Generative Multi-View Human Action Recognition

Lichen Wang[1], Zhengming Ding[2], Zhiqiang Tao[1], Yunyu Liu[1], Yun Fu[1]
[1]Northeastern University, USA
[2]Indiana University-Purdue University Indianapolis, USA
wanglichenxj@gmail.com, zd2@iu.edu, zqtao@ece.neu.edu,
liu.yuny@husky.neu.edu, yunfu@ece.neu.edu

Abstract

Multi-view action recognition targets to integrate complementary information from different views to improve classification performance. It is a challenging task due to the distinct gap between heterogeneous feature domains. Moreover, most existing methods neglect to consider the incomplete multi-view data, which limits their potential compatibility in real-world applications. In this work, we propose a Generative Multi-View Action Recognition (GMVAR) framework to address the challenges above. The adversarial generative network is leveraged to generate one view conditioning on the other view, which fully explores the latent connections in both intra-view and cross-view aspects. Our approach enhances the model robustness by employing adversarial training, and naturally handles the incomplete view case by imputing the missing data. Moreover, an effective View Correlation Discovery Network (VCDN) is proposed to further fuse the multi-view information in a higher-level label space. Extensive experiments demonstrate the effectiveness of our proposed approach by comparing with state-of-the-art algorithms[1].

1. Introduction

Multi-view approaches [58, 30, 29, 62, 40] explore the complementary information among different views, where the views refer to various feature representations, modalities or sensors. Most existing methods focus on analyzing static multi-view data (*e.g.*, image, description, and attributes), while recently, multi-view action recognition [10, 4, 18, 22] has become attractive and urgent as the increasing multi-modal sensors are widely deployed in a great number of real-world applications.

There are two categories in the multi-view action recognition scenario. The first category explores action sequences captured by multiple sensors which belonging to the same visual modality (*e.g.*, surveillance system usually captures videos with RGB-only cameras). These methods assume that actions recorded by different viewpoints (*e.g.*, front, back, and top) or distances could provide distinctive aspects for recognition tasks [4, 18, 22]. The second category methods analyze action sequences captured from different types of sensors (*e.g.*, RGB, depth, skeleton, acceleration, trajectory, 3D, and electromyography [26, 32, 34, 50, 52, 53]) and attempt to integrate the complementary information among various modalities. For instance, Kinect sensor [61, 33] provides high-quality RGB, depth, and skeleton sequences simultaneously, where both depth [1, 54] and skeleton [39, 44, 59] modalities have been demonstrated to provide effective and unique motion knowledge for action recognition. Electromyography (EMG) signal which reflects the electrical activity produced by skeletal muscles is utilized for action/motion analysis [3, 50]. Acoustical and acceleration are also utilized for multi-view event detection and action recognition tasks [11, 9].

In this study, we focus on the second category. As shown in Figure 1, both RGB and depth views are available in the

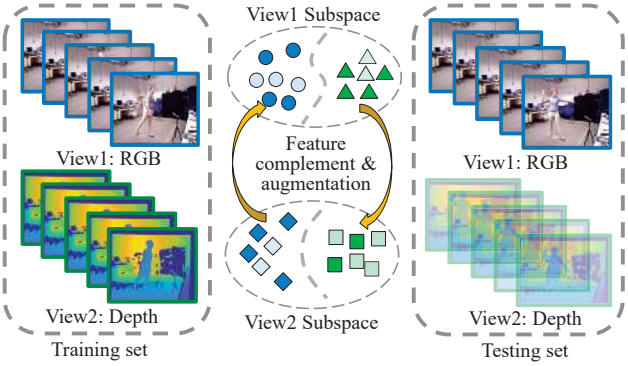

Figure 1. Illustration of our GMVAR approach, which is trained on both RGB and depth views. However, in the test stage, GMVAR is capable of dealing with different scenarios including complete multi-view, partially missing view, or even single-view. It is due to the generative mechanism in our model which significantly extends the potential applications of our approach.

[1]Code is available on: https://github.com/wanglichenxj/Generative-Multi-View-Human-Action-Recognition

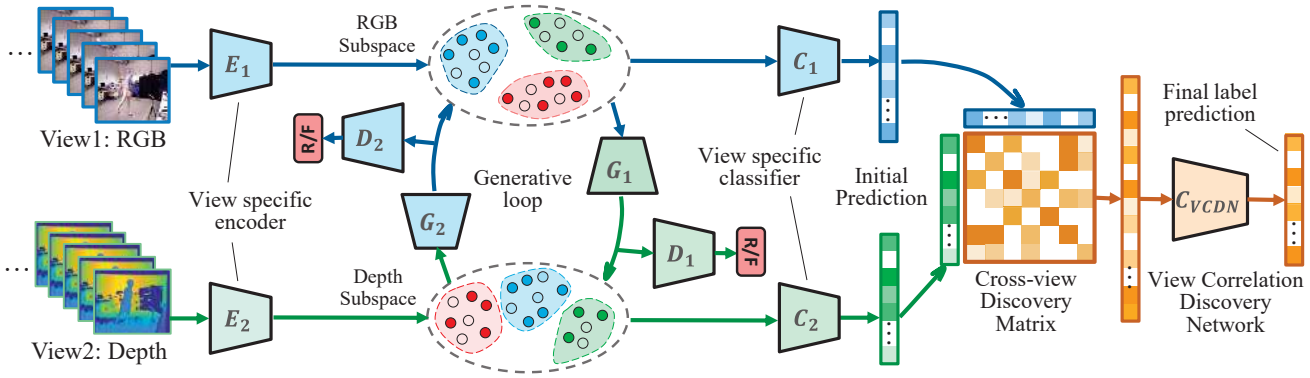

Figure 2. Framework of our proposed model. The RGB and depth views first go through the feature encoders $E_1(\cdot)$ and $E_2(\cdot)$ respectively to obtain more distinctive representations in the latent subspaces Z_1 and Z_2. Two generators $G_1(\cdot)$ and $G_2(\cdot)$ generate representations conditionally based on the other subspace. This generative mechanism fully explores the feature distribution across Z_1 and Z_2. Two view-specific classifiers $C_1(\cdot)$ and $C_2(\cdot)$ are trained to obtain initial recognition prediction from each view, then the proposed View Correlation Discovery Network (VCDN), $C_{VCDN}(\cdot)$, is utilized to further enhance the multi-view final prediction. Our model fully reveals the latent cross-view connection by the generative model in latent subspaces, and further explores the high-level view-correlation knowledge in label space. Due to the generative model, our model is compatible for both multi-view and single-view scenarios.

