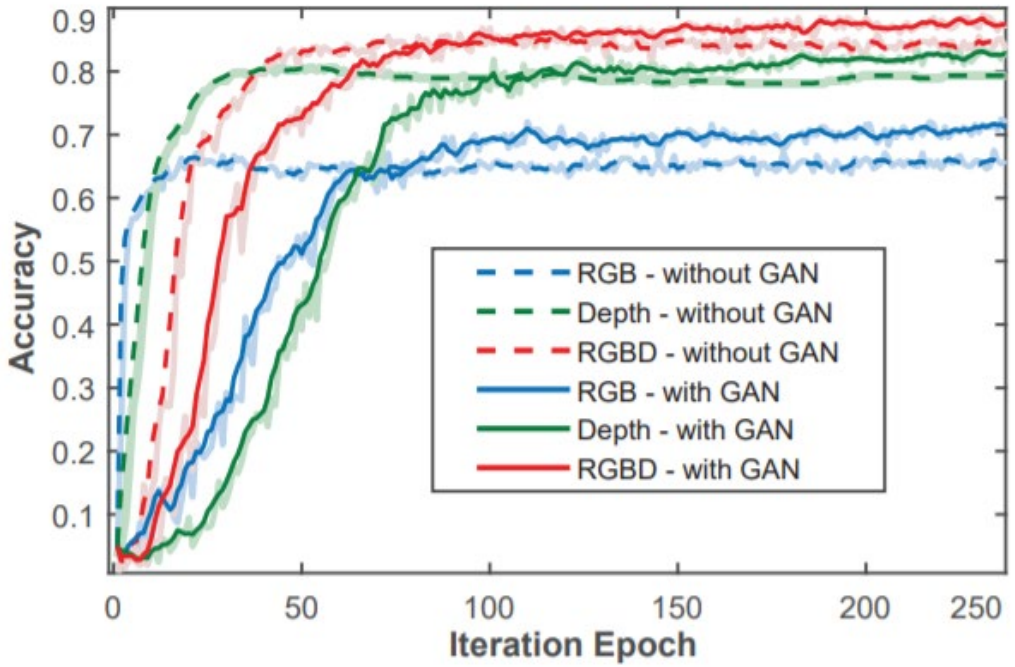
- VCDN studies
  - Regular neural networks

| Dataset | 1-layer | 2-layer | 3-layer | 4-layer | VCDN |
|---|---|---|---|---|---|
| UWA | 74.31 | 74.70 | 73.52 | 75.10 | **76.28** |
| MHAD | 97.83 | 97.88 | 96.47 | 95.76 | **98.94** |
| DHA | 86.01 | 87.24 | 85.19 | 82.72 | **88.72** |

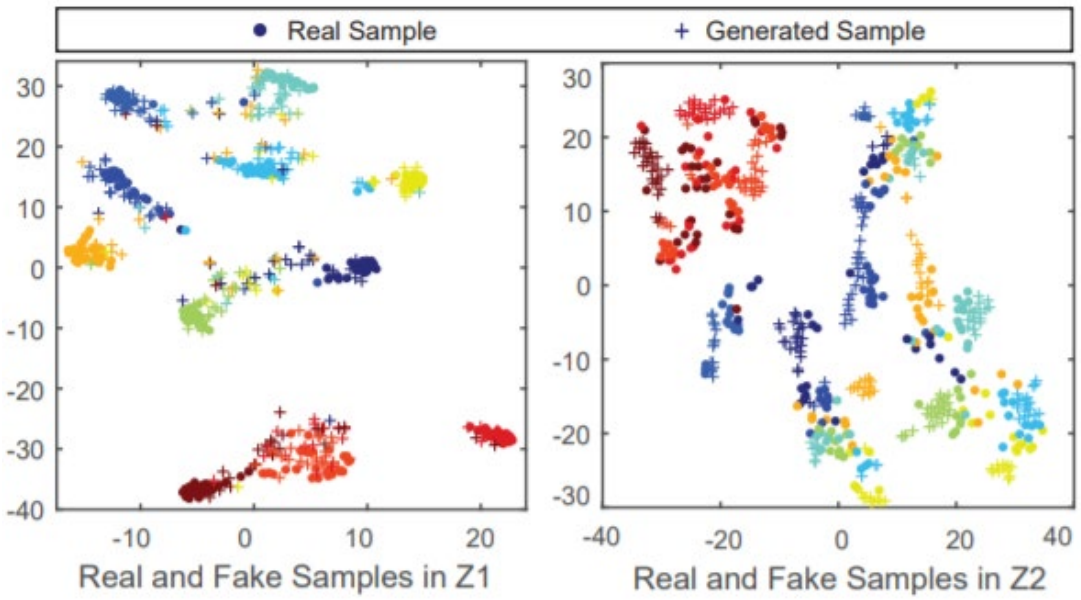
# Ablation Studies

- GAN studies

Performance (DHA)




t-SNE visualization



# Contributions and conclusion

- GMVAR can handle complete-view, partial-view, and missing-view scenarios
- Generative adversarial training enhances the accuracy and robustness of the model
- VCDN learns the intra-view and cross-view label correlations in the higher-level label space and improves the model performance
- GMVAR is an effective, accurate, robust framework, and compatible with a wide range of multi-view action recognition tasks

# Thank you!



https://github.com/wanglichenxj/Generative-Multi-View-Human-Action-Recognition

- **Lichen Wang** - https://sites.google.com/site/lichenwang123/
- **Zhengming Ding** - http://allanding.net/
- **Zhiqiang Tao** - http://ztao.cc/
- **Yunyu Liu** - https://wenwen0319.github.io/
- **Yun Raymond Fu** - http://www1.ece.neu.edu/~yunfu/