Generative Multi-View Human Action Recognition

Lichen Wang[1], Zhengming Ding[2], Zhiqiang Tao[1], Yunyu Liu[1], and Yun Fu[1]

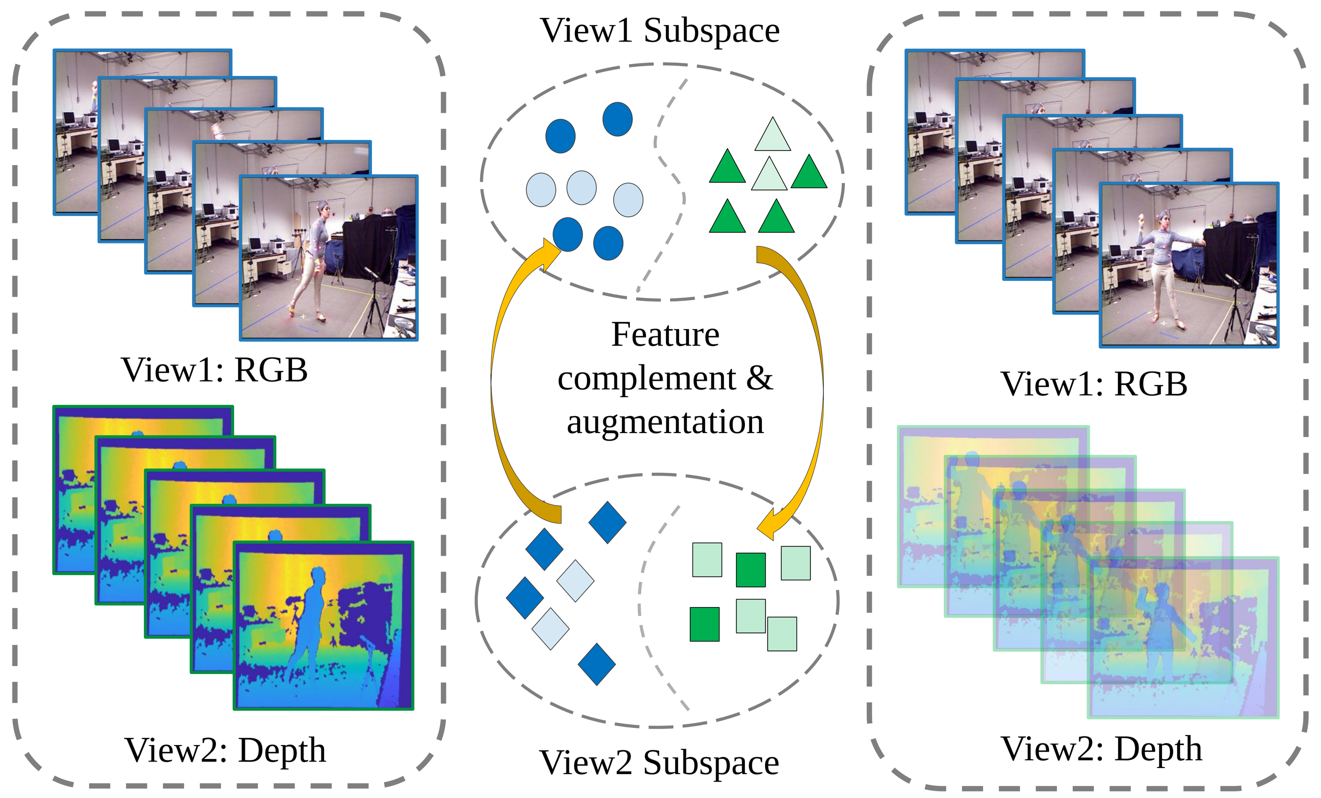
wanglichenxj@gmail.com

[1]Northeastern University, [2]Indiana University-Purdue University Indianapolis

Introduction

❖ Multi-view Action Recognition

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



Concept of multi-view human action recognition

❖ Challenges

- Heterogeneous (significantly different) multi-view feature domains (e.g., RGB and depth, RGB and electronic signal)
- Incomplete/missing view sequences (e.g., only one view is available)
- Inconsistent view-specific prediction (e.g., RGB and depth have different prediction results)

