Cross-Subject and Cross-Modal Transfer for Generalized Abnormal Gait Pattern Recognition

Xiao Gu, Yao Guo, Fani Deligianni, Benny Lo, and Guang-Zhong Yang*, *Fellow, IEEE*

***Abstract*—For abnormal gait recognition, pattern-specific features indicating abnormalities are interleaved with the subject-specific differences representing biometric traits. Deep representations are therefore prone to overfitting and the models derived cannot generalize well to new subjects. Furthermore, there is limited availability of abnormal gait data obtained from precise Motion Capture (Mocap) systems because of regulatory issues and slow adaptation of new technologies in health care. On the other hand, data captured from markerless vision sensors or wearable sensors can be obtained in home environments but noises from such devices may prevent effective extraction of relevant features. To address these challenges, we propose a cascade of deep architectures that can encode cross-modal and cross-subject transfer for abnormal gait recognition. Cross-modal transfer maps noisy data obtained from RGBD and wearable sensors to accurate four-dimensional (4D) representations of the lower limb and joints obtained from the Mocap system. Subsequently, cross-subject transfer allows to disentangle subject-specific from abnormal pattern-specific gait features based on a multi-encoder autoencoder architecture. To validate the proposed methodology, we obtained multi-modal gait data based on a multi-camera motion capture system along with synchronized recordings of Electromyography (EMG) data and 4D skeleton data extracted from a single RGBD camera. Classification accuracy was improved significantly in both Mocap and noisy modalities.**

***Index Terms*—gait analysis, model generalization, body sensor network, multi-modal representation.**

I. Introduction

ONE of the key characteristics of Deep Neural Networks (DNN) is their ability to automatically extract relevant features from a large amount of complex data. This is important in the analysis of health informatics as conventional machine learning algorithms depend on explicit features for related applications [1], [2]. Potentially meaningful information is ignored and the ability to learn and generalize from abstract data is limited.

However, in health informatics, datasets are usually small and complex [1]. In particular, the four-dimensional (4D) human motion data (xyz-t) encodes information in both the spatial and temporal dimensions. This complex information reflects both subject-specific as well as motion-specific/pathological information. Human motion analysis and in particular gait analysis play an important role in several biomedical applications, including pathological gait detection [3], [4], [5], rehabilitation assistance [6] and emotion recognition [7]. Training DNNs for abnormal gait recognition requires precise estimation of the lower limbs and joint kinematics, which normally is only available via highly specialized Motion Capture (Mocap) Systems. These systems are state-of-the-art technologies for tracking 3D human skeleton by recording highly precise 3D locations of reflective markers attached to some key points of the human body. The encoded kinematic data, containing subtle changes of the lower limbs, can provide abnormality indicators in healthcare applications [8], [9].

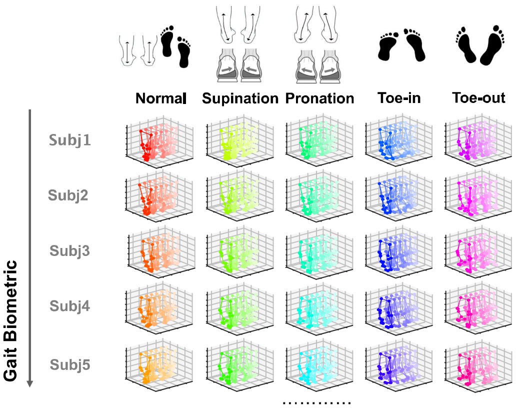


Fig. 1. A gait cycle across subjects and walking patterns. Disentangling pattern characteristics from gait biometrics is a major challenge in abnormal gait detection.

Furthermore, automated abnormal gait recognition based on wearable and vision sensors can play a prominent role in the so-called connected health model because it facilitates objective gait assessment in home environments. Those sensing technologies provide a more convenient and comfortable solution for pervasive monitoring [10]. However, these modalities result in noisy representations, which can largely reduce the generalization of DNNs in new subjects [4], [5].

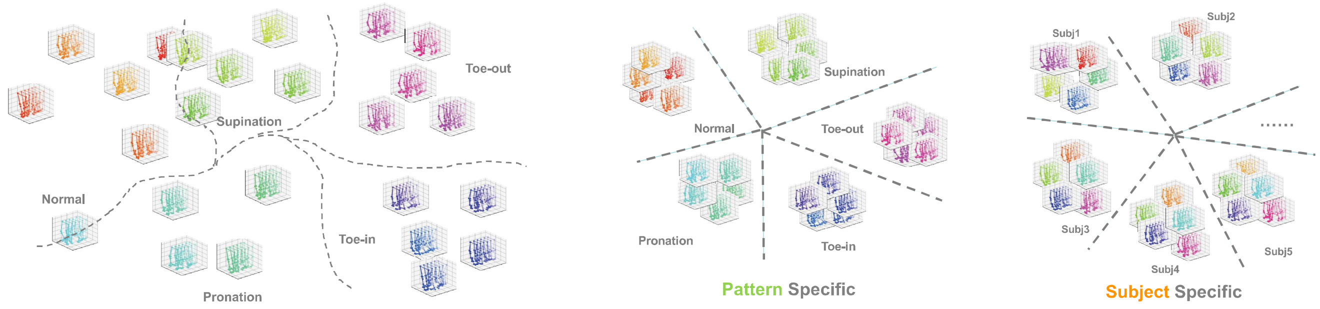
One of the main challenges in abnormal gait recognition is that the pattern-specific features representing abnormal gait are often more subtle compared to the subject-specific differences. Subject-specific differences highlight the biometric traits existing in the gait data representations [11], [12], [13]. As an example[1] shown in Fig. 1, each subject tends to exhibit unique walking characteristics due to their age, gender, weight, bone

X. Gu, Y. Guo, and B. Lo are with the Hamlyn Centre, Institute of Global Health Innovation, Imperial College London, London SW7 2AZ, UK (e-mail: {xiao.gu17, yao.guo, benny.lo}@imperial.ac.uk).

F. Deligianni is with the School of Computing Science, University of Glasgow, Glasgow G12 8RZ, UK (e-mail: fani.deligianni@glasgow.ac.uk).

G.-Z. Yang is with the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: gzyang@sjtu.edu.cn).

[1] The visualization is based on data in Section IV. The animation effect can be viewed on https://guxiao0822.github.io/GAGR



(a) Entangled Representations

(b) Disentangled Representations

Fig. 2. Schematics of entangled representations and disentangled representations for normal and four abnormal gait patterns.

length proportion, and geometric shape of lower extremities. Hence, the factors affecting the way a person walks involve both gait abnormality and inherent biometric traits. These two specific data representations, as shown in Fig. 2(a), would lend information to each other when training a classifier for identity recognition or abnormal detection, thus impeding the generalization of the classifier to new subjects. This is reflected by the fact that the accuracy of intra-subject classification is high, whereas the accuracy is largely decreased in cross-subject classification [5]. Although, the subject-specific differences could be eliminated automatically if the training dataset is large, the acquisition of large-scale healthcare data is usually limited due to several reasons, including rare abnormalities, patient compliance, ethics, and slow adaptation of emerging technologies in healthcare [1].

For accurate abnormal gait classification, this paper firstly proposes a multi-encoder autoencoder network to automatically split the raw 3D kinematic data into two independent parts, i.e., the **pattern-specific** and the **subject-specific** representations (shown in Fig. 2(b)). This kind of architecture enables the disentanglement of the pattern-specific and subject-specific representations by minimizing a cross reconstruction loss when transferring the patterns across two different subjects. The normal and abnormal walking patterns can be converted while preserving the identity characteristics during the **cross-subject transfer**. Subsequently, the pattern-specific features that are subject-invariant, are extracted for abnormal pattern recognition. Results demonstrate significantly improved performance on cross-subject validation compared to conventional deep architectures, highlighting the good generalization ability of the trained model to new subjects.

Furthermore, we explore **cross-modal transfer** enabling the data representation transfer across modalities, not only to address the limited availability of accurate pathological human pose data but also to derive kinematic information from noisy skeleton estimated by RGBD sensing data and electrocardiogram (EMG). The principle behind cross-modal common representation learning is to retrieve the common representation that links two modalities. Here, modalities reflect sensing technologies applied to gait analysis [9], [10]. We aim to transfer the clean representation knowledge from accurate 4D human pose Mocap data to noisy RGBD and EMG data.

There is a large heterogeneity gap between the Mocap skeleton and the EMG signals. EMG records the individual muscle response that reflect the intention of someone to walk. Previous work [14] has demonstrated its ability to estimate the lower limb kinematics. However, raw EMG signals across subjects are inevitably affected by several factors, including motion artifacts, skin-electrode interface and cross-talk, which results in the large contamination [15]. Furthermore, individual differences are enlarged by the different displacement of sensors and muscle conditions across trials and subjects.

On the other hand, RGBD data can model abnormal human gait based on human pose reconstruction. Recent efforts aim to improve accuracy of 3D skeleton tracking of the lower limbs based on a single RGBD camera [5]. However, their accuracy and robustness is notably lower than the infrared-based motion capture systems (Mocap) that are available in specialized clinics and labs. Extracting 3D human pose information from a single RGB camera is an under-determined problem [16]. RGBD sensors provide additional depth information but it has low resolution and large noise that is also distance dependent [17]. Furthermore, the human pose estimation model [18] utilized in most literature is different from the model applied in Mocap settings, in terms of keypoint locations and numbers. Therefore, it is not ideal or possible to directly apply the model trained on Mocap data to the skeleton extracted from RGBD cameras.

The heterogeneity gap between different modalities also suffers from domain shift problems. In our case, the association between modalities of the training subjects may well be different from the testing subjects, where each subject can be viewed as a domain. However, most of existing works are only focused on either across-subject or multi-modality problems [19]. The combination of those two issues is a relatively new area. We propose a multi-modal representation learning method to enable cross-subject and cross-modal transfer. With the accurate Mocap data as the target to be reconstructed, the skeletal data estimated from RGBD cameras and EMG signals are mapped to this clean representation. In this way, we have developed a mechanism to filter noises from EMG and RGBD data and enhance abnormal gait classification performance.

