Pattern Recognition 118 (2021) 108043

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Arbitrary-view human action recognition via novel-view action generation☆

Kumie Gedamu, Yanli Ji*, Yang Yang, LingLing Gao, Heng Tao Shen

Center for Future Media, School of Computer Science and Engineering, UESTC, Chengdu, China

ARTICLE INFO

Article history:
Received 2 June 2020
Revised 25 January 2021
Accepted 11 May 2021
Available online 20 May 2021

Keywords:
Arbitrary-view action recognition
Novel-view action generation
View domain generalization

ABSTRACT

Arbitrary-view human action recognition is still a big challenge due to the view changes. A possible solution is to enlarge the view range of action samples in the training set. Therefore, we propose a Two-Branch Novel-View action Generation approach based on auxiliary conditional GAN, which generates a novel-view action sample for arbitrary-view human action recognition. The generated sample enlarge the view range of action samples for training. Furthermore, to narrow the representation of actions in different views, we propose a view-domain generalization model that improves the recognition performance of arbitrary-view human action recognition. Our approach is evaluated on three large-scale RGB+D skeleton datasets including UESTC varying-view RGB+D dataset, NTU RGB+D 60, and NTU RGB+D 120 datasets, with two types of view-invariant evaluations, i.e., the cross-view, and arbitrary-view recognition. The proposed approach achieves outstanding performance in human action recognition.

1. Introduction

Human action recognition has attracted great attention in recent years due to its wide range of applications such as video understanding and surveillance, human-robot interaction, healthcare, sport analysis [1], and many more. Most of these real-world applications are required to overcome the view gap for correct recognition of human action in arbitrary-views, e.g., surveillance with multiple cameras, and human-robot interactions. Arbitrary-view human action recognition became a challenging problem due to complicated volatile human body poses limited viewpoints and the non-rigid human body as seen in Fig. 1. Besides, since 3D skeleton data is not real 3D, the geometric transformation wouldn't help the model to learn robust features against viewpoint changes. Arbitrary-view human action recognition researches [2–8] make an attempt to overcome view variation problems for robust action recognition. These works require either 1) training data under many views or 2) designing view-invariant feature representations using handcrafted feature extraction strategies. However, they didn't consider the view variant feature representation and failed to learn discriminative view-invariant features.

In recent years, several multi-view action datasets were collected to evaluate multi-view human action recognition [9–11] using visual information approaches. Nevertheless, their action representation approaches heavily depends on the viewpoints of cameras. One possible solution is to collect action datasets, which contain viewpoint of action samples as large as possible. However, there is no dataset covering large enough views for efficient arbitrary-view human action recognition due to various difficulties in data collection. To solve the problem of view changes in multi-view human action recognition, [12–14] have introduced domain adaptation and transfer learning. These approaches reduced the action representation difference in different views. However, these methods suffer difficulties in novel views for arbitrary-view human action recognition. One feasible solution is to generate synthetic action samples in novel views, which is an effective and efficient solution than capturing samples. Therefore, we present a two-branch action sample generation approach for arbitrary-view human action recognition, that enables to enlarge the view range of action samples in the training set. In the proposed approach, we generate novel view action samples and use them as a training set to solve the aforementioned problems in arbitrary-view human action recognition.

Several variants of GAN such as CGAN [15], InfoGAN [16], AC-GAN [17], VAE-GAN [18], Least Square GAN [19], Wasserstein GAN [20], CVAE-GAN [21], and CR-GAN [22] have shown promising results in generating samples. Even though GANs have a common goal of preserving identity, they usually have a single-pathway i.e. an encoder-generator network followed by a discriminator network. The encoder maps input images into a latent space, where the embedding is first manipulated and then fed into the generator

☆ This research is supported by the National Key Research and Development Program of China (No. 2018AAA0102200).

* Corresponding author.
E-mail address: yanliji@uestc.edu.cn (Y. Ji).

https://doi.org/10.1016/j.patcog.2021.108043
0031-3203/

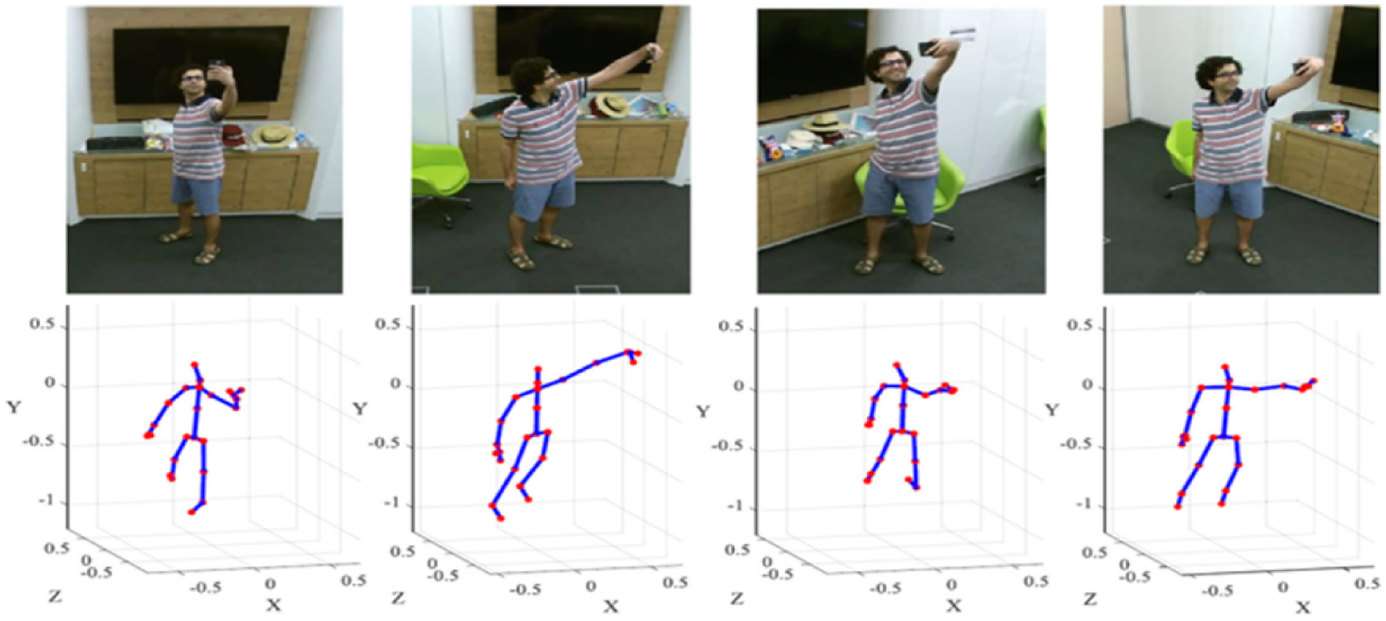

Fig. 1. The challenging problem of view changes in arbitrary-view human action recognition. The same action performed by the same person may be visually different from one view to another view.

to generate synthetic samples. These approaches can learn only incomplete representations which lead to limited generalization ability on unseen or unconstrained samples [22]. The proposed two-branch generation model, therefore, can address the above issues. The main idea in the two-branch generation model is to create view-specific samples from embeddings that are randomly sampled from noise in the generation branch in addition to the reconstruction branch. Beyond this, better control of synthetic samples and stable convergence is also required. With this CGAN [15] and WGAN-GP [20] are known for their ability to control the attribute of synthetic samples and stable convergence, respectively.