training stage while either complete or incomplete views is available in the test stage. RGB-D action recognition is one of the most important research directions due to the popularity of depth/3D sensors and the corresponding applications [61, 23, 19]. It is a challenging task due to the distinct properties among heterogeneous modalities. Naively fusing multi-view features (*e.g.*, concatenation or summation) could induce a negative effect and hurt the performance. Previous research efforts [6, 60, 25, 20, 5] mainly utilize effective feature extraction approaches to obtain view-specific representation first, then deploy fusion mechanism to integrate these representations together. However, these methods assume data are accessible for all the views, yet without considering the practical and common incomplete view scenarios (*e.g.*, sensor malfunction, equipment deficiency, and signal loss in data transformation). Hence, their performances inevitably degrade when dealing with partial multi-view data. Moreover, different views could provide class-level unique distinctiveness, and it is crucial to explore the correlation across action classes and views to further improve the learning performance.

In this work, we propose a Generative Multi-view Action Recognition (GMVAR) framework, which adopts generative adversarial training as well as a simple yet effective View Correlation Discovery Network (VCDN) to address the above challenges. Particularly, two generative networks are developed to learn the instance-level pairwise cross-view connection knowledge, which could fully leverage the complementary information among views. More specifically, each view's generator is trained to reproduce its own latent representation, conditioning on the other view's information. By this way, our approach is able to effectively enrich the multi-view representations, and handle the missing modality case. Moreover, a View Correlation Discovery Network (VCDN) is designed to learn the higher-level cross-view correlation in the label space, which further explores the class-level distinctiveness of the views. Experimental results on three RGB-D video datasets demonstrate the superiority of our model. The main contributions of our approach are listed below:

- We proposed a generative multi-view action recognition framework, which can simultaneously handle complete-view, partial-view, and missing-view scenarios using a unified strategy.
- The adversarial training is encapsulated into our model to explore the complementary information shared by different modalities, which works as a regularizer to enhance the accuracy and robustness of our model.
- A simple yet effective View Correlation Discovery Network (VCDN) is proposed to learn the intra-view and cross-view label correlations in the higher-level label space. It further explores the label information and significantly improves model performance.

2. Related work

2.1. Multi-view Action Recognition

Multi-view action recognition uses data taken from multiple views/resources to achieve higher performance. It assumes different views are complementary which provide extra information and help to distinguish actions. DA-Net [42] obtains both view-independent and view-specific representations and utilizes a view classifier to combine the classification score from each view. PM-GANs [49] deploys generative and feature fusion strategies for inferred

action recognition. [38] proposed a shared-specific feature factorization network which effectively fuses RGB and depth information. [20] presents a joint learning model to simultaneously explore the shared and feature-specific components to improve learning performance. [17] achieves modality hallucination through shared weights neural network for image classification. [41] proposed a cascaded residual autoencoder to handle missing view scenario. [4] fuses the action descriptors by utilizing a Multi-view Super Vector. [18] designs a novel approach for combining optical flow into enhanced 3D motion vector fields to achieve feature fusion. [13] proposes a first-person hand action recognition baseline based on 3D hand pose and RGB view. [63, 43] explores a view-invariant feature extraction approach which is robust for actions captured from different views. Depth view is considered in [1, 54] and there are skeleton based recognition approaches [39, 44, 59] for action recognition.

Compared with existing methods, our approach is different in the following two aspects. First, it is a general multi-view action recognition approach which could handle complete-view, partial-view, and missing-view scenarios in a unified framework; second, instead of fusing views in feature space, our approach explores the correlations residing in the high-level label space which could deliver more accurate recognition results.

2.2. Generative Adversarial Network (GAN)

GAN [15] consists of two networks: the generator and the discriminator. The generator is trained to make generated samples while the discriminator tries to differentiate the samples. Competition strategy drives both networks to enhance their abilities. Many GAN variants are recently proposed. Mode-Regularized GAN [8] introduces ways to dramatically stabilize the training process. Conditional GAN (CGAN) [28] extends GAN model by adding extra conditional information (*e.g.*, label knowledge) to regularize generation process. Auxiliary Classifier GAN (ACGAN) [31] combines an auxiliary classifier with CGAN for image synthesis applications. Ding et al. explore two-stage conditional generative model for zero-shot learning [12]. Small Object Detection GAN (SOD-MTGAN) [2] generates high resolution small objects to improve multi-class detection performance. [55] deploys generative strategy to handle missing view clustering task, and [40] uses ensemble strategy to achieve final clustering result. Cycle GAN [64] utilizes the generative approach and its inverse direction to achieve unpaired image style translation. However, current models are mainly (*e.g.*, GAN, CGAN) designed to subjectively diversify images and utilize the human perceptual aspect (*e.g.*, MS-SSIM [56]) to evaluate the diversity; while we want to generate representations from one view to another view to solve the multi-view, partial-view, and missing view problems.

