Learning to Recognize Human Actions From Noisy Skeleton Data Via Noise Adaptation

Sijie Song, *Student Member, IEEE*, Jiaying Liu, *Senior Member, IEEE*, Lilang Lin, and Zongming Guo, *Member, IEEE*

***Abstract*—Recent studies have made great progress on skeleton-based action recognition. However, most of them are developed with relatively clean skeletons without the presence of intensive noise. We argue that the models learned from relatively clean data are not well generalizable to handle noisy skeletons commonly appeared in the real world. In this paper, we address the challenge of recognizing human actions from noisy skeletons, which is seldom explored by previous methods. Beyond exploring the new problem, we further take a new perspective to address it, *i.e.*, noise adaptation, which gets rid of explicit skeleton noise modeling and reliance on skeleton ground truths. Specifically, we develop regression-based and generation-based adaptation models according to whether pairs of noisy skeletons are available. The regression-based model aims to learn noise-suppressed intrinsic feature representations by mapping pairs of noisy skeletons into a noise-robust space. When only unpaired skeletons are accessible, the generation-based model aims to adapt the features from noisy skeletons to a low-noise space by adversarial learning. To verify our proposed model and facilitate research on noisy skeletons, we collect a new dataset Noisy Skeleton Dataset (NSD), the skeletons of which are with much noise and more similar to daily-life data than previous datasets. Extensive experiments are conducted on the NSD, VV-RGBD and N-UCLA datasets, and results consistently show the outstanding performance of our proposed model.**

***Index Terms*—Action recognition, noisy skeletons, regression model, generative model, noise adaptation.**

I. Introduction

HUMAN action recognition has been extensively studied in recent years. It plays an important role in computer vision with broad applications in human-machine interaction, video surveillance and robotics.

In the past decades, many efforts have been devoted to action recognition [1]. One important branch of this area focuses on recognition based on RGB videos, while skeleton-based action recognition attracts much research attention more recently. As a high-level human representation [2], skeleton data is essentially 2D/3D human joint coordinates, and invariant to background and viewpoints. The robustness of skeletons and low data dimensions make it an ideal source to support real-time action recognition algorithms. Moreover, with the prevalence of depth cameras (*e.g.*, Microsoft Kinect [3]) and the advance of pose estimation technologies from RGB frames [4], skeleton data is more accessible nowadays.

The key to the success of skeleton-based action recognition lies on how to capture robust and discriminative features embedded in the spatial configuration and temporal dynamics. There have been many attempts leveraging the strength of deep neural networks to achieve this task, including recurrent neural network (RNNs) [5]–[7], the convolutional neural network (CNNs) [8], [9], and graph neural networks [10]. Though these methods achieve promising results, they are developed and evaluated on the datasets with relatively clean skeletons[1] [11]–[14]. However, skeletons in real life are always with much noise for many reasons (*e.g.*, occlusion, environment, *etc.*), resulting in heavy degradation in skeleton-based human representations. It remains unclear whether these models can well adapt to deal with extremely noisy skeletons for action recognition. Though there are a few works taking skeleton noise into account for action recognition [15]–[17], they model it as an independent skeleton denoising problem as data preprocessing, and perform action recognition based on the denoised skeleton data. However, we argue that there might be two issues in the above pipeline:

- With various human skeleton capture conditions and devices, it is hard to design a uniform and general skeleton denoising method because how the ground-truth skeletons degrade to noisy ones is unknown.
- Independent skeleton denoising does not guarantee better action recognition performance, because filtering noise (*e.g.*, motion jittering) may eliminate critical action cues for recognition.

In this paper, we tackle action recognition for noisy skeletons from a new perspective. Instead of skeleton denoising, we regard it as a skeleton noise-adaptation problem. Considering the difficulties in collecting totally noise-free data, which requires expensive motion capture systems, we explore different models that can learn to extract noise-robust features with

Manuscript received February 26, 2021; revised July 9, 2021 and September 1, 2021; accepted October 8, 2021. Date of publication October 15, 2021; date of current version March 4, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102702, in part by the Fundamental Research Funds for the Central Universities, and in part by the National Natural Science Foundation of China under Grant 61772043. The Guest Editor coordinating the review of this manuscript and approving it for publication was Dr. Xian-Sheng Hua. *(Corresponding author: Jiaying Liu.)*

The authors are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China (e-mail: ssj940920@pku.edu.cn; liujiaying@pku.edu.cn; linlilang@pku.edu.cn; guozongming@pku.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3120631.

Digital Object Identifier 10.1109/TMM.2021.3120631

[1]Due to the accuracy limit of capture devices and inevitable random errors, skeleton data from experimental datasets is not totally clean and only called *relatively clean*.

paired or unpaired noisy skeletons. Specifically, we design a regression-based adaptation model for paired noisy skeletons, and a generation-based adaptation model for unpaired noisy skeletons, respectively. In our paper, *paired noisy skeletons* refer to pairs of two observed noisy skeletons of the same action sequence, *e.g.*, skeletons recorded from different viewpoints, *unpaired noisy skeletons* refer to two skeleton sequences at different noise levels, *e.g.*, skeletons from real world and lab environment, and they do not have to be observations of the same action sequence. The regression-based adaptation model aims to map the multiple observed skeletons for a certain action sequence to a noise-robust common space. In this process, the influence of random noise is suppressed while intrinsic feature representations of skeletons are extracted. The generation-based model is to model the distribution of skeletons with low-level noise (*e.g.*, lab-environment skeletons), and then adapt feature embeddings extracted from skeletons with high-level noise (*e.g.*, real-world skeletons) to a low-noise feature space by adversarial learning.

The main contributions of our paper can be summarized as follows:

- We propose to address action recognition from noisy skeletons from the perspective of noise adaptation. We explore regression-based and generation-based adaptation models respectively to take full advantage of accessible skeleton data.
- We collect a new dataset, Noisy Skeleton Dataset (NSD), containing simultaneous skeleton sequences captured by cameras set from surrounded viewpoints. The occlusion leads to extremely noisy skeletons, which are much closer to real-world data than previous datasets. Besides, we design a method to estimate the skeleton noise level.
- We make sufficient observations and analyses of our proposed models, as well as state-of-the-art methods, and try to provide some insights to inspire the community on recognition from noisy skeletons and boost applications in the real world.

The remainder of the paper is organized as follows. In Section II, we review the related works on skeleton-based action recognition, skeleton datasets, and domain adaptation. In Section III, we present more details of our newly-collected NSD dataset and the skeleton noise estimation method. In Section IV, we introduce the regression-based model for paired noisy skeletons and generation-based model for unpaired noisy skeletons, respectively. Experiments and analysis are presented in Section V. Conclusion remarks are finally given in Section VI.

II. Related Work

A. Skeleton-Based Action Recognition

Earlier works for skeleton-based action recognition are generally based on hand-crafted features [13], [18]–[22]. More recent works focus on models with deep neural networks [5], [7]–[9], [23], [24]. Some works leverage the merits of recurrent neural networks due to their powerful temporal modeling ability. Du *et al.* [5] achieved the pioneering work with a hierarchical RNN to process each body part. Song *et al.* [7] proposed an attention-based LSTM network to improve discrimination of skeleton features by automatically selecting important human joints and video frames. Zhang *et al.* [24] developed a view-invariant model, enabling the network to adapt to the most suitable observation viewpoints. To better handle the spatial-temporal features, convolutional neural networks (CNNs) equipped with excellent capacity in extracting high-level information are employed in skeleton-based action recognition [8], [9]. By transforming skeleton sequence into clips, a new representation was presented in [8], and the clip images are processed with CNNs. Li *et al.* [9] designed a hierarchical CNN model to learn joint co-occurrence and temporal evolutions. Given that human skeletons are naturally with graph structure, graph convolution neural networks are recently explored to capture spatial structural information [10], [25].