Therefore, our two-branch generation model for novel-view action sample generation is designed by taking the aforementioned advantages of CGAN and WGAN-GP. Furthermore, to weaken the representation difference of actions in different views, we propose a view domain generalization model. The architecture of the proposed approach is shown in Fig. 2. Given action samples in view v_i, the generation model is used to generate action samples in novel views v_j, $v_i \neq v_j$. The real and synthetic samples are combined to construct a new training set for arbitrary-view human action recognition. The major contributions of our work are summarized as follows:

- We propose a two-branch novel-view action generation approach for arbitrary-view action recognition.
- The two-branch generation model generates novel-view action samples, which enlarges the view range of action samples for classifier training.
- A view-domain generalization module is designed to weaken the difference of action representation in various views for the arbitrary-view action recognition.
- Extensive experiments and ablation studies are performed on three large-scale benchmarks, UESTC, NTU-60, and NTU-120 datasets.

2. Related work

2.1. Deep generative models

Goodfellow et al. [23] introduced GAN to estimate the target distribution via adversarial fashion. Conventional GAN had no control over the mode of generated samples. Therefore, CGAN [15] was proposed to address this problem. CGAN could generate multiple synthetic outputs by setting predefined controlling vectors, which had a significant advantage for action generation in multiple views. AC-GAN [17] extended the discriminator by containing an auxiliary decoder network to estimate class labels for the training samples. VAE-GAN [18] generated good results when input got mapped into the learning spaces. However, unseen or unconstrained samples might be mapped out of the learning spaces, leading to poor generation results. CVAE-GAN [21] presented a conditional variational generative adversarial network, which synthesized samples in fine-grained categories by varying the fine-grained conditional class label for controlling the generation. CR-GAN [22] utilized a self-supervised complete representation learning to improve synthesized samples. Although, it failed to generate photo-realistic results. The above multi-view image generation approaches cannot deal with unseen or unconstrained samples leading to generate either poor results or easily collapse.

In contrast, our proposed approach can learn complete representations using a two-branch network, which promises the generation of high-quality visual samples even for unconstrained input. Major barriers to recognize actions in arbitrary-view are lack of action samples in various views. Considering the difficulty to collect a large number of action samples per arbitrary-view, we design a two-branch generation model for novel-view action generation. Taking the advantage of AC-GAN and VAE-GAN fine-grained conditions, we generate action samples in novel views and combine them with the original training samples to enlarge the view range of the training set.

2.2. Arbitrary-view action recognition

Arbitrary-view human action recognition is a challenging task, besides occlusions caused by view changes [3,8,24], there exist few datasets for researchers to develop investigation on it. Most of the studies simulated arbitrary-view action recognition in fixed multi-view datasets [2,9,25]. Recently, many approaches had been proposed to address the problem of arbitrary-view human action recognition [5,6,12,26]. These approaches delivered promising results based on training with sufficient labeled datasets. However, due to various difficulties of collecting training data under many

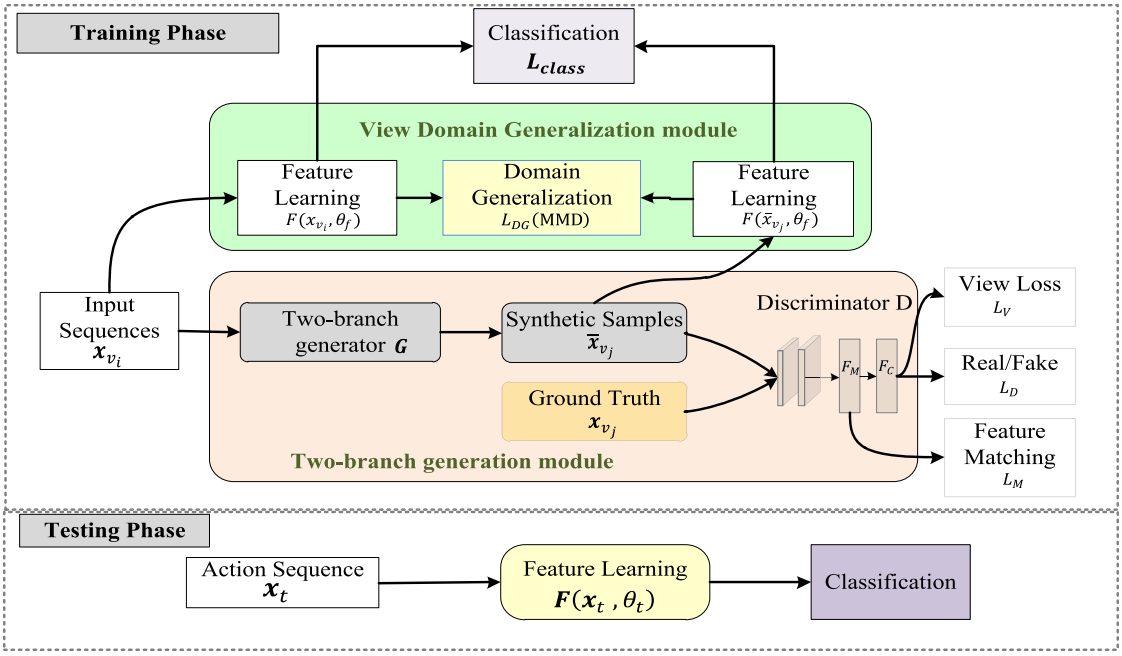

Fig. 2. The architecture of our proposed approach. It consists of two major modules, novel-view action sample generation via a two-branch generation module and the view-domain generalization. Here, x_{v_i} denotes training action sample in view v_i, $\bar{x}_{v_j}$ is output from the generator, which is a synthetic sample in view v_j. Here, x_{v_j} denotes the ground truth. F_M denotes features on an intermediate layer of the discriminator. For the generator training, three loss functions are designed, discriminator loss L_D, view classification loss L_V, and feature matching loss L_M. The loss L_{DG} is designed for the training of domain generalization module.

views, the proposed models did not achieve satisfactory results for arbitrary-view action recognition. Some researchers tried to create action samples in novel views and used these created samples to train a classifier. Ji et al. [27] presented the only varying-view RGB+D action dataset with full-circle views (360°) for arbitrary-view human action recognition. An attention transfer network [2] was presented to transfer learned attention from the front view to the arbitrary-views. On another side, paper [13] presented an approach that generated a dictionary of synthetic 3D data and pre-trained a sparse representation model to extract view-invariant features for action recognition.

