sensors

MDPI

Article

ConMLP: MLP-Based Self-Supervised Contrastive Learning for Skeleton Data Analysis and Action Recognition

Chuan Dai [1], Yajuan Wei [1,2], Zhijie Xu [1,*], Minsi Chen [1], Ying Liu [3] and Jiulun Fan [4]

[1] School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, UK
[2] School of Cyberspace Security, Xi'an University of Posts and Telecommunications, Xi'an 710061, China
[3] International Joint Research Center for Wireless Communication and Information Processing, Xi'an 710121, China
[4] School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710061, China
[*] Correspondence: z.xu@hud.ac.uk

Abstract: Human action recognition has drawn significant attention because of its importance in computer vision-based applications. Action recognition based on skeleton sequences has rapidly advanced in the last decade. Conventional deep learning-based approaches are based on extracting skeleton sequences through convolutional operations. Most of these architectures are implemented by learning spatial and temporal features through multiple streams. These studies have enlightened the action recognition endeavor from various algorithmic angles. However, three common issues are observed: (1) The models are usually complicated; therefore, they have a correspondingly higher computational complexity. (2) For supervised learning models, the reliance on labels during training is always a drawback. (3) Implementing large models is not beneficial to real-time applications. To address the above issues, in this paper, we propose a multi-layer perceptron (MLP)-based self-supervised learning framework with a contrastive learning loss function (ConMLP). ConMLP does not require a massive computational setup; it can effectively reduce the consumption of computational resources. Compared with supervised learning frameworks, ConMLP is friendly to the huge amount of unlabeled training data. In addition, it has low requirements for system configuration and is more conducive to being embedded in real-world applications. Extensive experiments show that ConMLP achieves the top one inference result of 96.9% on the NTU RGB+D dataset. This accuracy is higher than the state-of-the-art self-supervised learning method. Meanwhile, ConMLP is also evaluated in a supervised learning manner, which has achieved comparable performance to the state of the art of recognition accuracy.

Keywords: human action recognition; skeleton data; multi-layer perceptron; self-supervised learning

check for updates

Citation: Dai, C.; Wei, Y.; Xu, Z.; Chen, M.; Liu, Y.; Fan, J. ConMLP: MLP-Based Self-Supervised Contrastive Learning for Skeleton Data Analysis and Action Recognition. *Sensors* **2023**, *23*, 2452. https://doi.org/10.3390/s23052452

Academic Editors: Chao Liu, Wenfeng Zheng, Mingzhe Liu, Dan Wang and Kenan Li

Received: 16 January 2023
Revised: 20 February 2023
Accepted: 21 February 2023
Published: 22 February 2023

1. Introduction

The RGB videos provided by traditional datasets contain rich semantic information, which can be extracted for action recognition and classification, which are fundamental problems in computer vision. However, they also bring various forms of noise. For example, backgrounds unrelated to the actions, poor illumination, and object occlusions [1]. On the other hand, other video representations have emerged due to the development of video capture technologies. Skeleton data generated by depth sensors [2] are highly valuable modalities. They have a low amount of data and take fewer computing resources. Moreover, they are affected relatively little by interference factors, such as illumination and background [1].

Therefore, studies on recognizing and classifying actions through skeleton-based data have been extensively explored. Most of these studies are derived from image-based recognition [3,4]. Convolutional neural network (CNN)-based approaches express skeleton joint coordinates across multiple frames in terms of pseudo images, presenting this as an image processing challenge [3]. As recurrent neural network (RNN)-based methods, skeletons are encoded into a sequence of structured coordinate vectors. They exploit RNNs' ability to deal with time-series data [5], while graph convolutional network (GCN)-based approaches combine the hierarchical features of skeleton data to represent frame sequences as connections and updates in the interrelationships between graph vertices and edges [6,7].

Although a remarkable effect is achieved, several issues are observed: (1) Complex models can be quite expensive in terms of computational resources. The FLOPs (FLoating-point OPerations) of these models tend to be several orders of magnitude larger than those of naive architectures. (2) For supervised learning approaches, the need for massively labeled data is always a disadvantage. (3) The combination of complex algorithms and large models is difficult to truly deploy into production applications.

In contrast, ConMLP, which is proposed in this paper in conjunction with recent advances in MLP, is a simple framework. It does not contain any computationally expensive layers such as convolution operations or attention mechanisms. It is built on a naive MLP as a base encoder network combined with a contrastive loss function, which does not fall into the categories of CNN-, RNN-, or GCN-based approaches. It can be applied to both self-supervised learning and supervised learning. The inference results of ConMLP have a top one of 96.9% on the NTU RGB+D dataset. This accuracy is 0.2% higher than the state-of-the-art self-supervised learning approach [8], and it is on par with the best result of supervised learning [9]. The code of this research is available at: https://github.com/ChuanDai/ConMLP (accessed on 28 December 2022).

