

Received 5 December 2022, accepted 10 December 2022, date of publication 15 December 2022,  
date of current version 21 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3229478

## RESEARCH ARTICLE

# Graph Laplacian-Improved Convolutional Residual Autoencoder for Unsupervised Human Action and Emotion Recognition

GIANCARLO PAOLETTI<sup>1,2</sup>, CIGDEM BEYAN<sup>1,3</sup>, AND ALESSIO DEL BUE<sup>1</sup>, (Member, IEEE)

<sup>1</sup>Pattern Analysis and Computer Vision (PAVIS) Research Line, Fondazione Istituto Italiano di Tecnologia (IIT), 16152 Genoa, Italy

<sup>2</sup>Electrical, Electronics, and Telecommunication Engineering and Naval Architecture Department (DITEN), University of Genoa, 16145 Genoa, Italy

<sup>3</sup>Department of Information Engineering and Computer Science (DISI), University of Trento, 38123 Trento, Italy

Corresponding author: Cigdem Beyan (cigdem.beyan@unitn.it)

**ABSTRACT** This paper presents an unsupervised feature learning approach based on 3D-skeleton data for human action and human discrete emotion recognition. Relying on the time series of skeleton data analysis to perform such tasks is effective and important to preserve the individual's privacy better. Besides, such methods can represent a viable alternative to emotion recognition applications, in which most works use frontal or profile facial images disclosing the subject's appearance. On the other hand, current unsupervised methods are able to encode the high variety of contexts and nature of the data, but often at the expense of a higher model complexity or longer computational time. To lessen these shortcomings, this paper proposes a convolutional residual autoencoder that models the skeletal geometry across the temporal dynamics of the data without relying on computationally expensive recurrent architectures. Our approach also implements a Graph Laplacian Regularization leveraging upon the implicit skeleton joints connectivity, further improving the robustness of the feature embeddings learned without using action or emotion labels. It was validated on large-scale datasets, having variability in the domain, the input skeleton data (*e.g.* the number of joints, adjacency matrices), and sensor technology. The results show its effectiveness by notably surpassing the performance of the state-of-the-art unsupervised methods while also achieving better recognition scores compared to the several fully supervised approaches. Extensive experimental analysis proves the usefulness of the proposed method under various evaluation protocols with observed higher-quality feature representations, even if when it is trained with fewer data. The results highlight the proposed method's remarkable transfer-ability across various domains, and its faster inference time.


**INDEX TERMS** Action recognition, autoencoder, emotion recognition, full-body movement, graph Laplacian, skeletal data, unsupervised feature learning.

## I. INTRODUCTION

Human Action Recognition (HAR) and Human Emotion Recognition (HER) are ubiquitous tasks enabling applications ranging from smart video surveillance, human-robot interaction, and healthcare monitoring, to name a few. Given substantial signs of progress in the last years, HAR is still a challenging task because of the varying imaging conditions (*e.g.* camera viewpoints and lighting) and the complexity of human motion [1], [2]. One way to tackle these problems is to rely on the analysis of time-series of body joints (skeleton

data), which are proven to be effective in representing the actions [3], [4], and better preserve the individual privacy as they do not disclose the subject appearance [5]. On the other hand, the analysis of human emotions can include several modalities, such as text, physiological signals, acoustic data, facial landmarks, facial images, or full-body motion. Among all, recent HER works [6], [7], [8], [9], [10] highlight the effectiveness of processing full-body motion represented in terms of 3D-skeleton data.

The successes in skeleton-based HAR [11], [12], [13] and skeleton-based HER [6], [7], [8], [9], [10] primarily leverage on the supervised learning paradigm, which requires a significant amount of manually labeled data. However, data

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang .

annotation is notoriously expensive and time-consuming [14]. Moreover, action and/or emotion classes may vary significantly from dataset to dataset, while several methods lack enough generalization for application in different scenarios without requiring extra annotations.

As a (recent) alternative, unsupervised approaches for HAR [15], [16], [17], [18], [19], [20], [21], [22], [23] are tremendously increasing their impact in terms of action classification, while competing to reduce the performance gap with the fully supervised counterparts. On the other hand, there has been yet no attempt to apply unsupervised HER using skeleton data.

In this paper, we propose an *unsupervised feature learning* approach based on 3D-skeleton data. We show our method's effectiveness on two downstream tasks: *i*) human action recognition (**HAR**) and *ii*) human discrete emotion recognition (**HER**, see Figure 1). Using *not-labeled* 3D-skeleton sequences, we learn a feature representation that is then fed to an action or emotion recognition classifier, depending on the downstream task to be performed. Our feature learning step is based on a **Convolutional Residual Autoencoder** (shown as **CR-AE** for the rest of this paper). We demonstrate the benefits of performing residual convolutions to jointly learn representations with spatio-temporal convolutions instead of relying on more complex and/or memory-intense architectures, (e.g., [16] using recurrent neural networks) as well as having the fastest inference time compared to prior art. Another important aspect of the proposed convolutional residual autoencoder is the adoption of **graph Laplacian regularization** [24] (shown as **CR-AE-L** for the rest of this paper) to learn representations that are aware of the spatial configuration of the skeletal geometry. We observe that representing the full-body motion with skeletal joints is principled and rooted in cognitive perception [25], therefore we approximate a continuous human body moving in time with a collection of discrete trajectories. Although several works demonstrate the usefulness of using Laplacian regularization to encode the geometry of hand-crafted [24] or learned [26] feature space, we apply this regularization in the *reconstruction* space (*i.e.* the space induced by the last layer of the decoder). This work uses for the first time Laplacian Regularization within an unsupervised feature learning paradigm for action and emotion recognition together, injecting into the network the intrinsic knowledge of the connectivity patterns of the body whose skeleton is represented through 3D joints.

To validate our method, experiments were realized on three large-scale skeletal *action* datasets: NTU-60 (cross-subject and cross-view settings) [27], NTU-120 (cross-subject and cross-setup settings) [28] and Skeletics-152 [29] as well as two large-scale *emotion* datasets: Dance Motion Capture Emotion Database [30] and Emilya [9]. The ablation study, performed on each dataset, shows the positive contribution of residual layers of our autoencoder and the skeletal graph Laplacian regularization. We also evaluate our approach on several settings, *e.g.* with finetuning and end-to-end training protocol (see Section IV-F), unsupervised

training with fewer data (see Section IV-G), and transferability (see Section IV-H). When our *end-to-end* approach is tested for HAR, it outperforms all the state-of-the-art (SOTA) unsupervised skeleton-based methods and even surpasses a few SOTA-supervised approaches. For HER, the results show that the proposed method can achieve better performance than several supervised SOTA, even when evaluated using a 1-NN protocol, which is parameter-free. We adopted the SOTA U-HAR methods to work on the downstream task of emotion recognition and compared their performance with ours. The results show that the proposed method is able to perform better than all of them also for HER problem. Additionally, even if it is trained with less data, our method achieves better results compared to several approaches learned on more data. Thus, the proposed method is effective in case of limited training data. Importantly, when we train our model on one domain and test it on another, it demonstrates transferability by scoring on par or by achieving improved results for several cases.

This paper grounds on our earlier study [23], which presents preliminary results of the proposed method for HAR and particularly focuses on viewpoint-invariant HAR. Different from [23], in this study, we do not focus on viewpoint-invariant HAR, but instead, we perform the following new analysis bringing in new contributions to the relevant domains.

- This is the first time where our proposed method, *CR-AE-L*, is tested for the unsupervised full-body motion-based emotion recognition task. We benchmarked the SOTA unsupervised skeleton-based action recognition methods for HER and compared their performance with *CR-AE-L*. It is important to notice that there yet exists no skeleton-based HER method implementing unsupervised feature learning, which makes our work novel.
- We show that our *CR-AE-L* works well within different skeleton data representations (*e.g.* various numbers of joints, adjacency matrices, various sensors, and their relative sampling rates). All comparisons and ablation studies are performed on a larger number of datasets, allowing us to show that the effective performance of *CR-AE-L* generalizes well.
- We explore the performance of *CR-AE-L* within finetuning protocol, end-to-end training, and unsupervised training with fewer data. We also investigate its transferability of it. It performs better than several approaches even when it is trained with fewer data compared to others, also presenting its usefulness in case of limited training data. Moreover, when it is trained on one domain and tested on another, it is able to demonstrate the advantage of performing unsupervised pre-training by scoring on par or by achieving improved results for several cases.
- For both downstream tasks, *CR-AE-L* notably surpasses the unsupervised SOTA in standard evaluation protocols and datasets, with noticeable effectiveness compared to SOTA-supervised methods.

- The space complexity of *CR-AE-L* in terms of the number of parameters is less than some methods while its computational complexity in terms of the inference time is the least out of all prior art.

The rest of the paper is organized as follows. Section II summarizes SOTA unsupervised skeleton-based human action recognition approaches by highlighting their differences with the proposed method. We also discuss the literature on human emotion recognition from full-body movements and graph Laplacian regularization, in the same section. The proposed method is introduced in Section III, specifying the convolutional residual autoencoder, skeletal Laplacian regularization and inference phase. Section IV presents the experimental analysis, datasets, implementation details, and results. Finally, we conclude the paper with a summary and discussions in Section V.

## II. RELATED WORK

There have been several attempts regarding skeleton-based Human Action Recognition (HAR) and Human Emotion Recognition (HER) problems in supervised settings. However, this paper addresses the more complex and challenging scenario where no labeled data is available for feature learning. Only recently, unsupervised approaches have become popular in skeleton-based HAR, while they have not yet been inherited for HER (particularly full-body motion-based HER).

This section first reviews the state-of-the-art (SOTA) skeleton-based *unsupervised HAR (U-HAR)* and *HER of full-body movements* by explaining the differences between them and our method. Following that, we summarize the usage of Laplacian Regularization in the previous works.

### A. UNSUPERVISED SKELETON-BASED HUMAN ACTION RECOGNITION

To perform HAR, several modalities have been exploited, such as video frames (RGB) [31], video frames with depth information (RGB+D) [32], and/or skeleton data [20]. Skeleton-based HAR is advantageous due to its privacy-preserving properties since not a single RGB image needs to be stored. It is a representation easily given by off-the-shelf body pose detectors and potentially allows to perform HAR in real time. Most of the work follows a supervised learning framework where the set of actions should be pre-defined and annotated for training a model [3], [4], [33], [34], [35], [36]. Whereas unsupervised HAR (U-HAR) approaches are in general under-performing compared to their supervised counterparts, but *i)* they can provide a more robust adaptation to real-world applications as they do not need re-training when the scenario of application changes and *ii)* they eliminate the need for very expensive and time-consuming annotation efforts. Nevertheless, the number of skeleton-based U-HAR methods is limited compared to supervised approaches.

