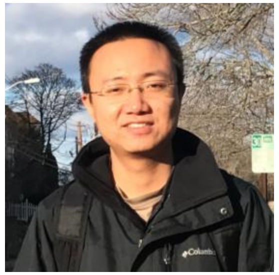
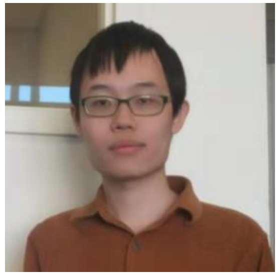
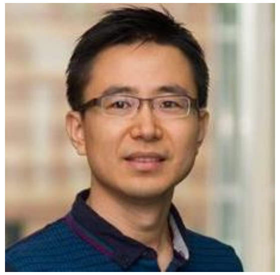
Generative Multi-View Human Action Recognition



Lichen Wang
wanglichenxj@gmail.com

Zhengming Ding
zd2@iu.edu

Zhiqiang Tao
zqtao@ece.neu.edu

Yunyu Liu
Liu.yuny@husky.neu.edu

Yun Fu
yunfu@ece.neu.edu

SMILE Lab
Electrical & Computer Engineering
Northeastern University

SmileLab website: https://web.northeastern.edu/smilelab/
Code will be available: https://github.com/wanglichenxj/Generative-Multi-View-Human-Action-Recognition

Introduction

Topic:

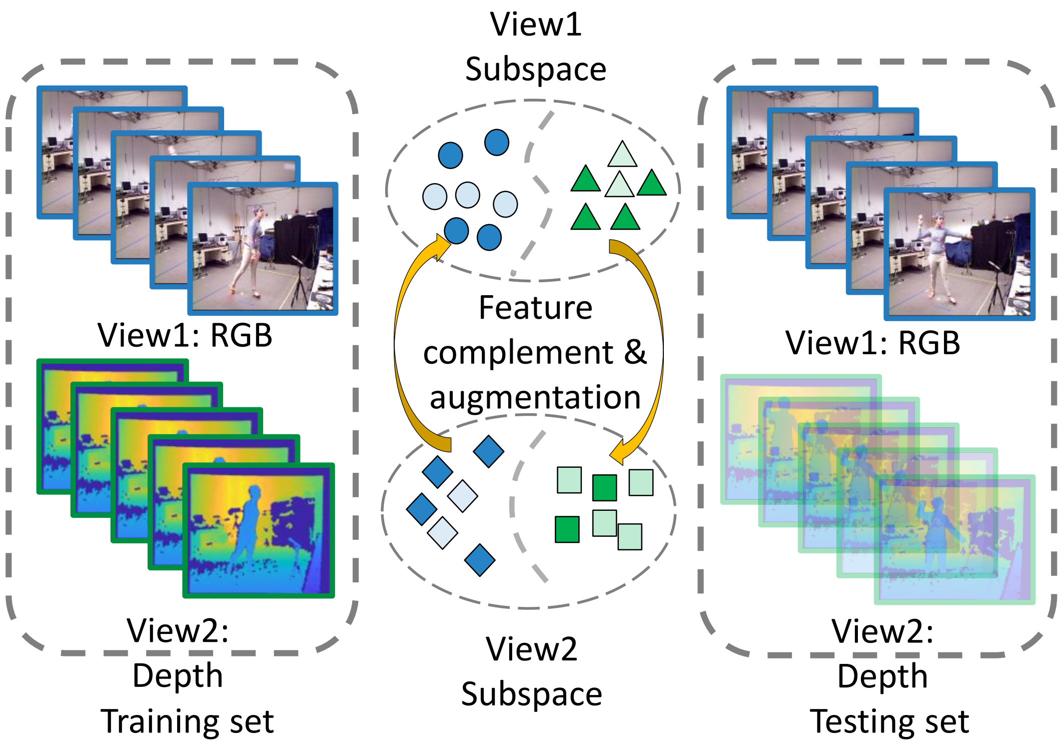
- Multi-view Action Recognition

Setting:

- Input: Multi-view action sequences
  (e.g., RGB + Depth)
- Output: Action prediction

Challenges:

- Heterogeneous multi-view feature domains
- Incomplete/missing view sequences
- Inconsistent view-specific predictions



Concept of multi-view action recognition

Motivation

Generative Multi-View Human Action Recognition (GMVAR)

Three major components to solve the challenges:

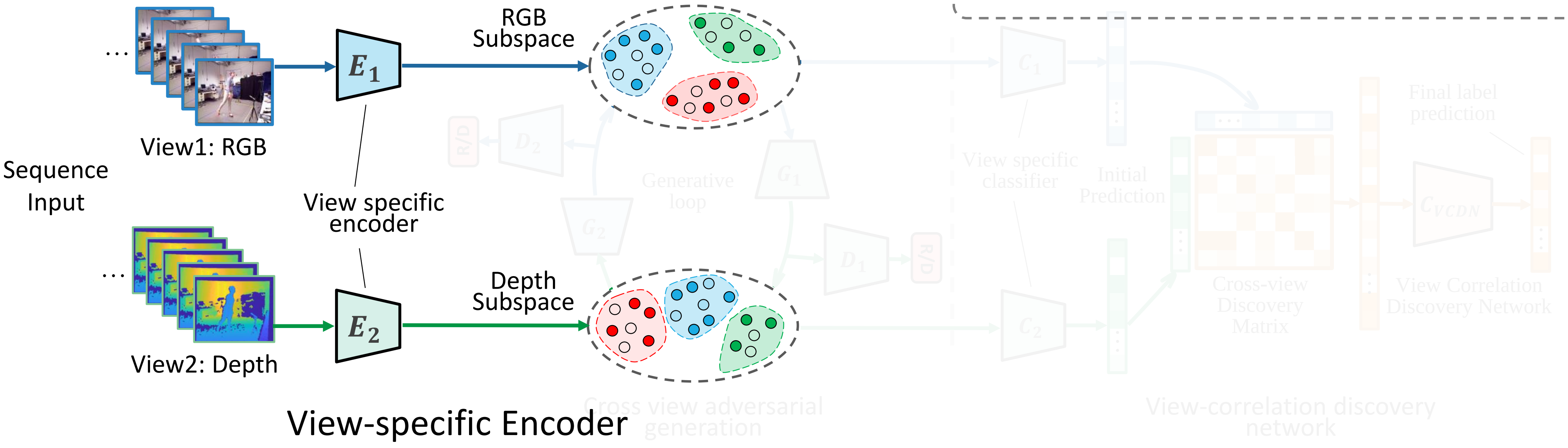
1. View-specific Encoders
2. Cross-view Adversarial Generation
3. View Correlation Discovery Network (VCDN)

1. View-specific Encoders

Mapping original feature to more distinctive subspaces

- Seek distinctive action representations in subspaces
- Label information + triplet loss objective:

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

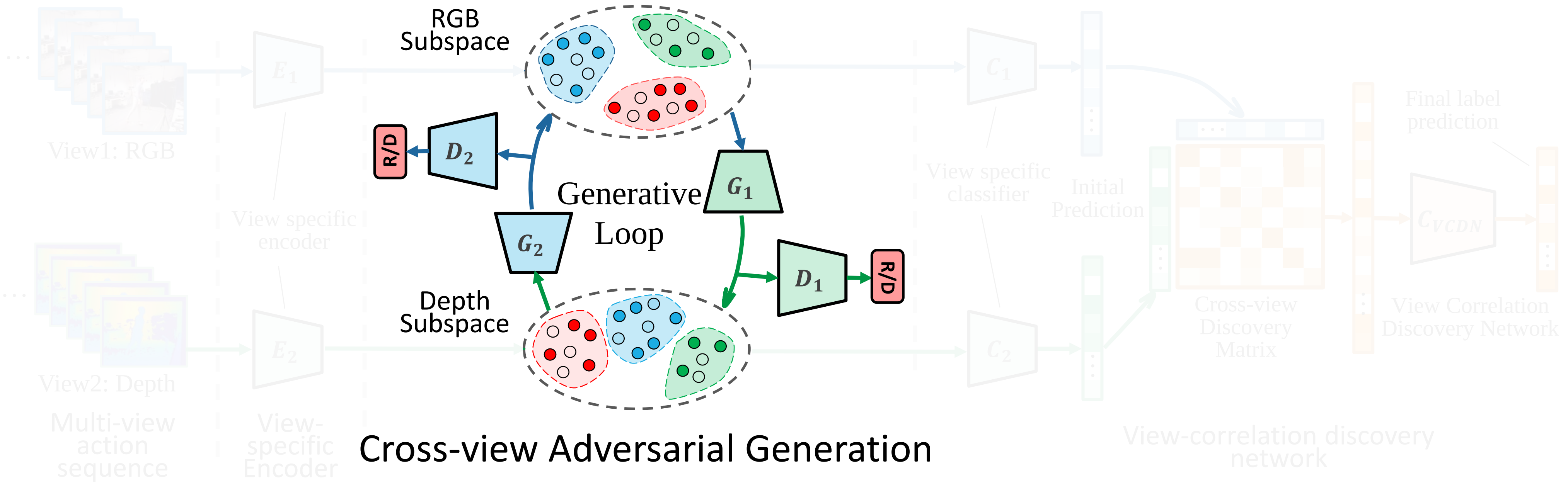


2. Cross-view Adversarial Generation

Generate one view conditioning on the other view

- Increase cross-view representation diversity
- Enhance model robustness
- Address missing/incomplete view sequences