The whole framework is visualized in Fig. 3. In summary, the main contribution of this paper is three-fold:

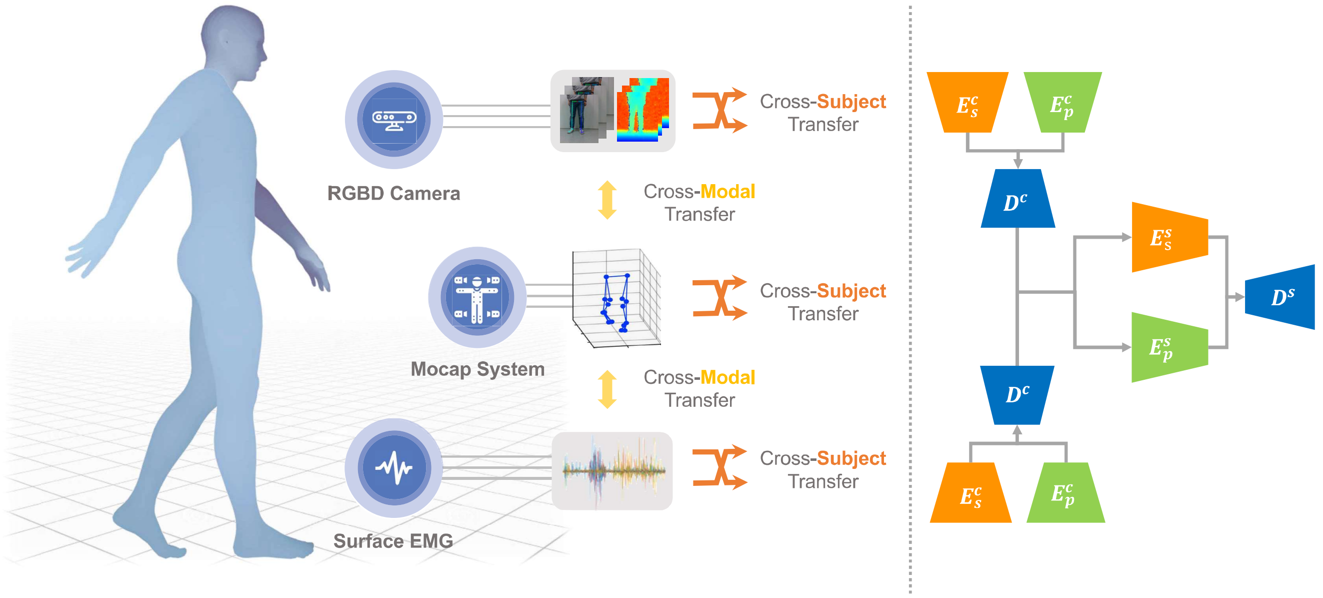


Fig. 3. Illustration of the cross-subject and cross-modal transfer framework in abnormal gait recognition, where E and D indicate the encoder and decoder. The subscripts s and p represent the encoder for encoding **s**ubject-specific and **p**attern-specific features, and the superscripts s and c indicate **s**ingle-modal and **c**ross-modal, respectively. This method is proposed to enhance the generalization of model representations through cross-subject transfer, and find common representations between 4D skeletal Mocap and noisy RGBD skeleton or surface EMG data through cross-modal transfer.

*a) **Single-Modal – Cross-Subject Transfer**:* To overcome the limited availability of labelled clinical gait datasets of the lower limb kinematics, we propose a cross-subject generalization architecture that can effectively extract the subject-invariant gait pattern features. This improves the classification performance of abnormal gait recognition with high-quality Mocap data.

*b) **Multi-Modal – Cross-Modal Transfer**:* To allow generalization of the above framework in noisy skeleton data obtained from a single RGBD sensor we propose a novel cross-modal common representation learning approach. This maps the noisy data to the Mocap kinematic data, thus facilitating a filtering process. We demonstrate that this, subsequently, enables subject-invariant features to be extracted from the mapped clean representations, successfully. Furthermore, we show that this framework can apply also to other gait sensing signals of inherently different nature, such as EMG.

*c) **Generalized Gait Analysis**:* A novel framework is proposed to ensure generalization of our method for gait analysis. The work represents the first attempt of subject-independent assessment of abnormal gait based on DNN.

II. Related Works

A. Deep Learning Based Gait Analysis

For pathological gait analysis, a diversity of methods and gait models have been explored to investigate effective feature extraction from the recorded skeletons, such as quantitative gait parameters [20], joint angles [5], joint motion history features [21], and other deep representations [22]. Conventional methods manually select important clinically-relevant features for classification. Different gait models with pre-defined features are accordingly proposed to facilitate the diagnosis of different diseases [8]. However, they are inherently limited as the pre-selected features, in most cases, cannot represent an adequate gait description and might ignore meaningful information from the abstract data [23]. With the advancement of deep neural networks (DNNs) both convolutional neural networks (CNN) and recurrent neural networks (RNN) have been proposed in gait analysis and they are briefly described below.

a) Convolutional Neural Network: In [24], the authors applied CNN on the data from wearable sensors to extract gait parameters. It directly translated the signals from wearable sensors to context-related features. Kim *et al.* [25] utilized the temporal CNN for the 3D motion recognition based on one-dimensional (1D) convolution. The proposed model performs convolution operation in the temporal axis while fully connecting the feature dimensions. A more advanced variant, the graph convolutional neural network (GCNN) has been applied in a recent gait analysis study [26]. The graph convolution focuses on the relationship within highly connected nodes and it can extract successfully the spatial information from human skeletons.

b) Recurrent Neural Network: RNN is capable of modeling and predicting time series signals, especially for signals of varied length. It has been applied on gait data with respect to various applications [27], [28], [29]. Kidziński *et al.* [30] adopted a Long-Short-Term-Memory (LSTM) network to develop an online gait phase estimation method based on the input joint angles and positions information. Liu *et al.* [31] also applied a similar LSTM to predict the occurrence of abnormal knee joint trajectory and then a wearable assistance tool was actuated for rehabilitation training. In [27], the authors implemented automatic gait recognition in the wild based on LSTM and embedded inertial sensors of smart phones. Compared to CNN, the recurrent neural network is harder and slower to train because of its folded architecture [32]. Although recent works on the further development or the combinations of RNN and CNN have demonstrated promising results [27], [29], their work is not suitable for cross subject generalization.

B. Cross-Subject Human Motion Retargeting

Disentangling subject-specific from pathology specific gait patterns is related to human motion synthesis and retargeting problems. Various kinds of architectures, like Conditional Restricted Boltzmann Machine [33], Encoder-Recurrent-Decoder [34] and Auto-Conditioned Recurrent Neural Networks [35] have been proposed for modeling, generating and predicting human motion data. As a research topic in computer graphics area, the human motion retargeting focuses on transferring movement across subjects (virtual or real) while optimizing the spatial-temporal information by preserving the subject-specific features.

Early works tried to apply the inverse kinematics on this issue [36]. However, it has been argued that the walking patterns cannot get well transferred across subjects by simply copying joint kinematics between each other because it causes unrealistic movements [37]. To address this, Villegas *et al.* [37] introduced a recurrent neural network with neural forward kinematics layer. It enables realistic joint rotations transformation across subjects through cycle consistency based adversarial training, and then the subject-specific motion is generated by the prior-known basic skeleton and the kinematics. Yan *et al.* [38] adopted a variational recurrent auto-encoder to perform the multi-modal motion transformation from one to another. The embedding feature was generated by the variational auto-encoder and it was further enhanced by motion retargeting. Recently, Aberman *et al.* [39] encoded the human motion data into a dynamic character-agnostic latent motion representation, along with static latent components. It is successful in generating character-agnostic motion. Another work [40] on image synthesis was able to generate novel images of human unseen poses. Our work is inspired by the above but more focused on effectively disentangling the pathology and biometrics related features by cross-subject pattern transfer/retargeting.

C. Cross-Modal Common Representation Learning

Given a specific task, different categories of modalities or media can be associated to that. Based on the association between different modalities, the common representation can be retrieved to enable the knowledge transfer across modalities. The general common representation learning methods can be classified into three groups, the joint representation, coordinated representation and encoder-decoder architectures [41]. The joint representation learning integrates the multi-modal data to draw a complete picture by fusing the complementary information from different modalities. Several advanced information fusion strategies have been proposed to address this [42]. The coordinated representation learning aims to enlarge the similarity or correlation of the representations of different modalities, so that the shared representation subspace can be derived. Salvador *et al.* [43] proposed the semantic regularization for a joint neural embedding model in order to get the embedding space between food images and cooking recipes. Finally, the encoder-decoder framework has gained popularity because it can capture the shared representations for a specific task by learning the mapping between two different modalities [44]. The multi-modal representation learning has been applied in several common natural language processing and computer vision applications in the literature [45], [46].

The heterogeneity gap between different modalities may also suffer from domain shift problem. In our case, the relationship between different sensing modalities would differ on each subject. To bridge the heterogeneity gap between different modalities and address the domain shift problems, Huang *et al.* [47] proposed a hybrid transfer network for transferring knowledge from single-modal source domain to cross-modal target domain, enabling knowledge sharing between texts and images for media retrieval.

Previous work [48] on gait recognition has successfully fused inertial and RGBD sensors together to achieve robust person identification. However, it is focused on the multi-modal sensor fusion in a feature level based on classical machine learning approaches. This is different from our work, where we introduce transfer knowledge across different modalities and thereby enhance the performance of a signal noisy modality for pathological gait analysis applications.

D. Domain Generalization

Our work reported here is related to domain generalization. It aims to aggregate the knowledge from multiple source domains and then learn to extract a shared common subspace that can be transferred to an unseen target domain. By taking each subject or modality as a specific domain, increasing the generalization ability across subjects or modalities can be viewed regarded as a domain generalization task [49], [50], [51]. This is different from and less explored than domain adaptation [52], which mostly focuses on one source domain and takes advantage of target domain information to align domain distribution.

Domain generalization does not require data from the ‘unseen’ target domain during training. Instead, it eliminates domain-specific bias from multiple ‘seen’ source domains. Muandet *et al.* [53] developed a kernel-based optimization method to transform data from different domains to a canonical space while preserving the functional relationship between the input and output. Li *et al.* [54] proposed a variational autoencoder based method, incorporating the Maximum Mean Discrepancy (MMD) loss to align the embedding domain distribution. In [55], the authors developed a Siamese architecture to learn an embedding subspace across domains that is discriminative. Li *et al.* [56] proposed a episodic training strategy for domain generalization, which mismatches the feature extractor and classifier across different domains to achieve robust performance.

III. Methods

A. Gait Representation

Gait is represented as repetitive cycles of motion denoted as $\{x_{s_i}^{p_j}\} \in \mathbb{R}^{T \times J_x}$, where $\mathbf{x}$ refers to the gait cycle data. s, p denotes the subject identity $s \in \{s_i\}_{i=1}^{N_s}$ and gait pattern categories $p \in \{p_j\}_{j=1}^{N_p}$, respectively. T denotes the resampling number in a gait cycle, and J_x reflects the dimensionality of the modality. For clarity, $\mathbf{x}_{s_i}^{p_j}$ is written as $\mathbf{x}_i^j$ in the following.

B. Problem Formulation

For the single-modal representation, we assume that gait cycle is composed of the biometric and abnormality characteristics along with noise n and we denote this as

$$\mathbf{x}_i^j = \mathcal{F}(s_i, p_j; n) \tag{1}$$

It is worth noting that the latent vectors e_{s_i} & e_{p_j} representing identity and pathology pattern are directly denoted as s_i and p_j respectively for clarity.