We review each unsupervised skeleton-based HAR methodology below and contextualize our approach's technical novelty compared to them. Prior methods mainly leverage

### Algorithm 1 Proposed Approach

- 1: Randomly initialize the Encoder  $\mathbf{E}_\varphi$  and Decoder  $\mathbf{D}_\theta$ .
- 2: Compute the skeletal graph Laplacian  $\mathbf{L}$  from adjacency matrix  $\mathbf{W}$ .
- 3: **while** not converged **do**
- 4:   Sample a mini-batch of training data  $\mathcal{B}$ .
- 5:   Do a forward pass through  $\mathbf{E}_\varphi$  and  $\mathbf{D}_\theta$ , obtaining  $\hat{\mathbf{X}}$ .
- 6:   Update  $\mathbf{E}_\varphi$ ,  $\mathbf{D}_\theta$  using the MSE loss  

$$\mathcal{L}_{MSE} = \frac{1}{2} \mathbb{E}_{\mathbf{X} \sim \mathcal{B}} [\|\mathbf{X} - \hat{\mathbf{X}}\|_F^2]$$
- 7:   Update  $\mathbf{E}_\varphi$ ,  $\mathbf{D}_\theta$  using the  $\mathcal{R}_{skel}$   

$$\mathcal{R}_{skel} = \mathbb{E}_{\mathbf{X} \sim \mathcal{B}} [\mathbb{E}_{t,d} [\hat{\mathbf{X}}^{(t,d)\top} \mathbf{L} \hat{\mathbf{X}}^{(t,d)}]]$$
- 8: **end while**
- 9: Freeze encoder parameters  $\mathbf{E}_\varphi$  and append a linear classifier (*LEP*) or a 1-Nearest Neighbor classifier (*1-NN*).

neural architectures composed of encoder-decoder recurrent architectures to perform HAR [15], [16], [17], [21], [22]. Zheng et al. [15] introduce an encoder-decoder architecture (called LongT-GAN) based on GRUs that learns how to represent skeletal body poses in time, while an adversarial loss supports an auxiliary inpainting task favorably helps the learning stage. *MS<sup>2</sup>L* [22] is also based on GRUs and benefits from contrastive learning, motion prediction, and jigsaw puzzle recognition. Kundu et al. [21] additionally include a GAN in their recurrent architecture and present the method called *EnGAN*. The method called *PCR* [20], builds upon a vanilla autoencoder trained to reconstruct the skeletal data using mean-squared error (MSE) loss. This vanilla model is boosted by an ad-hoc training mechanism based on expectation-maximization with learnable class prototypes.

Su et al. [16] present the Predict & Cluster (P&C) method based on an encoder-decoder Recurrent Neural Network that learns representations for HAR in an unsupervised manner from skeletal joints while solving action classification with a 1-Nearest Neighbor predictor. Another related work, AS-CAL [17], combines contrastive learning with momentum LSTM where the similarity between augmented instances and the input skeleton sequence is contrasted, and then a momentum-based LSTM encodes the long-term actions. Li et al. [37] also inherited contrastive learning. Instead, a Siamese denoising autoencoder – SeBiReNet [19] – is used with feature disentanglement, showing good performance across pose denoising and unsupervised cross-view HAR.

Unlike the studies mentioned above, we designed our autoencoder with residual convolutions. Our architecture has space complexity less than some prior art while it has the fastest inference time (see Table 11 for details), and it performs better than all prior art.

### B. HUMAN EMOTION RECOGNITION FROM FULL-BODY MOVEMENTS

Emotion recognition from full-body movement data is a complex task since the act of expressing and perceiving affect

differs a lot *w.r.t.* its context, and also their variety increases due to the interpersonal differences (*e.g.* personality, physical capacity, and personal experience) [38], [39].

Emotion recognition from full-body representation has been so far addressed by: *i*) processing single body pose (*e.g.* a forward head and chest bend express sadness in [40]), *ii*) recognizing specific gestures which are emblems of the emotions (*e.g.* raising arms and hands-on-hips are the gestures of pride according to [41] and [42]), or *iii*) processing the expressive quality of the movement [6], [9], [43], [44]. Out of these three possibilities, the second and the third use the temporal information of the data, while the first one performs only spatial processing.

In this work, we evaluate our proposed method for *processing the expressive quality of the movement* (*i.e.* category *iii*). The existing related datasets were curated with diverse motion capture (MoCap) systems and various numbers of markers. These datasets are relatively smaller than the HAR counterparts due to the effort needed to expertly collect and, most importantly *annotate* such data with high reliability. As annotations are more costly, it is crucial to develop unsupervised feature learning methods that can effectively apply to HER.

Earlier works define hand-crafted features and apply learning methods such as Support Vector Machines (SVM) and Random Forests [7], [45], [46], [47]. For instance, Castellano et al. [45] use motion quantity, velocity, and movement fluidity as the descriptors of movements and aggregate them in the temporal dimension to classify four-emotion classes. Instead, Piana et al. [7] extend the low-level features by adding high-level features (*e.g.* contraction index, impulsiveness) and applying an SVM classifier. On the other hand, Fourati et al. [46] show the importance of using temporal features (*e.g.* regularity of a motion profile, overall or single gesture phase impulsiveness) and multi-level body cues (*e.g.* based on Body Action and Posture Coding System) for emotions elicited during the daily-life actions. In [8], the 3D-skeleton data is represented in the Riemannian manifold and then processed with a covariance operator. This methodology was adapted by Kacem et al. [48], where the former applies a Nearest Neighbour classifier, and the latter uses a temporal warping and SVM. Both methods improved the emotion recognition from 3D-body movements results *w.r.t.* the prior art. As a different approach, Creen et al. [10] synthesize neutral motion by quantizing it with a cost function and then calculates the difference between the neutral class and the other emotions to decide the class 3D-body expression at inference.

Deep learning architectures *e.g.* Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) have been explored for skeleton-based HER in recent studies. For example, [49] used an RNN with 3-layers to perform emotion classification from MoCap data of daily activities: clapping, drinking, throwing, and waving, etc., associated with four emotions: happy, angry, sad, and neutral. Beyan et al. [6] present the joint training of two CNNs such that one of them

performs coarse-grained modeling while the other applies fine-grained modeling in the time. The inputs of this network are 8-bit RGB images obtained from 3D-skeleton data over time. This approach [6] achieves better performance compared to [9], [10], and [46], showing generalization properties over the diverse number of emotion classes and contexts.

All the approaches mentioned above apply supervised learning. There exists no work exploring the effectiveness of unsupervised feature learning to be used for the downstream task: HER from fully-body movements. This paper presents the first attempt by applying our proposed model which is based on convolutional residual autoencoders. The effectiveness of the proposed method is compared *w.r.t.* SOTA supervised methods as well as against SOTA U-HAR methods after adapting them for HER.

### C. GRAPH LAPLACIAN REGULARIZATION

The Laplacian of a graph is a well-known mathematical tool that supports graph theory, especially for problems related to connectivity. In the machine learning domain, it was adapted by Belkin et al. [24] with the idea of replacing the adjacency matrix of a Graph (that is used to compute the Laplacian) with a cross-examples or cross-features similarity score where the computation of the graph Laplacian remains unchanged. In this manner, one can regularize a kernel machine while modeling the implicit geometry of the feature space, regardless of the distribution of their labels. The same feature-based approach was pursued by an end-to-end trainable approach for image denoising [26]. On the other hand, there exist *supervised* HAR methods, *e.g.* [50] and [51] directly exploiting the “raw” adjacency matrix to encode skeletal connectivity.

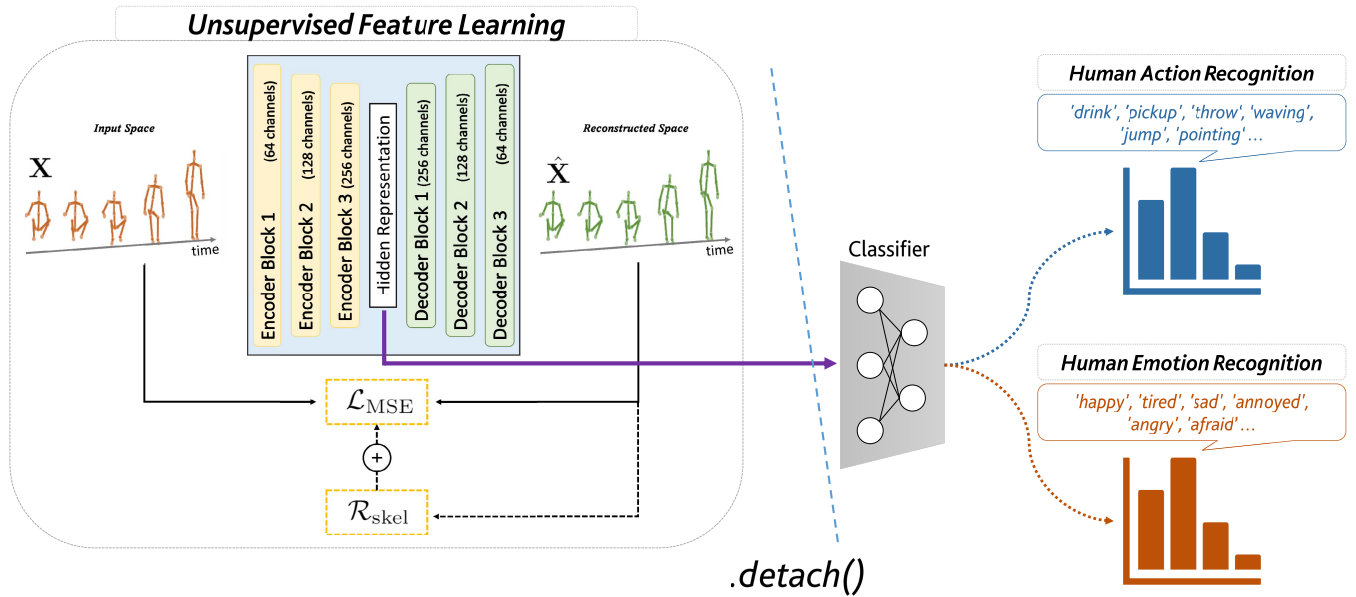
In this work, different from [24] and [26], we apply Laplacian regularization in space while our convolutional residual autoencoder learns to reconstruct input skeletal data, *i.e.* in *reconstruction* space. In this way, unlike [50] and [51], we utilize the graph Laplacian (*i.e.* not explicit adjacency matrices) to include the information of skeletal geometry into our model. Our approach differs from several unsupervised feature learning methods, *e.g.* [15] and [16] that rely on MSE-based feature reconstruction.

## III. PROPOSED METHOD

An overview of the proposed approach is given in Figure 1 and Algorithm 1. We first apply *unsupervised feature learning* when the input of our encoder-decoder architecture is the 3D-skeletal data. This architecture comprises of convolutional residual autoencoder (Section III-A) and the Laplacian regularization (Section III-B). Then, the downstream tasks of HAR and HER are performed from the unsupervised learned features (Section III-C).

### A. CONVOLUTIONAL RESIDUAL AUTOENCODER

The proposed Convolutional Residual Autoencoder (CR-AE) input is a temporal sequence of 3D-human body joints (skeletal data) extracted from either a video sequence or



**FIGURE 1.** The proposed method performs *unsupervised feature learning* with a *convolutional residual autoencoder* (which is a technical contribution and its details are given in Fig. 2) when the loss function is the *mean-squared error* shown as  $\mathcal{L}_{MSE}$  (see Eq. 1). In the reconstruction space, the information of skeletal geometry is injected by the *Laplacian Regularization* (another technical contribution and it is shown in the figure with  $\mathcal{R}_{skel}$ ) enriching the learned (hidden) feature representations with the skeletal geometry information. Our convolutional encoder and deconvolutional decoder blocks both exploit residual connections. In the inference stage, the hidden representations are fed to a classifier to perform either action recognition or discrete emotion recognition.

by a motion capture (MoCap) system of a subject performing an action or expressing an emotion.