Though previous methods achieve promising results, they are trained and evaluated on relatively clean skeletons. Recently, a few works have paid attention to dealing with action recognition from noisy skeletons [15]–[17]. All of them regard it as a skeleton-denoising problem, which first apply linear or non-linear transformation to filter skeleton noise and then perform action recognition. Suffering from the lack of ground-truth clean skeletons, some of them [15], [16] are based on the prior from the bio-constrained skeleton structure, and perform skeleton denoising with linear transformation. However, the linear transformation based on human prior may not be optimal and powereful enough for denoising. To tackle the above issues, Demisse *et al.* [17] first proposed a non-linear denoising transformation based on an autoencoder. However, it is problematic by only forcing the mean square loss between the noisy input and reconstructed output, because it can easily converge to a trivial solution (*i.e.*, identical mapping). Besides, the unknown skeleton degradation process leads to difficulties in finding a general denoising model and further facilitating action recognition. Instead, we regard action recognition from noisy skeletons as a noise-adaptation problem to get rid of an explicit skeleton noise modeling and reliance on the ground truth clean skeletons.

B. Dataset for 3D Skeletons

There are usually two sources for skeletons in datasets for 3D action analytics, *i.e.*, motion capture system (MoCap) and depth cameras. The skeletons from MoCap are quite accurate because they are obtained by sensors stuck on the human body. With the release of Microsoft Kinect [3], more datasets are collected by different research teams [11], [12], [14], [26]. The well-known NTU RGB+D dataset [11] recorded over 50 k skeleton sequences from multiple view points. The cameras in the NTU RGB+D dataset are set mainly in the front with slight view changes, and the actors perform different actions towards the cameras. Though noises exist in these datasets, they are far from those in the skeletons from real world, in which the actors can suffer from heavy occlusions. We simulate the scenarios in real-life, and the skeletons in our dataset are with much noise.

C. Domain Adaptation

The domain gap between relatively clean and noisy skeletons leads to the degradation of the models trained on relatively clean skeletons. This is essentially a domain adaptation problem. One of the major categories minimize Maximum Mean Discrepancy (MMD) to narrow the domain shifts. Long *et al.* [27] proposed a deep adaptation network (DAN) to project features in a kernel Hillbert space and minimize the MMD in the project space. In addition, Joint Maximum Mean Discrepancy (JMMD) [28] is adopted to learn transferable features by aligning the joint distributions of multiple layers. However, the work in [29] claims that focusing only on the shared features leads to the ignorance of individual characteristics.

Other methods use adversarial learning to perform domain adaptation. The adversarial losses have been applied in the embedding space [30], [31] and pixel space [32]–[35]. The main idea of such approaches is to use generative models such as GANs to perform cross-domain mapping. Inspired by these methods, we use adversarial learning to utilize relatively clean skeletons to facilitate the recognition from noisy ones. However, the methods above map a shared feature to a specific domain. They may ignore characteristics of each domain. Instead, we use a residual compensation network to extract robust features.

III. Noisy Skeleton Dataset

A. Dataset Settings

Our dataset aims to provide noisy skeletons that consistent with those in the real world. The noise in skeletons is largely due to heavy occlusion caused by viewpoints. Thus, we set Microsoft Kinect V2 cameras around the actors. The horizontal angles of each camera are -120° (side view 1), 0° (front view), and $+120^\circ$ (side view 2) with the height of 120 cm. Our dataset provide simultaneous color images, depth maps, 3D joints and IR frames. The data format is consistent with [11].

B. Dataset Details

We collect 1,009 untrimmed videos, each of which lasts about 1$\sim$2 minutes and contains about 7 action instances. In total, there are 6,952 trimmed action clips in 41 action categories. We invite 13 subjects and each subject takes part in 4 daily action videos. Some sample frames can be viewed in Fig. 1. The actors perform actions towards a random direction. Thus, in any case, the data from one of the cameras suffer from heavy occlusion and thus noisy skeletons.

C. Evaluation Protocol

We suggest two data splits (*i.e.*, cross-subject and cross-view) in our dataset.

Cross-Subject: Cross-subject evaluation aims to test the ability to handle intra-class variations among different actors. 10 subjects are chosen to be training samples and 3 for testing.

Cross-View: Cross-view evaluation aims to test the robustness in terms of transformation (*e.g.*, translation, rotation). The videos from Camera #1 and #2 are chosen as the training set, and those from Camera #3 are as the testing set.

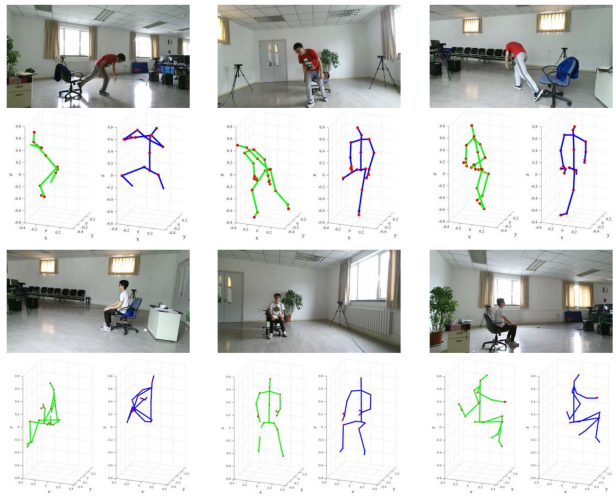


Fig. 1. From left to right, we show sample frames in the NSD dataset captured from the front view, side view 1, and side view 2, respectively. Skeletons are with much noise due to occlusion caused by viewpoints. We show the skeletons in their original coordinate system in green, and then manually transform them for better visualization in blue.

D. Action Taxonomy

Our dataset contains 41 action labels, including health related actions, home related actions, dressing related actions, interaction with items and human locomotion.

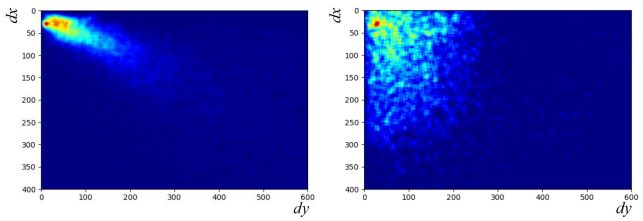
- **Health related (4):** touch head (headache), touch neck (neckache), touch back (backache), touch chest (stomachache/heart pain).
- **Home related (5):** brush teeth, comb hair, wipe face, drink water, eat meal/snack.
- **Dressing related (6):** put on glasses, put on jacket, put on a hat/cap, take off glasses, take off jacket, take off a hat/cap.
- **Interaction with items (14):** drop, write, read, pick up, take a selfie, tear up paper, type on a keyboard, play with phone/tablet, check time (from watch), use a fan (with hand or paper), make a phone call/answer phone, point to something with finger.
- **Human locomotion (14):** bow, clap, throw, salute, fall, hop (one foot jumping), sit down, stand up, cheer up, jump up, kick something, hand waving, rub two hands together, cross hands in front (say stop).