Compared with other generative models, our model builds connections across views and is designed to complement/boost the feature diversity for classification goal. Specifically, there are two major differences compared with other generative models: first, our approach is proposed to explore the generative strategy in the multi-view scenario. In addition, we deploy the generative strategy in latent subspace instead of raw feature space which hopefully explores the data structure and obtains more distinctive feature representations; second, a triplet loss is deployed to an autoencoder which fully utilizes the available supervision information to obtain high quality subspace.

3. Our approach

3.1. Preliminaries & Motivation

Given the multi-view training data X_{tr}^1 and X_{tr}^2, where $X_{tr}^1 \in \mathbb{R}^{d_1 \times n_{tr}}$ and $X_{tr}^2 \in \mathbb{R}^{d_2 \times n_{tr}}$ are the feature matrices of two views, where each column represents one instance, n_{tr} is the training instance number, and d_1, d_2 are the feature dimensions of view1 and view2. $Y_{tr} \in \mathbb{R}^{d_l \times n_{tr}}$ is the one-hot label matrix, where d_l is the dimension of the label space. Correspondingly, $X_{te}^1 \in \mathbb{R}^{d_1 \times n_{te}}$, $X_{te}^2 \in \mathbb{R}^{d_2 \times n_{te}}$, and $Y_{te} \in \mathbb{R}^{d_l \times n_{te}}$ are the test features and the label matrices. Considering some of the test samples only containing single-view data, thus, the goal of our approach is to predict the label matrix Y_{te}, when either only single-view (X_{te}^1 or X_{te}^2) or both views (X_{te}^1 and X_{te}^2) are available. Generally, the feature space is much more diverse than the label space especially in multi-view action recognition scenario. To this end, we aim to compensate the visual feature and mitigate the gap between the training and test samples especially when the other view is not available.

3.2. Subspace Conditional Feature Generation

Inspired by the idea of generative models [15, 28, 31], we propose the generative networks to synthesize one view conditioned on the other view. By this way, the generators learn the cross-view connections and also borrow shared motion components from other actions which effectively diversifies the generated representations. Moreover, considering the original visual feature contains high-level noise, and directly generating features conditioned on visual space could bring in negative influence to the label prediction [46, 47]. To this end, we further propose a subspace conditional generative mechanism to utilize the samples projected into the corresponding subspace for view complementing/augmentation. The framework of our proposed model is shown in Figure 2. Our approach contains two generators, $G_1(\cdot)$ and $G_2(\cdot)$, and their corresponding discriminators, $D_1(\cdot)$ and $D_2(\cdot)$, which are trained in inverse direction; meanwhile, two view-specific encoders $E_1(\cdot)$ and

$E_2(\cdot)$ are introduced to encode both views from original feature spaces to the latent subspaces Z_1 and Z_2, respectively. Moreover, in order to make the projected samples more distinctive across views, thus, the available label information associated with the triplet loss function [37] is utilized, where the goal of triplet loss is to make the projected representations closer to the samples of the same action than it is to any other actions. To this end, the objectives of $E_1(\cdot)$ and $E_2(\cdot)$ are introduced below:

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

where M means there are M semi-hard triplets in the given embeddings and labels, $m = \{1, 2\}$ indicates $E_1(\cdot)$ and $E_2(\cdot)$. $X^a_{tr_i}$, $X^p_{tr_i}$, and $X^n_{tr_i}$ represent the i-th training sample as *anchor*, *positive*, and *negative* respectively. α is a margin that is enforced between positive and negative pairs. By this way, the learned subspace could obtain more distinctive and robust feature representations in the corresponding subspace compared with the original feature space. Both $E_1(\cdot)$ and $E_2(\cdot)$ are implemented by a two-layer fully-connected network with the LeakyReLU activation [57] deployed in the first layer.

Then, two generative structures including $G_1(\cdot)$, $D_1(\cdot)$, $G_2(\cdot)$, and $D_2(\cdot)$, are designed for cross-view representation generation goal. Since the two networks are in symmetrical positions and have the same objective equations, thus, we only discuss $G_1(\cdot)$ and $D_1(\cdot)$ in this section. In our model, the first term is the competing approach with $D_1(\cdot)$ and makes the generated samples as real as possible:

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

where z is the noise matrix and $E_1(X^1_{tr})$ is the learned representation as the generation condition of $G_1(\cdot)$. Since the subspaces Z_1 and Z_2 are changed when encoders $E_1(\cdot)$ and $E_2(\cdot)$ are optimized, it is difficult to directly obtain stable generative results. Thus, we include similarity constraint which pulls the generated samples and real samples to be similar in subspace. The objective term is shown as follows:

$$L_{G_1 s} = E_{z\sim p_z(z)}\Big(\|G_1(z|E_1(X^1_{tr})) - E_2(X^2_{tr})\|_{\mathrm{F}}^2\Big). \tag{3}$$

To this end, the overall objective of $G_1(\cdot)$ is represented as $L_{G_1} = L_{G_1 d} + \lambda L_{G_1 s}$, where λ is the trade-off parameter to balance the scales across discriminator loss and similarity loss. $G_1(\cdot)$ is a three-layer neural network with a batch normalization layer [21] to normalize input vector and stabilize the training procedure. The goal of $D_1(\cdot)$ is to differentiate the generated samples and the real samples in subspace Z_2. And the objective function is shown below which manages to maximize L_{D_1}:

$$L_{D_1} = E_{X\sim p_X(X)} \log D_1\big(E_2(X^2_{tr})\big) + E_{z\sim p_z(z)} \log\Big(1 - D_1\big(G_1(z|E_1(X^1_{tr}))\big)\Big). \tag{4}$$