Given an input sequence of 3D-body joints  $\mathbf{X}$ , represented as a  $d \times m \times t$  tensor, containing the  $x, y, z$  coordinates (thus,  $d = 3$ ), the number of joints  $m$ , and the number of timestamps  $t$ , we fix each skeleton sequence to a given temporal length. We target to obtain feature representations by learning an autoencoder that reconstructs the input data  $\mathbf{X}$  using a Mean-Squared Error (MSE) loss such as:

$$\mathcal{L}_{MSE} = \frac{1}{2} \mathbb{E}_{\mathbf{X} \sim \mathcal{B}} [\|\mathbf{X} - \hat{\mathbf{X}}\|_F^2], \quad (1)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm, *i.e.* the Euclidean norm of the vector obtained after flattening the tensor  $\mathbf{X}$ . The MSE loss in (1) is minimized by using Adam optimizer over mini-batches  $\mathcal{B}$ . The reconstructed data are defined as follows:

$$\hat{\mathbf{X}} = \mathbf{D}_\theta \circ \mathbf{E}_\varphi(\mathbf{X}), \quad (2)$$

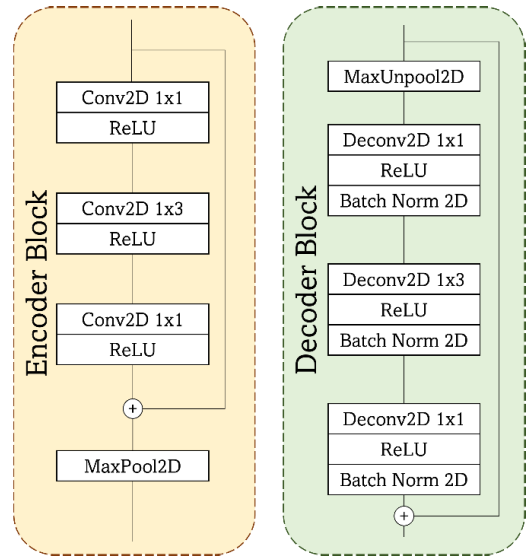
and computed using an encoder-decoder architecture, where  $\varphi$  denotes the learnable parameters of the encoder  $\mathbf{E}$  and  $\theta$  are the parameters for the decoder  $\mathbf{D}$ .

The mean-squared error loss  $\mathcal{L}_{MSE}$  depends upon the learnable parameters  $\theta, \varphi$  updated by mini-batch gradient descent after a forward pass, where we estimate

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{B}} [\mathcal{L}_{MSE}(\theta, \varphi)] = \mathbb{E}_{\mathbf{x} \sim \mathcal{B}} \left[ \|\mathbf{x} - \mathbf{D}_\theta(\mathbf{E}_\varphi(\mathbf{x}))\|_F^2 \right],$$

by averaging the MSE loss  $\mathcal{L}_{MSE}$  over the mini-batch  $\mathcal{B}$ .

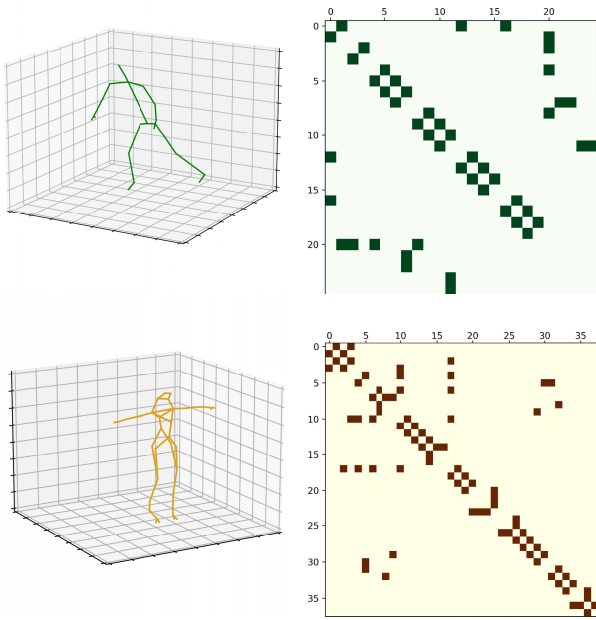
The complete architecture of our convolutional autoencoder is illustrated in Figure 1, where we specify the size of the 2D-convolutional kernel used, and the non-linearities (rectified linear units) with their relative channels. Our



**FIGURE 2.** The details of the architecture of our convolutional encoder ( $\mathbf{E}$ ) and deconvolutional decoder blocks ( $\mathbf{D}$ ), where we specify the size of the 2D-convolutional kernel used inside  $\mathbf{E}$  and  $\mathbf{D}$  (either  $1 \times 1$  or  $1 \times 3$ ) and the non-linearities (rectified linear units). Both  $\mathbf{E}$  and  $\mathbf{D}$  exploit residual connections while batch normalization is exclusive for the decoder.

**CR-AE** architecture stacks different *fully-residual blocks* for both  $\mathbf{E}$  and  $\mathbf{D}$  (three residual blocks for each), and each block is made of convolutions capable of jointly learning spatial representations of skeletal data in time by treating each sample  $\mathbf{X}$  as 2D convolutions, with fixed size kernels, either  $1 \times 1$  or  $1 \times 3$  (see Figure 2 for a graphical demonstration).

In detail, within the  $\mathbf{E}$  blocks, the residual layer is made of a series of three *2D-convolutional layers*, each with *ReLU*



**FIGURE 3.** The location of the skeletal joints in NTU-60 [27] (top-left), the corresponding binary adjacency matrix for NTU-60 [27] (top-right), the location of skeletal joints in DMCD [30] (bottom-left), and the corresponding binary adjacency matrix for DMCD (bottom-right).

activations, stacked together. On the other hand, the  $\mathbf{D}$  blocks share a similar structure but instead use  $2D$ -deconvolutional layers with the addition of  $2D$ -BatchNorm applied after each  $ReLU$  activation. We also apply a  $MaxPool$  layer at the end of each  $\mathbf{E}$  block, and a  $MaxUnpool$  layer at the beginning of each  $\mathbf{D}$  block. At the end of the  $\mathbf{E}$ , a fully-connected ( $FC$ ) layer represents the latent space  $\mathbf{z}$  of size 2048.

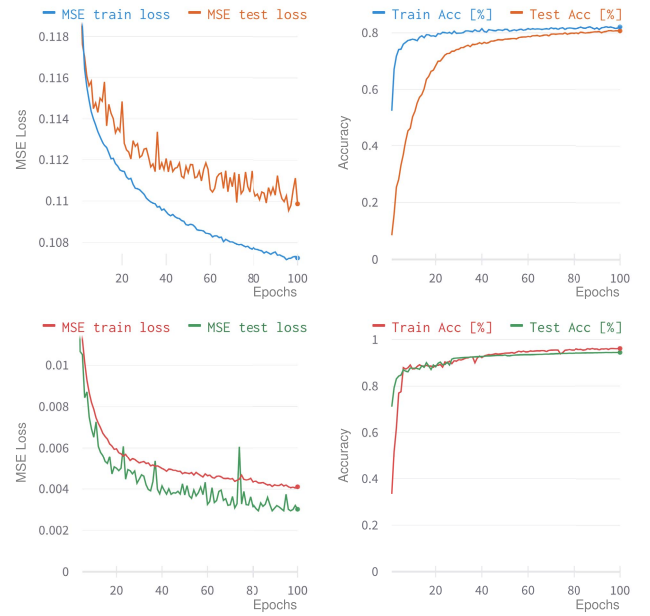
### B. SKELETAL LAPLACIAN REGULARIZATION

The graph Laplacian is defined as an adjacency matrix  $\mathbf{W}$ , whose entries  $W_{ij}$  are defined such that  $W_{ij} = 1$  if and only if the nodes  $i$  and  $j$  are connected through an edge. The not-normalized graph Laplacian  $\mathbf{L}$  is computed by  $\mathbf{W}$  as  $\mathbf{L} = \mathbf{D} - \mathbf{W}$ , where  $\mathbf{D}$  is the degree matrix, obtained as the diagonal matrix where its  $(i, i)$ -th element is  $D_{ii} = \sum_j W_{ij}$  [52].

The Laplacian regularizer:

$$\mathcal{R}(\mathbf{z}) = \sum_{i,j} W_{ij}(z_i - z_j)^2, \quad (3)$$

is applied to a hidden vectorial feature embedding  $\mathbf{z}$  to learn the geometry of the feature space where  $\mathbf{z}$  belongs to [24]. In detail, by having the weights  $W_{ij}$ , one can prioritize the alignment between the scalar components  $z_i$  and  $z_j$  by simply enforcing a stronger penalty between pairs of components that must be well aligned. Herein, we aim to apply this by considering the physical characteristics of the skeletal joints. For example, we define an edge between two joints, *e.g.* from the shoulder and the elbow joints, as they cannot be decorrelated to each other since those joints are close in space (*i.e.* connected by a bone). Differently, there can be joints, which are more distant in space (*e.g.* left foot and right hand, which are indeed not connected by a bone) that



**FIGURE 4.** The learning curves of our CR-AE model. We provide train/test MSE loss (left) and accuracy values (right) of CR-AE trained on NTU-60 xsub (top) and DMCD (bottom).

are allowed to be more independent. In this way, we can inject the knowledge of **skeletal geometry** into the feature representation learning through Laplacian regularization (see Figure 3). The results given in Table 2 also justify that such a setting improves the learning as compared to initializing  $\mathbf{W}$  in several other ways.

The  $\mathcal{R}$  in (3) is termed Laplacian regularizer because:

$$\mathcal{R}(\mathbf{z}) = 2\mathbf{z}^\top \mathbf{L}\mathbf{z}, \quad (4)$$

and  $\mathcal{R}(\mathbf{z})$  implements a “ $\mathbf{L}$ -weighted weight decay” as  $\mathcal{R}(\mathbf{z}) = \|\mathbf{Q}\mathbf{z}\|_2^2$  if we set  $\mathbf{Q} = \sqrt{\mathbf{L}}$ . Different from other methods [24], [26], we apply Laplacian regularization to the **reconstruction space** learned by our decoder, *i.e.* the space where  $\hat{\mathbf{X}}$  belongs to. Therefore, the proposed **skeletal Laplacian regularizer** is computed as:

$$\mathcal{R}_{skel} = \mathbb{E}_{\mathbf{X} \sim \mathcal{B}} \left[ \mathbb{E}_{t,d} \left[ \hat{\mathbf{x}}^{(t,d)\top} \mathbf{L} \hat{\mathbf{x}}^{(t,d)} \right] \right], \quad (5)$$

where  $\hat{\mathbf{x}}^{(t,d)}$  is the  $m$ -dimensional column vector stacking the scalar (abscissæ, ordinatæ or quotæ) coordinates along the dimension  $d$  obtained from the reconstructed sequence  $\hat{\mathbf{X}}$  at time  $t$ . In (5), the regularizer  $\mathcal{R}_{skel}$  is averaged over the mini-batch  $\mathcal{B}$ , considering the reconstructions produced by the convolutional autoencoder across coordinates and timestamps.

### C. INFERENCE: HUMAN ACTION & EMOTION RECOGNITION

Recalling that we target two downstream tasks in this paper: *i)* human action recognition and *ii)* human emotion recognition, we perform the following evaluation protocols:

- **Linear Evaluation Protocol (LEP):** This is the most standard evaluation protocol for unsupervised feature

learning [15], [17], [18], [19], [20], [21], [37]. A downstream task verifies the methods by attaching a linear classifier (a fully-connected layer followed by a softmax layer) to the *frozen* encoder (shown as **E** in Section III-A). Then, the linear classifier is trained by using the available labels.

- **1-Nearest Neighbor Predictor (1-NN):** Another standard evaluation protocol is applying a 1-nearest neighbor predictor [16]. In detail, the class inference of a test data  $\tilde{\mathbf{X}}$  is performed by applying a 1-nearest neighbor predictor, fed by  $\mathbf{E}_\varphi(\tilde{\mathbf{X}})$ , and exploiting a Euclidean Gram matrix computed over the whole training set, which, in turn, is obtained using the splits of the datasets.