In past decades, multi-view recognition has attracted heavy attention, and they explored possible solutions for arbitrary-view action recognition. One possible solution is to collect action data of numerous views to construct a dictionary [13,26], but it is an expensive and time-consuming task constructing a dictionary for each domain. It required a large number of samples and action categories in various views. Modeling geometric relations of human body parts/joints was another solution to weaken the effect of view changes [28,29], but it is difficult to solve the problem thoroughly. In addition, domain adaption [5] and transfer learning [30] achieved positive performance for multi-view action recognition, which provided a practical solution to overcome the view changes in arbitrary-view action recognition. Due to a lack of data, they were all evaluated for multi-view action recognition. To break the limitation of existing datasets, we propose an action generation approach for the arbitrary-view action recognition in this paper.

3. The proposed approach

As shown in Fig. 2, our approach consists of two major operations, novel-view action sample generation via a two-branch generation module and the view-domain generalization. The former one enriches source views of action samples, providing samples of sufficient views for model training. The latter operation weakens the difference of action representation in different views, which reduces confusion caused by the view change.

3.1. Two-branch generator

The detailed architecture of the two-branch generation module is shown in Fig. 3. Different from previous methods [16,18,21] which usually employed the synthesis function with a single encoder-generator network, we design a two-branch generator for novel-view action sample generation. During training, as shown in Fig. 3, one branch involves the generator G_1 and the discriminator, called generation branch, while the other branch is composed of an encoder E, the generator G_2 and the discriminator, named as reconstruction branch. In the two-branch generation module, two generators (G_1 and G_2) share the same network structure and parameters. They share a common input vector V which is used as a view guide for creating synthetic samples. In the upper branch, G_1 takes a random noise z_0 and a view label V as input to generate samples $\bar{\boldsymbol{x}}_{v_j}$ of view v_j. Next, we employ an encoder E to encode skeleton samples $\boldsymbol{x}_{v_i}$ and to output feature $z_e = E(\boldsymbol{x}_{v_i})$ which is fed into G_2 to reconstruct $\bar{\boldsymbol{x}}_{v_j}$ of view v_j. To improve the generation ability to generate unseen and unconstrained action samples, the two-branch generators work together and compete in a parameter-sharing manner (the learned representations in the generation branch will guide the reconstruction branch and vice versa).

To maximize the difference between training and synthetic instances, we use gradient penalty as in [31]. To circumvent tractability issues, we use soft version of the constraint gradient norm for random samples $\bar{\boldsymbol{x}} \sim p_{\bar{\boldsymbol{x}}}$ with loss $(||\nabla_{\bar{\boldsymbol{x}}} D(\bar{\boldsymbol{x}})||_2 - 1)^2$, as listed in Eq. (2). Here, $\bar{\boldsymbol{x}}$ refers to generated samples, and $D()$ represents the discriminator. The discriminator is composed of convolution layers of a ResNet and two FC layers. Features output from one middle FC layer (represented by $F_M()$) are used for feature matching, and output features of the final FC layer are input to loss functions for network training. We set a loss function L_V for view classification

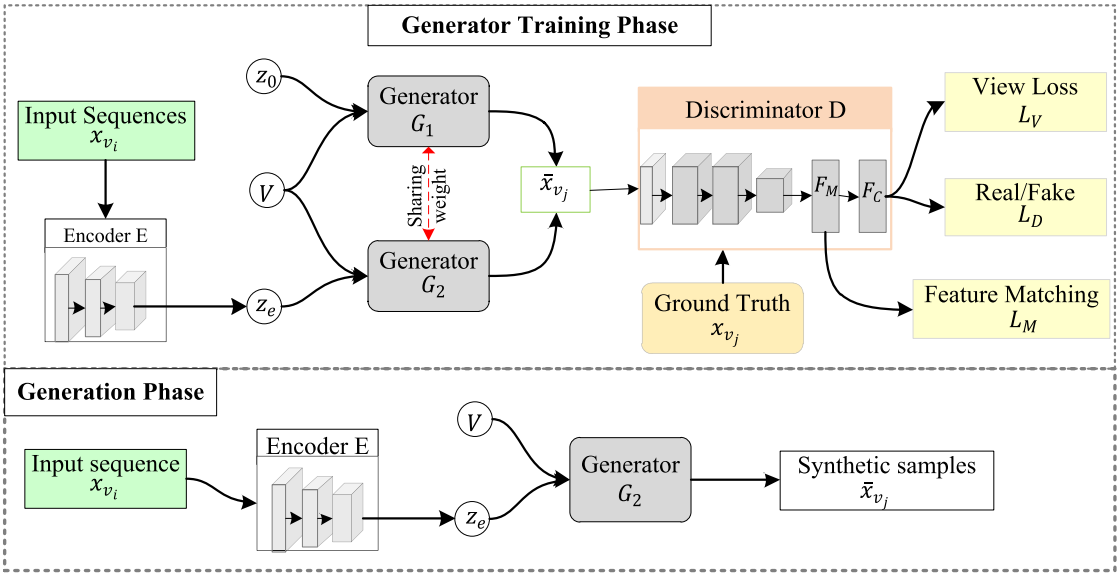

Fig. 3. The flowchart of our proposed two-branch generation module. Here, z_0 is a random noise variable, and V refers to a view vector that controls views of synthetic samples. z_e is encoded action feature, which is used as the latent variable in the reconstruction branch.

that trains the discriminator to distinguish the view of training and synthetic samples. The cross-entropy function is used, as illustrated in Eq. (1).

$$\begin{aligned}\hat{y}_v &= \frac{\exp(wD(\boldsymbol{x}_v, \theta_f))}{\sum_k \exp(wD(\boldsymbol{x}_k, \theta_f))}\\ \hat{\mathbf{y}}_V &= \{\hat{y}_1, \hat{y}_2 \ldots ., \hat{y}_V\}\\ L_V &= -\mathbf{y}_V log(\hat{\mathbf{y}}_V)\end{aligned} \tag{1}$$

Where, $\boldsymbol{x}_v$ and $\boldsymbol{x}_k$ can be either generated or ground truth samples, $\mathbf{y}_V$ refers to the ground truth of view categories, and $\hat{\mathbf{y}}_V$ represents predicted result of views.