E. Skeleton Noise Estimation

Though it is difficult to obtain skeleton ground-truths, we design a method to estimate the skeleton noise level for our dataset. With the development of pose estimation approaches, openpose [4] is able to accurately estimate 2D pose joints on RGB frames. Therefore, we regard the 2D pose estimation results as skeletal joint ground-truths, and then further estimate the skeleton noise level by projecting the 3D skeletons to the RGB frames. More specifically, for a skeleton coordinate $p = [x, y, z]^T$ in our dataset, we can get the corresponding coordinate $p' = [x,' y']^T$

TABLE I
SKELETON NOISE ESTIMATION RESULTS FOR THE NSD AND NTU DATASETS, RESPECTIVELY

	Front View		Side View 1		Side View 2		Overall	
	mean	std	mean	std	mean	std	mean	std
NTU-dx	0.261	0.054	0.260	0.073	0.264	0.076	0.258	0.064
NSD-dx	0.371	0.089	0.338	0.172	0.435	0.312	0.381	0.216
NTU-dy	0.177	0.059	0.183	0.064	0.188	0.057	0.181	0.058
NSD-dy	0.197	0.086	0.291	0.141	0.337	0.193	0.275	0.157



(a) NTU (b) NSD

Fig. 2. Offset maps for the NTU and NSD datasets, respectively. A brighter point in the offset map indicates a larger number of joints. Note the offsets are calculated based on pose coordinates without normalization.

in the RGB image with the camera parameter matrix $\mathbf{M}$

$$[a, b, c]^T = \mathbf{M}p, \tag{1}$$

$$x' = a/c, \tag{2}$$

$$y' = b/c. \tag{3}$$

We then calculate an offset (dx, dy) with the pose estimation result $\hat{p} = [\hat{x}, \hat{y}]$ for the given joint as

$$dx = ||N(x') - N(\hat{x})||_1, dy = ||N(y') - N(\hat{y})||_1, \tag{4}$$

where N is a normalization function that translates the body coordinate system with its origin on the torso joint, and normalizes the distance between the torso and head joints to 1. To estimate the noise level, we calculate the mean and standard variation for the offsets of all the joints, and compare them with those from the NTU dataset. The results for each view and the overall dataset are shown in Table I. Note that for the NTU dataset, we also regard the pose estimation results from [4] as ground-truths, and use the 2D coordinates provided by the dataset as 3D projected results. From Table I, we observe higher mean and standard variation values for the NSD dataset, indicating the skeletons are much noisier than those from NTU. It is also noticed that, for each dataset, the skeletons from the front view are less noisy compared to those from side views.

We further visualize the offset maps in Fig. 2 for the NTU and NSD datasets, respectively. We calculate the number of joints for each (dx, dy). It is observed that points from NSD distribute more evenly in the map, illustrating that the offsets towards ground-truths are larger and thus the skeletons from NSD are with more noise. It is also consistent with the results in Table I.

IV. NOISE ADAPTATION NETWORKS

A. Motivation

In this work, our goal is to mitigate the performance degradation in action recognition caused by skeleton noise. However, in reality it is challenging to model skeleton noise directly due to the unavailability of totally clean skeletons. In this work, we take a more accessible way, making full of available skeleton data (*i.e.*, paired noisy skeletons or unpaired skeletons at different noise levels) to mitigate the effect of captured skeleton noise for action recognition via noise adaptation. There are usually two branches to tackle noisy data. One lies in the regression model, which is able to converge to the expected value of multiple unreliable measurements for a true unknown target [36], [37], another is the generative model with adversarial learning, which is able to adapt data among different distributions [38], [39] and deal with noisy inputs [40] or labels [41]–[43]. Therefore, we are inspired to make efforts on noise adaptation for action recognition along the above two routes.

B. Regression-Based Adaptation Model

From the perspective of regression model, we aim to learn the noise-robust feature space from multiple skeleton measurements of a certain action sequence. Though it is unfeasible and expensive to collect ground-truths for noisy skeletons, we can easily have a set of unreliable skeleton measurements $\{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$ for a certain action sequence, *i.e.*, recording the sequences simultaneously with different cameras. Therefore, the training set is considered as $\mathcal{D} = \{(\mathbf{X}_1^1, \mathbf{X}_2^1, \mathbf{y}^1), \ldots, (\mathbf{X}_1^N, \mathbf{X}_2^N, \mathbf{y}^N)\}$, where $\mathbf{X}_1^i$ and $\mathbf{X}_2^i$ denote the i^{th} observed noisy skeleton pairs, and $\mathbf{y}^i \in \{0, 1\}^c$ is a one-hot vector indicating the ground-truth action label. To learn the common space, we constrain the l_2 distance of sequence-level feature embeddings from a feature encoder $E(\cdot)$ by minimizing

$$\min \sum_i ||E(\mathbf{X}_1^i) - E(\mathbf{X}_2^i)||_2^2, \tag{5}$$

The goal of learning the common space is to suppress skeleton noise in the feature representations and further achieve noise adaptation. More importantly, the learned feature space should be optimal for action recognition at the same time. With a classifier $C(\cdot)$, we have to minimize the cross-entropy loss $\mathcal{L}_{cls}$ for classification

$$\min \sum_i \mathbf{y}^i \log p(\mathbf{X}^i), \tag{6}$$

where $p(\mathbf{X}^i)$ is the output of classifier $C(\cdot)$ indicating the classification probability for the given sequence $\mathbf{X}^i$ over all classes.

Instantiation. Fig. 3(a) shows an instantiation for the regression-based noise adaptation network. To learn the noise-robust feature space, we constrain the l_2 distance of video-level feature embeddings from an encoder $E(\cdot)$. We leverage the merits of bidirectional GRU (BiGRU) networks in $E(\cdot)$ to process skeletons, obtaining $\mathbf{v}_1^l = \frac{1}{T_1}\sum_{t=1}^{T_1} \mathbf{h}_1^{l,t}$ and $\mathbf{v}_2^l = \frac{1}{T_2}\sum_{t=1}^{T_2} \mathbf{h}_2^{l,t}$ as video-level representations from the l^{th} BiGRU layer, where $\mathbf{h}_1^{l,t}, \mathbf{h}_2^{l,t}$ are hidden states from BiGRU layers. Then

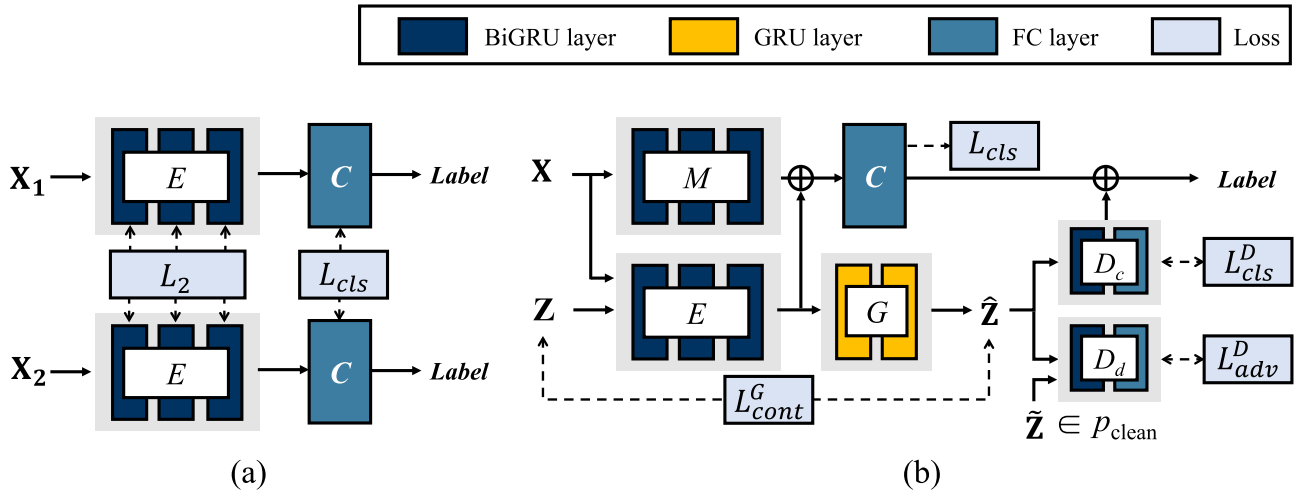


Fig. 3. Instantiations for action recognition from noisy skeletons. (a) Regression-based noise adaptation model. $\mathbf{X}_1$ and $\mathbf{X}_2$ are observed noisy skeletons for a certain action sequence. (b) Generation-based noise adaptation model. $\mathbf{X}$ and $\mathbf{Z}$ are noisy and relatively clean skeleton sequences, respectively. (Note we omit some losses for simplicity.).

we constrain the l_2 distance of video-level feature embeddings after each BiGRU layer. We adopt video-level features here for two reasons: (1) Video-level features are able to encode temporal dynamics and spatial relationship at the same time, which are both critical for action recognition. (2) Video-level features do not require frame-wise paired skeletons. Thus, we do not need to have $T_1 = T_2$, which is more practical and flexible. Then the classifier $C(\cdot)$ consisting of a single fully connected layer is learnt jointly with video-level features. Besides, we adopt a parameter γ on the l_2 distance term to balance its contribution.