In our implementation, $D_1(\cdot)$ is a three-layer network. The first layer is a fully connected layer with LeakyReLU activation [57]. The second layer is a mini-batch [35] layer, which increases the diversity of the fake samples. The activation functions of both layers are LeakyReLU and the last layer is the Sigmoid function to output the real-fake possibility of the input representations. After the generated representation is obtained in subspace, both the real and fake representations are forwarded to the view-specific classifiers $C_1(\cdot)$ and $C_2(\cdot)$ to obtain the initial label prediction. The objective function of the classifiers include two objectives. The first one is trained to let the classifier predict labels from real samples:

$$L_{C_{m_r}} = \|Y_{tr} - C_m(E_m(X^m_{tr}))\|_{\mathrm{F}}^2, \tag{5}$$

where $m = \{1, 2\}$ indicates the classifiers $C_1(\cdot)$, $C_2(\cdot)$ and the encoders $E_1(\cdot)$, $E_2(\cdot)$. The second one further obtains generated samples associated with the conditional subspace representations to improve the robustness and generalization of the classifier:

$$L_{C_{1_g}} = \|Y_{tr} - C_1(G_2(z|E_2(X^2_{tr}))\|_{\mathrm{F}}^2, \tag{6}$$

$$L_{C_{2_g}} = \|Y_{tr} - C_2(G_1(z|E_1(X^1_{tr}))\|_{\mathrm{F}}^2. \tag{7}$$

To this end, the objective function of $C_m(\cdot)$ is $L_{C_m} = \beta L_{C_{m_r}} + (1-\beta) L_{C_{m_g}}$, where β is the trade-off parameters and we always set $\beta = 0.5$ in our experiments. $C_m(\cdot)$ aims to minimize L_C based on both real and generated features by benefiting from the augmented features.

3.3. View Correlation Discovery Network (VCDN)

Existing multi-view classification methods [58, 29, 30] either learn the score weights of each view or try to fuse the multi-view features in low-level feature space. However, it is hard to well align various views and easy to cause negative influence. While, in multi-view action recognition scenario, we notice that some actions are distinctive in one view (*e.g.*, *Turning Around* in RGB view), and others are distinctive in the other view (*e.g.*, *Answering Phone* in depth view). Thus, simply learning the weights of each view cannot take the full advantage of the view-specific motion characteristics, while exploring the latent relation hidden inside the label [45, 48] is crucial to obtain higher performance.

To this end, we further propose a simple yet effective View Correlation Discovery Network (VCDN), $C_{VCDN}(\cdot)$, to refine the action prediction by exploring the label-level knowledge across views. Instead of naively averaging/weighting the view-specific classification scores,

VCDN explores the initial scores and discovers the latent correlations across different views. To this end, the final prediction is based on both the view-specific prediction and the learned across-view label-correlation knowledge.

The framework of $C_{VCDN}(\cdot)$ is shown in Figure 2. After the initial classification results are achieved by $y_i^1 = C_1(E_1(x_{tr_i}^1))$ and $y_i^2 = C_2(E_2(x_{tr_i}^2))$, where $y_{tr_i}^1 \in \mathbb{R}^{d_l}$ and $y_{tr_i}^2 \in \mathbb{R}^{d_l}$ are the initial predictions of the corresponding i-th sample from the two views, $x_{tr_i}^1$ and $x_{tr_i}^2$. We make a transformation from the two view predictions, $y_{tr_i}^1$ and $y_{tr_i}^2$, to obtain an cross-view label-level adjacency matrix $\mathbf{c}_i$ by multiplying $y_{tr_i}^2$ and the transpose of $y_{tr_i}^1$ as $\mathbf{c}_i = y_{tr_i}^2 \cdot y_{tr_i}^{1\top}$, where $\mathbf{c}_i \in \mathbb{R}^{d_l \times d_l}$ is the adjacency matrix. By this way, the elements in $\mathbf{c}_i$ are the multiplication of the pair-wise predicted scores. Then, the obtained $\mathbf{c}_i$ is reshaped to a d_l^2-dimensional vector and forwarded to $C_{VCDN}(\cdot)$ to predict the final prediction. To this end, $C_{VCDN}(\cdot)$ could reveal the latent correlation between the two views and help the model improve the learning performance. Since both label vectors are achieved from real samples, thus, the objective function can be written as:

$$L_{C_{VCDN}^{rr}} = \sum_{i=1}^{n_{tr}} \|y_i - C_{VCDN}(y_{tr_i}^2 \cdot y_{tr_i}^{1\top})\|_2^2, \quad (8)$$

where $y_i \in \mathbb{R}^{d_l}$ is the ground-truth label vector of i-th sample, and *rr* means *real-real* setting. Moreover, since $G_1(\cdot)$ and $G_2(\cdot)$ also contain effective cross-view structure information, thus, we also want this knowledge to be transferred to $C_{VCDN}(\cdot)$. To this end, we assign the predicted label vector of the fake representations $y_{fi}^1 = C_1(G_2(z|E_2(X_{tr_i}^2)))$, and $y_{fi}^2 = C_2(G_1(z|E_1(X_{tr_i}^1)))$ be utilized in the VCDN training procedure, where $y_{fi}^1 \in \mathbb{R}^{d_l}$, and $y_{fi}^2 \in \mathbb{R}^{d_l}$. We deploy both *real-fake* and *fake-real* combinations to design the objective functions:

$$L_{C_{VCDN}^{rf}} = \sum_{i=1}^{n_{tr}} \|y_i - C_{VCDN}(y_{f_i}^2 \cdot y_{tr_i}^{1\top})\|_2^2, \quad (9)$$

$$L_{C_{VCDN}^{fr}} = \sum_{i=1}^{n_{tr}} \|y_i - C_{VCDN}(y_{tr_i}^2 \cdot y_{f_i}^{1\top})\|_2^2. \quad (10)$$

Then, we obtain the final objective of $C_{VCDN}(\cdot)$:

$$L_{C_{VCDN}} = \gamma L_{C_{VCDN}^{rr}} + \tfrac{1-\gamma}{2}(L_{C_{VCDN}^{rf}} + L_{C_{VCDN}^{fr}}), \quad (11)$$

where γ is a trade-off parameter which balances the weights between real and fake label instances for training the classifiers. $C_{VCDN}(\cdot)$ is a two-layer fully connected network with Leak-ReLU activation in the first layer.