Our goal is to learn an inverse model $\mathcal{F}^{-1}$ that takes apart the subject-specific s_i and pattern-specific p_j from raw data representations, resolving the entanglement of these two independent factors. If the gait data $\mathbf{x}$ is a clean representation (e.g., the skeletal data captured by Mocap system), the noise can be ignored. Then (1) can be rewritten as

$$\mathbf{x}_i^j = \mathcal{F}(s_i, p_j) \tag{2}$$

Thus the derived pattern-related latent vector p_j is expected to be subject-invariant, which can additionally generalize on new subjects.

On the other hand, noisy data $\mathbf{z}$, such as the data derived from the RGBD camera and EMG data derived from wearable sensors, is inherently contaminated with several sources of noises. In these cases, the subject-specific and pattern-specific features could not be directly disentangled without considering the noise effects. Hence, the goal of cross-modal transfer learning is to derive the pattern-specific features from the noisy modality representation $\mathbf{z}$ with the help of the clean or less-noisy modality representation $\mathbf{x}$. This is formulated as

$$\mathcal{G}^x(\mathbf{z}_i^j) = \mathcal{F}(s_i, p_j) \tag{3}$$

where $\mathcal{G}^x(\cdot)$ represents a mapping from the noisy gait data representation to the clean/ground-truth data $\mathbf{x}$.

C. Multi-Encoder Autoencoder Network

An autoencoder like architecture is adapted to achieve cross-subject tranfer learning, as shown in Fig.4. Inspired by recent work introducing multiple encoders for disentangled representations learning [39], [57], we split the encoder part into the subject-specific and the pattern specific branches, respectively. The subject-specific branch E_s encodes the features s_i related to human biometrics, $E_s : \mathbf{x}_i \mapsto s_i$, and the patient-specific branch E_p encodes the features representing pathological patterns p_j, $E_p : \mathbf{x}^j \mapsto p_j$. In the decoding procedure, the s_i and p_j are concatenated together as the input of the decoder to perform the reconstruction of $\mathbf{x}_i^j$, $D : (s_i, p_j) \mapsto \hat{\mathbf{x}}_i^j$. This kind of architecture enables to draw a complete picture of human gait data by the two disentangled representations. In practice, the encoder and decoder were realized by 1D convolution across the time dimension [39].

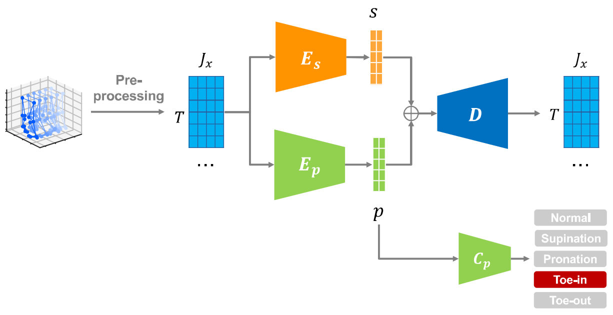


Fig. 4. Basic architecture of the multi-encoder autoencoder. The whole architecture can be directly applied on clean Mocap data.

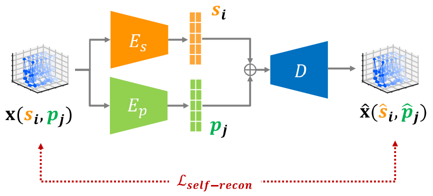

E_s
E_p
s_i
p_j
D
$\mathbf{x}(s_i, p_j)$
$\hat{\mathbf{x}}(\hat{s}_i, \hat{p}_j)$
$\mathcal{L}_{self-recon}$

Fig. 5. Illustration of self-reconstruction loss. The $\mathcal{L}_{self-recon}$ loss is minimized to ensure the reconstructing a given input by the autoencoder.

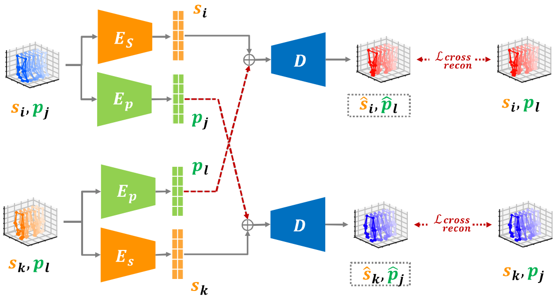


Fig. 6. Illustration of cross-subject reconstruction loss. Based on the given sample x_i^j and x_k^l, the latent representations s_i, p_j, s_k, p_l are extracted individually firstly. Then the patterns are transferred across subjects and the combinations s_i, p_l and s_k, p_l are put into the decoder separately to reconstruct $\hat{x}_i^l$ and $\hat{x}_k^j$. The x_i^l and x_k^j are indexed from existing data and the reconstruction loss between $\{x_i^l, \hat{x}_i^l\}$, $\{x_k^j, \hat{x}_k^j\}$ are minimized respectively.

D. Single-Modal: Self-Reconstruction and Cross-Subject Reconstruction

For a single clean modality representation, a multi-encoder autoencoder network is applied, as shown in Fig. 4. In order to learn the complete and disentangled subject and pattern representations, three sets of loss functions $\{\mathcal{L}_{self-recon}$; $\mathcal{L}_{cross-recon}$; $\mathcal{L}_{trip}^s$ & $\mathcal{L}_{trip}^p\}$ are minimized for optimization.

1) Self-Reconstruction: The objective of self-reconstruction is to realize the basic function of this multi-encoder autoencoder, the identity mapping. As illustrated in Fig. 5, the model

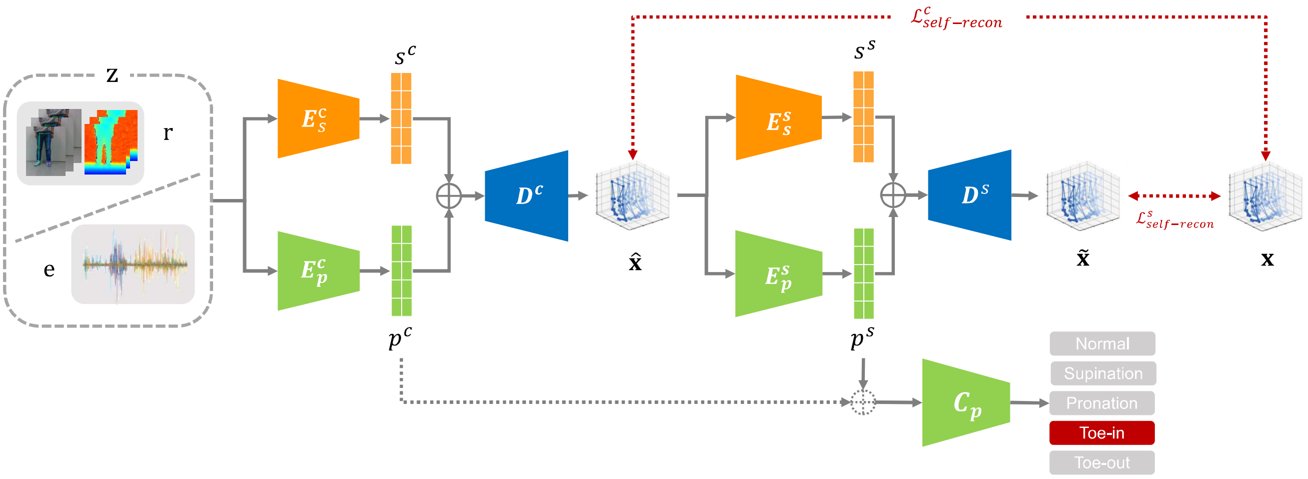


Fig. 7. Illustration of cross-modal cross-subject transfer model. The data from noisy modalities $\mathbf{z}$, namely the skeleton from RGBD ($\mathbf{r}$) or the EMG data ($\mathbf{e}$), is firstly converted to the Mocap representation $\hat{\mathbf{x}}$ to facilitate the data filtering. This first cross-modal generation part is structured as a multi-encoder autoencoder, composed of $\{E^c_s, E^c_p \text{ \& } D^c\}$. The superscript c refers to cross-modal. The encoders transform data $\mathbf{z}$ into subject-specific s^c and pattern-specific p^c features respectively. p^c represents the associated pattern-specific feature between $\mathbf{z}$ and $\mathbf{x}$, which might get contaminated by noises. Subsequently the generated sample $\hat{\mathbf{x}}$ goes through the second multi-encoder autoencoder network $\{E^s_s, E^s_p \text{ \& } D^s\}$, being reconstructed to $\tilde{\mathbf{x}}$. The superscript s refers to single-modal. This further refines the generated samples $\hat{\mathbf{x}}$ by minimizing the loss between $\tilde{\mathbf{x}}$ and $\mathbf{x}$. The related encoders divide the embedding features into the subject-specific s^s and pattern-specific p^s ones as well. p^s represents subject-invariant pattern-specific features from a cleaner data representation. Subsequently, the pattern feature p^s (optionally concatenated with p^c) is fed into a classifier C_p for abnormal gait recognition. (The cross reconstruction is not sketched for simplification.)

is expected to learn how to reconstruct the input given a gait cycle sample $\mathbf{x}^j_i$. The loss function is formulated as follows,

$$\mathcal{L}_{self-recon} = \mathbb{E}[\|\mathbf{x}^j_i - D(s_i, p_j)\|] \tag{4}$$

2) Cross-Subject Reconstruction: However, features encoded by E_s and E_p cannot be well split by only using a simple autoencoder. To separate subject-specific and pattern-specific data representations, the cross reconstruction procedure is introduced during training, as shown in Fig. 6. During each training step, the training group of samples $\{\mathbf{x}^j_i, \mathbf{x}^l_k, \mathbf{x}^l_i\}$ $(i \neq k)$ is selected, and the pattern p_l in subj k ($\mathbf{x}^l_k$) is transferred to subj i to constitute $\hat{\mathbf{x}}^l_i$ in our cross reconstruction strategy. The loss function is as follows,

$$\mathcal{L}_{cross-recon} = \mathbb{E}[\|\mathbf{x}^l_i - D(E_s(\mathbf{x}^j_i), E_p(\mathbf{x}^l_k))\|] \tag{5}$$

In addition to the cross reconstruction loss, similar to [39], we adopted the triplet loss $\mathcal{L}^s_{trip}$ & $\mathcal{L}^p_{trip}$ to help the s_i and p_j get clustered to their own categories more tightly, where α is the margin constraining intra-class features have smaller distance ($\leqslant \alpha$) than inter-class features. The loss functions are formulated as follows.

$$\begin{aligned}\mathcal{L}^s_{trip} &= \mathbb{E}[\|E_s(\mathbf{x}^j_i) - E_s(\mathbf{x}^m_i)\| - \|E_s(\mathbf{x}^j_i) - E_s(\mathbf{x}^n_k)\| + \alpha] \\ \mathcal{L}^p_{trip} &= \mathbb{E}[\|E_p(\mathbf{x}^j_i) - E_p(\mathbf{x}^j_m)\| - \|E_p(\mathbf{x}^j_i) - E_p(\mathbf{x}^k_n)\| + \alpha]\end{aligned} \tag{6}$$

3) Classifier: In the recognition stage, the parameters of the autoencoder model are fixed and the pattern features are extracted from E_p. This feature vector is then fed into a classifier C_p to be trained for abnormality recognition. The loss function $\mathcal{L}_C$ applied here is the Cross Entropy Loss.

$$\mathcal{L}_C = -\sum_{\mathcal{O}} \sum_{j}^{N_p} y_j \log(C_p(p_j)) \tag{7}$$

where $\mathcal{O}$ represents each observation during training, and j refers to each class. $y_j \in \{0, 1\}$ indicates whether belonging to class j and $\log(C_p(p_j))$ is the predicted probability.