C. Generation-Based Adaptation Model

In the regression model, paired noisy skeletons are required. To go a step further, we explore to learn a noise adaptation network from unpaired skeletons at different noise levels. We employ the generative model with adversarial learning to model the distribution of relatively clean skeletons, and then adapt feature embeddings corrupted by high-level noise towards those from relatively clean data. In this scenario, the training data is a set of noisy skeleton sequences with labels, and relatively clean sequences. It can be formulated as $\mathcal{D} = \mathcal{D}_{\mathbf{X}} \cup \mathcal{D}_{\mathbf{Z}}$, where $\mathcal{D}_{\mathbf{X}} = \{(\mathbf{X}^1, \mathbf{y}^1), \ldots, (\mathbf{X}^{N_n}, \mathbf{y}^{N_n})\}$, and $\mathcal{D}_{\mathbf{Z}} = \{\mathbf{Z}^1, \ldots, \mathbf{Z}^{N_c}\}$, N_n and N_c are the number of samples for noisy and relatively clean skeleton sequences, respectively. Note that we do not require labels for relatively clean data in this scenario. We build upon a standard action recognition model $C \circ M$, where $M(\cdot)$ is an encoder to extract features from $\mathbf{X}$ for the classifier $C(\cdot)$ to make the prediction. To incorporate the adapted feature embeddings, we further use $E(\mathbf{X}) \sim p_{clean}$ to compensate $M(\mathbf{X})$ by $h(M(\mathbf{X}), E(\mathbf{X}))$, where h is a pre-defined compensation function.

Instantiation. Fig. 3(b) shows an instantiation of our idea for action recognition with a generative adversarial network. It includes a main network $M(\cdot)$, a classifier $C(\cdot)$, an encoder $E(\cdot)$, a decoder $G(\cdot)$, discriminators $D_d(\cdot)$ and $D_c(\cdot)$. This model adapts skeleton features with high-level noise into the low-level noise space by adversarial learning, which is achieved with $E(\cdot)$, $G(\cdot)$, $D_d(\cdot)$ and $D_c(\cdot)$. The encoder $E(\cdot)$ is encouraged to learn an embedding robust to noise, by generating skeletons with $G(\cdot)$ to confuse discriminators. There are two components in the discriminator: $D_d(\cdot)$ is a binary classifier to indicate the probability of its input being relatively clean skeletons, and the other is a C-way classifier $D_c(\cdot)$ that models the class probability distribution of its input, forcing $E(\cdot)$ and $G(\cdot)$ encoding action information in the generated skeletons to avoid mode collapse. To adapt the high-level noisy features to the low-level noisy space, the encoder $E(\cdot)$ and the decoder $G(\cdot)$ are trained to generate clean skeletons from high-level noisy input to confuse the discriminator $D_d(\cdot)$.

For feature compensation, we have $h = M \oplus E$, where the original feature from $M(\cdot)$ is compensated as $M(\mathbf{X}) + E(\mathbf{X})$. To further boost action recognition performance in testing, we integrate the results from $C(\cdot)$ and $D_c(\cdot)$, and the final result is given by $C(M(\mathbf{X}) + E(\mathbf{X})) + D_c(G(E(\mathbf{X})))$.

During training, the network is optimized iteratively due to their mutual influence of each part. The optimization procedure is described as follows:

Step 1: We optimize the binary classifier D_d, which is for judging the input being *real* or *fake*, by maximizing a least square adversarial loss [44] $\mathcal{L}_{adv}^D$

$$\mathcal{L}_{adv}^D = \max_{D_d} \mathbb{E}_{\mathbf{Z}\sim p_{clean}(\mathbf{Z})}[1 - D_d(\mathbf{Z})]^2 + \mathbb{E}_{\mathbf{X}\sim p_{noisy}(\mathbf{X})}(D_d(G(E(\mathbf{X}))))^2. \tag{7}$$

We optimize D_c by minimizing the cross-entropy loss $\mathcal{L}_{cls}^D$ for reconstructed $\mathbf{X}$

$$\mathcal{L}_{cls}^D = \min_{D_c} -\log(D_c(G(E(\mathbf{X})))_{\mathbf{y}}). \tag{8}$$

Thus, the objective function for discriminators can be formulated as

$$\mathcal{L}_D = \mathcal{L}_{adv}^D + \mathcal{L}_{cls}^D. \tag{9}$$

Algorithm 1:Training details of the integrated framework.

Require: relatively clean skeleton sequences for training $\{\mathbf{Z}^i\}$, noisy skeletons for training $\{(\mathbf{X}^i, \mathbf{y}^i)\}$, main network $M(\cdot)$, classifier $C(\cdot)$, encoder $E(\cdot)$, decoder $G(\cdot)$, discriminators $D_c(\cdot)$ and $D_d(\cdot)$, training iterations N, learning rate lr
FOR t = 1,..., N **DO**
Forward the network
Update $D_c(\cdot)$ and $D_d(\cdot)$ by minimizing (9) with learning rate $0.5 \times lr$.
Update $E(\cdot)$ and $G(\cdot)$ by minimizing (13) with learning rate lr.
Forward the network
Update $E(\cdot)$, $M(\cdot)$ and $C(\cdot)$ by minimizing (14). Note $M(\cdot)$ and $C(\cdot)$ are updated with learning rate $10 \times lr$, while $E(\cdot)$ is updated by a learning rate lr.
END

Step 2: With the gradient from $D_d(\cdot)$ and $D_c(\cdot)$, we update $G(\cdot)$ and $E(\cdot)$ with an adversarial loss to produce realistic skeletons.

$$\mathcal{L}_{adv}^{G} = \min_{G} \mathbb{E}_{\mathbf{X}\sim p_{noisy}}[1 - D_d(G(E(\mathbf{X})))]^2. \quad (10)$$

To avoid mode collapse, that the decoder always generates mean skeleton sequences, we employ a content loss between the relatively clean input and its reconstruction from $G(\cdot)$.

$$\mathcal{L}_{cont}^{G} = ||G(E(\mathbf{Z})) - \mathbf{Z}||_2^2. \quad (11)$$

Besides, we add a classification loss to produce class consistent skeletons with the input noisy skeletons.

$$\mathcal{L}_{cls}^{G} = \min_{G,E} -\log(D_c(G(E(\mathbf{X})))_{\mathbf{y}}). \quad (12)$$

Therefore, the objective function for $G(\cdot)$ is given by

$$\mathcal{L}_{G} = \min_{G,E} \mathcal{L}_{adv}^{G} + \mathcal{L}_{cls}^{G} + \mathcal{L}_{cont}^{G}. \quad (13)$$

Step 3: Finally we update $M(\cdot)$, $E(\cdot)$ and $C(\cdot)$ in a supervised manner for noisy skeletons.

$$\mathcal{L}_{cls} = \min_{E,M,C} -\log(C(M(\mathbf{X}) + E(\mathbf{X}))_{\mathbf{y}}). \quad (14)$$

The training details of the integrated framework are summarized in Algorithm 1 to avoid fast convergence of discriminators.