In the upper generation branch of Fig. 3, G_1 and D compose an AC-GAN, so we follow the training rule of AC-GAN to train the generator. Given a view label V and a random noise z_0, G_1 is trained to generate a visual samples $\bar{\boldsymbol{x}}_{v_j}$. D is trained to distinguish real samples and synthetic samples generated by G_1, which minimizes,

$$\begin{aligned}L_D = & E_{z\sim p_z}[D(\bar{\boldsymbol{x}}_{v_j})] - E_{\boldsymbol{x}\sim p_x}[D(\boldsymbol{x}_{v_j})]\\ & + \lambda_1 E_{\bar{\boldsymbol{x}}\sim p_{\bar{x}}}[(||\nabla_{\bar{x}} D(\bar{\boldsymbol{x}})||_2 - 1)^2] - \lambda_2 L_V\end{aligned} \tag{2}$$

Where, $p_{\bar{x}}$ refers to distribution of synthetic samples, and p_x is the true sample distribution, p_z is the noise uniform distribution. Then G_1 tries to fool D, which maximizes,

$$L_{G_1} = E_{z\sim p_z}[D(\bar{\boldsymbol{x}}_{v_j})] + \lambda_2 L_V \tag{3}$$

In the reconstruction branch, we propose a cross-reconstruction task to reconstruct action samples in various views. Giving network parameters of G_1 to G_2, we train the encoder E and discriminator D. For the training, action samples, $(\boldsymbol{x}_{v_i}, \boldsymbol{x}_{v_j})$, which belong to the same action category but in different views, $(v_i \neq v_j)$, are adopted. To reconstruct $\boldsymbol{x}_{v_j}$ from the input $\boldsymbol{x}_{v_i}$, encoder E takes $\boldsymbol{x}_{v_i}$ as input and outputs an identity-preserved representation $z_e = E(\boldsymbol{x}_{v_i})$. After obtaining a mapping from $E(\boldsymbol{x}_{v_i})$ to z_e, we obtain the generated sample $\bar{\boldsymbol{x}}_{v_j}$ from G_2. For the network training, a $L2$-norm loss and a pair-wise feature matching loss are calculated between $\boldsymbol{x}_{v_j}$ and $\bar{\boldsymbol{x}}_{v_j}$, which compose the reconstruction loss function, as illustrated in Eq. (4). The reconstruction loss L_M drives the generator to generate synthesized samples $\bar{\boldsymbol{x}}_{v_j}$ which approach to the center of ground truth samples $\boldsymbol{x}_{v_j}$.

$$L_M = ||\boldsymbol{x}_{v_j} - \bar{\boldsymbol{x}}_{v_j}||_2^2 + \frac{1}{2}||F_M(\boldsymbol{x}_{v_j}) - F_M(\bar{\boldsymbol{x}}_{v_j})||_2^2 \tag{4}$$

As described in the upper generation branch, G_1 is an auxiliary generator whose weights are utilized by generator G_2, and G_2 learns to reconstruct from the latent representation of encoder E. In the proposed generation model, two generators (G_2 and G_2) share the same network parameter setting. The parameter sharing helps the model to accelerate the convergence rate during training, and generates effective synthetic samples in novel views. The encoder E helps G_2 to generate high quality samples by maximizing,

$$E_{\boldsymbol{x}_{v_i}, \boldsymbol{x}_{v_j}\sim p_x}[D(\bar{\boldsymbol{x}}_{v_j}) + \lambda_2 L_V - \lambda_3 L_M] \tag{5}$$

3.2. View domain generalization

Through novel-view action generation, we extend the view range of action samples in the training set, which provides sufficient view samples in the arbitrary-view action recognition. For action recognition, action samples that belong to the same category are expected to have a common feature distribution. Therefore, we propose a view-domain generalization model that weakens the difference of action representation in different views to improve the recognition performance.

In view domain generalization (Fig. 2) module, at the training phase, we combine both real and synthetic action samples in multiple views to construct a new training set. For intuition, we use real samples in view v_i, $\boldsymbol{x}_{v_i}$, and synthetic samples in view v_j, $\bar{\boldsymbol{x}}_{v_j}$ as inputs of the domain generalization module to learn a generalized feature representation from different views. We employ the Maximum Mean Discrepancy (MMD) [32] to set a loss function L_{DG} to enforce the view generalization of action representation. The loss L_{DG} is defined in Eq. (6)).

$$\begin{aligned}L_{DG}(\boldsymbol{x}_{v_i}, \bar{\boldsymbol{x}}_{v_j}) &= MMD(F(\boldsymbol{x}_{v_i}), F(\bar{\boldsymbol{x}}_{v_j}))\\ &= E_{\boldsymbol{x},\boldsymbol{x}'}[K(\boldsymbol{x}, \boldsymbol{x}')] + E_{\bar{\boldsymbol{x}},\bar{\boldsymbol{x}}'}[K(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}')] - 2E_{\boldsymbol{x},\bar{\boldsymbol{x}}}[K(\boldsymbol{x}, \bar{\boldsymbol{x}})]\end{aligned} \tag{6}$$

Where $\boldsymbol{x}$ and $\boldsymbol{x}'$ represent two random variables subject to real train sample $\boldsymbol{x}_{v_i}$, and $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{x}}'$ represent random variables subject to generated synthetic samples $\bar{\boldsymbol{x}}_{v_j}$. K is the corresponding kernel function.

3.3. Arbitrary-view action recognition

For arbitrary-view action recognition, both synthetic and original training samples compose a new training set to train the domain generalization model and an action classifier. We build the domain generalization model using the ResNet as a backbone for feature learning. The classifier consists of three fully connected layers and one softmax layer. Let x_t denotes a test sample, we can obtain a representation feature $F(x_t, \theta_f)$ by the trained domain generalization model. The cross-entropy function is adopted for the classifier training, as listed in Eq. (7).

$$\begin{aligned} \hat{y}_c &= \frac{\exp(wF(x_t,\theta_f))}{\sum_j \exp(wF(x_j,\theta_f))} \\ \hat{\mathbf{y}} &= \{\hat{y}_1, \hat{y}_2 \ldots\ldots\ldots, \hat{y}_c\} \\ L_{class} &= -\mathbf{y} log(\hat{\mathbf{y}}). \end{aligned} \tag{7}$$

where, $\mathbf{y}$ refers to the ground truth of action categories, and $\hat{\mathbf{y}}$ represents predicted results of action categories.

3.4. Training and testing

We summarize the novel-view sample generation using our proposed two-branch generation approach in Algorithm 1 . We first

Algorithm 1: The pipeline of the novel-view generation.

Input: $\boldsymbol{x}_{v_i}$ are action samples in view v_i, $\boldsymbol{x}_{v_j}$ are real action samples in view v_j, N refers to maximum number of epoch. B refers to the batch size.