Our model is an end-to-end model and all networks are trained simultaneously. It can also be easily deployed to a wide range of applications. There are two major differences compared with other methods: first, a generative mechanism is utilized to synthesize view information from the other view, which fully explores the latent connection across the views; second, a View Correlation Discovery Network (VCDN) is proposed to fully explore the cross-view label correlations and improve the learning performance. This strategy is effective due to the high correlation of actions across different views.

4. Experiments

4.1. Multi-View Action Datasets

Berkeley Multimodal Human Action Database (MHAD) [32] is a comprehensive multimodal human action dataset. It contains RGB, depth, skeleton, acceleration, and audio views. MHAD contains 11 actions performed by 12 subjects for 5 repetitions of each action, yielding 660 action sequences in total.

UWA3D Multiview Activity (UWA) [34] is a multi-view dataset collected by Kinect sensors. There are 10 subjects performed 30 human activities in a continuous manner without breaks or pauses. The dataset is challenging because of varying viewpoints, self-occlusion and high similarity among activities.

Depth-included Human Action dataset (DHA) [26] is an RGB-D multi-model dataset which contains 23 categories performed by 21 subjects, and there are 483 video clips in total for training and test. Each actions has RGB images, human masks and depth data.

In our experiments, we utilize roughly half of the available samples for training and another half for test. Specifically, there are 254 samples for training and 253 for test in UWA dataset. 244 samples for training and 283 for test in MHAD dataset. 240 samples for training and the rest 243 for test in DHA dataset. In the training procedure, both RGB and depth features are utilized. In the test procedure, there are three settings including single-view (RGB or depth) and multi-view (RGB-D) scenarios.

4.2. Multi-view Recognition Baselines

We test our approach in multi-view (RGB-D) scenarios. In each setting, we also deploy the state-of-the-art methods to demonstrate the effectiveness of our model. Comparison baselines are briefly introduced below. **Least Square Regression (LSR)** is a straightforward linear regression model. The multi-view features are concatenated together and LSR learns a linear mapping between the feature and label spaces. **Support Vector Machine (SVM)** [36] is a classical and robust classifier which constructs one hyperplane or multiple hyperplanes in high-dimensional space to achieve classification, regression, or other tasks. We utilize the implementation from [7] for our baseline. **Action Vector of Local Aggregated Descriptor (VLAD)** [14] is an effective action representation that aggregates local convolutional features and the video spatio-temporal content by an

extension of Net-VLAD layer. It integrates two-stream networks and is trainable in an end-to-end framework. **Temporal Segment Networks (TSN)** [51] proposes a strategy that combines a sparse temporal sampling with video-level supervision. In this way, the entire video was learned effectively while it still achieves accurate and stable performance. **Weighted Depth Motion Maps (WDMM)** [1] aims to recognize human gestures from depth views, which is based on linear aggregation of spatio-temporal information. It proposed a video summarization procedure for hierarchical representation, which results in increasing intra-class similarity and also effectively reduces the inter-class similarities. **Auto-Weight Multiple Graph Learning (AMGL)** [30] is a multi-view classification methods. It learns the optimal weight for each graph automatically without introducing any additive parameters, which is convex and easy to get the global optimal result in a semi-supervised learning scenario. **Multi-view Learning with Adaptive Neighbours (MLAN)** [29] designes an adaptive graph-based method which performs semi-supervised and local structure learning simultaneously. It learns the ideal weight for each view without any parameter tuning. **Partial-modal Generative Adversarial Networks (PM-GANs)** [49] learns a full-modal representation based on partial modalities and implements feature-level fusion for infrared action classification tasks.

4.3. Implementation

We deploy the TSN [51] structure to extract RGB features. Each video is divided into 5 segments. A snippet is randomly chosen from each segment. The ResNet-101 [16] with weights pre-trained on ImageNet produces class scores for each snippet. After the training procedure, we sample 3 snippets from each video instead of 25 which is utilized in TSN since we did not observe significant improvements (less than 0.5%) between these two configurations. We obtain the final features by concatenating the output of the last layer. To this end, each video is represented in a 6144-dimensional feature vector. We utilize WDMM [1] to extract depth feature. WDMM samples each video in three three projection views. After that, HOG and LBP are used to extract the features associated with VLAD and PCA for feature dimension reduction. We follow a similar scheme to WDMM [1] and obtain 110-dimensional feature vectors. As shown in Figure 2, the label vector concatenated with random noise is set as input to $G_1(\cdot)$ and $G_2(\cdot)$. We set the batch size to 64. The Adam optimizer [24] is used for optimization and the learning rates are set to 0.00002, 0.0001, and 0.0002 for $C_m(\cdot)$, $D_{1/2}(\cdot)$, and $G_{1/2}(\cdot)$ respectively. λ limits the feature similarity scales which is set to 0.1. In the training procedure, $D_{1/2}(\cdot)$ and $G_{1/2}(\cdot)$ are pre-trained to obtain stable initialization, while $G_{1/2}(\cdot)$ is optimized by minimizing $L_{G_{1/2}s}$ without including $L_{G_{1/2}d}$ at first, and after 50 epochs, we switch $L_{G_{1/2}}$ back and train $D_{1/2}(\cdot)$ simultaneously with the other networks. The model is implemented using TensorFlow with GPU acceleration.