E. Multi-Modal: Cross-Modal Generation & Self-Modal Reconstruction

On top of the single-modal clean representation $\mathbf{x}$ of the gait, namely the skeleton acquired from the Mocap system, vision and wearable technologies can enable mobile monitoring outside the laboratory settings. In our study, the noisy data $\mathbf{z}$, i.e., RGBD based skeleton data ($\mathbf{r}$) and EMG ($\mathbf{e}$), are considered to demonstrate the implementation of generalized cross-modal transfer representation learning.

Based on the paired modalities $\mathbf{x}$ and $\mathbf{z}$, we introduce a novel cascaded architecture of two multi-encoder autoencoders shown in Fig. 7. This involves two steps, the cross-modal generation and self-modal reconstruction. The former maps the noisy modality $\mathbf{z}$ to the clean representation $\hat{\mathbf{x}}$, while the latter further filters the generated sample to $\tilde{\mathbf{x}}$. The pattern-specific features are extracted from relevant encoders for classification. Further details are given below.

1) Cross-Modal Generation: Firstly, we adopt the model in Section III-C to facilitate the cross-modal generation, which is composed of E^c_p, E^c_s, D^c (superscript c refers to cross modal). It generates clean representation data $\hat{\mathbf{x}}$ from the noisy modality $\mathbf{z}$. Meanwhile, the multiple encoders in this part can disentangle the associated common representations to be subject-specific s^c and pattern-specific p^c. The mapping across different modalities is learned by considering the unique pattern along with the subject characteristics based on cross-subject reconstruction. This training strategy, which is similar to the single-modality cross-subject transfer learning, has the advantage to avoid over-fitting. The loss functions are formulated as below,

$$\begin{aligned}\mathcal{L}^{z2x}_{self-gen} &= \mathbb{E}[\|\mathbf{x}^j_i - D^c(E^c_s(\mathbf{z}^j_i), E^c_p(\mathbf{z}^j_i))\|] \\ \mathcal{L}^{z2x}_{cross-gen} &= \mathbb{E}[\|\mathbf{x}^l_i - D^c(E^c_s(\mathbf{z}^j_i), E^c_s(\mathbf{z}^l_k))\|]\end{aligned} \tag{8}$$

Besides, similar to the loss involved in Section III-D2, we also applied the triplet loss.

2) Self-Modal Reconstruction: The cross-modal generation part maps the noisy representations into a clean representation, thus facilitating denoising. To this end, the pattern-specific p^c and subject-specific s_c features extracted from the cross-modal generation part may suffer from the interference of the noise embedded inside the raw data. To provide further filtering of the generated sample $\hat{\mathbf{x}}$, the second multi-encoder autoencoder is concatenated based on E^s_p, E^s_s, D^s (superscript s refers to single modal). It reconstructs $\widetilde{\mathbf{x}}$ from the generated sample $\hat{\mathbf{x}}$, minimizing the distance between $\widetilde{\mathbf{x}}$ and $\mathbf{x}$. The second part also extracts the disentangled features from the generated samples $\hat{\mathbf{x}}$. These features are subject-specific s^s and pattern-specific p^s, and are extracted from a cleaner representation $\hat{\mathbf{x}}$ compared to $\mathbf{z}$. The training loss functions are the same as in Section V-B.

a) Classifier: The whole model except for the classifier is trained end-to-end. Subsequently, the classifier for abnormality recognition is applied for training, similarly to Section III-D3. The pattern-specific feature from the feature from the reconstruction part p^s is provided as the only input to the classifier as it is less influenced by noises. Optionally, we also concatenate the first generation part p^c. It encodes some complementary information from the original modality, which might gets lost during the cross-modal generation.

IV. Abnormal Gait Database

A. Data Collection

In our abnormal gait database[2], 18 healthy volunteers (16 male and 2 female) were recruited and instructed to walk normally and imitate four pathological gait patterns (i.e., toe-in, toe-out, supination, and pronation), following the settings of previous simulation based works [58], [59], [60]. The recruited subjects were with no lower-limb injury history and did not undergo any joint instability during the course of the past six months. Among pathological gait categories as shown in Fig. 1, the supination refers to the outward roll of ankle joint while the pronation indicates the inward rotation. The toe-in and toe-out are the inward and outward of the foot forward direction during walking [5]. To avoid exaggerated imitation, specialized correction insoles were placed inside the shoes to resemble supination and pronation characteristics. For each pattern, subjects were walking within a straight pathway inside the laboratory (four meters, in four diagonal directions). Details can be found in the Supplementary Material.

During the capturing process, multi-modal gait data as listed in Fig. 8 was collected by Mocap system, RGBD camera, and wearable surface EMG sensors. For the Mocap data, 16 markers as illustrated in Fig. 9(a), were attached on the key joints of the lower limb. They were recorded by the Vicon multi-camera Mocap system, with similar settings as in [5]. The sampling rate was 120 Hz. A single RGBD camera was placed at the corner of the sensing area and 4D skeletal data were extracted from color and depth images based on previous work [5]. The Mocap data detect 16 joints that are shown in Fig. 9(a), whereas the skeleton data extracted from the RGBD sensor detect in real-time twelve joints as shown in Fig. 9(b). The sampling rate of the RGBD sensor is 30 Hz. The RGBD camera and Vicon system data streams were synchronized by the pre-calibrated time stamps of each system. As shown in

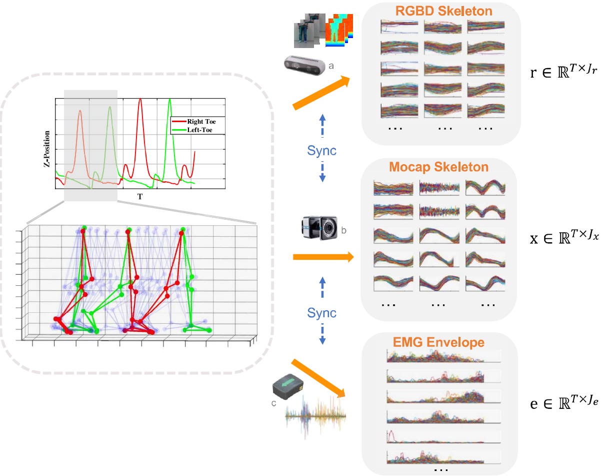


Fig. 8. Gait cycle representation. The Mocap data $\mathbf{x}$ was collected by the motion capture system (Vicon Motion System Ltd., Oxford, UK). RGBD camera used is RealSense D435 (Intel Corporation, California, US) and the skeleton data $\mathbf{r}$ was extracted by [5]. The surface EMG data $\mathbf{e}$ was collected by Trigno Avanti wireless EMG System (Delsys Incorporated, Massachusetts, US).
device images sources:
a:https://www.intelrealsense.com/depth-camera-d435/
b:https://www.vicon.com/hardware/cameras/
c:https://www.delsys.com/trigno/sensors/

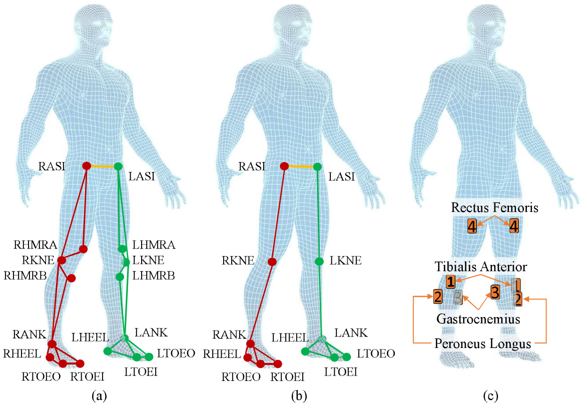


Fig. 9. Experimental setup during the multi-modal data collection. (a) Marker positions of ground truth markers recorded with Vicon motion capture system [5]; (b) Keypoint positions of human pose estimation model [18]; (c) Positions of eight wireless EMG sensors attached to different muscles.

[2] This experiment was approved by Imperial College Research Ethics Committee under Reference No. 18IC4915.

Fig. 9(c), eight wireless surface EMG sensors[3] were attached on the lower limb to measure the muscle response of Tibialis Anterior, Peroneus Longus, Gastrocnemius, and Rectus Femoris, respectively [61]. The EMG data has sampling rate 1200 Hz and it was synchronized with the Mocap data via the Vicon Lock Lab.

B. Preprocessing

For each walking sequence, data was segmented along the time axis to extract each gait cycle, as shown in Fig. 8. This involves detecting the right toe-off phase automatically based on the Mocap data. Subsequently, we resample each cycle to T frames with the toe-off phase of the left leg fixed at 50%. The global displacement was removed by subtracting the x and y positions of the root joint (i.e., the average of LASI and RASI) and the rotation around the axis vertical to the ground was removed as well. Then each joint trajectory across time is Z-Normalized by substracting the average value and deviding with the standard deviation along the sequence. Gait cycles contain the turning around point during walking were excluded. As mentioned in Section IV-A, synchronized skeleton data based on a single RGBD camera and EMG data were recorded. The RGBD camera was upsampled to the same sampling rate of the Mocap system. The EMG data was on-chip filtered, and then the extracted envelop was downsampled to the frame rate of the Mocap data.

V. Results

A. Implementation Details

1) Datasets and Parameters: Based on the dataset we introduced in Section IV, in our case, the resampling size of the gait cycle is T=128, whereas the dimensionality of the Mocap, RGBD and EMG data is J_x=3×16, J_r=3×12, J_e=8, respectively. J_x and J_r represent the number of 3D joints coordinates (each contains x,y,z) of the skeleton data, whereas J_e represents the number of EMG channels. The number of subjects is N_s=18 and the number of walking patterns is N_p=5.

2) Experimental Settings: The DNN architecture was implemented by Pytorch and trained with Titan XP. The architecture details are listed in the Supplementary Material. The loss functions $\{\mathcal{L}_{self-recon}, \mathcal{L}_{cross-recon}, \mathcal{L}^{z2x}_{self-gen}, \mathcal{L}^{z2x}_{cross-gen}, \mathcal{L}^{s/p}_{trip}\}$ are summed by weights {1,1,1,1,0.5} when used, respectively. The training for the multi-encoder autoencoder E_s, E_p, D and the classifier C_p were done separately. The adaptive learning rate optimization algorithm (ADAM) was applied for training with learning rate initialized as 0.002 and β as {0.9, 0.999}. For the multi-encoder autoencoder, the learning rate was decayed by 0.5 every 200 iterations until reaching 1000 iterations. For the classifier, the learning rate was decayed by 0.5 every 100 iterations until reaching 300 iterations. During training, 10% of data was randomly sliced out from the training set for validation. All the quantitative results related experiments are done five times independently and the average value is reported.