#### IV. EXPERIMENTAL ANALYSIS

We conducted extensive experiments on three human action recognition (HAR) and two human discrete emotion recognition (HER) datasets to validate the proposed method. The experimental analysis of these benchmarks shows the potential of the proposed solution, *i.e.* **learning feature representation in an unsupervised manner by using convolutional residual autoencoder and Laplacian regularization, no matter there exists variability in the input skeleton data in terms of the number of joints, adjacency matrices, the sensors used to capture them and sensor sampling rates.**

In detail, first, we addressed the problem of HAR from depth sensors, *i.e.* Microsoft Kinect, in which the 3D skeleton landmarks correspond to the joints of the body skeleton, extracted from the RGB-Depth images. Then, we tested our method for the skeleton datasets curated using the off-the-shelf 3D-pose estimators. It is important to notice that, in such data collection procedures, the position of the joints can be noisy. Next, we addressed the relatively new emerging problem of finding the relationship between emotions and the full-body movements represented by 3D-skeletal data over time (referred shortly as HER for the rest of the paper). For HER, the full-body landmarks correspond to markers placed on the body, tracked with a high temporal frequency by a MoCap system, which might result in less noisy 3D-position data. However, the number and the location of the markers lying on the body change depending on the context of the HER datasets (*e.g.* dance, daily-life actions). Also, the *discrete emotions* *e.g.* anger, disgust, fear, joy, *etc.*, possess distinctive neurophysiological, physiognomic, motivational, and phenomenological properties. Even though there exist common classes among emotion datasets, there are also novel classes in a dataset *w.r.t.* to another.

The readers can see Section IV-A for the detailed description of each dataset used. We describe the implementation details of the proposed method in Section IV-B. Following that, the results of the ablation study (Section IV-C), performance comparisons against SOTA (Section IV-D), qualitative analysis (Section IV-E), results of applying finetuning protocol and end-to-end training (Section IV-F), linear evaluation with fewer data (Section IV-G), and the transfer-ability of

the proposed method (Section IV-H) are given, respectively. In Section 11, we also report the inference time and space complexity of our model and the other methods.

#### A. DATASETS

The utilized benchmark datasets are different in terms of the action and emotion class labels, the number and position of the joints, and the way of obtaining the 3D-skeleton data. Below, we summarize the characteristics of each.

##### 1) NTU-60 ACTION RECOGNITION DATASET

NTU-60 [27] contains 60 action classes performed by 40 subjects and was captured with Microsoft Kinect v2. The corresponding 3D-skeleton data is composed of 25 joints. We evaluated the proposed method on the NTU-60 dataset for cross-view (*xview*, including the second and third views as the training split, and using the first view as test set) and cross-subject (*xsub*, including 20 subjects as training and 20 other subjects as test set) settings. The standard evaluation metric of this dataset is accuracy.

##### 2) NTU-120 ACTION RECOGNITION DATASET

NTU-120 [28] encompasses 120 action classes of 106 subjects while there are 32 different setups in total, referring to, *e.g.* different backgrounds or locations where the data is captured. The *xsub* setting includes 53 subjects in the training set and 53 subjects in the testing set. The *xsetup* setting includes 16 setups in the training split and the other 16 setups in the test split. For this dataset, we used accuracy as the evaluation metric in line with the prior art.

##### 3) SKELETICS-152 ACTION RECOGNITION IN-THE-WILD DATASET

Skeletics-152 [29] was made from the Kinetics-700 dataset [53] by discarding some of the Kinetics-700 dataset's data due to being unfeasible and/or irrelevant to skeleton-based HAR. For example, videos containing occluded poses, egocentric videos, and videos composed of object interactions were omitted by [29]. Afterwards, VIBE [54] algorithm, and some post-processing steps were applied, resulting in 125621 3D-skeleton sequences corresponding to 152 action classes. As NTU datasets, Skeletics-152 has also defined training and testing splits, which we use as provided. The standard evaluation metric of this dataset is accuracy.

##### 4) DANCE MOTION CAPTURE EMOTION DATABASE

DMCD [30] consists of various dance performance recorded with PhaseSpace Impulse X2 MoCap system. The contemporary dance sequences were performed by six participants having different dance-related backgrounds. Each choreography performed by the artists is associated with one of 12 emotions: excited, happy, pleased, satisfied, relaxed, tired, bored, sad, miserable, annoyed, angry, and afraid. There are in total 108 performances (12 emotions  $\times$  9 as 3 artists performed two trials per emotion) corresponding to 614898 3D points captured with 38 markers. We followed the cross-validation

**TABLE 1.** Hyper-parameter tuning for our *CR-AE-L* trained on NTU-60 [27] dataset. Each row reports the Accuracy (%) when a single hyperparameter is changed and the others are kept the same as the final model. The hyperparameter values given in bold correspond to the final model.

	NTU-60 [27]	
	<i>xsub</i>	<i>xview</i>
# of Residual blocks		
1	28.2	37.0
2	51.7	69.8
<b>3</b>	69.9	85.4
# of Convolutional layers		
1	32.1	42.9
2	57.5	71.6
<b>3</b>	69.9	85.4
# of Convolutional channels		
[16, 32, 64]	49.9	65.2
[32, 64, 128]	61.7	78.8
<b>[64, 128, 256]</b>	69.9	85.4
Convolutional kernel size		
1 × 1	59.8	70.1
3 × 3	66.2	81.8
3 × 5	62.2	77.4
<b>1 × 3</b>	69.9	85.4
Latent space <b>z</b> size		
128	42.2	53.9
512	59.4	74.7
<b>2048</b>	69.9	85.4
Learning rate		
1e-5	56.6	79.1
1e-4	67.2	82.2
<b>1e-3</b>	69.9	85.4

setting applied in [6] for all the experimental analyses performed. We used the standard evaluation metric of this dataset: the F1-score, to perform fair comparisons with the SOTA.

##### 5) EMILYA EMOTIONAL BODY EXPRESSIONS DATASET

Emilya is a 3D-MoCap dataset [9] of emotional body expressions during eight daily actions: simple walking, walking with an object in hands, moving books on a table, knocking, sitting down, being seated, lifting, and throwing. The dataset was collected with 28 markers from 12 people who performed the actions mentioned earlier associated with eight emotional states: anxiety, pride, joy, sadness, panic, fear, shame, anger, and neutral. Prior papers have applied two types of cross-validation on the Emilya dataset. We followed both of them and indicated each in the table caption where the experimental results are declared (*i.e.* Table 7). We used the evaluation metric, *i.e.* accuracy, in line with the prior art.

## B. IMPLEMENTATION DETAILS

The experimental analysis reported in this paper was obtained when our model was trained for 100 epochs using Adam optimizer with a learning rate of  $10^{-3}$  and the batch size was 128. The hidden representation layer of our model was a fully-connected layer with a size of 2048. Figure 4 shows the learning curves of our model after applying z-normalization. As seen in that figure, our model achieves a stable performance at the testing time across training epochs. This is an affirmative characteristic, also showing that we are able to learn representations without over-training. Besides,

Table 1 shows hyper-parameter tuning for our model which was applied on NTU-60 [27] dataset. The final model, whose hyperparameters were defined above, has the best performance out of all combinations we tried on NTU-60, and importantly these hyperparameters were kept the same for all analyses with all datasets.

To correctly compute the Graph Laplacian regularization for HAR and HER, each dataset uses its corresponding adjacency matrix as a fixed weight matrix. As pre-processing of each dataset, we mostly inherit the procedure in P&C [16]. We normalize each skeleton *w.r.t.* bone-length in  $[-1, 1]$  range, regularizing the temporal length of each sample by setting it up 100 time-frames (cutting frames of longer samples, replicating frames for shorter samples, and discarding the missing frames). For HER datasets, on par with [6], we applied 25-frames overlapping time-patches while still retaining the temporal length of 100 frames.

## C. ABLATION STUDY

To demonstrate the effectiveness of our *CR-AE-L*, we compare it with architectures: *CR-AE* (*i.e.* proposed method without Laplacian regularization) and *C-AE* (*i.e.* proposed method without Laplacian regularization and without residual layers). Then, we investigate the effect of initializing the Graph Laplacian weight matrix  $\mathbf{W}$  *w.r.t.* the performance of *CR-AE-L*. As mentioned earlier, our *CR-AE-L* promotes the alignment of skeletal joints, connected through a bone (*i.e.* an edge exists if and only if the joints are connected), and in this way, we aim to inject the knowledge of skeletal geometry while learning feature representations in an unsupervised manner.

Our proposal is called *Fixed W*, a binary and symmetric  $n \times n$  skeleton adjacency matrix, including the connectivity between pairs of skeletal joints (see Figure 3).  $n$  is equal to the number of joints of each skeleton (*i.e.* 25 joints for NTU-60 [27], NTU-120 [28], and Skeletics-152 [29] while it is 38 and 28 markers for DMCD [30] and [9], respectively.). The  $W_{ij}$  entries of  $\mathbf{W}$  are defined such that  $W_{ij} = 1$  if and only if the joints  $i$  and  $j$  are connected through an edge (*i.e.* a *bone*); otherwise,  $W_{ij} = 0$ . As an alternative to our proposal, we randomly initialize the weight matrix  $\mathbf{W}$  ( $n \times n$ ). This setting is called *Random W*, and the range of  $W_{ij}$  is  $[0, 1]$ . The corresponding results are given in Table 2.

The ablation study (Table 2) shows that *CR-AE-L* improves the performance of the *CR-AE* model, demonstrating the advantages of using Laplacian regularization in all datasets and all evaluation protocols. Especially for HER datasets, this improvement is remarkable, *i.e.* for DMCD dataset [30], the increase is 17.5% in *I-NN* and 10.9% in *LEP*, and for Emilya dataset [9], the increase is 3.4% in *I-NN* and 7.8% in *LEP*. For HAR datasets, the usage of Laplacian regularization brings in +1.8% in *I-NN* +0.7% in *LEP* for NTU-60 *xsub* [27] and +1.4% in *I-NN* +2% in *LEP* for NTU-120 *xsub* [28]. For *xview* and *xsetup*, the boosts are 2.1% in *I-NN* 0.3% in *LEP* for NTU-60 and 0.2% in *I-NN* 0.3% in *LEP* for NTU-120. Similarly, for skeletics dataset [29], *CR-AE-L* achieves



+0.5% performance in *I-NN* and +5.6% performance in *LEP* performance as compared to *CR-AE*.

Results also show that *CR-AE* is preferable to *C-AE*, *i.e.* using residual layers in our design contributes positively to all datasets and evaluation protocols. In detail, *CR-AE* performs +2.2% in *I-NN* +0.7% in *LEP* for NTU-60 *xsub*, and +0.8% in *I-NN* +0.7% in *LEP* for NTU-120 *xsub*. Similar improvements are observed for *xsetup*, *xsub*, and *skeletics-152* [29]. Having residual layers also improves the results of *CR-AE* when it is tested on HER. The obtained improvement of *CR-AE w.r.t. C-AE* is within a margin of 2.5-4.8%.

The comparisons among initializing the Graph Laplacian weight matrix  $\mathbf{W}$  in the proposed way (*i.e. Fixed W*) versus initializing it randomly (*Random W*) show that *fixed W* achieves better performance independent of the number and the position of the joints in the skeletal data, showing that injecting the skeletal geometry into the regularization is useful. The better performance is within a margin of 1.3-2.8% for all HAR datasets and 3.4-7.8% for all settings of HER datasets.

#### D. COMPARISONS AGAINST THE STATE-OF-THE-ART

Herein, we compare our *CR-AE-L* against the state-of-the-art (SOTA) unsupervised and supervised learning methods. It is important to highlight that our main competitors are the methods performing unsupervised feature learning. Still, we include the fully supervised methods in our comparisons to show each dataset's current upper bound performance, and also the gap between unsupervised and supervised methods. The corresponding results are given in Table 3 to 7 for NTU-60 [27], NTU-120 [28], *skeletics-152* [29], DMCD [30], and *Emilya* [9] datasets, respectively.