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**

Table 1. Action recognition performance on UWA dataset [34]

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**

Table 2. Action recognition performance on MHAD dataset [32]

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**

Table 3. Action recognition performance on DHA dataset [26]

Since VLAD and TSN are specifically designed for action recognition in RGB view (single view), thus, we follow the same protocol to pre-process the action data and run the code provided by the authors and report the highest performance. The same strategy is also used to evaluate WDMM in depth view. For general classification algorithms, we utilize the RGB features extracted from TSN, and depth features from WDMM since these methods are new and achieve high performance in RGB and depth representation learning respectively. To evaluate the SVM and LSR performance in multi-view scenario, we concatenate both RGB and depth features after normalization and achieve a single feature vector for classification. Since AMGL and MLAN are designed for multi-view learning, thus, we input RGB and depth features separately and evaluate the performance. PM-GANs utilizes one view to complement another view for classification in the test stage, and we follow the same

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**

Table 4. Recognition performance of our model and the modified fusion strategies in both low-level feature space and high-level label space. It demonstrates the effectiveness of the VCDN framework which considerably improves the performance. (Please note that the performance is lower than our complete model since we removed the generative module for a fair comparison.)

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**

Table 5. Classification performance of our VCDN model compared with the multi-layer neural networks.

setting and evaluation in our experiments.

4.4. Performance Analysis

The experimental results are shown in Table 1, Table 2, and Table 3, where *RGB*, *Depth*, and *R+D* indicate the classification accuracy of single RGB view, single depth view, and RGB-D views respectively. Since our model conditionally generates another view based on the available view, thus we show *R*→*D* and *D*→*R* which indicate these settings (*e.g.*, *R*→*D* means the depth view is conditionally generated by RGB view).To prove the effectiveness of the generated view, we deploy the pseudo feature which is the average feature from the training samples as the "generated" view of and forward to SVM, AMGL, and MLAN baselines. The results are also shown in the same column of the tables.

From the results, we observe that in the single-view scenario, our model achieves the highest performance. In *D*→*R* scenario, our generative strategy gains averagely 3% improvements in all baseline datasets. For other pseudo feature baselines, only parts of the results have slight improvements (*e.g.*, 0.5%) while others are even lower than the single-view scenario. Therefore, the consistent pseudo feature cannot provide any extra distinctive information for improving classification performance, and concatenating the available and generated features directly (with/without normalization) could even hurt the data structure and diminish final recognition performance. These results demonstrate the effectiveness of the generative strategy of our model.

For the multi-view recognition scenario, which means both the RGB and depth views are available, the generative strategy further augments the feature distribution which

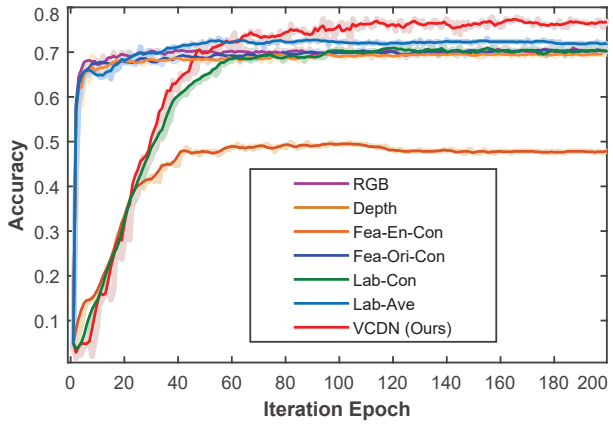

Figure 3. Recognition performance as the training epoch increases in UWA3D dataset [34]. The shadow lines indicate the exact performances per iteration. It shows that our VCDN framework achieves the highest performance after tens of iterations and keeps stable eventually. It demonstrates the robustness and stability of VCDN in this multi-view scenario.

helps both view-specific classifier and the VCDN framework. The results shown in column *R+D* illustrate that our model further improves the accuracy which is considerably higher than any single view scenario.

4.5. Ablation Study

To prove the effectiveness of VCDN, we utilize several feature/label fusion strategies to achieve multi-view classification. In addition, to avoid the influence of the augmented samples from the generative components, we first evaluate our model without including any generated samples. The result is shown in Table 4. The first two lines show the single-view baseline performance from the view-specific classifier $C_1(\cdot)$ and $C_2(\cdot)$; *RGBD-Fea-Ori-Con* indicates the performance when the straightforward feature concatenation approach is processed; *RGBD-Fea-En-Con* indicates the obtained features are concatenated together from $E_1(\cdot)$, $E_2(\cdot)$ and then goes through a network which has the same structure as $C_{VCDN}(\cdot)$; while *RGBD-Lab-Con* denotes the concatenated labels from $C_1(\cdot)$, $C_2(\cdot)$ and also goes through the same structure classifier as $C_{VCDN}(\cdot)$; meanwhile, *RGBD-Lab-Con* shows the performance when the obtained labels from $C_1(\cdot)$ and $C_2(\cdot)$ are averaged; in addition, *RGBD-Lab-Wei* shows the weighted sum of $C_1(\cdot)$ and $C_2(\cdot)$ where the weight is learned simultaneously in the training process; and the last line is our VCDN model. In this experiment, we show the performance of fusing view information in both low-level (*e.g.*, *RGBD-Fea-En-Con* and *RGBD-Fea-Ori-Con*) and high-level (*e.g.*, *RGBD-Lab-Con* and *RGBD-Lab-Con*). To further prove the effectiveness of VCDN, we concatenate the outputs and forward to a deeper network (*i.e.*, 2,3,4-layer structures). The results (Table 5) show 2-layer structure tends to be enough. However, it still works worse than our VCDN. The result indicates that mul-