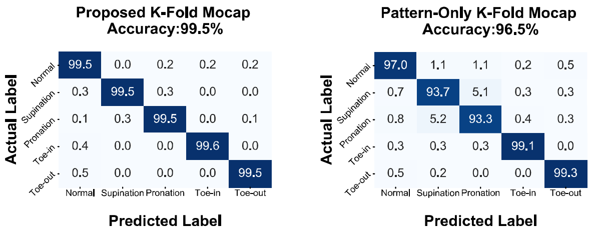


Fig. 10. Comparison results of K-Fold cross validation on Mocap data. Left: Cross-subject method; Right: Pattern Only variant.

B. Single-Modal Cross-Subject Transfer: Mocap

1) K-Fold Cross Validation: First of all, to evaluate the effectiveness of our model, a K-Fold (K=10) cross validation was first conducted on our data. The entire dataset was divided into nine subsets, and during each time, one subset was selected as the testing set while the remaining nine subsets were utilized for training. We also compared the results by the common single-encoder autoencoder architecture and the same training procedure but without the cross-reconstruction part, called as **Pattern Only**. As shown in Fig. 10, both our method and the ablated model of only pattern branch show good performance, and our two-branch model shows better classification accuracy than the ablated one.

2) Leave-One-Subject-Out Cross Validation: The good performance shown in K-Fold cross validation part may be caused by overfitting, as the model was trained on a mixture of all subjects and tested on the remaining set of the same subjects. No new subjects appear in the testing procedure. Therefore, this approach cannot demonstrate the generalization ability of the trained models. To address this, we applied more challenging Leave-One-Subject-Out (LOSO) cross validation. In each session, one subject was selected for testing, and the model is trained on the subset of the other subjects.

a) Qualitative results of cross-subject transfer: With LOSO, the representative qualitative results of cross-subject reconstruction is shown in Fig. 11. This cross reconstruction part facilitates learning subject-invariant gait features across training subjects. As the changes between normal, supination and pronation are relatively small to be observed from the skeleton data, we only present herewith the pattern transfer between normal, toe-in and toe-out. It can be seen from Fig. 11 that the pattern is successfully retargeted across subjects without introducing unrealistic movements, while preserving the subject-specific geometric characteristics.

b) Abnormal gait recognition results: During LOSO cross validation on the entire 18 subjects, the performance of each subject and each pattern is evaluated in terms of the following three metrics, Precision (Pre), Recall (Rec), and F1. The overall accuracy (Acc) is also reported. These results are presented in Table I. The boxplots represent the accuracy across each held-out subject under LOSO validation and it is also shown in Fig. 12. These plots demonstrate the overall consistency in generalization performance across subjects. Furthermore, we applied paired t-test to confirm that differences in performance across methods are statistically significant. The corresponding p values are displayed on the

[3] The EMG data with Subjs 6 & 17 and the RGBD data with Subjs 12&16 were not recorded.

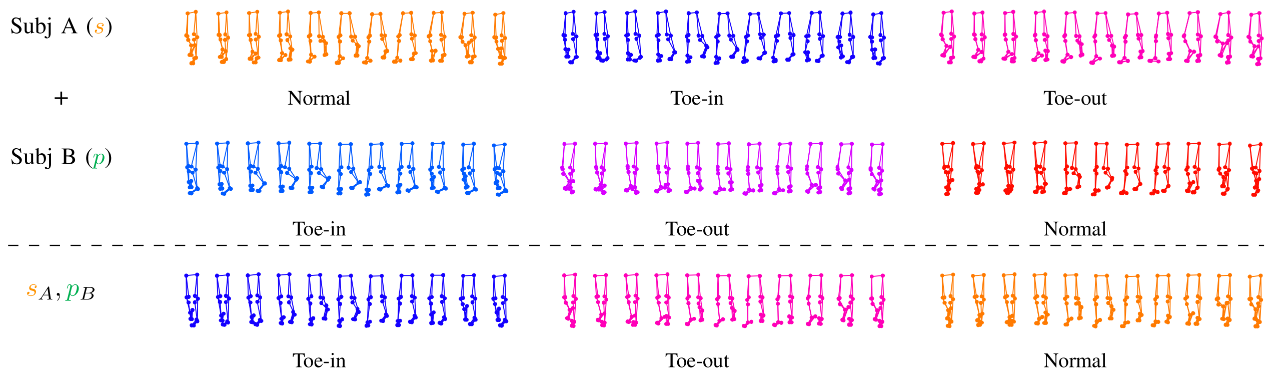


Fig. 11. Selected qualitative results of cross-subject transfer of different walking patterns. 11 skeleton samples are downsampled from the whole gait cycle and visualized horizontally. Other results and the animation effect can be viewed on https://guxiao0822.github.io/GAGR.

TABLE I
QUANTITATIVE RESULTS (%) UNDER LOSO VALIDATION FOR SINGLE-MODAL MOCAP DATA.

Modality	Methods	Normal			Supination			Pronation			Toe-in			Toe-out			Acc
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	
Mocap	SVM-angle[5]	70.14	72.91	71.50	74.75	70.72	72.68	88.33	87.27	87.80	94.48	98.13	96.27	96.12	95.88	96.00	85.45
	LSTM-angle[5]	78.90	75.14	76.97	81.14	78.38	79.74	88.72	90.43	89.56	96.41	97.73	97.06	94.35	97.90	96.09	89.51
	LSTM-xyz[5]	84.40	90.44	87.32	88.24	85.83	87.02	92.26	88.82	90.51	97.84	98.59	98.21	97.15	97.03	97.09	91.90
	Pattern Only	74.85	84.59	79.42	82.78	73.44	77.83	78.16	79.17	78.66	95.61	95.28	95.45	94.29	93.93	94.11	85.34
	CCSA [55]	91.48	88.03	89.72	92.07	87.72	89.84	88.63	90.88	89.74	97.88	98.15	98.01	93.91	98.25	96.03	92.81
	Epi-FCR [56]	85.52	90.98	88.17	90.41	85.29	87.78	89.97	90.21	90.09	98.57	97.73	98.15	97.03	97.88	97.45	92.50
	Proposed	**94.98**	**93.11**	**94.04**	**95.55**	**92.71**	**94.11**	**93.41**	**95.04**	**94.22**	**98.71**	**98.65**	**98.43**	**97.20**	**99.00**	**97.48**	**95.72**

score in bold indicates the best.

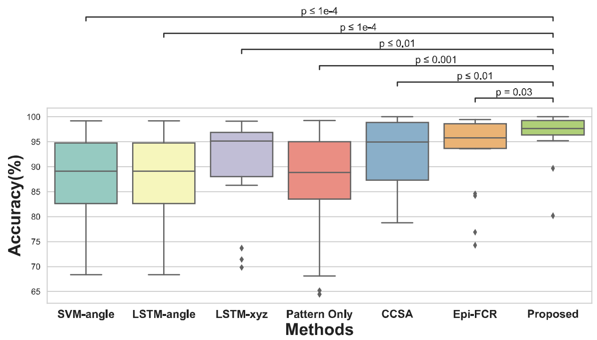


Fig. 12. Classification accuracy across different methods represented as boxplots that encompasses the accuracy of each held-out subject in the LOSO cross validation. The results of paired t-test between our proposed method and those compared are annotated on the top.

top of Fig. 12.

For detailed comparison, in addition to the proposed method and ablated **Pattern Only**, the accuracy of methods in our previous work [5] is shown as well. These include the SVM (Support Vector Machine) trained with angle features (**SVM-angle**), LSTM trained with angle features (**LSTM-angle**), and LSTM trained directly with 3D positions input (**LSTM-xyz**, the same input as our proposed model) and they are viewed as baseline abnormal gait recognition methods. As shown in Table I, compared with these methods, overall our proposed model shows superior performance. Furthermore, we notice that the LSTM models trained with direct 3D joints skeleton information outperform the model trained with the extracted joint angles. This shows that the features related to gait abnormalities are not limited to the joint angles extracted in our previous work [5], and therefore directly dropping out the original 3D data would cause information loss.

To highlight the robustness of our method, we also tested two deep learning based domain generalization methods. These are the Classification and Contrastive Semantic Alignment (**CCSA** [55])[4], and the Episodic Domain Generalization (**Epi-FCR** [56])[5]. Our method outperforms these approaches as well. Pre, Rec, F1 scores and overall accuracy are shown in Table I. In addition, Fig. 12 demonstrates that the proposed method outperforms these state-of-the-art domain generalization methods as well and the difference in accuracy is statistically significant ($p \leq 0.05$).

3) Leave-Multiple-Subject-Out Validation:

a) Feature visualization during the training procedure: To further validate the learned disentanglement of both subject-specific and pattern-specific features, we selected the first 14 subjects for training and the remaining 4 subjects for testing. To visualize those two features, we applied the t-Distributed Stochastic Neighbor Embedding (t-SNE)[6] to show the extracted s and p separately. The t-SNE is a high-dimensional data visualization tool that is able to effectively reduce feature dimensions while clustering the data of similar distributions. The t-SNE plot of s, p on training set and testing set are presented in Fig. 13. It can be observed that with the increment of iterations, the pattern feature space and subject feature space are disentangled and each sample is clustered into their individual categories, gradually. Similar

[4] Based on the codes from https://github.com/samotiian/CCSA

[5] Based on the codes from https://github.com/HAHA-DL/Episodic-DG

[6] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE

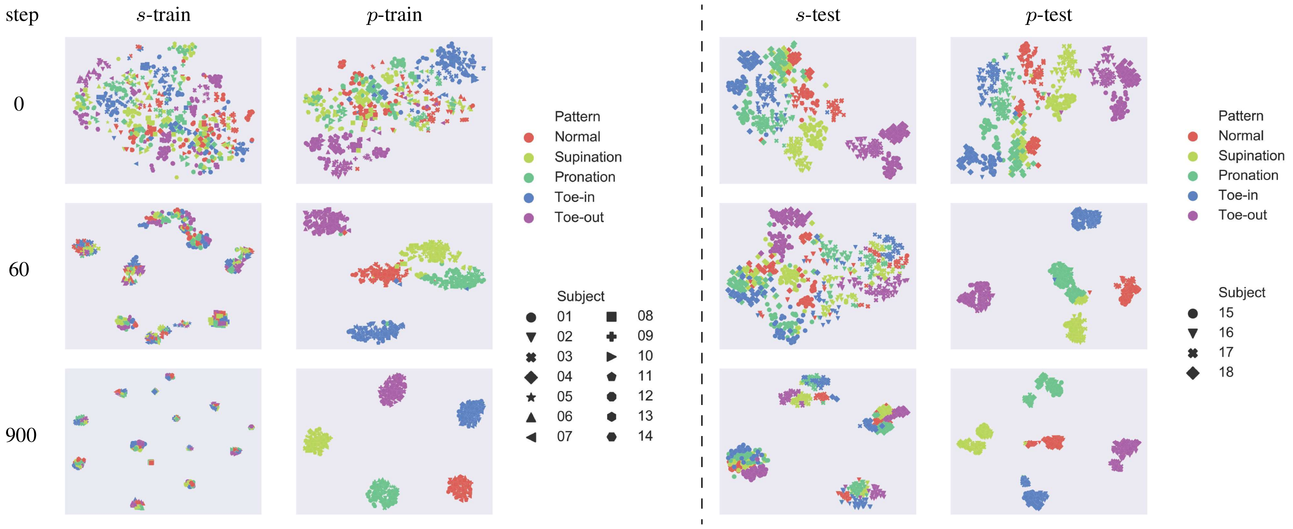


Fig. 13. The t-SNE visualization of s and p on training set and testing set based on the cross-subject transfer on Mocap data. 14 subjects are used for training while the held-out subjects are used for testing.