##### 1) NTU-60 [27]

For NTU-60 *xsub*, the learned features of *CR-AE-L* are superior to any other unsupervised feature learning SOTA. For example, the improvements supplied by the *CR-AE-L* compared to P&C [16] are +3.5% and +3.4%. While exploiting *LEP CR-AE-L* performs better than the approaches based on RNNs [17], [21], performing +11.4% better than AS-CAL [17] and +8.7% than MM-AE [18]. It surpasses VAE-PoseRNN [21], EnGAN-PoseRNN [21] and SkeletonCLR joint [37] by +13.5%, +1.3%, +1.6%, respectively. It also achieves better performance than MS<sup>2</sup>L [22] (+17.4%), which benefits from contrastive learning, motion prediction, and jigsaw puzzle recognition. For NTU-60 [27] *xview*, *CR-AE-L* performs better than all unsupervised counterparts. In detail, it improves the performance by +6.8% and +7.0% over P&C FS [16] and P&C FW [16], respectively within the *I-NN* Protocol. In *LEP* the superiority of *CR-AE-L* is much visible such that it notably exceeds LongT GAN [15] (+37.3%), PCR [20] (+21.9%), AS-CAL [17] (+20.8%), VAE-PoseRNN [21] (+21.6%), MM-AE [18] (+15.2%), EnGAN-PoseRNN [21] (+7.6%) and SkeletonCLR joint [37] (+9%).

We also compare the performance of our *CR-AE-L* with SOTA-supervised skeleton-based HAR approaches, although they are not our direct competitors. This comparison includes kernel-based methods [55], [56] and the methods realizing feature learning [11], [12], [27], [50], [57], [58], [59], [60] with several different deep learning architectures, *e.g.* RNNs, LSTMs, CNNs, and Graph Convolutional Networks (GCNs). Although based on unsupervised learning, *CR-AE-L* can achieve better performance than the fully supervised kernel-based methods [55], [56], with a +7.2% to +19.8% improvement in *xsub* and a +22% to +32.6% improvement in *xview* setting.

It also outperforms several fully supervised deep architectural methods: H-RNN [57] (providing an increase of 10.8% in *xsub* and up to 21.4% in *xview*), Spatial-Temporal LSTM [50] (resulting in a boost of +0.7% in *xsub* and up to +7.7% in *xview*) and part-aware LSTM [27] (achieving an improvement of +7% in *xsub* and up to +15.1% in *xview*) while performing better than temporal CNN (TCN) [58] (up to +2.3%) in *xview* setting. These results show that our unsupervised residual convolutions with Laplacian regularization exceed even supervised GRUs, RNNs, and LSTMs (and variants) for HAR.

Besides the mentioned favorable results of *CR-AE-L* it is important to note that fully supervised techniques, *e.g.* [11], [12], [59], and [60], perform better than *CR-AE-L*. These supervised methods mostly implement GCNs, and some of them additionally adapt LSTMs [61] or a variable temporal dense block [62]. As expected, the best performing method for this dataset is [60], with 92.4% and 96.8% in *xsub* and *xview*, respectively.

##### 2) NTU-120 [28]

For NTU-120 *xsub*, *CR-AE-L* once again performs better than all unsupervised SOTA. This corresponds to +0.7% improvement compared to P&C [16] when *I-NN* is applied, +10.5% increase in the performance *w.r.t.* both AS-CAL [17], and +17.4% improvement compared to PCR [20] in *LEP* On NTU-120 [28] *xsetup*, *CR-AE-L* is the best out of all unsupervised SOTA. It performs better than P&C within the *I-NN* (+2.0%), and in *LEP* it achieves better performance than AS-CAL [17] and PCR [20] by margins of +13.2% and +17.3%, respectively.

The performance gap between the unsupervised and supervised learning methods is bigger in NTU-120 [28] *xsub* and *xsetup* splits compared to the NTU-60 dataset. Still, *CR-AE-L* is able to achieve better performance than other more complex methods, *e.g.* [27], [50], and [63], which rely on variations of LSTM and RNNs. On the other hand, similar to the NTU-60 dataset's results, the best performance achieved in NTU-120 is also based on GCNs (*e.g.* [64] and [60]).

##### 3) SKELETICS-152 [29]

We compare the performance of *CR-AE-L* against supervised and unsupervised SOTA methods. Especially in *LEP*, *CR-AE-*

**TABLE 2.** Ablation study and the effect of Graph Laplacian Weight Matrix ( $W$ ) initialization.  $C$ ,  $R$ ,  $AE$ , and  $L$  stand for convolution, residual layer, autoencoder, and graph Laplacian regularization, respectively. The proposed method: Convolutional Residual AutoEncoder with Graph Laplacian Regularization, is shown as CR-AE-L w/ Fixed  $W$ . The first results were obtained through  $I$ -NN, and the second results were obtained through  $LEP$ . This is shown as  $I$ -NN /  $LEP$ . All the scores are in terms of accuracy (%) except the F1-scores (%) given for DMCD dataset [30].

	NTU-60 [27]		NTU-120 [28]		Skeletics-152 [29]	DMCD [30]	Emilya [9]
	$x_{sub}$	$x_{view}$	$x_{sub}$	$x_{setup}$			
C-AE	50.1 / 68.5	80.4 / 84.3	40.2 / 56.4	44.3 / 60.3	46.2 / 45.0	75.3 / 81.4	52.8 / 55.7
CR-AE	52.3 / 69.2	81.0 / 85.1	41.0 / 57.4	44.5 / 61.8	48.5 / 46.4	78.9 / 86.2	55.3 / 58.1
CR-AE-L w/ Random $W$	52.6 / 69.0	80.3 / 84.8	40.9 / 57.1	42.8 / 60.9	47.7 / 50.3	90.8 / 92.5	71.8 / 74.5
CR-AE-L w/ Fixed $W$ (Ours)	<b>54.1 / 69.9</b>	<b>83.1 / 85.4</b>	<b>42.4 / 59.1</b>	<b>44.7 / 62.4</b>	<b>49.0 / 52.0</b>	<b>96.4 / 97.1</b>	<b>75.2 / 82.3</b>

**TABLE 3.** Performance comparisons on NTU-60 [27] in terms of accuracy (%). Our results are in *ITALIC*. Underlined scores are the ones the proposed method surpasses. \*FS and FW stand for a decoder with “fixed states” and “fixed weights”, respectively [16]. Refer to [65] for the full list of supervised benchmark results. Herein, we only list a few example approaches that our method surpasses as well as the top scorers.

Method	Feature Learning	$x_{sub}$	$x_{view}$
Lie Group [55]	supervised	50.1	<u>52.8</u>
Cavazza <i>et al.</i> [56]	supervised	<u>60.9</u>	63.4
H-RNN [57]	supervised	59.1	64.0
Spatio-Temporal LSTM [50]	supervised	<u>69.2</u>	<u>77.7</u>
Part-Aware LSTM [27]	supervised	<u>62.9</u>	<u>70.3</u>
TCN [58]	supervised	74.3	<u>83.1</u>
VA-LSTM [12]	supervised	79.2	87.7
DGNN [59]	supervised	89.9	96.1
4s-ShiftGCN [11]	supervised	90.7	96.5
CTR-GCN [60]	supervised	92.4	96.8
<b>I-NN</b>			
P&C FS* [16]	unsupervised	50.6	<u>76.3</u>
P&C FW* [16]	unsupervised	<u>50.7</u>	76.1
<b>Ours</b>	unsupervised	<u>54.1</u>	<u>83.1</u>
<b>LEP</b>			
LongT GAN [15]	unsupervised	39.1	<u>48.1</u>
MS <sup>2</sup> L [22]	unsupervised	<u>52.5</u>	–
PCRP [20]	unsupervised	53.9	63.5
VAE-PoseRNN [21]	unsupervised	56.4	<u>63.8</u>
AS-CAL [17]	unsupervised	<u>58.5</u>	64.6
MM-AE [18]	unsupervised	<u>61.2</u>	<u>70.2</u>
EnGAN-PoseRNN [21]	unsupervised	68.6	77.8
SkeletonCLR joint [37]	unsupervised	<u>68.3</u>	76.4
<b>Ours</b>	unsupervised	<u>69.9</u>	<u>85.4</u>

$L$  has promising results *w.r.t.* the supervised SOTA, which performs only 4.1% and 4.4% less than 4s-ShiftGCN [11] and MS-G3D [70], respectively. It is important to notice that 4s-ShiftGCN [11] and MS-G3D [70] are based on multiple numbers of spatial-temporal graph convolutional blocks, *i.e.* more complex than our architecture also requiring fully annotated large-scale training data. The performance gaps between our  $CR$ -AE- $L$  and 4s-ShiftGCN [11] and MS-G3D [70] decreased in this dataset compared to the NTU-60 dataset. As for unsupervised results,  $CR$ -AE- $L$  performs better than  $I$ -NN competitors (+3.9% over P&C FS [16], and +1.6% over P&C FW [16]), exceeding  $LEP$  competitors as well (MS<sup>2</sup>L [22] +31.6%, PCRP [20] +30.9%, AS-CAL [17] +26.1%, LongT GAN [15] +21.3%, and SkeletonCLR joint [37] +14.7%).

#### 4) DMCD [30]

Performance of our proposed  $CR$ -AE- $L$  greatly outperforms both supervised (+22.4% over Beyan *et al.* [6]) and unsupervised counterparts: exceeding P&C FS [16] (+21.3%),

**TABLE 4.** Performance comparisons on NTU-120 [28] in terms of accuracy (%). Our results are in *ITALIC*. Underlined scores are the ones the proposed method surpasses. †Taken from PCRP [20]. Refer to [69] for the full list of supervised benchmark results. Herein, we only list a few example approaches that our method surpasses as well as the top scorers.

Method	Feature Learning	$x_{sub}$	$x_{setup}$
Part-Aware LSTM [27]	supervised	<u>25.5</u>	<u>26.3</u>
Soft RNN [66]	supervised	36.3	44.9
Dynamic Skeletons [63]	supervised	<u>50.8</u>	<u>54.7</u>
Spatio-Temporal LSTM [50]	supervised	<u>55.7</u>	<u>57.9</u>
Internal Feature Fusion [50]	supervised	<u>58.2</u>	<u>60.9</u>
Qihong <i>et al.</i> [67]	supervised	58.4	<u>57.9</u>
DualHead-Net [68]	supervised	88.2	89.3
EfficientGCN-B4 [64]	supervised	88.7	89.1
CTR-GCN [60]	supervised	88.9	90.6
<b>I-NN</b>			
P&C† [16]	unsupervised	41.7	42.7
<b>Ours</b>	unsupervised	<u>42.4</u>	<u>44.7</u>
<b>LEP</b>			
PCRP [20]	unsupervised	41.7	<u>45.1</u>
AS-CAL [17]	unsupervised	48.6	49.2
<b>Ours</b>	unsupervised	<u>59.1</u>	<u>62.4</u>

P&C FW [16] (+11.9%), MS<sup>2</sup>L [22] (+69.8%), PCRP [20] (+66.2%), AS-CAL [17] (+54.4%), LongT GAN [15] (+21.4%), and SkeletonCLR joint [37] (+9.3%).

#### 5) EMILYA [9]

Our  $CR$ -AE- $L$  obtains comparable results *w.r.t.* supervised counterpart Crenn *et al.* [10] and even outperforming Fourati *et al.* [9] (+7.3%). As for comparisons against unsupervised SOTA, our  $CR$ -AE- $L$  is superior than P&C FS [16] (+10.2%), MS<sup>2</sup>L [22] (+49.9%), PCRP [20] (+50.2%), AS-CAL [17] (+35.6%), LongT GAN [15] (+11.6%), and SkeletonCLR joint [37] (+2.1%), showing once again its effectiveness for HER.