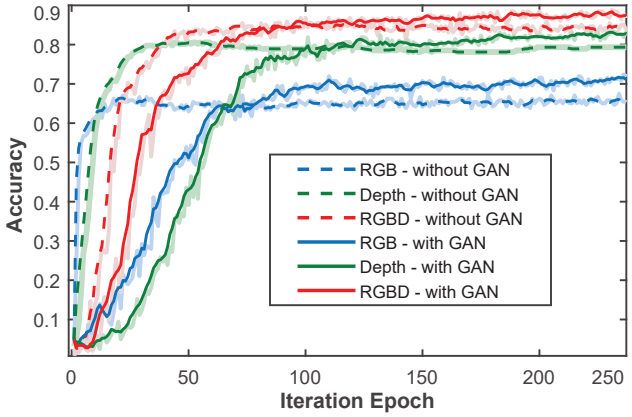

Figure 4. Performance of our GMVAR approach with (solid lines) and without (dashed lines) the generative strategy in DHA dataset. Different colors indicate different settings. The shadow lines indicate the exact performances per iteration. It demonstrates that the generative model does learn the cross-view connection knowledge and further improves the recognition performance.

tiple views knowledge does provide extra distinctive features for action recognition; while high-level fusion performs better than low-level fusion due to the significant difference across views, and our VCDN achieves the best performance since it fully explores the label correlations.

Following the previous experimental setting, we further visualize the recognition performance as the training epoch increases, and the result is shown in Figure 3, where we observe that most fusion strategies cannot outperform the highest single-view classification performance. We assume that simply feature level fusion cannot provide clear distinctive clue for classifier, and it is too difficult to capture the correlation by itself; while label average approach achieves slight improvement which indicates the high-level fusion performs well in multi-view action scenario; meanwhile, our approach achieves the highest performance and keeps stable after around 100 epoch which further demonstrates the effectiveness of the VCDN model.

We evaluate our GMVAR with and without the generative strategy to prove its effectiveness in our model. Figure 4 shows the recognition accuracy of GMVAR with and without the generative model in single-view (RGB and depth) and multi-view (RGB-D) settings on DHA dataset. From the results, we observe that the generative strategy indeed improves the performance of all settings considerably. Moreover, we changed the GAN module to a mapping module for further comparison. In this case, one modality is a mapping of the other, and the obtained performance (*i.e.*, UWA: 74.52%, MHAD: 98.23%, DHA: 88.07%) is lower than the model with the generative model. We assume GAN captures better feature distribution and diversifies the training space to achieve higher performance.

Furthermore, we visualize the distribution of the real and generated representations of the test samples in Z_1 and Z_2

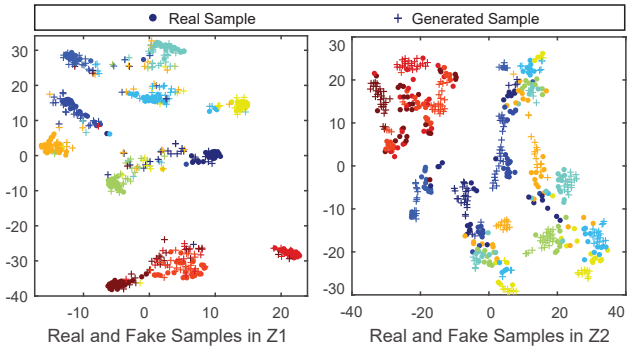

Figure 5. t-SNE [27] visualization results of the real and the generated test sample representations in Z_1 and Z_2 respectively. The solid circles and the cross marks indicate the real and generated representations, and different colors denote different action categories. We observe that real and generated representations which belong to the same category are close to each other. It illustrates that the generative model is capable to "recover" one view conditioned on the other view. And it further demonstrates the effectiveness of the generative strategy in this multi-view scenario.

by t-SNE [27] method respectively. The results are in Figure 5 which illustrate that the real and generated representations which belong to the same action category are close to each other and vice-versa. It indicates that this generative approach effectively learns the across-view correlations in the subspace which can accurately generate similar representations to complement/augment the other view. And the view-specific classifiers associated with the proposed VCDN further utilizes the knowledge to improve the action recognition performance.

5. Conclusion

We proposed a novel Generative Multi-View Action Recognition (GMVAR) framework in this paper. A generative mechanism is designed to generate one view conditioned on the other view. By this way, the comprehensive cross-view motion structure knowledge can be revealed. Due to this generative strategy, our model works well in single-view and missing-view scenarios which are difficult for other multi-view approaches. Moreover, we proposed an effective View Correlation Discovery Network (VCDN) which further explores the cross-view correlation in high-level label space and obtains more accurate classification results. Evaluation of three multi-view action datasets and extensive ablation studies show the effectiveness of both generative model and VCDN framework. All experimental results illustrate that our GMVAR is an effective, accurate, robust framework, and compatible with a wide range of multi-view action recognition tasks.

Acknowledgments: This research is supported in part by the NSF IIS award 1651902 and U.S. Army Research Office Award W911NF-17-1-0367.

References

[1] Reza Azad, Maryam Asadi-Aghbolaghi, Shohreh Kasaei, and Sergio Escalera. Dynamic 3d hand gesture recognition by learning weighted depth motion maps. *IEEE TCSVT*, 2018.

[2] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proc. ECCV*, 2018.

[3] Nan Bu, Masaru Okamoto, and Toshio Tsuji. A hybrid motion classification approach for emg-based human–robot interfaces using bayesian and neural networks. *IEEE Trans. on Robotics*, 25(3):502–511, 2009.

[4] Zhuowei Cai, Limin Wang, Xiaojiang Peng, and Yu Qiao. Multi-view super vector for action recognition. In *Proc. IEEE CVPR*, pages 596–603, 2014.

[5] Alexandros Chaaraoui, Jose Padilla-Lopez, and Francisco Flórez-Revuelta. Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. In *Proc. IEEE CVPR workshop*, pages 91–97, 2013.