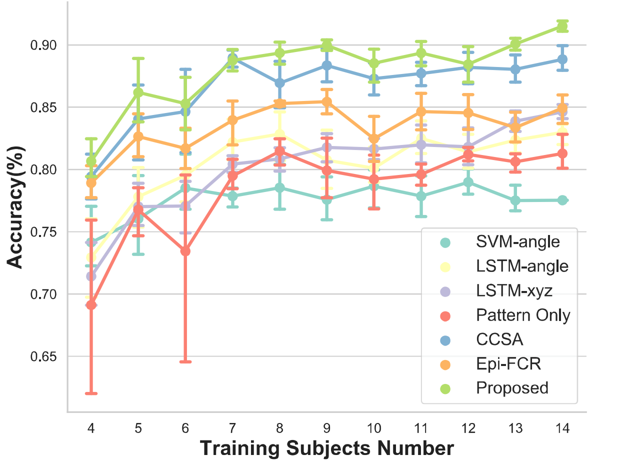


Fig. 14. Comparison of results under random-leave-subject-out (sampling with replacement) evaluation. In each training session, four subjects are randomly selected and substituted in the training set from the remaining subjects. For each training number, the random selection is repeated 10 times.

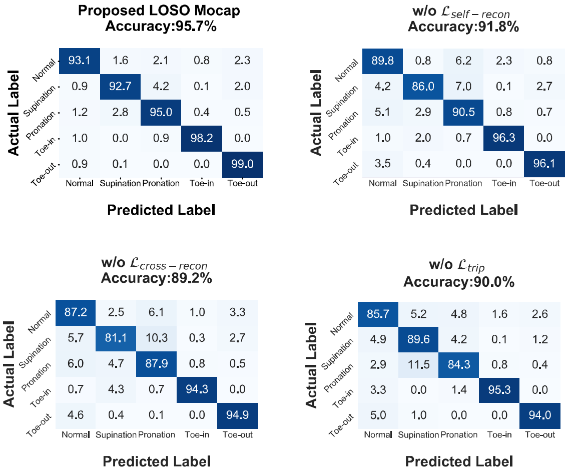


Fig. 15. Results of ablation study for single-modal cross-subject transfer on Mocap data.

results are also shown on the pattern and subject feature space extracted from the testing set. This qualitatively demonstrates that the pattern encoding model is of good generalization ability on 'unseen' new subject.

b) Random-leave-subject-out validation (Sampling with replacement): In addition, to highlight the generalization ability of our method when the training subject number is limited, we implemented a random-leave-subject-out strategy to further explore the robustness in terms of the number of training subjects. We randomly select four subjects as the test dataset, whereas in the training set, we substitute them with a random subsampling from the remaining set of subjects. The training subject number increases from 4 to 14. The sampling process is repeated 10 times for each number of the training subjects. The results are presented in Fig. 14. It is shown that our proposed method can achieve consistently superior performance compared to all other methods.

4) Ablation Study: Apart from merging the multi-encoder into a single one (namely **Pattern Only**), we also applied an ablation study on the effectiveness of each loss function ($\mathcal{L}_{self-recon}$, $\mathcal{L}_{cross-recon}$ and $\mathcal{L}_{trip}$), as shown in Fig. 15. It can be observed that the classification performance decreases to some extent in the ablated ones, demonstrating the effectiveness of each loss function.

C. Cross-Modal Cross-Subject Transfer: RGBD & EMG

Our goal is firstly to learn a mapping from a noisy sensing modality $\mathbf{z}$ to a clean representation $\hat{\mathbf{x}}$. This is achieved with the first multi-encoder decoder network $\{E_p^c, E_s^c, D^c\}$. Subsequently, we reconstruct the generated $\hat{\mathbf{x}}$ from $\tilde{\mathbf{x}}$ with the second multi-encoder decoder network $\{E_p^s, E_s^s, D^s\}$. Therefore, the pattern-related features are extracted from the E_p^c and E_p^s and they are subject-invariant. This cascaded architecture enables progressively converting the noisy representations to the associated clean representation.

1) Comparison of Generated and Reconstructed Samples:

a) Visualization of different samples: First of all, we show the qualitative results of generated and reconstructed

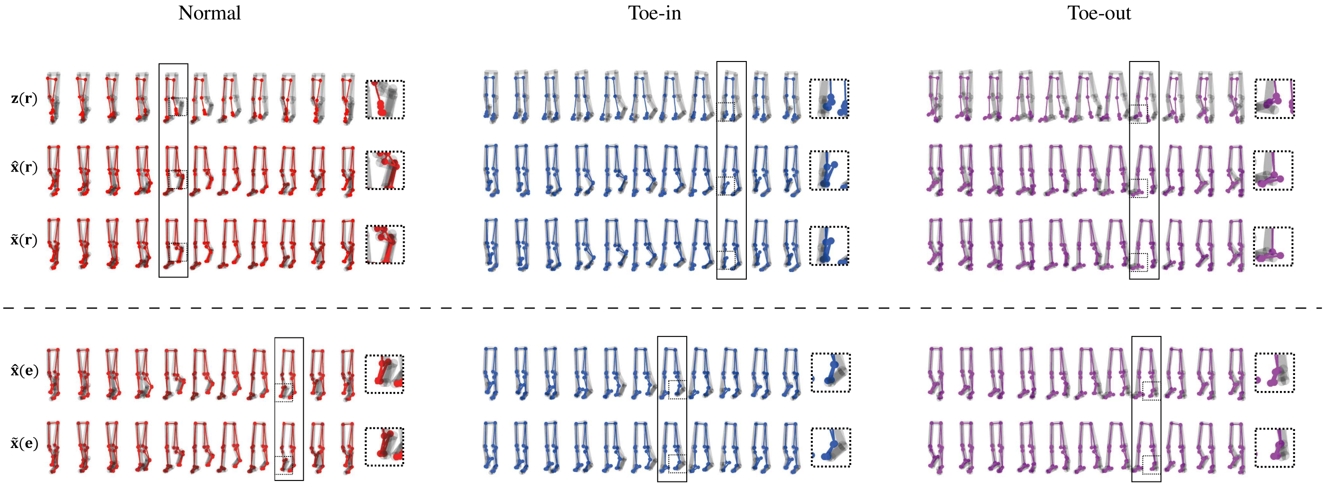
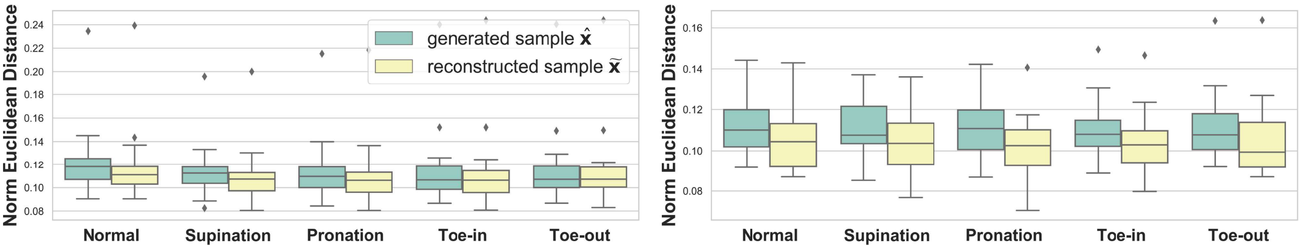


Fig. 16. Comparison of the original noisy representation $\mathbf{z}$, generated $\hat{\mathbf{x}}$, and reconstructed $\widetilde{\mathbf{x}}$ samples in the cross-modal transfer framework. The skeletons in grey color are the reference Mocap data. The skeletons selected with black rectangle highlight the compared difference, and the close-up is shown on the right side of each sample. Top: RGBD Skeleton; Bottom: EMG.



(a) RGBD $\mathbf{r}$

(b) EMG $\mathbf{e}$

Fig. 17. Boxplot of the normalized Euclidean distance between generated $\hat{\mathbf{x}}$ and reconstructed $\widetilde{\mathbf{x}}$ samples compared to ground truth Mocap samples $\mathbf{x}$, respectively.

samples of {Normal, Toe-in, Toe-out} in Fig. 16. As the difference between Normal and Pronation/Supination is hard to observe by human eyes during walking, only Normal and Toe-in, Toe-out samples are shown. The skeletons in Fig. 16 are their corresponding Mocap data. On the top RGBD part of Fig. 16, the comparison between the original RGBD samples $\mathbf{r}$ and the Mocap samples $\mathbf{x}$ are also displayed. The observed difference between these two modalities reflect the noises from the depth sensors and 3D pose estimation errors, in addition to the heterogeneity gap of the localization between human keypoints and attached markers. For both modalities, favorable performance can be observed in reconstructed $\widetilde{\mathbf{x}}$ compared to generated $\hat{\mathbf{x}}$.

b) Quantitative results: In Fig. 17, the boxplot of the normalized Euclidean distance between $\widetilde{\mathbf{x}}$/$\hat{\mathbf{x}}$ samples and $\mathbf{x}$ is shown. They are displayed by each pattern group separately. As shown in Fig. 17, the reconstructed samples are further refined. Quantitative results in Fig. 17, along with the qualitative results in Fig. 16 demonstrate the effectiveness of our cascaded architecture.

2) Abnormal Gait Recognition Results:

a) General results: To compare with previous methods, we applied four methods and two architecture variants, including **LSTM** [5], **Pattern Only**, **CCSA**, **Epi-FCR** and Single-Modal Cross-Subject (**SM-CS**), Cross-Modal Direct-Mapping (**CM-DM**). Among them, **SM-CS** adopts the structure for Mocap data, one multi-encoder autoencoder. **CM-DM** changed the first cross-modal generation part, which only minimizes the self-generation loss $\mathcal{L}^{z2x}_{self-gen}$ during training. It directly maps one modality to another in the first generation part, the same as the settings of a conventional autoencoder. Furthermore, different pattern features extracted from our proposed model were also explored. p^c, p^s, and $p^c + p^s$ are extracted separately and then used for training the classifier. We call these variant architectures as **Proposed-C**, **Proposed-S**, **Proposed-CS**, respectively.