## E. QUALITATIVE RESULTS

We show the feature embeddings of  $CR$ -AE- $L$  learned during unsupervised training of it by using t-SNE [71] for the epochs 2, 20, 60, and 100, in Figure 5. For NTU-60 [27], NTU-120 [28], and Skeletics-152 [29] datasets, we randomly selected 10 action classes. For NTU-60 [27], these are: “drink water”, “pickup”, “throw”, “wear jacket”, “hand waving”, “jump up”, “pointing to something with finger”, “put the palms together”, “falling”, and “touch back (backache)”. The actions for NTU-120 [27] are: “tennis bat swing”, “toss a coin”, “move heavy objects”, “shake fist”, “throw up

**TABLE 5.** Performance comparisons on Skeletics-152 [29] in terms of accuracy (%). Our results are in *ITALIC*. Underlined scores are the ones the proposed method surpasses. \*FS and FW stand for a decoder with “fixed states” and “fixed weights”, respectively [16].

Method	Feature Learning	ACC
4s-ShiftGCN [11]	supervised	56.1
MS-G3D [70]	supervised	56.4
<b>1-NN</b>		
P&C FS* [16]	unsupervised	<u>45.1</u>
P&C FW* [16]	unsupervised	<u>47.4</u>
<b>Ours</b>	unsupervised	<u>49.0</u>
<b>LEP</b>		
MS <sup>2</sup> L [22]	unsupervised	20.4
PCRP [20]	unsupervised	21.1
AS-CAL [17]	unsupervised	25.9
LongT GAN [15]	unsupervised	30.7
SkeletonCLR joint [37]	unsupervised	37.3
<b>Ours</b>	unsupervised	<u>52.0</u>

**TABLE 6.** Performance comparisons on DMCD [30] in terms of F1-score. Our results are in *ITALIC*. Underlined scores are the ones the proposed method surpasses. \*FS and FW stand for a decoder with “fixed states” and “fixed weights”, respectively [16].

Method	Feature Learning	F1-score
Beyan <i>et al.</i> [6]	supervised	74.7
<b>1-NN</b>		
P&C FS* [16]	unsupervised	<u>75.1</u>
P&C FW* [16]	unsupervised	<u>84.5</u>
<b>Ours</b>	unsupervised	<u>96.4</u>
<b>LEP</b>		
MS <sup>2</sup> L [22]	unsupervised	27.3
PCRP [20]	unsupervised	30.9
AS-CAL [17]	unsupervised	42.7
LongT GAN [15]	unsupervised	75.7
SkeletonCLR joint [37]	unsupervised	87.8
<b>Ours</b>	unsupervised	<u>97.1</u>

cap/hat”, “cross arms”, “arm circles”, “running on the spot”, “side kick”, and “stretch oneself”. For Skeletics-152 [29] the selected actions are: “robot dancing”, “dancing gangnam style”, “chopping wood”, “jumping into pool”, “moon walking”, “archery”, “sword fighting”, “belly dancing”, “salsa dancing”, “using a sledge hammer”. Feature embeddings of *CR-AE-L* are more clustered in NTU-60 compared to NTU-120 [27] and Skeletics-152 [29] dataset. This is in line with the quantitative results of *CR-AE-L* in which it performs numerically better in NTU-60 (see Section IV-D). On the other hand, one can observe more compact and less overlapping clusters after the epoch of 20 for all datasets.

#### F. FINETUNE PROTOCOL AND END-TO-END TRAINING

As an additional investigation, we also analyzed the performance of the *CR-AE-L* with the following settings.

- **Finetune Protocol [37]:** This refers to first end-to-end pre-training of our *CR-AE-L* in an unsupervised way. Then append a linear classifier to the encoder and finetune the whole model for the target task (in our case, it is either HAR or HER). Therefore, this protocol is *supervised*.
- **End-to-end Training:** This refers to *fully supervised* learning of our *CR-AE-L* from scratch using the class labels of the training data.

**TABLE 7.** Performance comparisons on Emilya [9] in terms of accuracy (%). Our results are in *ITALIC*. Underlined scores are the ones the proposed method surpasses. \*FS and FW stand for a decoder with “fixed states” and “fixed weights”, respectively [16].  $\diamond$  and  $\nabla$  stand for the cross-validation set-up applied in [10], and [9], respectively.

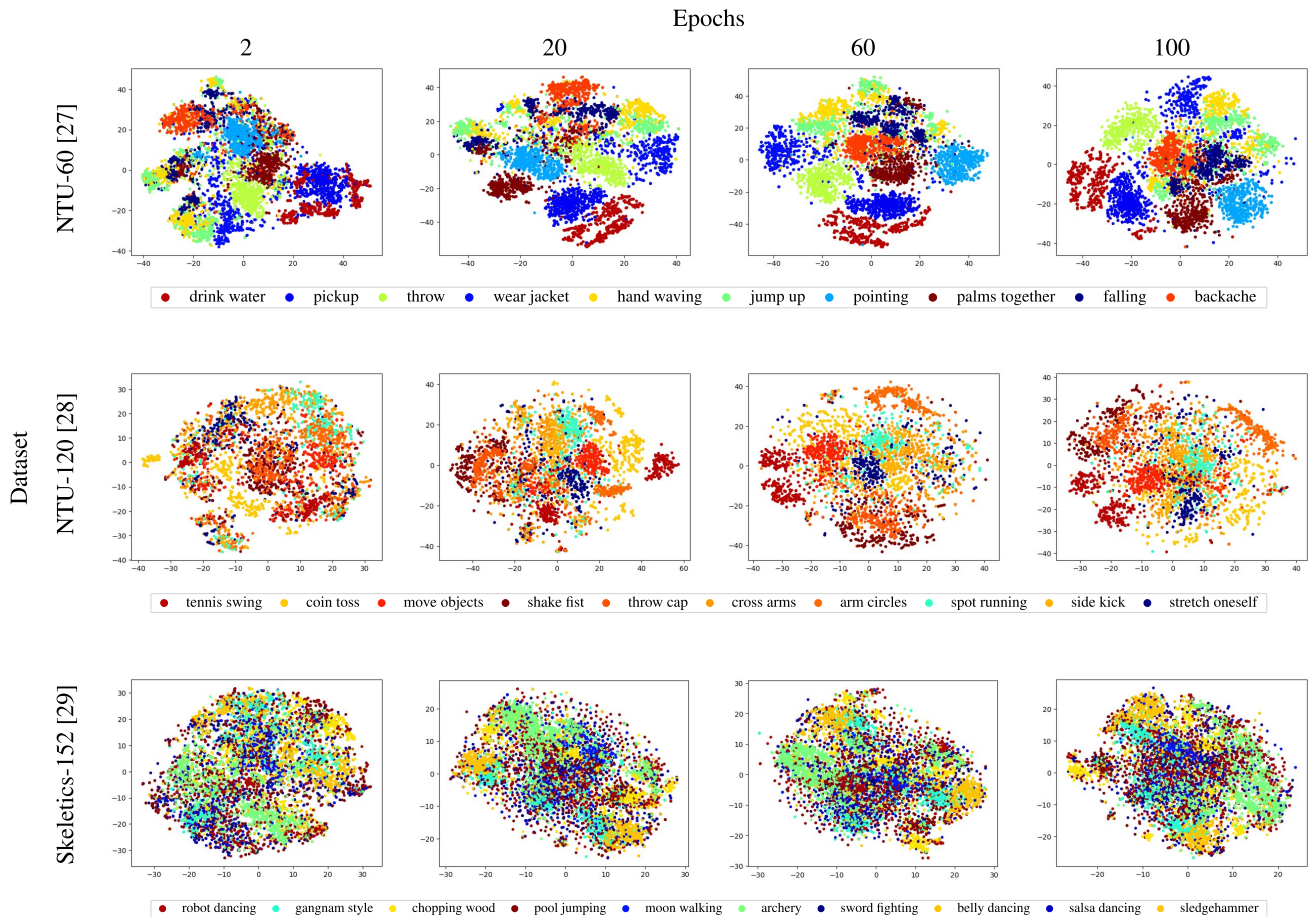
Method	Feature Learning	ACC
Fourati <i>et al.</i> [9] $\nabla$	supervised	75.0
Beyan <i>et al.</i> [6] $\nabla$	supervised	90.5
Crenn <i>et al.</i> [10] $\diamond$	supervised	82.2
Beyan <i>et al.</i> [6] $\diamond$	supervised	91.3
<b>1-NN</b>		
P&C FS* [16] $\diamond$	unsupervised	65.0
P&C FW* [16] $\diamond$	unsupervised	76.8
<b>Ours</b> $\nabla$	unsupervised	<u>71.8</u>
<b>Ours</b> $\diamond$	unsupervised	<u>75.2</u>
<b>LEP</b>		
MS <sup>2</sup> L [22] $\diamond$	unsupervised	32.4
PCRP [20] $\diamond$	unsupervised	32.1
AS-CAL [17] $\diamond$	unsupervised	46.7
LongT GAN [15] $\diamond$	unsupervised	70.7
SkeletonCLR joint [37] $\diamond$	unsupervised	80.2
<b>Ours</b> $\nabla$	unsupervised	<u>76.4</u>
<b>Ours</b> $\diamond$	unsupervised	<u>82.3</u>

Notice that, when applying *LEP* and *1-NN* (see Section III-C for the definitions), the encoder is frozen (*i.e.* the encoder is *detached*) and the feature learning is *unsupervised*. Besides, *1-NN* does not learn any classifier but relies only on a distance metric. On the other hand, the encoder is *not frozen* in the application of the finetune protocol and the end-to-end training, *i.e.* it is *learnable* and, the proposed *CR-AE-L* is no longer unsupervised as aimed in this paper. We inherited these evaluation protocols in line with the recent SOTA, *e.g.* [37] to show that *CR-AE-L* is flexible to adjust between supervised and unsupervised settings. The corresponding results are given in Table 8. It is important to highlight that for this set of experiments, we did not optimize the training procedure (*e.g.* by adjusting the hyper-parameters of *CR-AE-L*). Instead, we kept all implementation settings as it was used in unsupervised training (Section IV-D), to supply direct comparisons with *LEP*. In some cases the finetune and end-to-end protocol results are lower than the *w.r.t.* *LEP* performance (*e.g.* NTU-60 *xview* [27], Emilya [9]), while still achieving **better scores** than several **supervised SOTA**. We argue that these results can be improved by performing a hyperparameter search on the validation sets.

#### G. LINEAR EVALUATION PROTOCOL WITH FEWER TRAINING DATA

To better examine the learning capability of our *CR-AE-L*, we first train them in an unsupervised way (as described in Section III) with all training data. During inference, we follow the *LEP*, but the linear classifier is trained with only 1%, 25%, 50%, and 75% randomly selected data, while keeping the class balance the same as the original datasets. Also, we did not perform a hyper-parameter search for these experiments and kept all settings as in Section IV-D.

The results in Table 9 show that, for all cases, when the percentage of the training data is increased, the performance



**FIGURE 5.** The t-SNE visualization of feature embeddings at different epochs when training CR-AE-L. 10 random categories are sampled and visualized with different colors. As epochs, we select epoch 2, 20, 60, and 100 respectively. For each dataset we report the corresponding action labels w.r.t. cluster colours.

**TABLE 8.** Performance of the proposed method when the finetune protocol and the end-to-end training are applied. All the scores are in terms of accuracy (%) except the F1-scores (%) given for the DMCD dataset [30].  $\uparrow$   $\downarrow$  and  $\leftrightarrow$  stand for the performance improvement, decrease and no-change, respectively with respect to LEP results obtained for the proposed method.