$$
\begin{aligned}
L_{G_1 d} &= -E_{z\sim p_z(z)} \log\left(1 - D_1(G_1(z|E_1(X^1_{tr})))\right)\\
L_{G_1 s} &= E_{z\sim p_z(z)}\left(\|G_1(z|E_1(X^1_{tr})) - E_2(X^2_{tr})\|^2_{\mathrm{F}}\right)\\
L_{D_1} &= E_{X\sim p_X(X)} \log D_1(E_2(X^2_{tr}))\\
&\quad + E_{z\sim p_z(z)} \log\left(1 - D_1(G_1(z|E_1(X^1_{tr})))\right)
\end{aligned}
$$

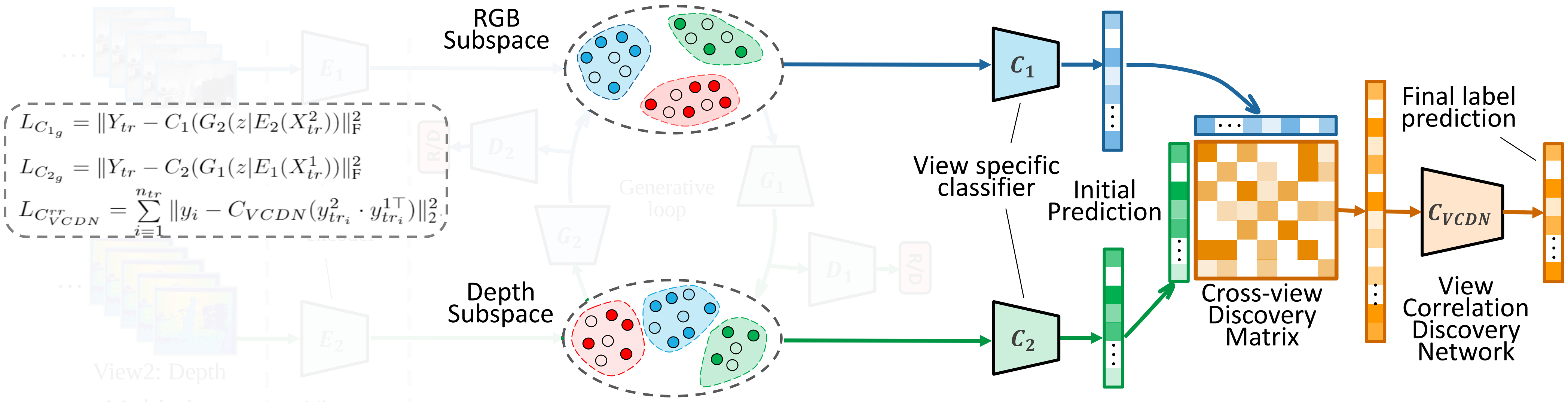


Cross-view Adversarial Generation

3. View Correlation Discovery Network (VCDN)

Explore high-level label correlations across different views

- View-specific initial classification is firstly obtained
- Pair-wise label correlation matrix is generated
- VCDN fully explore the latent high-level label correlation for higher performance