For RGBD skeleton and EMG modality, we performed the LOSO validation on the entire dataset (16 subjects), respectively. Similarly to the metrics used in Section V-B, Pre, Rec, F1 of each gait pattern and the total accuracy are reported in Table II. The distribution of accuracy for all held-out subjects are visualized in Fig. 18 with statistical annotation. These results have shown superior performance of our proposed method (**Proposed-S** & **Proposed-CS**) compared to others.

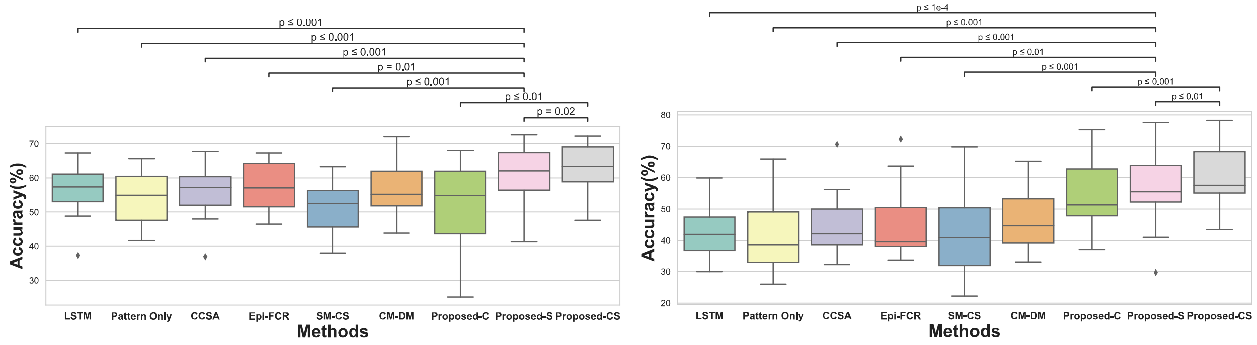
It can be observed that for **SM-CS**, the improvement of the Mocap data cannot be reproduced on the RGBD/EMG data by a similar architecture and training strategy. Perhaps this indicates that the single-modal model cannot well deal with the noise in the RGBD/EMG data representations. Therefore, the cross-subject reconstruction strategy alone fails to disentangle subject-specific and pattern-specific information in the presence of noise. Moreover, the result of **CM-DM** is poor,

TABLE II
QUANTITATIVE RESULTS (%) UNDER LOSO VALIDATION FOR RGBD & EMG DATA.

Modality	Methods	Normal			Supination			Pronation			Toe-in			Toe-out			Acc
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	
RGBD	LSTM [5]	47.31	48.57	47.93	43.23	40.40	41.77	41.64	43.14	42.38	66.59	65.08	65.82	63.00	64.67	63.82	52.55
	Pattern Only	48.48	49.59	49.03	45.79	43.32	44.52	42.24	40.88	41.55	65.57	65.85	65.71	59.67	63.17	61.37	52.67
	CCSA [55]	**64.15**	43.37	51.75	43.70	46.82	45.21	40.67	**56.52**	47.30	76.03	64.39	69.73	70.61	67.19	68.86	56.20
	Epi-FCR [56]	51.30	**55.87**	53.49	**52.39**	42.48	46.92	43.44	43.17	43.30	67.41	75.04	71.02	67.70	68.76	68.22	57.07
	SM-CS*	53.16	41.48	46.60	35.23	43.26	38.83	40.13	49.53	44.34	74.95	65.22	69.75	71.36	60.77	65.64	52.42
	CM-DM♯	47.23	46.78	47.00	39.72	40.16	39.94	41.74	49.84	45.44	83.30	71.38	76.88	76.97	72.47	74.65	56.56
	Proposed-C	49.88	39.77	44.26	35.29	36.43	35.91	35.92	36.65	36.28	68.35	62.90	65.51	58.25	68.47	62.95	49.34
	Proposed-S	58.87	53.41	56.01	42.36	**62.79**	50.59	47.11	43.01	44.97	**87.30**	73.21	79.64	**82.26**	69.47	75.33	60.63
	Proposed-CS	57.66	54.17	**55.86**	46.22	57.83	**51.38**	**48.98**	48.29	**48.63**	87.10	**76.37**	**81.38**	79.20	**73.89**	**76.46**	**62.42**
EMG	LSTM [5]	21.73	15.26	17.93	43.27	32.92	37.39	39.72	44.24	41.86	53.44	60.91	56.93	48.92	59.97	53.88	43.67
	Pattern Only	19.10	15.62	17.19	37.92	32.92	35.25	43.48	49.16	46.14	51.16	52.04	51.59	54.65	61.44	57.85	43.45
	CCSA [55]	26.82	30.51	28.55	44.67	38.35	41.27	43.22	**52.38**	47.36	61.78	42.32	50.23	49.17	54.66	51.77	44.30
	Epi-FCR [56]	27.87	21.88	24.51	43.62	35.56	39.18	43.65	51.77	47.36	55.04	53.92	54.47	50.60	59.75	54.79	45.59
	SM-CS*	22.84	21.88	22.35	35.81	60.71	45.05	51.90	33.49	40.71	61.42	48.90	54.45	54.50	49.58	51.92	43.67
	CM-DM♯	33.39	36.95	35.08	40.59	31.83	35.68	44.17	48.85	46.39	62.52	56.74	59.49	55.58	61.16	58.24	47.69
	Proposed-C	37.98	34.56	36.19	46.91	49.53	48.19	51.68	47.31	49.40	69.01	70.85	69.91	64.56	69.21	66.80	55.16
	Proposed-S	41.40	**47.79**	44.37	47.69	**61.02**	53.54	54.08	41.78	47.14	**75.83**	68.34	71.89	**72.21**	66.81	69.41	57.58
	Proposed-CS	**46.9**9	43.01	**44.91**	**51.77**	56.88	**54.11**	**60.20**	47.16	**52.89**	71.37	**76.18**	**73.69**	68.65	**76.69**	**72.45**	**60.75**

* SM-CS: Single Modal; Cross Subject
♯ CM-DM: Cross Modal; Direct Mapping

score in bold indicates the best.



(a) RGBD **r** (b) EMG **e**

Fig. 18. Classification accuracy across different methods represented as boxplots that encompasses the accuracy of each held-out subject in the LOSO cross validation. The results of paired t-test between our proposed method and those compared are annotated on the top.

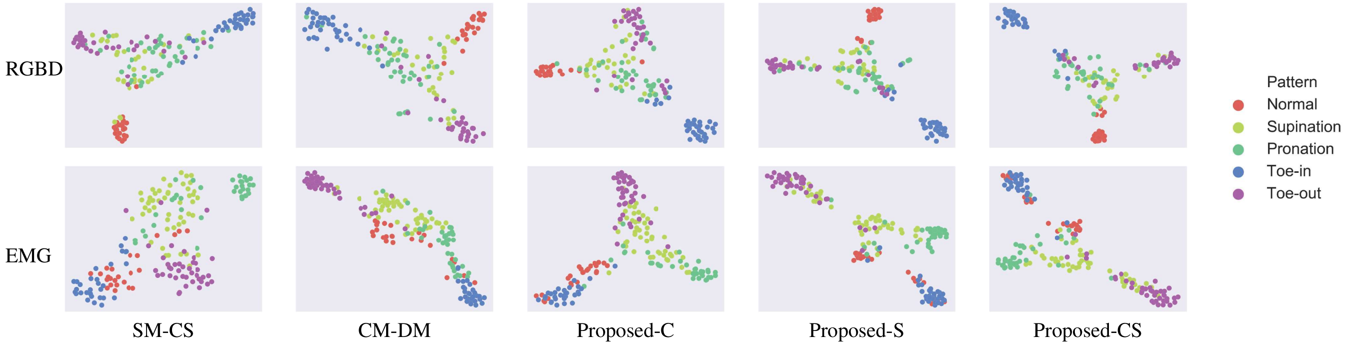


Fig. 19. The t-SNE visualization of the pattern feature p extracted from different variants and our proposed method under LOSO.

indicating that the direct mapping method can cause overfitting inside the first generation part and it does not generalize well.

3) Pattern Feature Comparison: Here we further discuss the difference between **Proposed-C**, **Proposed-S**, and **Proposed-CS**. **Proposed-C** utilizes p^c, the associated pattern-specific feature during mapping from $\mathbf{z}$ to $\hat{\mathbf{x}}$, while **Proposed-S** takes p^s as the input of classification, pattern specific feature extracted from the process of reconstructing $\widetilde{\mathbf{x}}$ from $\hat{\mathbf{x}}$. As p^s is from a cleaner representation $\hat{\mathbf{x}}$ than p^c, **Proposed-S** outperforms **Proposed-C** as shown in Table II. We also observe a slightly better performance after concatenating p^c and p^s together. This is probably due to the complementary information provided by p^c from the original sensing modality.

Moreover, the utilized pattern feature p from different methods $\{$**CM-CS**, **CM-DM**, **Proposed-C**, **Proposed-S**, **Proposed-CS**$\}$ are visualized by t-SNE and the results are shown in Fig. 19. It is shown that the clustering of each group is better in **Proposed-S** and **Proposed-CS**. This demonstrates that our framework helps to transform noisy data to cleaner and more discriminative representations.

4) Fine-tuning of Mocap-EMG: The EMG is a type of physiological signal that records the electrical activities of muscles, while the skeleton data records the kinematics in-

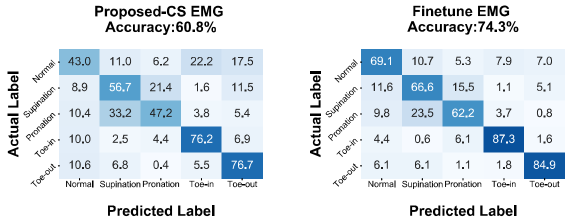


Fig. 20. Results based on fine tuned self-reconstruction loss on surface EMG.

formation. As the muscle distribution and strength varies largely across subjects, the mapping between EMG and Mocap skeletal information might be affected the stochastic nature of EMG signals. To examine this, we split the data of test subjects in two halves. For each half, we fine-tune the model by minimizing the self-generation loss $\mathcal{L}_{self-gen}^{z2x}$. Then the fine-tuned model is applied on the left half test set. The results are shown in Fig. 20. It is observed that the accuracy can achieve 74.3% with this strategy. However, there might be some unique information from Mocap skeleton to EMG by such a countermeasure, and in most cases, the ground truth clean kinematic data obtained from the Mocap system cannot be acquired outside the lab. Further work would partially focus on how to get more robust and generalized mapping between human skeleton and EMG.