	Finetune Protocol [37]	End-to-end Training
NTU-60 [27] <i>xsub</i>	69.9 $\leftrightarrow$	70.5 $\uparrow$
NTU-60 [27] <i>xview</i>	83.7 $\downarrow$	83.8 $\downarrow$
NTU-120 [28] <i>xsub</i>	57.1 $\downarrow$	57.5 $\downarrow$
NTU-120 [28] <i>xsetup</i>	59.6 $\downarrow$	61.1 $\downarrow$
Skeletics-152 [29]	45.5 $\downarrow$	54.3 $\uparrow$
DMCD [30]	97.2 $\uparrow$	97.4 $\uparrow$
Emilya [9]	80.7 $\downarrow$	76.9 $\downarrow$

of the proposed method also improves. Moreover, our CR-AE-L is able to surpass several SOTA when it is trained on much less data (e.g. 25%, 50%) compared to the amount of the data SOTA is trained on (i.e. 100%). In detail,

- **NTU-60 [27] *xsub*.** By using 25% of the data, our CR-AE-L is able to achieve better results compared to Lie Group [55], Cavazza et al. [56], H-RNN [57],

**TABLE 9.** Performance of the proposed method when the Linear Evaluation Protocol is applied with fewer labels. All the scores are in terms of accuracy (%) except the F1-scores (%) given for DMCD dataset [30].

	Data Percentage			
	1%	25%	50%	75%
NTU-60 [27] <i>xsub</i>	32.3	60.9	65.0	66.4
NTU-60 [27] <i>xview</i>	26.5	70.5	76.0	78.6
NTU-120 [28] <i>xsub</i>	16.1	47.3	50.7	51.6
NTU-120 [28] <i>xsetup</i>	18.9	49.7	53.7	55.1
Skeletics-152 [29]	8.8	25.0	31.6	36.5
DMCD [30]	23.0	71.7	81.3	86.5
Emilya [9]	26.4	61.5	68.7	71.9

P&C [16], LongT GAN [15], MS<sup>2</sup>L [22], PCRP [20], VAE-PoseRNN [21] and AS-CAL [17], whose are trained with 100% of the data.

- **NTU-60 [27] *xview*.** When we use 50% of the training data, our CR-AE-L surpasses the performance of Lie Group [55], Cavazza et al. [56], H-RNN [57], LongT GAN [15], MS<sup>2</sup>L [22], PCRP [20], VAE-PoseRNN [21], AS-CAL [17] and MM-AE [18] trained with the whole training data.

**TABLE 10.** The transfer-ability of our *CR-AE-L* across different datasets. Unsupervised pre-training is performed *w.r.t.* each dataset's training/testing split (except DMCD and Emilya, in which cross-validation is applied as in [6]). NTU 61~120 refers to using only the action classes from 61 to 120. The darker colors perform better than the lighter colors in the same column.

	Tested on									
	NTU-60 [27]		NTU-120 [28]		NTU-61~120 [28]		Skeletics-152 [29]	DMCD [30]	Emilya [9]	
	<i>xsub</i>	<i>xview</i>	<i>xsub</i>	<i>xsetup</i>	<i>xsub</i>	<i>xsetup</i>				
Pre-trained on	NTU-60 [27] <i>xsub</i>	54.1	82.2	42.1	46.4	46.6	43.4	48.9	75.1	76.4
	NTU-60 [27] <i>xview</i>	54.6	83.1	42.0	46.2	45.8	45.1	49.1	92.7	75.1
	NTU-120 [28] <i>xsub</i>	52.0	81.0	42.4	44.3	44.0	45.2	48.0	92.6	74.7
	NTU-120 [28] <i>xsetup</i>	52.3	81.2	38.9	44.7	44.0	45.4	47.9	92.7	74.7
	NTU-61~120 [28] <i>xsub</i>	55.6	52.1	39.4	43.8	45.1	46.4	48.9	92.7	76.4
	NTU-61~120 [28] <i>xsetup</i>	54.3	53.3	39.9	44.4	45.2	46.1	48.1	92.6	75.1
	Skeletics-152 [29]	47.8	71.3	35.4	39.1	42.2	44.0	49.0	92.7	74.7
	DMCD [30]	51.3	70.4	39.1	44.5	44.9	45.2	47.6	96.4	75.0
	Emilya [9]	48.3	70.8	38.5	43.7	44.6	45.0	47.1	82.7	75.2

**TABLE 11.** Space (in terms of the number of parameters) and time (in terms of inference time of one epoch in seconds) complexity of our proposed *CR-AE-L* and unsupervised counterparts. All experiments were performed on the machine equipped with an Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz, 64GB RAM, and a single NVIDIA RTX2080 GPU. The lower space and time complexity is preferable (the best out of all shown in bold). \*FS and FW stand for a decoder with "fixed states" and "fixed weights", respectively [16].

	Space Complexity		Time Complexity					
	# of Parameters	NTU-60 [27]	NTU-120 [28]	Skeletics-152 [29]	DMCD [30]	Emilya [9]		
		<i>xsub</i>	<i>xview</i>	<i>xsub</i>	<i>xsetup</i>			
P&C FS* [16]	57.7M	35.06 s	43.06 s	104.57 s	123.58 s	24.68 s	29.07 s	28.75 s
P&C FW* [16]	57.7M	35.35 s	40.58 s	104.10 s	123.74 s	24.12 s	28.90 s	27.46 s
MS <sup>2</sup> L [22]	11.2M	24.66 s	29.81 s	64.63 s	69.59 s	17.14 s	24.37 s	17.92 s
SkeletonCLR joint [37]	3.6M	16.96 s	19.36 s	51.53 s	58.99 s	11.52 s	20.25 s	15.20 s
LongT GAN [15]	10.2M	15.84 s	18.85 s	64.56 s	80.32 s	11.40 s	14.05 s	12.47 s
PCRP [20]	19.4M	14.30 s	16.44 s	41.97 s	48.97 s	10.39 s	10.84 s	11.27 s
AS-CAL [17]	<b>340K</b>	9.42 s	10.37 s	28.45 s	33.54 s	6.71 s	10.19 s	8.08 s
<b>Ours</b>	38.5M	<b>3.41 s</b>	<b>3.91 s</b>	<b>9.91 s</b>	<b>11.91 s</b>	<b>2.52 s</b>	<b>3.84 s</b>	<b>3.08 s</b>

- **NTU-120 [28] *xsub*.** By training our *CR-AE-L* with the 50% of the training data, we achieve better results compared to Part-Aware LSTM [27], Soft RNN [66], P&C [16], PCRP [20] and AS-CAL [17] trained by using 100% of the data.
- **NTU-120 [28] *xsetup*.** By using 50% of the training data, our *CR-AE-L* is able to achieve better results compared to Part-Aware LSTM [27], Soft RNN [66], P&C<sup>†</sup> [16], PCRP [20] and AS-CAL [17] whose model are learned with the whole training data.
- **Skeletics-152 [29].** By being trained with the 50% of the training data, our *CR-AE-L* surpasses the methods: MS<sup>2</sup>L [22], PCRP [20], AS-CAL [17] and LongT GAN [15], all trained with the 100% of the data.
- **DMCD [30].** Our *CR-AE-L* trained on 50% of the training data, achieves better performance compared to Beyan et al. [6], P&C [16], MS<sup>2</sup>L [22], PCRP [20], AS-CAL [17] and LongT GAN [15] trained with the whole training data.
- **Emilya [9]** By using 50% of the training data, our *CR-AE-L* surpasses the methods: P&C FS\* [16], MS<sup>2</sup>L [22], PCRP [20] and AS-CAL [17] trained on whole dataset.

#### H. TRANSFER-ABILITY

In this section, we test the transfer-ability of *CR-AE-L* across different datasets. The unsupervised pre-training is considered to be useful in a practical scenario in which (in our case) action and/or emotion classes are varying and labelling new

data is expensive. Herein, we test the transfer-ability of our models across different datasets, when *a*) in the unsupervised training and inference the same task but a different set of classes exist (*e.g.* pre-training on *action* dataset NTU-60 [27] *xsub* → transfer learning on *action* dataset NTU-120 [28] *xsetup*) and *b*) different tasks during unsupervised training and inference are being addressed (*e.g.* pre-training on *action* dataset Skeletics-152 [29] → transfer learning on *emotion* dataset Emilya [9]). The corresponding results are given in Table 10 in terms of *I-NN* protocol.

Overall, due to the domain gap between datasets (*e.g.* variety in actions and emotions), a drop in performance can be expected. Still, results show the effectiveness of our approach in dampening this phenomenon. In many cases, the performance even surpasses their same-dataset baseline. For example, in case of actions → actions, a boost in performance can be observed when NTU-60 [27] *xsub* is tested with a model pre-trained with NTU-61~120 [28] *xsub* and NTU-60 [27] *xview* (+1.5% and +0.5%, respectively); NTU-120 [28] *xsetup* is tested with a model pre-trained on NTU-60 [27] *xsub* and NTU-60 [27] *xview* (+1.7% and +1.5%, respectively); and NTU-61~120 [28] *xsub* is classified by a model pre-trained on NTU-60 [27] *xsub* and NTU-60 [27] *xview* (+1.5% and +0.7%, respectively). On the other hand, for actions → emotions, there are performance improvements (up to +1.2%) when Emilya dataset [9] is recognized by a model pre-trained on NTU-60 [27] *xsub* or NTU-61~120 [28] *xsub*. Overall, the proposed method's

transfer-ability is noticeable, showing its potential to process effectively when implemented in real life.

### I. TIME AND SPACE COMPLEXITY

In Table 11, we report the time complexity of our proposed CR-AE-L and the most prominent unsupervised competitors in terms of the inference time of one epoch using the testing split of both HAR and HER datasets. All analyses were performed with the machine equipped with an Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz, 64GB of RAM, and a single NVIDIA RTX2080 GPU. In the same table, we also declare the space complexity of our model and our counterparts in terms of the number of parameters. Despite our model having higher (or comparable) space complexity in terms of the number of parameters *w.r.t.* to some other architectures, we achieve the lowest per-epoch inference time, proving the effectiveness of using residual convolutional layers instead of relying on contrastive-based approaches, GANs, gated networks, or recurrent networks. It is also noticeable that our method has a low space complexity compared to P&C [16], which is based on recurrent networks.

### V. CONCLUSION

We have introduced a novel unsupervised feature learning method based on convolutional residual autoencoder and adapting Laplacian regularization to capture the skeletal geometry in time. Our method is validated on various large-scale action and emotion datasets. It generalizes well to result in effective feature representations from the input 3D-skeleton sequences whose labels, number of joints, and connections, are significantly varying from one dataset to another. This paper is the first attempt to tackle unsupervised full-body motion-based emotion recognition (HER). We have presented baselines for unsupervised human emotion recognition tasks by benchmarking the SOTA methods, willing to foster future research on this topic.

The proposed method notably achieves better results compared to the unsupervised counterparts in standard evaluation protocols and demonstrates remarkable effectiveness against the supervised SOTA. It performs better than several approaches even when it is trained with fewer data compared to others, also presenting its usefulness in case of limited training data. Importantly, when it is trained on one domain and tested on another, it is able to demonstrate the advantage of performing unsupervised pre-training by scoring on par or by achieving improved results for several cases. The proposed method's faster inference time compared to its counterparts is also notable. The future work will focus on enforcing the spatio-temporal connectivity through regularization over time, and the adaptation of the proposed method for online unsupervised learning.

### ACKNOWLEDGMENT

The research in this article uses the Emilya Database collected by Nesrine Fourati and Catherine Pelachaud (CNRS-ISIR and Sorbonne University, France). The authors warmly

thank them for providing the dataset. They are also grateful to Andreas Aristidou and Yiorgos Chrysanthou (University of Cyprus) for supplying the Dance Motion Capture Database.