V. Experiments

A. Experimental Settings

We evaluate our instantiated models with the Noisy Skeleton Dataset (NSD) collected by ourselves, Varying-View RGB-D action dataset (VV-RGBD) [45] and Northwestern-UCLA dataset (N-UCLA) [26].

Noisy Skeleton Dataset (NSD). The detailed configuration for NSD can be found in Section III. For the evaluation protocol, we suggest cross-subject (CS) and cross-view (CV) splits for our dataset.

Varying-View RGB-D action dataset (VV-RGBD). The VV-RGBD dataset consists of 25,600 videos observed from 8 fixed viewpoints and the entire 360° view angles. There are 40 action categories performed by 118 actors. Each body has 25 skeletal joints in 3D coordinates. We follow the protocols of cross-subject (CS) and cross-view II (CV) defined in [45] in our experiments.

Northwestern-UCLA dataset (N-UCLA). This dataset includes 1,494 videos in 10 action categories performed by 10 subjects. Each body has 20 skeletal joints in 3D coordinates. We use videos from the first two views for training and those from the third view for testing following [26].

All of above datasets provide synchronous skeletons captured from large varied viewpoints, providing multiple observed skeletons for the regression-based adaptation model. For the generation-based model, we use the skeletons from the front view of well-known NTU RGB-D dataset [11] as relatively clean data. On the one hand, there is less occlusion from the front view in recording process so we have less noisy skeletons. On the other hand, the large scale of the NTU dataset makes it a reliable source as an estimation of clean data distribution, even with some random noise.

Implementation details. During training, we use the Adam optimizer [46] to adjust the learning rate, which is initially set as 0.0002. The batch size is set as 16, 128, 16 for the NSD, VV-RGBD and N-UCLA datasets, respectively. To eliminate the viewpoint variation [6], [7], we conduct frame-level normalization by manually rotating each skeleton to the front view and translate it to the body coordinate system with its origin on the 'torso' joint. The number of neurons for each BiGRU and GRU layer in Fig. 3 is 100×2 and 100, respectively.

B. Results and Comparisons

In this section, we evaluate and compare our proposed model with several state-of-the-art methods on each dataset.

- **Baseline.** A simple network with three BiGRU layers and a fully connected layer without any noise-adaptation feature learning.
- **R-NAN.** Regression-based noise adaptation network instantiated in Fig. 3(a).
- **G-NAN.** Generation-based noise adaptation network instantiated in Fig. 3(b).

Noisy Skeleton Dataset (NSD). Table II shows the action recognition results in terms of accuracy. In the regression-based model, through mapping multiple observed skeletons into a common space, R-NAN effectively improves the baseline model by 4.6% and 5.9%. In the generation-based model, G-NAN improves the baseline results by 4.8% and 1.7%, by learning the distribution of low-level skeleton noise. We also implement several state-of-the-art skeleton-based action recognition methods to see whether these models can handle extremely noisy data, the results of which can be viewed in Table II. It is challenging for attention-based action recognition methods (*i.e.*, STA-LSTM [7], ST-GCN [10]) to learn attention patterns from noisy skeletons, since noise could mislead the determination on the attention weights. TPN [47] and VA-LSTM [24] also fail

TABLE II
PERFORMANCE COMPARISON IN TERMS OF ACCURACY (%) ON THE NSD DATASET

Methods	CS	CV
STA-LSTM [7]	44.3	28.6
TPN [47]	46.9	29.7
VA-LSTM [24]	50.0	34.5
ST-GCN [10]	48.2	35.8
Denoised-LSTM [17]	38.1	26.1
2S-AGCN (Joint) [48]	46.4	32.6
2S-AGCN (Bone) [48]	46.2	30.0
2S-AGCN (Joint + Bone) [48]	50.8	35.4
SGN [49]	51.2	34.5
Baseline	50.7	34.6
R-NAN	55.3	**40.5**
G-NAN	**55.5**	36.3

TABLE III
PERFORMANCE COMPARISON IN TERMS OF ACCURACY (%) ON THE VV-RGBD DATASET

Methods	CS	CV
TCN [50]	55.6	42.8
Res-TCN [51]	62.8	47.6
LSTM [11]	55.6	30.8
P-LSTM [11]	60.5	33.0
SK-CNN [16]	58.5	68.0
ST-GCN [10]	71.0	56.1
VS-CNN [45]	76.1	70.5
Denoised-LSTM [17]	84.4	76.9
SGN [49]	**95.4**	**89.9**
Baseline	88.6	79.4
R-NAN	91.8	87.6
G-NAN	91.6	84.2

to extract discriminative features due to the interference from noise. Though Denoised-LSTM [17] performs explicit skeleton denoising using pose-encoding auto-encoders, they oversmooth skeletons and fail to preserve action details as aforementioned. Even compared with these state-of-the-art methods [48], [49], both regression-based and generation-based adaptation models achieve significant improvement on action recognition from noisy skeleton data.

Varying-View RGB-D action dataset (VV-RGBD). We present action recognition results on the VV-RGBD dataset in Table III. In regression-based adaptation model, R-NAN achieves gains of 2.2% and 8.2% over the baseline results with CS and CV splits, respectively. In generation-based adaptation model, G-NAN achieves gains of 3.0% and 4.8% over the baseline results with CS and CV splits, respectively. Moreover, R-NAN and G-NAN outperform most of the state-of-the-art methods, indicating that the noise-adaptation models can better deal with skeleton noise in action recognition. It is observed that SGN [49] shows the best performance. We argue that SGN benefits from score fusion during inference.

Northwestern-UCLA dataset (N-UCLA). Table IV shows the results for action recognition in terms of accuracy on the N-UCLA dataset. R-NAN and G-NAN achieve better results than the baseline model by 2.2% and 2.4%, respectively, which suggests that the networks better extract discriminative features for action recognition. We also achieve remarkable performance in comparison with other state-of-the-arts.

TABLE IV
PERFORMANCE COMPARISON IN TERMS OF ACCURACY (%) ON THE N-UCLA DATASET

Methods	Acc. (%)
Actionlet ensemble [19]	76.0
Lie Group [18]	74.2
HBRNN-L [5]	78.5
SK-CNN [16]	86.1
VA-LSTM [24]	70.7
Denoised-LSTM [17]	80.3
2S-AGCN (Joint) [48]	75.1
2S-AGCN (Bone) [48]	79.0
2S-AGCN (Joint + Bone) [48]	81.2
Baseline	84.6
R-NAN	86.8
G-NAN	**87.0**

From Table II, Table III and Table IV, it is observed that R-NAN outperforms G-NAN in most cases. We explain it as that, R-NAN maps paired data into the common feature space to suppress the noise in the feature representations. However, G-NAN tries to adapts noisy skeleton features into the relatively clean skeleton feature space with unpaired data. Compared to G-NAN, R-NAN achieves more specific noise adaptation in a finer granularity, and therefore better action recognition performance.

C. Model Analysis

1) Comparisons With Two-Stage Solutions: We first illustrate the superiority of our proposed noise adaptation models compared with two-stage solutions (*i.e.*, performing skeleton denoising and then action recognition) through the experiments on the NSD dataset. According to the accessibility of noisy skeletons, we explore the following skeleton denoising methods. The detailed structure and loss functions can be found in Fig. 4. For paired noisy skeletons, we adopt a regression-based skeleton denoising method:

• **Regression-based Skeleton Denoising (R-SD).** Borrowing the idea from [36], we regress the feature embeddings from two measurements of the same action sequence, and reconstruct input skeletons to avoid mode collapse.