Output: Trained encoder E, generator G_1/G_2, and discriminator D.

for $n \leftarrow 1$ **to** N **do**

  for $b \leftarrow 1$ **to** B **do**

    1. Given samples $\boldsymbol{x}_{v_j}$ in view v_j and a random noise z_0,
    2. Using samples $\boldsymbol{x}_{v_j}$ as ground truth, train G_1 with loss functions L_{G_1}, L_V and L_D,
    3. $\bar{\boldsymbol{x}}_{v_j} \leftarrow G_1(V, z_0)$;
    4. Given samples $\boldsymbol{x}_{v_i}$ in view v_i,
    5. $z_e \leftarrow E(\boldsymbol{x}_{v_i})$;
    6. Train G_2 with loss functions L_M, L_V and L_D;
    7. $\bar{\boldsymbol{x}}_{v_j} \leftarrow G_2(V, z_e)$;
    8. Update discriminator D and encoder E;

  end

end

train the generator to generate samples in various views, then we use synthetic samples and original samples to train the view domain generalization module and the action classifier. The obtained view-invariant feature extractor $F()$ and the trained classifier is adopted for feature learning and arbitrary-view action recognition.

Novel-view sample generation: With trained generator, new action samples are generated using the encoder E and generator G_2, the generator G_2 loads a train weight with z_e and view vector V as input to produce a synthetic sample $\bar{x}_{v_j}$, as shown in Fig. 3. The synthetic action sample is further used for view domain generalization model training and arbitrary-view human action recognition.

Training for view domain generalization: Using the residual network as a backbone, we train the view domain generalization model to weaken the feature difference of actions in different views using the MMD loss function that is defined in Eq. (6). During training, we make sure training and synthetic action samples belong to the same action category but in different views.

Testing phase for action classification: The diagram of the testing phase for arbitrary-view human action recognition is shown in Fig. 2. In the testing phase, we obtain feature representation $F(x_t, \theta_f)$ and put it into the trained classifier for arbitrary-view human action recognition.

4. Datasets and evaluation settings

4.1. Datasets

We evaluate our approach in three large-scale datasets. Skeleton data is used for evaluations in our experiments.

UESTC varying-view RGB+D dataset (UESTC) [27] is captured for view-invariant human action recognition. The action dataset is collected using Microsoft Kinect v2 sensors, and it contains RGB videos, depth images, and skeleton sequences. The dataset contains sample actions captured in 8 fixed views and varying-view sequences covering the entire 360° view angles. There are 40 action categories, acted by 118 subjects. This dataset is the only action dataset containing full-circle view (360°) for arbitrary view human action recognition.

NTU RGB+D 60 dataset (NTU-60) [33] is a kinetic acquired dataset which consists of 56, 880 videos. This dataset includes 60 actions performed by 40 subjects. Each action is captured by 3 cameras at the same height but from different horizontal angles (-45^0, 0^0 and 45^0). The training set in this benchmark contains 37,920 videos that are captured by cameras 2 and 3, and the testing set contains 18,960 videos that are captured by camera 1.

NTU RGB+D 120 dataset (NTU-120) [34] extends the NTU-60 by adding 60 additional action classes to the existing one. It is the largest RGB+D dataset for 3D action recognition, containing 114,480 skeleton sequences performed by 106 distinct human subjects. It is a challenging dataset for human action recognition.

4.2. Evaluation settings

Cross-view evaluation (X-view) evaluates the performance of action recognition in cross views. Following [27] and [34], we separate action samples to the training and test set according to views. In the UESTC dataset, four views of 8 fixed views which connect a square crossing shape are grouped together, grouping into two view sets, $S_{tr} = (v_0, v_2, v_4, v_6)$ as a training set and $S_{te} = (v_1, v_3, v_5, v_7)$ as a test set. We further separate S_{te} into two subsets averagely, the generation set S_{tv} and testing set S_{tt}. We train the generator G_1, G_2 using S_{tr} and S_{tv} two sets. Cross-view action classification test is performed in S_{tt}.

In the NTU-60 dataset, similar to [34], action samples in views of camera 2 and 3 are separated into training set S_{trn}, and samples captured by camera 1 are defined as the test set S_{ten}. S_{ten} is further separated into a generation sample set S_{tvn} and a testing set S_{ttn}. We train the generation module using the generation set S_{tvn} and the training set S_{trn}. Test is performed in the set S_{ttn}. The same evaluation setting is performed for the NTU-120 dataset.

Arbitrary-view evaluation (A-view) is performed on varying-view action sequences in the UESTC dataset. We use action samples of 8 fixed views $(v_0, \ldots, v_7)$ for model training, and testing is performed on full-circle view sequences, which is similar to A-view I in [27]. In this experiment, we generate action samples in 8 neighborhood views of 8 fixed views $(v_0, \ldots, v_7)$, as shown in Fig. 4. After we obtain generation samples, real samples of 8 fixed views and synthetic action samples are all used to train the view domain generalization module and the action classifier. The remaining varying view action sequences are used for testing.

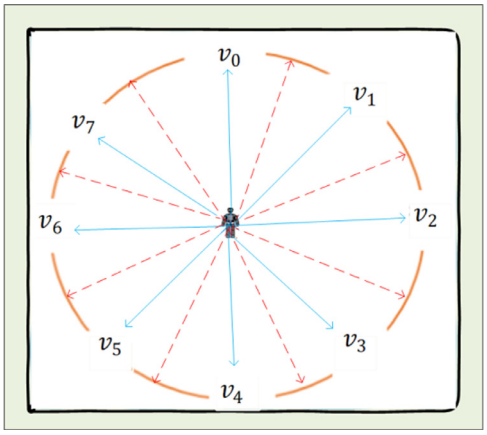

Fig. 4. Illustration of view setting in arbitrary-view action recognition. In the arbitrary-view recognition, 8 fixed views $(v_0, \ldots, v_7)$ which are marked with blue are original in the UESTC dataset. Using action samples of 8 fixed views, we generate synthetic samples in the other 8 views, marked with dashed red arrows. Ture data and synthetic data are used for the recognition model training, and finally, the trained model is applied to recognize full-circle view action sequences.

4.3. Implementation details

During the training of the generation module, we set V to be one hot vector with M_V dimensions, and it is used to control the view of generated action samples. In the UESTC dataset, M_V is set to 4 for the cross-view recognition and 8 to realize the arbitrary-view recognition. In the NTU datasets, action samples in views of camera 2 and 3 are all used for training, only samples in the view of camera 1 are used for testing, thus vector V has only one dimension. We use the Adam optimizer to optimize the generator network [35]. The network is trained with an initial learning rate of $1 \times e^{-4}$, the batch size is set to 16 and epochs are set to 35, and momentum is of $[0, 0.9]$ in both the UESTC and NTU datasets. In Eqs. (3), (4), (5), the weight coefficients of loss functions are set to $\lambda_1 = 10, \lambda_2 = \lambda_3 = 1$.

In the view domain generalization, we use residual networks as a backbone for feature learning. We also use the Adam optimizer algorithm [35] to optimize the network. The network is trained with an initial learning rate of $1 \times e^{-4}$, and the batch size is set to 16, and epochs are set to 100 for all datasets.

5. Ablation study

To certify the necessity of the generation module and the view domain generalization module, we provide a detailed ablation study in the UESTC and NTU datasets.