### REFERENCES

- [1] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent Kinect-based action recognition algorithms," *IEEE Trans. Image Process.*, vol. 29, pp. 15–28, 2020.
- [2] N. Carissimi, P. Rota, C. Beyan, and V. Murino, "Filling the gaps: Predicting missing joints of human poses using denoising autoencoders," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 1–16.
- [3] K. Cheng, Y. Zhang, X. He, J. Cheng, and H. Lu, "Extremely lightweight skeleton-based action recognition with ShiftGCN++," *IEEE Trans. Image Process.*, vol. 30, pp. 7333–7348, 2021.
- [4] H. Yang, D. Yan, L. Zhang, Y. Sun, D. Li, and S. J. Maybank, "Feedback graph convolutional network for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 164–175, 2021.
- [5] M. Cristani, A. D. Bue, V. Murino, F. Setti, and A. Vinciarelli, "The visual social distancing problem," *IEEE Access*, vol. 8, pp. 126876–126886, 2020.
- [6] C. Beyan, S. Karumuri, G. Volpe, A. Camurri, and R. Niewiadomski, "Modeling multiple temporal scales of full-body movements for emotion classification," *IEEE Trans. Affect. Comput.*, early access, Jul. 7, 2021, doi: 10.1109/TAFFC.2021.3095425.
- [7] S. Piana, A. Staglianò, F. Odone, and A. Camurri, "Adaptive body gesture representation for automatic emotion recognition," *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 1, pp. 1–31, May 2016.
- [8] M. Daoudi, S. Berretti, P. Pala, Y. Delevoe, and A. D. Bimbo, "Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2017, pp. 550–560.
- [9] N. Fourati and C. Pelachaud, "Perception of emotions and body movement in the emilya database," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 90–101, Jan. 2018.
- [10] A. Crenn, A. Meyer, H. Konik, R. A. Khan, and S. Bouakaz, "Generic body expression recognition based on synthesis of realistic neutral motion," *IEEE Access*, vol. 8, pp. 207758–207767, 2020.
- [11] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 183–192.
- [12] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2117–2126.
- [13] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–10.
- [14] G. Paoletti, J. Cavazza, C. Beyan, and A. Del Bue, "Subspace clustering for action recognition with covariance representations and temporal pruning," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 6035–6042.
- [15] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised representation learning with long-term dynamics for skeleton based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2644–2651.
- [16] K. Su, X. Liu, and E. Shlizerman, "PREDICT & CLUSTER: Unsupervised skeleton based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9631–9640.
- [17] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition," *Inf. Sci.*, vol. 569, pp. 90–109, Aug. 2021.
- [18] D. Holden, J. Saito, T. Komura, and T. Joyce, "Learning motion manifolds with convolutional autoencoders," in *Proc. SIGGRAPH Asia Tech. Briefs*, Nov. 2015, pp. 1–4.
- [19] Q. Nie, Z. Liu, and Y. Liu, "Unsupervised human 3D pose representation with viewpoint and pose disentanglement," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer, 2020, pp. 102–118.
- [20] S. Xu, H. Rao, X. Hu, J. Cheng, and B. Hu, "Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition," *IEEE Trans. Multimedia*, early access, Nov. 22, 2021, doi: 10.1109/TMM.2021.3129616.

- [21] J. N. Kundu, M. Gor, P. K. Uppala, and V. B. Radhakrishnan, "Unsupervised feature learning of human actions as trajectories in pose embedding manifold," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1459–1467.
- [22] L. Lin, S. Song, W. Yang, and J. Liu, "MS2L: Multi-task self-supervised learning for skeleton based action recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2490–2498.
- [23] G. Paoletti, J. Cavazza, C. Beyan, and A. Del Bue, "Unsupervised human action recognition with skeletal graph Laplacian and self-supervised view-points invariance," in *Proc. 32nd Brit. Mach. Vis. Conf.*, Nov. 2021, pp. 1–13.
- [24] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, no. 11, pp. 1–36, 2006.
- [25] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception Psychophys.*, vol. 14, no. 2, pp. 201–211, 1973.
- [26] J. Pang and G. Cheung, "Graph Laplacian regularization for image denoising: Analysis in the continuous domain," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1770–1785, Apr. 2017.
- [27] 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. (CVPR)*, June 2016, pp. 1010–1019.
- [28] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.
- [29] P. Gupta, A. Thatipelli, A. Aggarwal, S. Maheshwari, N. Trivedi, S. Das, and R. K. Sarvadevabhatla, "Quo vadis, skeleton action recognition?" *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2097–2112, Jul. 2021.
- [30] DMCD. (2021). *Dance Motion Capture Database*. [Online]. Available: <http://dancedb.eu/>
- [31] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.
- [32] N. C. Garcia, P. Morerio, and V. Murino, "Learning with privileged information via adversarial discriminative modality distillation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2581–2593, Jul. 2020.
- [33] H. Ramirez, S. A. Velastin, I. Meza, E. Fabregas, D. Makris, and G. Farias, "Fall detection and activity recognition using human skeleton features," *IEEE Access*, vol. 9, pp. 33532–33542, 2021.
- [34] Q. Wang, K. Zhang, and M. A. Asghar, "Skeleton-based ST-GCN for human action recognition with extended skeleton graph and partitioning strategy," *IEEE Access*, vol. 10, pp. 41403–41410, 2022.
- [35] R. Li, H. Fu, W.-L. Lo, Z. Chi, Z. Song, and D. Wen, "Skeleton-based action recognition with key-segment descriptor and temporal step matrix model," *IEEE Access*, vol. 7, pp. 169782–169795, 2019.
- [36] J. Cha, M. Saqlain, D. Kim, S. Lee, S. Lee, and S. Baek, "Learning 3D skeletal representation from transformer for action recognition," *IEEE Access*, vol. 10, pp. 67541–67550, 2022.
- [37] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3D human action representation learning via cross-view consistency pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4741–4750.
- [38] M. Karg, A.-A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic, "Body movements for affective expression: A survey of automatic recognition and generation," *IEEE Trans. Affective Comput.*, vol. 4, no. 4, pp. 341–359, Oct. 2013.
- [39] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 505–523, Apr./Jun. 2021.
- [40] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *J. Nonverbal Behav.*, vol. 28, no. 2, pp. 117–139, 2004.
- [41] J. L. Tracy and R. W. Robins, "Show your pride: Evidence for a discrete emotion expression," *Psychol. Sci.*, vol. 15, no. 3, pp. 194–197, 2004.
- [42] R. Niewiadomski, S. J. Hyniewska, and C. Pelachaud, "Constraint-based model for synthesis of multimodal sequential expressions of emotions," *IEEE Trans. Affect. Comput.*, vol. 2, no. 3, pp. 134–146, Jul. 2011.
- [43] N. Dael, M. Goudbeek, and K. R. Scherer, "Perceived gesture dynamics in nonverbal expression of emotion," *Perception*, vol. 42, no. 6, pp. 642–657, Jun. 2013.
- [44] R. Niewiadomski, M. Mancini, S. Piana, P. Alborno, G. Volpe, and A. Camurri, "Low-intrusive recognition of expressive movement qualities," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 230–237.
- [45] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Cham, Switzerland: Springer, 2007, pp. 71–82.
- [46] N. Fourati, C. Pelachaud, and P. Darmon, "Contribution of temporal and multi-level body cues to emotion classification," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 116–122.
- [47] G. Cimen, H. Ilhan, T. Capin, and H. Gurcay, "Classification of human motion based on affective state descriptors," *Comput. Animation Virtual Worlds*, vol. 24, nos. 3–4, pp. 355–363, May 2013.
- [48] A. Kacem, M. Daoudi, B. B. Amor, S. Berretti, and J. C. Alvarez-Paiva, "A novel geometric framework on Gram matrix trajectories for human behavior understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 1–14, Jan. 2020.
- [49] M. R. Loghmani, S. Rovetta, and G. Venture, "Emotional intelligence in robots: Recognizing human emotions from daily-life gestures," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1677–1684.
- [50] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3d human action recognition," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2016, pp. 816–833.
- [51] D. Yang, M. M. Li, H. Fu, J. Fan, and H. Leung, "Centrality graph convolutional networks for skeleton-based action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Mar. 2020, pp. 1–18.
- [52] N. Deo, *Graph Theory With Applications to Engineering and Computer Science*. New York, NY, USA: Dover, 2017.
- [53] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A short note on the kinetics-700–2020 human action dataset," 2020, *arXiv:2010.10864*.
- [54] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5253–5263.
- [55] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2430–2443, Dec. 2016.
- [56] J. Cavazza, P. Morerio, and V. Murino, "Scalable and compact 3D action recognition with approximated RBF kernel machines," *Pattern Recognit.*, vol. 93, pp. 25–35, Sep. 2019.
- [57] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.
- [58] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 20–28.
- [59] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7912–7921.
- [60] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13359–13368.
- [61] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1227–1236.
- [62] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. Xia, "Graph CNNs with motif and variable temporal block for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8989–8996.
- [63] 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. (CVPR)*, Jun. 2015, pp. 5344–5352.
- [64] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "EfficientGCN: Constructing stronger and faster baselines for skeleton-based action recognition," 2021, *arXiv:2106.15125*.
- [65] P. W. Code. (2022). *Skeleton Based Action Recognition Benchmarks on NTU RGB+D 60*. [Online]. Available: <https://paperswithcode.com/sota/skeleton-based-action-recognition-on-ntu-rgb-d>
- [66] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, J. Lai, and J. Zhang, "Early action prediction by soft regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2568–2583, Nov. 2019.

- [67] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3288–3297.
- [68] T. Chen, D. Zhou, J. Wang, S. Wang, Y. Guan, X. He, and E. Ding, "Learning multi-granular spatio-temporal graph network for skeleton-based action recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4334–4342.
- [69] P. W. Code. (2022). *Skeleton Based Action Recognition Benchmarks on NTU RGB+D 120*. [Online]. Available: <https://paperswithcode.com/sota/skeleton-based-action-recognition-on-ntu-rgb-d-1>
- [70] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 143–152.
- [71] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.



interests include unsupervised deep learning, skeleton-based human action, and emotion recognition.

**GIANCARLO PAOLETTI** received the B.S. degree in psychology from the University of Urbino, in 2016, and the M.S. degree in psychology from the University of Turin, Italy, in 2019. He is currently pursuing the Ph.D. degree with the Istituto Italiano di Tecnologia, Department of Pattern Analysis and Computer Vision with the collaboration of the Department of Naval, Electrical, Electronic, and Telecommunications Engineering, University of Genoa, Italy. His current research



**CIGDEM BEYAN** received the Ph.D. degree in informatics from the University of Edinburgh, U.K., in 2015. She is currently an Assistant Professor at the Department of Information Engineering and Computer Science, University of Trento, and an Affiliated Researcher with the Pattern Analysis and Computer Vision Research Line, Italian Institute of Technology. She has coauthored over 50 papers published in peer-reviewed journals and international conferences. Her main research interests include human behavior understanding, social signal processing, and multimodal data analysis. She is a member of ELLIS. She is also a Reviewer of several journals, including IEEE TRANSACTIONS and IEEE/ACM conferences, such as CVPR, ECCV, BMVC, and ACM MM. She was a Guest Co-Editor of *Frontiers in Robotics and AI*. She is on the Editorial Board of *ICES Journal of Marine Science* covering the area of applications of computer vision and machine learning.



**ALESSIO DEL BUE** (Member, IEEE) received the Ph.D. degree from the Department of Computer Science, Queen Mary University of London. He is currently a Tenured Senior Researcher leading the Pattern Analysis and Computer Vision Research Line, Italian Institute of Technology, Genoa, Italy. Previously, he was a Researcher at the Institute for Systems and Robotics, Instituto Superior Técnico (IST), Lisbon, Portugal. He is the coauthor of more than 100 scientific publications, in refereed journals and international conferences, and a member of the technical committees of important computer vision conferences (CVPR, ICCV, ECCV, and BMVC). His current research interest includes 3D scene understanding from multi-modal input (images, depth, and audio) to support the development of assistive AI systems. He is an ELLIS Member of the Genoa Unit. He serves as an Associate Editor for *Pattern Recognition* and *Computer Vision and Image Understanding* journals.

• • •

Open Access funding provided by 'Istituto Italiano di Tecnologia' within the CRUI CARE Agreement