For unpaired noisy skeletons, we investigate two different models:

• **Pose-Encoding Auto-Encoders (PE-AE).** This work in [17] models skeleton denoising as a non-linear pose variation, and solves it with an autoencoder by minimizing a reconstruction error.

• **Generation-based Skeleton Denoising (G-SD).** Another alternative for unpaired skeleton denoising is to employ a generative model with adversarial learning. We adopt the bottom stream in Fig. 3(b) to filter skeleton noise.

Note that after each skeleton denoising method, we apply the same recurrent network with three BiGRU layers to achieve action recognition. Each layer has 100 neurons.

The results for two-stage solutions are shown in Table V. It is interesting to see that, after skeleton denoising, the performance of action recognition drops a lot under both CS and CV settings. We also employ stronger action classifiers (*i.e.*, STA-LSTM [7],

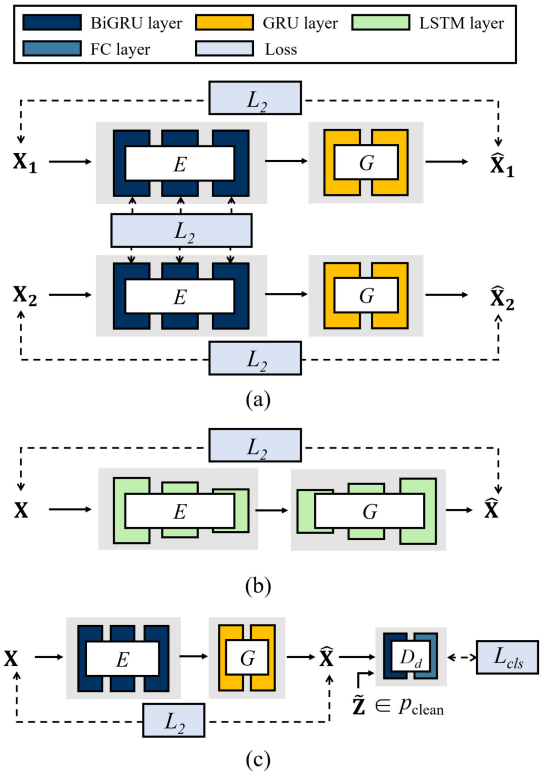
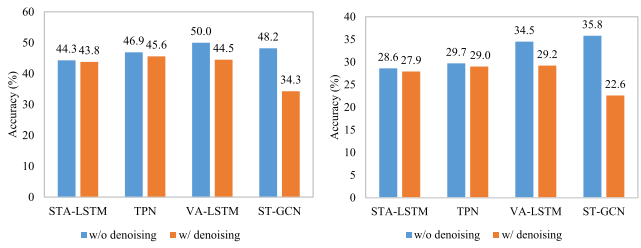


Fig. 4. Detailed structures for the independent skeleton denoising models: (a) R-SD, (b) PE-AE [17], (c) G-SD.

TABLE V
PERFORMANCE COMPARISONS IN TERMS OF ACCURACY (%) ON THE NSD DATASET WITH TWO-STAGE SOLUTIONS FOR ACTION RECOGNITION

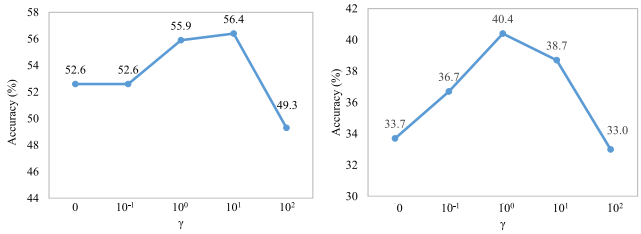
Methods	CS	CV
Baseline	50.7	34.6
BiGRUs w/ R-SD	29.8	20.2
BiGRUs w/ PE-AE [17]	38.1	26.1
BiGRUs w/ G-SD	45.3	27.4
Baseline + BiGRUs w/ G-SD	51.7	33.9
R-NAN	55.3	**40.5**
G-NAN	**55.5**	36.3



(a) Cross-subject (b) Cross-view

Fig. 5. Performance of state-of-the-art methods on the NSD dataset with skeletons denoised by G-SD.

TPN [47], VA-LSTM [24], ST-GCN [10]) with the same skeleton denoising method G-SD. The action recognition results are shown in Fig. 5 and the same phenomenon is observed. It can be explained through visualization of the denoised skeletons in



(a) Cross-subject (b) Cross-view

Fig. 6. Parameter sensitivity analysis on the NSD dataset.

Fig. 7. The results shown in Fig. 7(c)–(e) illustrate that independent skeleton denoising methods (R-SD, PE-AE [17], and G-SD) are able to correct the noisy joints, which are marked with green arrows in Fig. 7(b). However, the denoised results could lose some important cues for action recognition. For example, they filter motion jittering on the hand and elbow for the action of *hand waving* (3 rd row in Fig. 7(c)–(e)), making it similar with *taking a selfie* (4th row in Fig. 7(c)–(e)), which explains the degraded performance on action recognition.

Compared with two-stage solutions, the superior action recognition results in Table V from R-NAN and G-NAN indicate noise-adaptation learning is more effective. R-NAN is able to learn a noise-robust feature space by regularizing feature embeddings from different measurements of a certain action sequence. For G-NAN which is instantiated as Fig. 3(b), the upper stream is able to keep the original information for action recognition and the bottom stream suppresses irrelevant skeleton noises with adversarial learning. Therefore, the poorly represented features from noisy skeletons can be well compensated and then boost action recognition performance. Fig. 7(f) visualizes the adapted skeletons generated from the decoder of G-NAN in Fig. 3(b). It is observed that G-NAN generates more discriminative skeleton representations for action recognition, compared with independent skeleton denoising methods. As marked by red arrows in Fig. 7(f), it magnifies the hand motion for *rubbing two hands*, and preserves the details between one's neck and hand for *touching head*. Besides, it successfully distinguishes *taking a selfie* and *waving hands* by the details on the hand and elbow, that skeletons have the horizontal hand to take a selfie and bend the elbow to wave.

2) Model Structure Analysis: We now analyze our regression-based and generation-based model structures for noise adaptation, respectively. In Fig. 6, we analyze the parameter sensitivity for R-NAN, in which γ controls the regularization of feature embeddings. Smaller γ may not be powerful enough to regularize feature spaces from multiple observed skeletons, while a larger γ would focus more on feature space regularization but ignore preserving action details in feature embeddings. By showing validation accuracy with different values of γ on the NSD dataset, we conclude that our result is not sensitive to γ for a range and set γ as 10 in our experiments. For G-NAN, to confirm that our improvement comes from noise adaptation rather than the change of model structure from baseline, we provide fusion results (Baseline + BiGRUs w/ G-SD) in Table V, and the fusion model has the