[6] Alexandros Andre Chaaraoui, José Ramón Padilla-López, Pau Climent-Pérez, and Francisco Flórez-Revuelta. Evolutionary joint selection to improve human action recognition with rgb-d devices. *Expert systems with applications*, 41(3):786–794, 2014.

[7] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM TIST*, 2(3):27, 2011.

[8] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *arXiv:1612.02136*, 2016.

[9] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proc. IEEE ICIP*, pages 168–172, 2015.

[10] 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.

[11] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, and others. Guide to the carnegie mellon university multimodal activity (CMU-MMAC) database. *Robotics Institute*, page 135, 2008.

[12] Zhengming Ding, Ming Shao, and Yun Fu. Generative zero-shot learning via low-rank embedded semantic dictionary. *IEEE TPAMI*, 2018.

[13] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proc. IEEE CVPR*, pages 409–419, 2018.

[14] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proc. IEEE CVPR*, volume 2, page 3, 2017.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NIPS*, pages 2672–2680. 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE CVPR*, pages 770–778, 2016.

[17] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proc. IEEE CVPR*, pages 826–834, 2016.

[18] Michael B Holte, Thomas B Moeslund, Nikos Nikolaidis, and Ioannis Pitas. 3D human action recognition for multi-view camera systems. In *Proc. Inte. Conf. on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 342–349, 2011.

[19] Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clément Ménier. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine vision and applications*, 27(7):1005–1020, 2016.

[20] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proc. IEEE CVPR*, pages 5344–5352, 2015.

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015.

[22] Xiaofei Ji, Ce Wang, and Yibo Li. A view-invariant action recognition based on multi-view space hidden markov models. *International Journal of Humanoid Robotics*, 11(01):1450011, 2014.

[23] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In *Proc. IEEE CVPR workshop*, pages 1–10, 2017.

[24] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[25] Jinna Lei, Xiaofeng Ren, and Dieter Fox. Fine-grained kitchen activity recognition using rgb-d. In *Proc. ACM Conference on Ubiquitous Computing*, pages 208–211, 2012.

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

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008.

[28] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

[29] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *Proc. AAAI*, 2017.

[30] Feiping Nie, Jing Li, Xuelong Li, et al. Parameter-free auto-weighted multiple graph learning: A framework for multi-view clustering and semi-supervised classification. In *Proc. IJCAI*, pages 1881–1887, 2016.

[31] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv:1610.09585*, 2016.

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

[33] Diana Pagliari and Livio Pinto. Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors. *Sensors*, 15:27569–27589, 10 2015.

[34] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE Trans. PAMI*, 38(12):2430–2443, 2016.

[35] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proc. NIPS*, pages 2234–2242, 2016.

[36] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE CVPR*, pages 815–823, 2015.

[38] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep multimodal feature analysis for action recognition in rgb+ d videos. *IEEE Trans. PAMI*, 40(5):1045–1058, 2018.

[39] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *Proc. IEEE CVPR*, pages 20–28, 2017.

[40] Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. From ensemble clustering to multi-view clustering. In *Proc. IJCAI*, pages 2843–2849, 2017.

[41] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proc. IEEE CVPR*, pages 1405–1414, 2017.

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

[43] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013.

[44] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proc. IEEE CVPR*, 2017.

[45] Lichen Wang, Zhengming Ding, and Yun Fu. Adaptive graph guided embedding for multi-label annotation. In *Proc. IJCAI*, pages 2798–2804, 2018.

[46] Lichen Wang, Zhengming Ding, and Yun Fu. Learning transferable subspace for human motion segmentation. In *Proc. AAAI*, pages 4198–4202, 2018.

[47] Lichen Wang, Zhengming Ding, and Yun Fu. Low-rank transfer human motion segmentation. *IEEE Trans. on Image Processing*, 28(2):1023–1034, 2019.

[48] Lichen Wang, Zhengming Ding, Seungju Han, Jae-Joon Han, Changkyu Choi, and Yun Fu. Generative correlation discovery network for multi-label learning. In *Proc. ICDM*, 2019.

[49] 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.

[50] 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.

[51] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. ECCV*, pages 20–36, 2016.

[52] Lichen Wang, Aimin Zhang, Chujia Guo, Bhan Pervez, and Tian Yan. Modified multi-target recognition based on camcom. In *Proc. IEEE Chinese Control Conference*, pages 5367–5373, 2015.

[53] Lichen Wang, Aimin Zhang, Chujia Guo, Songyun Zhao, and Pervez Bhan. 3-d reconstruction for smt solder joint based on joint shadow. In *Proc. IEEE Chinese Control and Decision Conference*, pages 5766–5772, 2015.

[54] Pichao Wang, Wanqing Li, Zhimin Gao, Jing Zhang, Chang Tang, and Philip O Ogunbona. Action recognition from depth maps using deep convolutional neural networks. *IEEE THMS*, 46(4):498–509, 2016.

[55] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Partial multi-view clustering via consistent GAN. In *Proc. ICDM*, pages 1290–1295, 2018.

[56] Zhou Wang et al. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. IP*, 2004.

[57] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv:1505.00853*, 2015.

[58] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[59] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proc. AAAI*, 2018.

[60] Mengyang Yu, Li Liu, and Ling Shao. Structure-preserving binary representations for rgb-d action recognition. *IEEE Trans. PAMI*, 38(8):1651–1664, 2016.

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

[62] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *Proc. AAAI*, 2017.

[63] Jingjing Zheng, Zhuolin Jiang, and Rama Chellappa. Cross-view action recognition via transferable dictionary learning. *IEEE Trans. Image Processing*, 25(6):2542–2556, 2016.

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE CVPR*, pages 2223–2232, 2017.