5.1. Generation assessment and evaluation

Structural SIMilarity (SSIM): We employ the Structural SIMilarity (SSIM) [36] to evaluate the quality of generated action samples. The SSIM metric compares corresponding joints and their neighborhood between the real and synthetic samples in three terms: luminance $(\mathcal{I})$, contrast $(\mathcal{C})$, and structure $(\mathcal{S})$. The calculation of SSIM score is defined as,

$$\mathbf{SSIM}(\boldsymbol{x}_{v_j}, \bar{\boldsymbol{x}}_{v_j}) = \frac{1}{N}\sum_{i=1}^{N} \frac{(2\mu_{x_{v_j}}\mu_{\bar{x}_{v_j}} + c_1)(2\sigma_{x_{v_j}\bar{x}_{v_j}} + c_2)}{(\mu^2_{x_{v_j}}\mu^2_{\bar{x}_{v_j}} + c_1)(\sigma^2_{x_{v_j}} + \sigma^2_{\bar{x}_{v_j}} + c_2)} \quad (8)$$

where $\boldsymbol{x}_{v_j}$ and $\bar{\boldsymbol{x}}_{v_j}$ denote the real and synthetic samples, respectively. The variable $\mu_{x_{v_j}}$ and $\mu_{\bar{x}_{v_j}}$ denote mean intensity, and $\sigma_{x_{v_j}}$, $\sigma_{\bar{x}_{v_j}}$ denote the standard deviation of the intensity for $\boldsymbol{x}_{v_j}$ and $\bar{\boldsymbol{x}}_{v_j}$. The constant c_1 and c_2 are small values which are added for numerical stability. The maximum value 1 is achieved only if $\boldsymbol{x}_{v_j}$ and $\bar{\boldsymbol{x}}_{v_j}$ are identical. We perform the generation assessment by calculating the similarity between real and synthetic samples, which belong to the same action category and the same view, without considering subjects.

Table 1
Generation results comparison of one-to-many views in the UESTC dataset.

Source View		v_0			
Generated View		v_1	v_3	v_5	v_7
SSIM	CVAE-GAN [21]	0.92	0.88	0.87	0.89
	CR-GAN[22]	0.84	0.73	0.78	0.78
	Ours	**0.92**	**0.87**	**0.90**	**0.86**

Table 2
Generation results comparison of many-to-one views in the NTU datasets.

Dataset		NTU-60	NTU-120
Source View		v_2, v_3	v_2, v_3
Generated View		v_1	v_1
SSIM	CVAE-GAN [21]	0.81	0.78
	CR-GAN [22]	0.84	0.75
	Ours	**0.88**	**0.81**

We obtain novel-view synthetic action samples through one-to-many-view generation. The SSIM is employed to measure the performance of generation in the UESTC and NTU datasets, we illustrate experiment results in Tables 1 and 2. These results indicate that our proposed approach outperforms other methods. As shown in Tables 1 and 2, the SSIM scores of CR-GAN[1] and CVAE-GAN[2] are less than our approach in both the UESTC and NTU datasets. Therefore, our proposed two-branch generation model improves the generation quality of synthetic action samples for arbitrary-view human action recognition.

Visualization of generated results: In two NTU datasets, for the action generation, we train the generation model using samples of the 2nd and the 3rd cameras and generate synthetic action samples of the first camera setting. Here we illustrate some generation results in the NTU-60 dataset.

Using samples of camera 2 in the NTU-60 dataset as input, we obtain generated action samples in camera 1. The generated results are visualized in Fig. 5. We compare samples generated by our two-branch generation model (as shown in row 3) and generated by the CR-GAN approach (as shown in row 2). Our approach outperforms the CR-GAN. We also compare samples generated by our two-branch generation model with the ground truth (as shown in row 4). As shown, our results are more similar to ground truth samples. According to the comparison in Fig. 5, our two-branch generation model obtains better performance in cross-view action sample generation. Our proposed approach can generate high-quality samples.

5.2. Recognition comparison with related generation approaches

We compare the performance of the proposed two-branch generation module with the single branch generator G_2 and related generation models, CVAE-GAN [21] and CR-GAN [22]. For comparison, we employ CVAE-GAN, CR-GAN, single branch generator G_2 for novel-view sample generation, and combine generated samples and original training samples for the cross-view and arbitrary-view action recognition in the UESTC and two NTU datasets. Recognition results are shown in Table 3.

[1] We run the code found in Github: https://github.com/bluer555/CR-GAN
[2] We run the code found in Github: https://github.com/tkazusa/CVAE-GAN

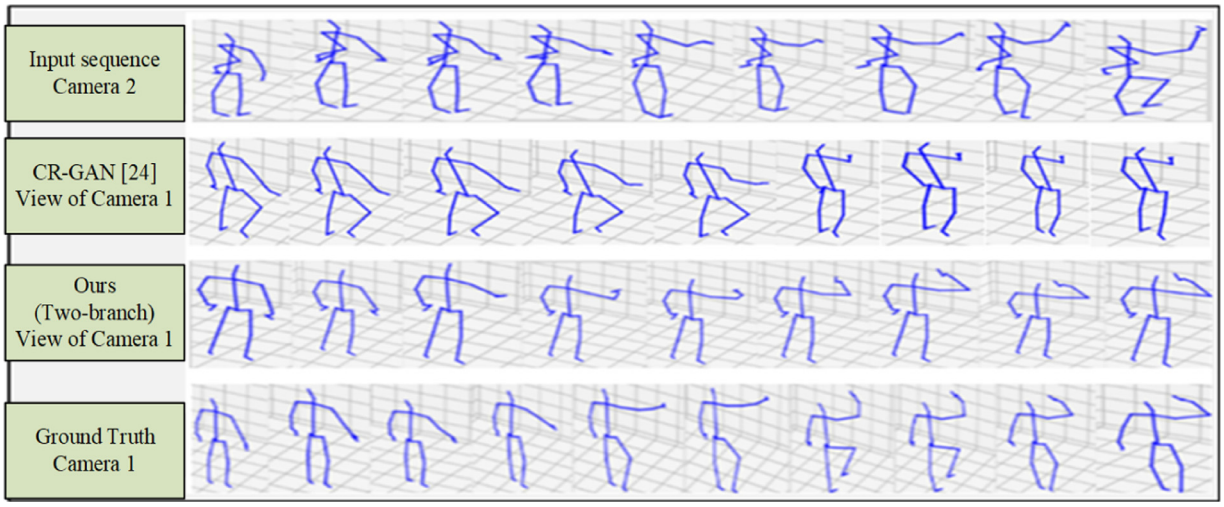

Fig. 5. Visualization of generation results in the NTU-60 dataset. Using samples of camera 2 as input, we generate action samples in camera 1 using our two-branch generation module (row 3). We compare our results with results generated by the CR-GAN approach(row 2) and the ground truth (row 4). The comparison indicates that our proposed approach generates high-quality cross-view samples.