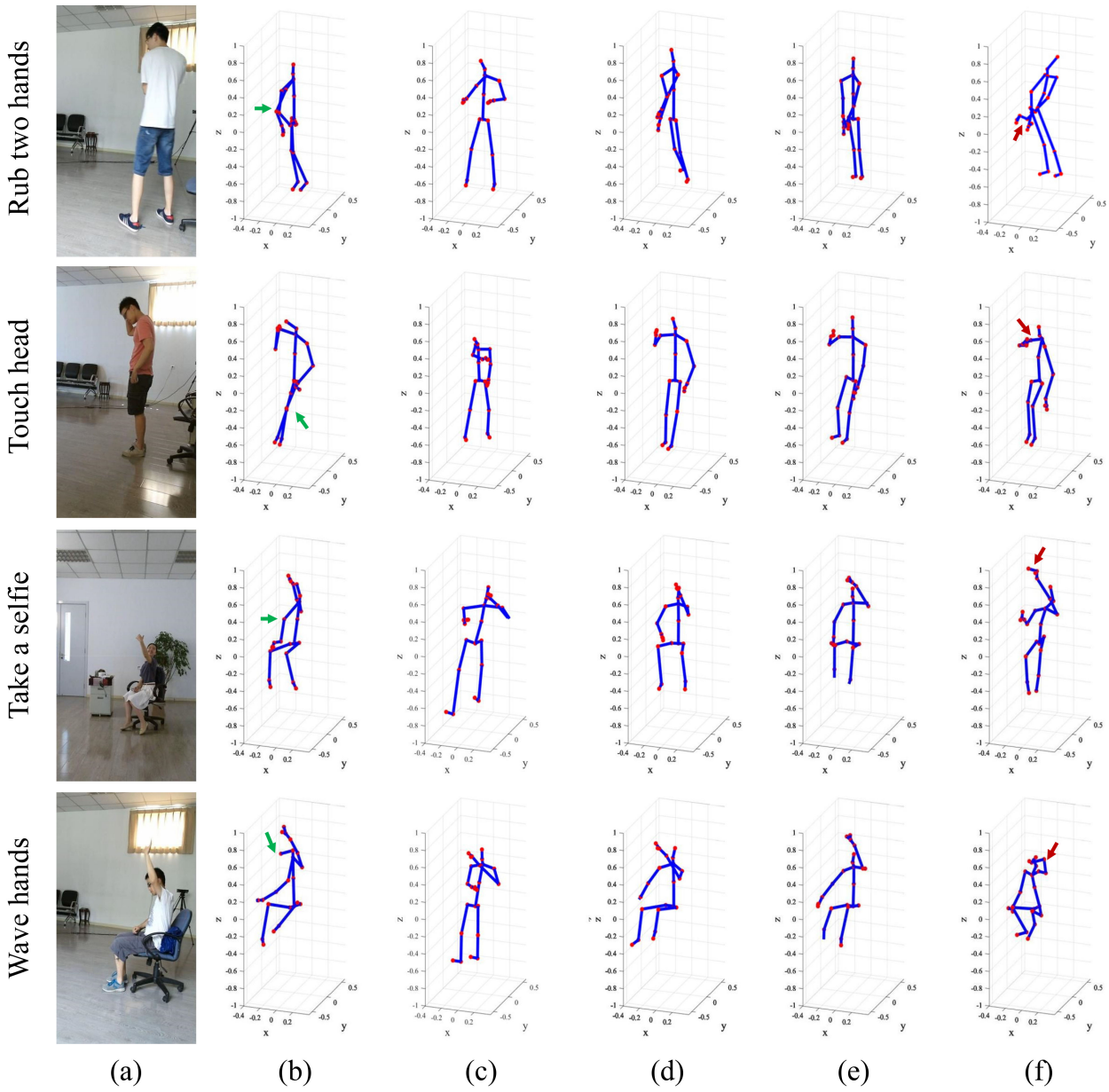


Fig. 7. (a) RGB images, (b) original noisy skeletons, (c) denoised results by R-SD, (d) denoised results by PE-AE [17], (e) denoised results G-SD, (f) adapted results by G-NAN. The skeletons are slightly rotated around their torsos for better visualization. Though R-SD, PE-AE and G-SD correct noisy skeletal joints marked by green arrows, our adapted results show more discriminative representations for action recognition marked by red arrows.

same number of parameters as G-NAN. The higher performance from G-NAN illustrates the effectiveness of noise adaptation.

In addition, experiments with more complex structures for R-NAN are conducted. We replace the GRU unit in Fig. 3(a) with the graph convolutional network (GCN) [10]. The results are shown in Table VI. It is observed with more complex structure, we obtain better performance, illustrating that our noise adaptation is generalizable and flexible to other structures.

TABLE VI
COMPARISONS WITH DIFFERENT STRUCTURES FOR R-NAN

Model Structure	Method	CS	CV
GRU	Baseline	50.7	34.6
	R-NAN	55.3	40.5
GCN	Baseline	53.6	39.4
	R-NAN	**56.0**	**41.8**

3) Noise Adaptation vs. *More Accurate Skeletons:* We further explore how much noise adaptation can compensate for the performance degradation caused by skeleton noise, compared with the action recognition performance when we have more accurate skeletons or even ground-truths. To investigate the problem, we analyze the action recognition performance from each viewpoint under the cross-subject setting of our NSD dataset. Note that the skeletons from different viewpoints are captured simultaneously and thus they can be regarded as different measurements of certain action sequences. Since there is less occlusion when recorded from the front view, the skeletons are more accurate than those from the side views. It is also

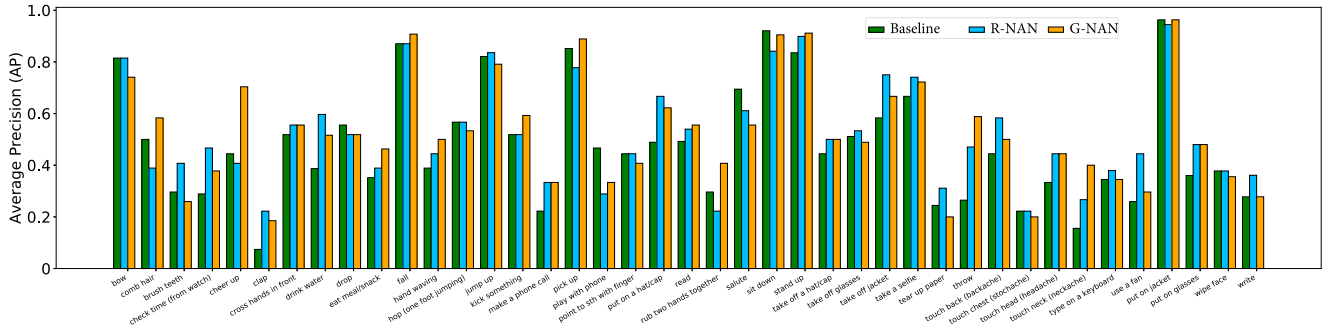


Fig. 8. Average precision on each action category of baseline, R-NAN and G-NAN on the NSD (CS) dataset.

TABLE VII
PERFORMANCE COMPARISONS IN TERMS OF ACCURACY (%) FOR DIFFERENT VIEWPOINTS IN NSD-CS

Methods	FV	SV-1	SV-2
Baseline	60.9	47.1	44.2
R-NAN	62.9	51.4	54.7
G-NAN	63.1	52.9	50.3

consistent with a higher baseline performance from front-view data in Table VII. With noise adaptation, we obtain larger improvement for action recognition from side-view data, that the performance is improved by 4.3%-10.5% for side views and 2.0%-2.2% for front view. It indicates that, though more accurate skeletons will definitely lead to higher action recognition, our proposed noise adaptation is able to largely compensate for feature embeddings encoded from noisy skeletons, and then mitigate the performance degradation caused by skeleton noise.

4) Classification Analysis: We analyze our adaptation by investigating their performance on each action category. We show the average precision on each action category of baseline, R-NAN, and G-NAN on the NSD dataset under the cross-subject split in Fig. 8. R-NAN and G-NAN are able to enhance the recognition for most of the action categories. For R-NAN, we found recognition performance degradation for some action categories, such as *rubbing two hands together* and *playing with phone*. It is mainly because these two actions look similar, especially when there is noise. R-NAN tends to encode them into similar feature embeddings after regularizing the feature spaces but makes the classifier confused. The performance of *combing hair* and *saluting* is degraded for the same reason. For G-NAN, it is found that the performance of recognizing actions *brushing teeth* and *touching chest* is worse than the baseline. The actors would be occluded by themselves whenever performing these actions. With adversarial learning, the network is prone to regard these skeletons as noisy ones and then adapt them. As a consequence, the action patterns are not well preserved for these actions.