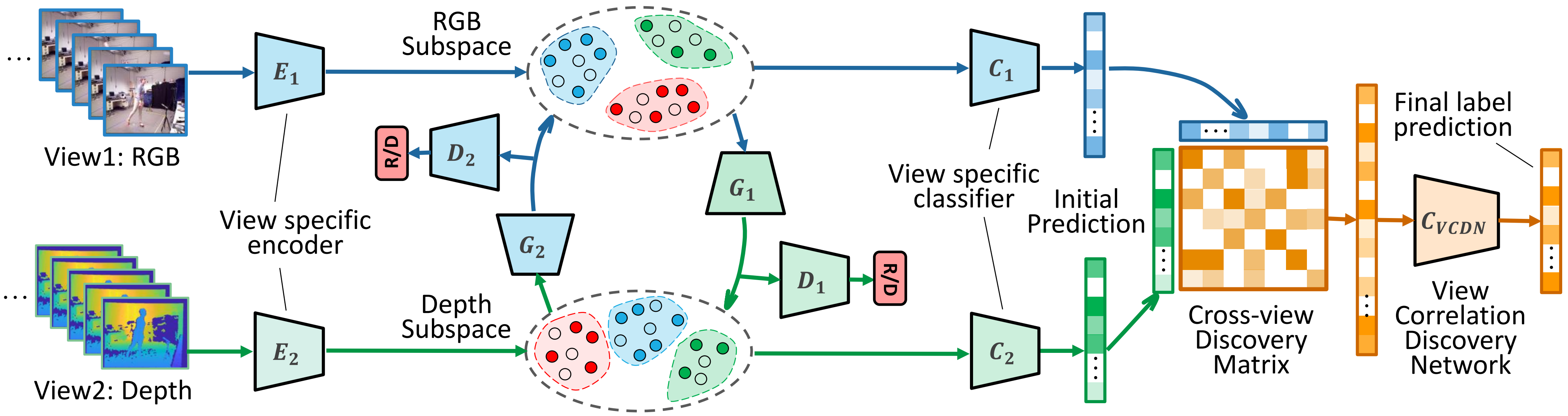


View-correlation discovery network

Our model

Generative Multi-View Human Action Recognition (GMVAR)

- Three components work together
- Jointly trained in end-to-end manner



Framework of Generative Multi-view Action Recognition

Experiments

Action recognition:

- Datasets: UWA[1], MHAD[2], and DHA[3]
- Multi-view action recognition
- Missing/incomplete multi-view (i.e., single-view) action recognition

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**

UWA

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**

MHAD

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**

DHA

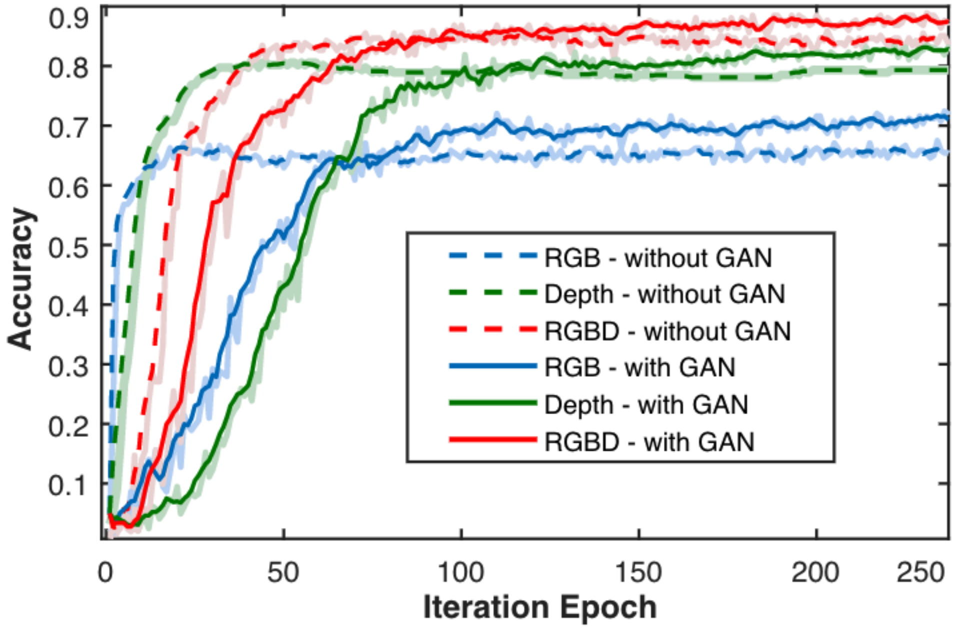
Performance on three multi-view action datasets

[1] Hossein Rahmani, et al. Histogram of oriented principal components for cross-view action recognition.IEEE Trans. PAMI, 38(12):2430–2443, 2016
[2] Ferda Ofli, et al. Berkeley mhad: A comprehensive mul-timodal human action database. In Proc. IEEE WACV, pages53–60, 2013.
[3] Yan-Ching Lin, et al. Human action recog-nition and retrieval using sole depth information. In Proc.ACM MM, pages 1053–1056, 2012.