VI. Conclusions

The applications of deep learning in health informatics have grown rapidly because of the ability of the deep models to not only outperform classical methods but also reach performance levels that fulfill clinical requirements. One of their major strength is to automatically extract relevant features in a hierarchical way. However, in applications where datasets are limited, subject-specific characteristics are interleaved with pathological characteristics, and this results in overfitting and lack of generalization to 'unseen' subjects. Various physiological signals contain both biometric information as well as pathological characteristics. Therefore, the development of robust inter-subject classification algorithms is of paramount importance.

In particular, abnormal gait recognition constitutes an interesting problem that can be formulated as a four-dimensional motion analysis. In fact, gait motion encodes both biometric characteristics, which reflect subject-specific information, as well as abnormal pattern-specific characteristics. In the absence of large datasets, we first propose a cross-subject reconstruction strategy that disentangles subject-specific from abnormal pattern-specific information. The proposed architecture improves classification performance with statistically significant difference compared to others.

Although, this strategy works well when the training kinematic datasets are precise (i.e., obtained by a Mocap system), it fails when the training datasets are noisy. In connected health applications, there is a need to perform abnormal gait recognition based on gait data obtained from a small number of vision-based or wearable sensors that would be able to work in home environments. To address this issue, we propose a novel cascaded architecture of a cross-modal, cross-subject transfer that allows to map noisy data obtained from a single RGBD camera or a set of eight EMG sensors to accurate Mocap kinematic estimations. This operation acts as filtering and it enables the successful application of the subsequent cross-subject transfer.

We have extensively validated our approach and demonstrated that it outperforms both state-of-the-art deep neural network architectures with statistically significant improvement for both the RGBD and the EMG data. Further work should aim to validate the robustness of the proposed model in clinical settings with longitudinal datasets from large cohort patient studies.

VII. Acknowledgement

We would like to express our gratitude to the subjects volunteering in the experiments. Also, we would like to acknowledge Imperial College Research Ethics Committee for the ethics approval (No. 18IC4915) of our experiments.

References

[1] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2016.

[2] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, p. 719, 2018.

[3] F. Deligianni, C. Wong, B. Lo, and G.-Z. Yang, "A fusion framework to estimate plantar ground force distributions and ankle dynamics," *Information Fusion*, vol. 41, pp. 255–263, 2018.

[4] X. Gu, F. Deligianni, B. Lo, W. Chen, and G.-Z. Yang, "Markerless gait analysis based on a single rgb camera," in *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2018, pp. 42–45.

[5] Y. Guo, F. Deligianni, X. Gu, and G.-Z. Yang, "3-d canonical pose estimation and abnormal gait recognition with a single rgb-d camera," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3617–3624, 2019.

[6] S. Bao, S. Yin, H. Chen, and W. Chen, "A wearable multimode system with soft sensors for lower limb activity evaluation and rehabilitation," in *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2018, pp. 1–6.

[7] F. Deligianni, Y. Guo, and G. Yang, "From emotions to mood disorders: A survey on gait analysis methodology." *IEEE journal of biomedical and health informatics*, 2019.

[8] H. Kainz, D. Graham, J. Edwards, *et al.*, "Reliability of four models for clinical gait analysis," *Gait & posture*, vol. 54, pp. 325–331, 2017.

[9] S. Chen, J. Lach, B. Lo, and G.-Z. Yang, "Toward pervasive gait analysis with wearable sensors: a systematic review," *IEEE journal of biomedical and health informatics*, vol. 20, no. 6, pp. 1521–1537, 2016.

[10] G.-Z. Yang, Ed., *Body Sensor Networks*. London: Springer London, 2014. [Online]. Available: http://link.springer.com/10.1007/978-1-4471-6374-9

[11] G. Yang, W. Tan, H. Jin, T. Zhao, and L. Tu, "Review wearable sensing system for gait recognition," *Cluster Computing*, pp. 1–9, 2018.

[12] A. Phinyomark, S. T. Osis, B. A. Hettinga, D. Kobsar, and R. Ferber, "Gender differences in gait kinematics for patients with knee osteoarthritis," *BMC musculoskeletal disorders*, vol. 17, no. 1, p. 157, 2016.

[13] Y. Sun and B. Lo, "An artificial neural network framework for gait-based biometrics," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 987–998, 2018.

[14] J. Chen, X. Zhang, Y. Cheng, and N. Xi, "Surface emg based continuous estimation of human lower limb joint angles by using deep belief networks," *Biomedical Signal Processing and Control*, vol. 40, pp. 335–342, 2018.

[15] C. J. De Luca, L. D. Gilmore, M. Kuznetsov, and S. H. Roy, "Filtering the surface emg signal: Movement artifact and baseline noise contamination," *Journal of biomechanics*, vol. 43, no. 8, pp. 1573–1579, 2010.

[16] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3d human pose estimation: A review of the literature and analysis of covariates," *Computer Vision and Image Understanding*, vol. 152, pp. 1–20, 2016.

[17] G. Halmetschlager-Funek, M. Suchi, M. Kampel, and M. Vincze, "An empirical evaluation of ten depth cameras: Bias, precision, lateral noise, different lighting conditions and materials, and multiple sensor setups in indoor environments," *IEEE Robotics & Automation Magazine*, no. 99, pp. 1–1, 2018.

[18] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[19] S. Ma, D. McDuff, and Y. Song, "M3d-gan: Multi-modal multi-domain translation with universal attention," *arXiv preprint arXiv:1907.04378*, 2019.

[20] S. Bei, Z. Zhen, Z. Xing *et al.*, "Movement disorder detection via adaptively fused gait analysis based on kinect sensors," *IEEE Sensors Journal*, vol. 18, no. 17, pp. 7305–7314, 2018.

[21] A. A. Chaaraoui, J. R. Padilla-López, and F. Flórez-Revuelta, "Abnormal gait detection with rgb-d devices using joint motion history features," in *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, vol. 7. IEEE, 2015, pp. 1–6.

[22] C. Prakash, R. Kumar, and N. Mittal, "Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges," *Artificial Intelligence Review*, vol. 49, no. 1, pp. 1–40, 2018.

[23] F. Horst, S. Lapuschkin, W. Samek, K.-R. Müller, and W. I. Schöllhorn, "Explaining the unique nature of individual gait patterns with deep learning," *Scientific reports*, vol. 9, no. 1, p. 2391, 2019.

[24] J. Hannink, T. Kautz, C. F. Pasluosta, K.-G. Gaßmann, J. Klucken, and B. M. Eskofier, "Sensor-based gait parameter extraction with deep convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 85–93, 2016.

[25] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE, 2017, pp. 1623–1631.

[26] K. Hu, Z. Wang, S. Mei, *et al.*, "Vision-based freezing of gait detection with anatomic directed graph representation," *IEEE journal of biomedical and health informatics*, 2019.

[27] Q. Zou, Y. Wang, Q. Wang, Y. Zhao, and Q. Li, "Deep learning based gait recognition using smartphones in the wild," *arXiv preprint arXiv:1811.00338*, 2018.

[28] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using lstm," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 325–330.

[29] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[30] Ł. Kidziński, S. Delp, and M. Schwartz, "Automatic real-time gait event detection in children using deep neural networks," *PloS one*, vol. 14, no. 1, p. e0211466, 2019.

[31] D.-X. Liu, W. Du, X. Wu, C. Wang, and Y. Qiao, "Deep rehabilitation gait learning for modeling knee joints of lower-limb exoskeleton," in *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2016, pp. 1058–1063.

[32] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, 2013, pp. 1310–1318.

[33] G. W. Taylor, G. E. Hinton, *et al.*, "Two distributed-state models for generating high-dimensional time series," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 1025–1068, 2011.

[34] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4346–4354.

[35] Z. Li, Y. Zhou, S. Xiao, C. He, Z. Huang, and H. Li, "Auto-conditioned recurrent networks for extended complex human motion synthesis," *arXiv preprint arXiv:1707.05363*, 2017.

[36] S. Tak and H.-S. Ko, "A physically-based motion retargeting filter," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 1, pp. 98–117, 2005.

[37] R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural kinematic networks for unsupervised motion retargetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8639–8648.

[38] X. Yan, A. Rastogi, R. Villegas, K. Sunkavalli, E. Shechtman, S. Hadap, E. Yumer, and H. Lee, "Mt-vae: Learning motion transformations to generate multimodal human dynamics," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 265–281.

[39] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or, "Learning character-agnostic motion for motion retargeting in 2d," *arXiv preprint arXiv:1905.01680*, 2019.

[40] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multi-level variational autoencoder: Learning disentangled representations from grouped observations," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[41] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.

[42] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[43] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3020–3028.

[44] N. Mor, L. Wolf, A. Polyak, and Y. Taigman, "A universal music translation network," *arXiv preprint arXiv:1805.07848*, 2018.

[45] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: Multimodal stochastic rnns for video captioning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 10, pp. 3047–3058, 2018.

[46] H. Chen, Y. Li, and D. Su, "Discriminative cross-modal transfer learning and densely cross-level feedback fusion for rgb-d salient object detection," *IEEE transactions on cybernetics*, 2019.

[47] X. Huang, Y. Peng, and M. Yuan, "Cross-modal common representation learning by hybrid transfer network," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 1893–1900.

[48] Q. Zou, L. Ni, Q. Wang, Q. Li, and S. Wang, "Robust gait recognition by integrating inertial and rgbd sensors," *IEEE transactions on cybernetics*, vol. 48, no. 4, pp. 1136–1150, 2017.

[49] B.-Q. Ma, H. Li, Y. Luo, and B.-L. Lu, "Depersonalized cross-subject vigilance estimation with adversarial domain generalization," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[50] X. Jiang, K. Xu, and W. Chen, "Transfer component analysis to reduce individual difference of eeg characteristics for automated seizure detection," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2019, pp. 1–4.

[51] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in *Advances in Neural Information Processing Systems*, 2019, pp. 6447–6458.

[52] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 53–69, 2015.

[53] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*, 2013, pp. 10–18.

[54] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409.

[55] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5715–5725.

[56] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1446–1455.

[57] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2138–2147.

[58] I. Mahmood, U. Martinez-Hernandez, and A. A. Dehghani-Sanij, "Evaluation of gait transitional phases using neuromechanical outputs and somatosensory inputs in an overground walk," *Human Movement Science*, vol. 69, p. 102558, 2020.

[59] W. Cui, C. Wang, W. Chen, Y. Guo, Y. Jia, W. Du, and C. Wang, "Effects of toe-out and toe-in gaits on lower-extremity kinematics, dynamics, and electromyography," *Applied Sciences*, vol. 9, no. 23, p. 5245, 2019.

[60] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5363–5371.

[61] D. H. Sutherland, "The evolution of clinical gait analysis part l: kinesiological emg," *Gait & posture*, vol. 14, no. 1, pp. 61–70, 2001.