5) Results on Relatively Clean Skeletons: Finally, to further illustrate the ability of our model, we also test our method on a general action recognition dataset, the NTU RGB-D dataset (NTU) [11]. We use 256 units for each BiGRU layer. Table VIII shows the results. Thanks to noise adaptation, our models consistently outperform the baseline results for both settings. We also compare with other state-of-the-art methods. With comparable number of parameters, our models show superiority in the action recognition performance even for relatively clean skeleton data.

TABLE VIII
PERFORMANCE COMPARISON IN TERMS OF ACCURACY (%) ON THE NTU DATASET

Methods	#Params	CS	CV
STA-LSTM [7]	0.5M	73.4	81.2
VA-LSTM [24]	0.5M	79.4	87.6
ST-GCN [10]	3.1M	81.5	88.3
SR-TSL [25]	19.1M	**84.8**	**92.4**
Baseline	1.8M	79.4	86.1
R-NAN	1.8M	81.9	88.2
G-NAN	5.6M	81.0	86.3

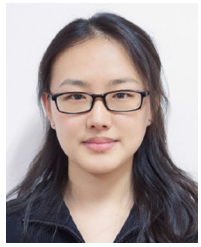
VI. CONCLUSION

In this paper, we study the problem of action recognition from noisy skeleton data, which is seldom explored by previous methods. We propose noise adaptation networks (NAN) to get rid of explicit noise modeling and reliance on ground-truths. To mitigate the performance degradation caused by skeleton noise, we explore the regression-based adaptation model for paired noisy skeletons and the generation-based adaptation model for unpaired noisy skeletons, respectively. The regression model aims to learn noise-robust feature representations by mapping the paired noisy skeletons into a common space. The generation-based model aims to suppress noise into a low-noise feature space by adversarial learning. We analyze our model by conducting comprehensive experiments on the NSD dataset collected by us, the VV-RGBD dataset and the N-UCLA dataset, respectively. Experiments show our proposed models consistently and significantly outperform other approaches.

REFERENCES

[1] D. Weinland, R. Ronfard, and E. Boyerc, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.
[2] G. Johansson, "Visual perception of biological motion and a model for it is analysis," *Percep. Psychophys.*, vol. 14, no. 2, pp. 201–211, 1973.
[3] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.
[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.
[5] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1110–1118.

[6] W. Zhu *et al.*, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3697–3703.
[7] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4263–4270.
[8] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3 d action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4570–4579.
[9] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 786–792.
[10] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.
[11] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB d: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1010–1019.
[12] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 28–35.
[13] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 20–27.
[14] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5344–5352.
[15] Q. Nie, J. Wang, X. Wang, and Y. Liu, "View-invariant human action recognition based on a 3D bio-constrained skeleton model," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3959–3972, Aug. 2019.
[16] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, 2017.
[17] G. G. Demisse, K. Papadopoulos, D. Aouada, and B. Ottersten, "Pose encoding for robust skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 188–194.
[18] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 588–595.
[19] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1290–1297.
[20] R. Vemulapalli and R. Chellapa, "Rolling rotations for recognizing human actions from 3D skeletal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4471–4479.
[21] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Comput. Vis. Image Understanding*, vol. 158, pp. 85–105, 2017.
[22] X. Cai, W. Zhou, L. Wu, J. Luo, and H. Li, "Effective active skeleton representation for low latency human action recognition," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 141–154, Feb. 2016.
[23] Z. Fan, X. Zhao, T. Lin, and H. Su, "Attention-based multiview re-observation fusion network for skeletal action recognition," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 363–374, Feb. 2019.
[24] P. Zhang *et al.*, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2117–2126.
[25] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 103–118.
[26] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2649–2656.
[27] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.
[28] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.
[29] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 343–351.
[30] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[31] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.
[32] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.
[33] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8503–8512.
[34] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 469–477.
[35] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–15.
[36] J. Lehtinen *et al.*, "Noise2noise: Learning image restoration without clean data," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2971–2980.
[37] N. Moran, D. Schmidt, Y. Zhong, and P. Coady, "Noisier2noise: Learning to denoise from unpaired noisy data," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, 2020, pp. 12064–12072.
[38] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2223–2232.
[40] J. Chen, J. Chen, H. Chao, and M. Yang, "Image blind denoising with generative adversarial network based noise modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3155–3164.
[41] K. K. Thekumparampil, A. Khetan, Z. Lin, and S. Oh, "Robustness of conditional GANs to noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10271–10282.
[42] H. Zhou, J. Sun, Y. Yacoob, and D. W. Jacobs, "Label denoising adversarial network (LDAN) for inverse lighting of faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6238–6247.
[43] T. Kaneko, Y. Ushiku, and T. Harada, "Label-noise robust generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2467–2476.
[44] X. Mao *et al.*, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2794–2802.
[45] Y. Ji *et al.*, "A large-scale varying-view RGB-D action dataset for arbitrary-view human action recognition," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1510–1518.
[46] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.
[47] Y. Hu, C. Liu, Y. Li, S. Song, and J. Liu, "Temporal perceptive network for skeleton-based action recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.
[48] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12 026–12 035.
[49] P. Zhang *et al.*, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1112–1121.
[50] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 156–165.
[51] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1623–1631.



Sijie Song (Student Member, IEEE) received the B.S. and Ph.D. degrees (Hons.) in computer science from Peking University, Beijing, China, in 2016 and 2021, respectively. Her research interests include computer vision and image processing.

Jiaying Liu (Senior Member, IEEE) received the Ph.D. degree (Hons.) in computer science from Peking University, Beijing China, in 2010. She is currently an Associate Professor, Peking University Boya Young Fellow with the Wangxuan Institute of Computer Technology, Peking University. From 2007 to 2008, she was a Visiting Scholar with the University of Southern California, Los Angeles, Los Angeles, CA, USA. She has authored more than 100 technical articles in refereed journals and proceedings, and holds 50 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She was a Visiting Researcher with the Microsoft Research Asia in 2015 supported by the Star Track Young Faculties Award. She was a Member of Multimedia Systems & Applications Technical Committee (MSA TC), and Visual Signal Processing and Communications Technical Committee (VSPC TC) in IEEE Circuits and Systems Society. She was the recipient of the IEEE ICME-2020 Best Paper Awards and IEEE MMSP-2015 Top10% Paper Awards. She was also an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *Elsevier Journal of Visual Communication and Image Representation*, the Technical Program Chair of IEEE ICME-2021/ACM ICMR-2021, the Publicity Chair of IEEE ICME-2020/ICIP-2019, and the Area Chair of CVPR-2021/ECCV-2020/ICCV-2019. During 2016–2017, she was the APSIPA Distinguished Lecturer. Dr. Liu is a Senior Member of CSIG and CCF.

Lilang Lin received the B.S. degree in 2021 in data science from Peking University, Beijing, China, where he is currently working toward the Ph.D. degree with the Wangxuan Institute of Computer Technology. His research interests include computer vision and self-supervised learning.

Zongming Guo (Member, IEEE) received the B.S. degree in mathematics, and the M.S. and Ph.D. degrees in computer science from Peking University, Beijing, China, in 1987, 1990, and 1994, respectively. He is currently a Professor with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include video coding, processing, and communication. Dr. Guo is the Executive Member of the China-Society of Motion Picture and Television Engineers. He was the recipient of the First Prize of the State Administration of Radio Film and Television Award in 2004, the First Prize of the Ministry of Education Science and Technology Progress Award in 2006, the Second Prize of the National Science and Technology Award in 2007, the Wang Xuan News Technology Award and the Chia Tai Teaching Award in 2008, the Government Allowance granted by the State Council in 2009, and the Distinguished Doctoral Dissertation Advisor Award of Peking University in 2012 and 2013.