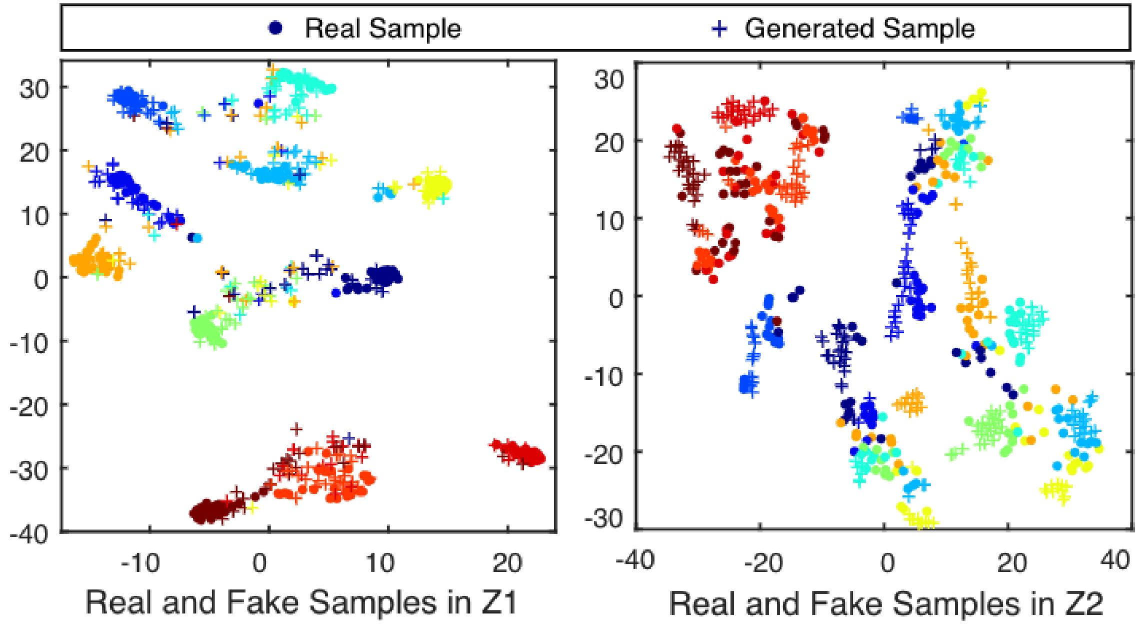
Experiments

Ablation Study for GAN:

- Performance with/without generative model
- t-SNE[1] visualization of real and fake samples



Performance with & without GAN



t-SNE[1] visualization of real & fake samples

[1] SL.J.P. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 15(Oct):3221-3245, 2014.

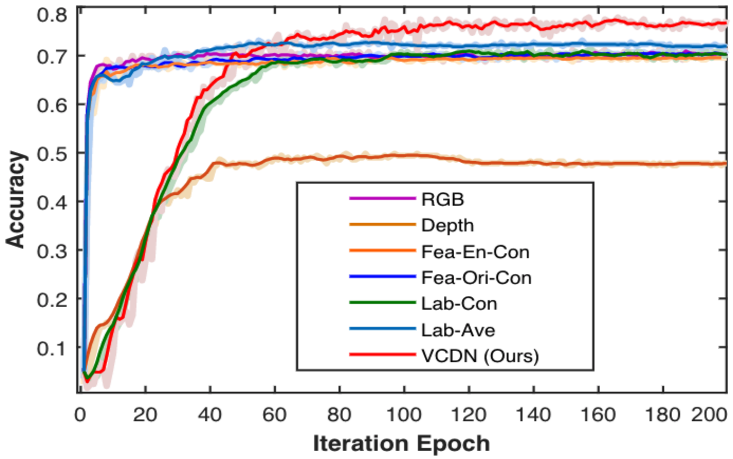
Experiments

Ablation Study for VCDN:

- VCDN compared with different label fusion/correlation learning models
  - Feature/label concatenation & label average/weighted fusion
- VCDN compared with 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**

Classification performance of VCDN compared with simple NN.



Performance with different label fusion modules

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**

Performance with different label fusion modules

Summary

Generative Multi-View Human Action Recognition:

- Multi-view Encoder → Distinctive representation
- Cross-view Generation → Diversify samples & Enhance generalization
- VCDN → Explore view correlations in high-level label space

Conclusion:

- Proposed modules are effective
- Obtain considerable classification improvements

Thank you!

Poster #25: Oct. 31, 15:30, Poster 3.2 (Hall B)

Please contact: wanglichenxj@gmail.com for questions.

SMILE Lab

Electrical & Computer Engineering

Northeastern University