❖ Motivations

- Obtain more distinctive feature representations
- Explore cross-view feature relations
- Explore the multi-view prediction results in high-level label space

Our model

❖ Three modules are proposed for solving the challenges

- **View-specific encoders**
  - Seek distinctive action representations in subspaces
  - Label information + triplet loss objective
- **Cross-view Adversarial Generation**
  - Increase cross-view representation diversity
  - Enhance model robustness
  - Address missing/incomplete view sequences
- **View Correlation Discovery Network**
  - 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

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

$$L_{G_1 d} = -E_{z\sim p_z(z)} \log\Big(1 - D_1\big(G_1(z|E_1(X^1_{tr}))\big)\Big)$$

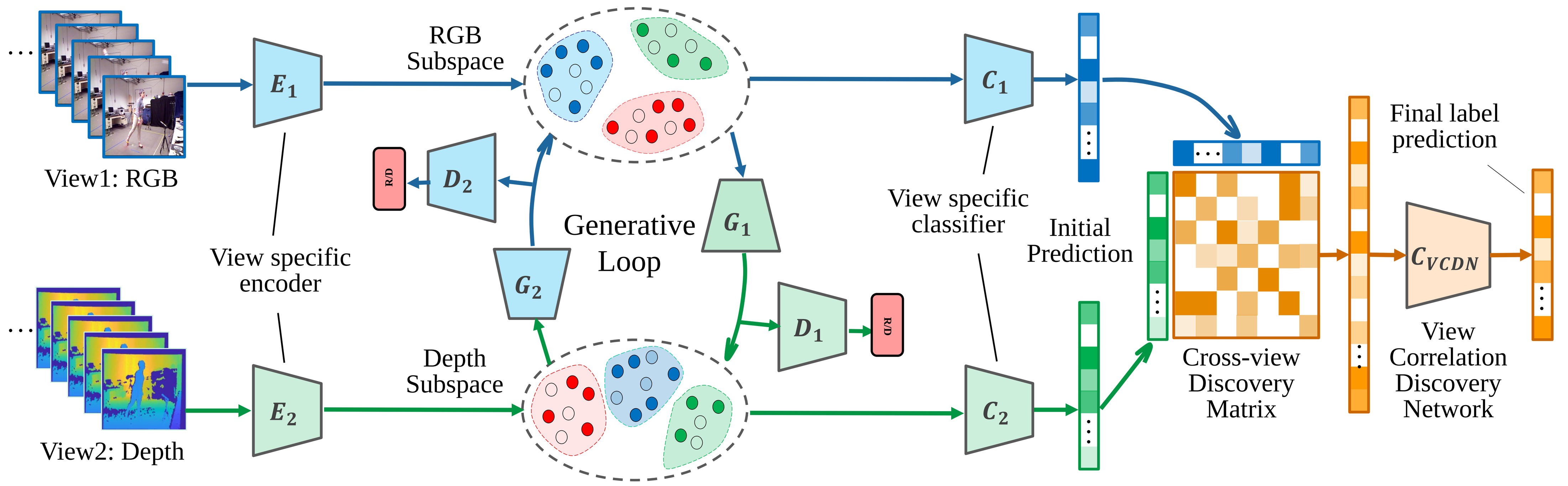
$$L_{G_1 s} = E_{z\sim p_z(z)}\Big(\|G_1(z|E_1(X^1_{tr})) - E_2(X^2_{tr})\|_{\mathrm{F}}^2\Big)$$

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

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

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

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



Framework of our generative multi-view human action recognition

Experiments

- Conventional multi-view action recognition setting.
- Single-view action recognition setting, where another view is considered as missing view which is generated by cross-view generation strategy.
- Ablation study for cross-view generation module and view correlation discovery module