Table 3
Comparison of recognition results with other generation models in UESTC & NTU datasets (%)

Approaches	UESTC		NTU(X-view)	
	X-view	A-view	NTU-60	NTU-120
CVAE-GAN [21]	90.2	86.2	86.1	75.5
CR-GAN [22]	86.9	85.9	87.7	74.5
Ours(G_2)	90.0	85.9	89.2	75.5
Ours (w/o generalization)	87.9	84.8	83.9	73.2
Ours	**91.9**	**87.9**	**93.3**	**81.3**

Table 4
Recognition result comparison with state-of-the-art in the UESTC dataset (%).

Approaches	X-view	A-view
JOULE [39]	60.5	35.4
SK-CNN [38]	68.5	43.3
VS-CNN [27]	71.7	57.2
ST-GCN [37]	85.6	70.9
ANT [2]	89.4	71.7
Ours	**91.9**	**87.9**

Compared the single branch generator G_2 with CVAE-GAN and CR-GAN, the G_2 performs better than two generation models. Compared with the single branch generator (G_2), our two-branch generation module (**Ours**) improves the recognition accuracy of 1.9% for the X-view evaluation and 2% for the A-view evaluation in the UESTC dataset. Similarly, in two NTU datasets, our two-branch generation module (**Ours**) improves the recognition accuracy of 4.1% and 5.8% for the X-view evaluation in the NTU-60 and NTU-120 dataset, respectively. It certifies that our two-branch generation module has better performance in generating samples with a complex feature distribution.

5.3. Effectiveness of view domain generalization

In the Table 3, we also list the experiment results without view generalization (Ours w/o generalization) which removes the view domain generalization from our proposed approach, combining $F(x_{v_i}, \theta_f), F(\bar{x}_{v_j}, \theta_f)$ directly to train the action classifier. Comparing the recognition result obtained using the whole proposed approach (**Ours**) with the Ours w/o generalization, we observe that the whole proposed approach improved accuracy by 4% for X-view recognition and 3.1% for A-view recognition in the UESTC dataset. Similarly, the whole proposed approach (with domain generalization) brings the performance improvement of 9.4% and 8.1% for X-view recognition in the NTU-60 and NTU-120 datasets, respectively. The operation of view domain generalization weakens the effect from the variation of views. Therefore, the view domain generalization is certified to be effective for view-invariant action recognition.

Table 5
Recognition result comparison with state-of-the-art in NTU datasets (%)

Approaches	NTU-60 (X-view)	NTU-120 (X-view)
Two-stream Attention [40]	84.0	63.3
ST-GCN [37]	88.3	71.3
GVFE+ST-GAN [41]	88.0	74.2
GVFE+ AS-GCN [41]	92.9	79.8
ANT [2]	93.5	-
AS-GCN [28]	**94.2**	78.9
VA-CNN [25]	**94.2**	78.9
Ours	93.3	**81.3**

6. Comparison with state-of-the-art approaches

6.1. Cross-view action recognition (X-view)

Based on our experiment setting in Section 4.2, we record recognition results for UESTC and NTU datasets, and list them in Tables 4 and 5. As shown in two tables, our proposed approach achieves higher recognition accuracy compared with the ANT [2], ST-GCN [37], SK-CNN [38] and other approaches. In Table 4, our proposed approach achieves a recognition accuracy of 91.9% for the X-view evaluation in the UESTC dataset, which is 2.5% higher improvement than the ANT method [2]. The comparison in two NTU datasets also demonstrates the effectiveness of the proposed approach (two-branch generation model) in generating novel-view action samples. Also, view domain generalization has improved the performance of view-invariant human action recognition by minimizing the difference of action representation in various views. According to these experiment results, our approach has an outstand-

ing performance in learning view-invariant features for arbitrary-view human action recognition.

6.2. Arbitrary-view action recognition

Following the explanation in Section 4.2 and the paper [27], we perform the arbitrary-view action recognition in the UESTC dataset. Using 8 fixed-view sequences as training data, we train the two branch generation module to generate action samples in the other 8 views (marked with blue color) which are around the neighborhood of those 8 fixed views (marked with red color), as shown in Fig. 4. Finally, training samples and generated samples are adopted for view domain generalization and arbitrary-view action classification. We illustrate recognition results in Table 4. Our approach achieves higher recognition accuracy of 87.9% in the A-view evaluation which improves the recognition accuracy by 16.2% compared with the ANT approach. Observing the results listed in Table 4, our proposed approach achieves state-of-the-art performance. It certifies the superiority of our proposed approach.

As shown in Table 5, our model achieve state-of-the-art performance compared to other approaches with a clear margin 1.5% improvement in cross-view recognition in NTU-120 dataset. However, we achieve a slightly lower but nearly competitive performance of 93.3% in the NTU-60 dataset. Therefore, compared with the state-of-the-art performance, our approach has better cross-view human action recognition accuracy in NTU-120 and competitive accuracy in the NTU-60 dataset.

7. Conclusion

In this paper, we presented a two-branch novel-view sample generation approach for arbitrary-view human action recognition, which enlarged the view ranges of action samples in the training set. Since it was difficult to collect/capture a large number of action samples in various views, the two-branch generation module made our approach satisfy many real application scenarios. For view-invariant action recognition, the view domain generalization was presented to decline variation of action feature representation in various views. We performed extensive experiments and ablation studies to evaluate the performance of our proposed approach, e.g. the cross-view, arbitrary-view evaluations. Experiment results demonstrated that the proposed approach consistently improved the recognition performance in three challenging benchmarks (UESTC, NTU-60, and NTU-120) and achieved state-of-the-art performance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. Ji, Y. Zhan, Y. Yang, X. Xu, F. Shen, H.T. Shen, A context knowledge map guided coarse-to-fine action recognition, IEEE Trans. Image Process. 29 (2020) 2742–2752.

[2] Y. Ji, Y. Yang, N. Xie, H.T. Shen, T. Harada, Attention transfer (ANT) network for view-invariant action recognition, ACM MM, 2019.

[3] J. Feng, J. Xiao, View-invariant human action recognition via robust locally adaptive multi-view learning, Front. Inf. Technol. Electron.Eng. 16 (11) (2015) 917–929.

[4] F.I. Bashir, A.A. Khokhar, D. Schonfeld, View-invariant motion trajectory-based activity classification and recognition, Multimed. Syst. 12 (1) (2006) 45–54.

[5] J. Wang, X. Nie, Y. Xia, Y. Wu, S.-C. Zhu, Cross-view action modeling, learning and recognition, CVPR, 2014.

[6] J. Liu, M. Shah, B. Kuipers, S. Savarese, Cross-view action recognition via view knowledge transfer, CVPR, 2011.

[7] A. Gupta, J. Martinez, J.J. Little, R.J. Woodham, 3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding, CVPR, 2014.

[8] H. Rahmani, A. Mian, 3d action recognition from novel viewpoints, in: Proc. IEEE Conf. Comput.Vis. Pattern Recognit., 2016, pp. 1506–1515.

[9] A. Iosifidis, A. Tefas, I. Pitas, View-invariant action recognition based on artificial neural networks, IEEE Trans. Neural Netw. Learn.Syst. 23 (3) (2012) 412–424.

[10] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Comput. Vis. Image Underst. 104 (2–3) (2006) 249–257.

[11] P. Wei, Y. Zhao, N. Zheng, S.-C. Zhu, Modeling 4D human-object interactions for event and object recognition, ICCV, 2013.

[12] I.N. Junejo, E. Dexter, I. Laptev, P. Perez, View-independent action recognition from temporal self-similarities, IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2011) 172–185.

[13] J. Zhang, H.P. Shum, J. Han, L. Shao, Action recognition from arbitrary views using transferable dictionary learning, IEEE Trans. Image Process. 27 (10) (2018) 4709–4723.

[14] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, M. Kankanhalli, Benchmarking a multimodal and multiview and interactive dataset for human action recognition, IEEE Trans. Cyberne. 47 (7) (2016) 1781–1794.

[15] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv:1411.1784 (2014).

[16] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, InfoGAN: interpretable representation learning by information maximizing generative adversarial nets, NeurIPS, 2016.

[17] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier GANs, ICML, 2017.

[18] A.B.L. Larsen, S.K. Sønderby, H. Larochelle, O. Winther, Autoencoding beyond pixels using a learned similarity metric, arXiv:1512.09300 (2015).

[19] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, Multi-class generative adversarial networks with the l2 loss function, arXiv:1611.04076 (2016).

[20] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein GANs, NeurIPS, 2017.

[21] J. Bao, D. Chen, F. Wen, H. Li, G. Hua, CVAE-GAN: fine-grained image generation through asymmetric training, ICCV, 2017.

[22] Y. Tian, X. Peng, L. Zhao, S. Zhang, D. Metaxas, CR-GAN: learning complete representations for multi-view generation, IJCAI, 2018.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, NeurIPS, 2014.

[24] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3D exemplars, ICCV, 2007.

[25] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, View adaptive neural networks for high performance skeleton-based human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2019) 1963–1978.

[26] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, C. Shi, Cross-view action recognition via a continuous virtual path, CVPR, 2013.

[27] Y. Ji Yanli Yang, F. Shen, H.T. Shen, W.-S. Zheng, A large-scale RGB-D database for arbitrary-view human action recognition, ACM MM, 2018.

[28] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-structural graph convolutional networks for skeleton-based action recognition, CVPR, 2019.

[29] Y. Yang, Z. Ma, Y. Yang, F. Nie, H.T. Shen, Multitask spectral clustering by exploring intertask correlation, IEEE Trans. Cybern. 45 (5) (2015) 1083–1094.

[30] M. Long, Y. Cao, J. Wang, M.I. Jordan, Learning transferable features with deep adaptation networks, ICML, 2015.

[31] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein GANs, NeurIPS, 2017.

[32] D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu, Equivalence of distance-based and RKHS-based statistics in hypothesis testing, Ann. Stat. 41 (5) (2013) 2263–2291.

[33] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, NTU RGB+ D: a large scale dataset for 3D human activity analysis, CVPR, 2016.

[34] J. Liu, A. Shahroudy, M.L. Perez, G. Wang, L.-Y. Duan, A.K. Chichung, NTU RGB+ D 120: a large-scale benchmark for 3d human activity understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (10) (2020) 2684–2701.

[35] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, ICLR, 2015.

[36] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[37] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, AAAI, 2018.

[38] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recognit. 68 (2017) 346–362.

[39] J.-F. Hu, W.-S. Zheng, J. Lai, J. Zhang, Jointly learning heterogeneous features for RGB-D activity recognition, IEEE Trans. Patter Anal. Mach. Intell. 39 (11) (2017) 2186–2200.

[40] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, A.C. Kot, Skeleton-based human action recognition with global context-aware attention LSTM networks, IEEE Trans. Image Process. 27 (4) (2017) 1586–1599.

[41] K. Papadopoulos, E. Ghorbel, D. Aouada, B. Ottersten, Vertex feature encoding and hierarchical temporal modeling in a spatio-temporal graph convolutional network for action recognition, ICPR, 2020.

Kumie Gedamu received his BSc degree in Information Science from Haramaya University, Ethiopia in 2010, and his MSc degree from Andhra University College of Engineering, India in 2015. From 2010 to 2018 he was working as a full lecturer and researcher at Adama Science & Technology University. Currently, he is pursuing his Ph.D. degree in the School of Computer Science and Engineering at UESTC and engaged in the Center for Future Media lab. His major research interests include computer vision problems related to action recognition.

Yanli Ji is currently an Associate Professor in the University of Electronic Science and Technology of China (UESTC). She obtained her Ph.D. degree from the Department of Advanced Information Technology, Kyushu University, Japan in Sep. 2012. Her research interests include Human-Robot Interaction related topics, human activity recognition, emotion analysis, and multimedia understanding.

Yang Yang is currently with University of Electronic Science and Technology of China. He was a Research Fellow under the supervision of Prof.Tat-Seng Chua in National University of Singapore during 2012-2014. He was conferred his Ph.D. Degree (2012) from The University of Queensland, Australia. During the Ph.D. study, Yang Yang was supervised by Prof. HengTao Shen and Prof. Xiaofang Zhou. He obtained Master Degree (2009) and Bachelor Degree (2006) from Peking University and Jilin University, respectively. His research interests include multimedia content analysis, computer vision, and social media analytics.

Lingling Gao is currently pursuing her MSc degree in the School of Computer Science and Engineering at University of Electronic Science and Technology of China. She had received her BSc degree in the University of Electronic Science and Technology of China in 2019. Her major research interests include action recognition.

Heng Tao Shen received the B.Sc. (First Class Hons.) and Ph.D. degrees in computer science from the Department of Computer Science at National University of Singapore in 2000 and 2004 respectively. He is currently a Professor, the Dean of School of Computer Science and Engineering, the executive Dean of AI Research Institute, and the Director of Center for Future Media at University of Electronic Science and Technology of China. His current research interests include multimedia search, computer vision, artificial intelligence, and big data management. He has published 250+peer-reviewed papers and received 7 best paper awards from international conferences, including the Best Paper Award from ACM Multimedia 2017and Best Paper Award-Honourable Mention from ACM SIGIR 2017. He has served as General Cochair for ACM Multimedia 2021 and TPC Co-Chair for ACM Multimedia 2015 and is an Associate Editor of ACM Transactions of data Science (TDS), IEEE Transactions on Image Processing (TIP), IEEETransactions on Multimedia (TMM), and IEEE Transactions on Knowledge and Data Engineering (TKDE).