❖ Single-view & multi-view action classification

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 dataset

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**

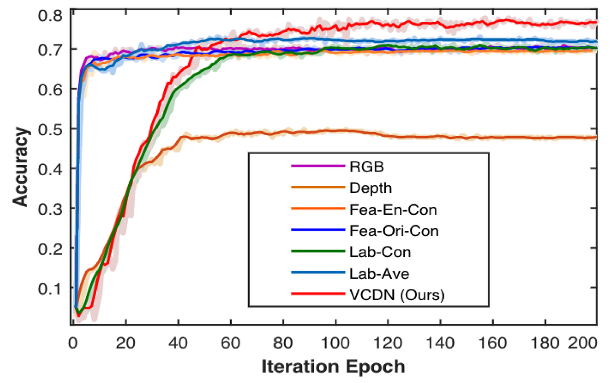
MHAD dataset

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 dataset

❖ Ablation study

- Intentionally remove or modify the View Correlation Discovery Network



Performance with different label fusion strategies

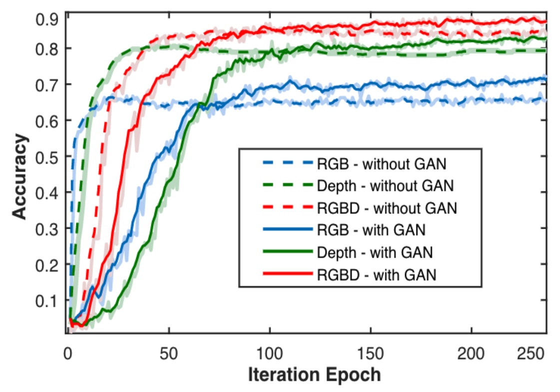
Setting	UWA	MHAD	DHA
RGB-C_1	69.18	96.42	68.15
Depth-C_2	45.28	63.05	79.79
RGBD-Fea-En-Con	68.78	96.82	70.85
RGBD-Fea-Ori-Con	69.22	97.32	70.83
RGBD-Lab-Con	70.38	96.28	80.95
RGBD-Lab-Ave	71.84	97.56	83.28
RGBD-Lab-Wei	71.15	97.17	83.95
RGBD-VCDN (Ours)	**74.07**	**98.06**	**84.32**

Performance with different label fusion strategies

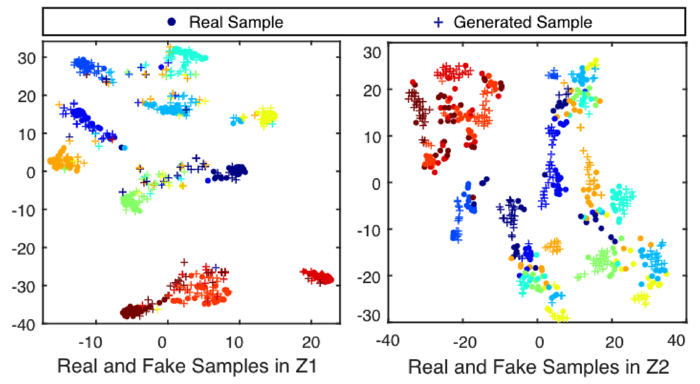
Dataset	1-layer	2-layer	3-layer	4-layer	VCDN
UWA	74.31	74.70	73.52	75.10	**76.28**
MHAD	97.83	97.88	96.47	95.76	**98.94**
DHA	86.01	87.24	85.19	82.72	**88.72**

Performance with different label fusion strategies

- The performance with and without the generative model



Performance with and without the GAN module



t-SNE visualization of real and generated samples

Conclusion

We proposed three modules to address the challenges of multi-view human action recognition. View-specific encoder learns distinctive action representations. Cross-view generation extend the representation distributions. View correlation discovery network explore the high-level correlations in label space. All modules are trained simultaneously to achieve the best performance.