Our main contributions are summarized below:

1. A novel MLP-based self-supervised learning framework is proposed, which significantly reduces computational complexity while achieving state-of-the-art performance. It saves more than 95% of FLOPs compared with the ResNet50-based encoder network;
2. The generic contrastive learning metrics of image classification [10] have been transformed to action recognition on skeleton sequences, which demonstrated its flexibility and feasibility for wider applications;
3. For validating the devised framework, ConMLP is also evaluated against benchmarking supervised learning approaches with superior performances recorded.

2. Related Work

2.1. Visual Representation Learning

Visual representation learning methods generally fall into the following three categories: handcrafted pretext tasks, pixel-level generation, and contrastive learning.

2.1.1. Handcrafted Pretext Tasks

To provide reasonable and efficient representations for downstream tasks, such methods need to use pretext tasks for pre-training. However, the formation of such pretext tasks is usually based on heuristic inspiration [11,12]. Even if good results can be reached by a larger network and a long training time, such designs typically have poor generalization [13].

2.1.2. Pixel-Level Generation

In general, adversarial models are used for pixel-level generation. The primary goal of adversarial models is to reproduce data distribution as effectively as is feasible to help recognize objects [14]. The main criticism of this kind of method is the intensive computational burden; additionally, whether it is necessary for representation learning is also worth discussing [15].

2.1.3. Contrastive Learning

Compared with the above two kinds of methods, contrastive learning approaches achieve the best performance, and their interpretability is also recognized [10]. The architecture of combining a base encoder network and a projection network can be supported by the theoretical basis of such methods [14]. Triplet [16] is a supervised contrastive-learning-based model. Apart from the anchor, each mini-batch contains only one positive and one negative sample. This means only one negative sample was compared in a mini-batch, while other classes of negatives were ignored. Meanwhile, it was necessary to exploit hard-negative mining, which was computationally expensive. N-pair [17] is an extended version of Triplet loss. All negative samples in a mini-batch participated in the computation, and more reasonable representations were learned. However, the number of positive samples does not change. SimCLR [18] is characterized by self-supervised loss with an optimized combination of schemes, demonstrating the importance of augmented data. Positive samples are generated by data augmentations, while negative samples are made up of the remaining $2(N-1)$ samples in two copies of a mini-batch. It can achieve much higher performance than Triplet, even if not using hard-negative mining. Additionally, SupCon [14] extended the loss proposed in SimCLR to also adapt supervised contrastive learning.

2.2. Skeleton-Based Action Recognition

2.2.1. Self-Supervised Methods

VaRe [8] is a GCN-based framework combined with a view-normalization generative adversarial network (VN-GAN) and subject-independent network (SINet). This framework could recognize actions without the knowledge of view- and subject-specific habits. SRCL [19] consisted of two networks, an online and a target network, and used the distribution of scores of inter-instance similarity as a relational metric to introduce relational consistency learning. MG-AL [20] treated the self-supervised for action representation learning as a self-attention problem, and it did not involve any data augmentations. CRRL [21] included a two-stage architecture for learning and representation fusing. Additionally, a new data augmentation scheme called velocity was proposed. In [22], a framework that combined the attention mechanism and contrastive learning was proposed. Although no GCN structures were employed, data augmentation still played a critical role. SKT [23] is also characterized by a contrastive learning strategy. It used a Barlow Twins objective function to minimize the agreement between similar samples without negative samples being required. GLTA-GCN [24] is a self-attention-equipped framework with intensive GCN structures. It introduced two losses to serve the framework of a multi-task process. SEMN [25] introduced a skeleton modality called skeleton edge motion and has a loss function to help perform self-supervised learning. In [26], the authors introduced a loss to estimate noise by contrastive learning and used several spatial–temporal based data augmentation schemes. MCAE [27] utilizes the modeling process as two levels, namely a lower level and a higher level for dividing and aggregating the spatial–temporal signal, respectively.

Most of these methods chose GCN as the backbone network [8,19,20,23,24,26]. On the other hand, only a few methods did not employ contrastive loss functions [8,20,25]. This shows that the idea of using contrastive learning to tackle self-supervised learning is mainstream thinking. Moreover, only a few methods did not use data augmentations [20,24,25], which also shows that it is indispensable for contrastive learning. The proposed ConMLP in this paper is a learning framework that combines a contrastive function as the loss with generating pairs of data through augmentation.

In addition to action recognition on skeleton data, self-supervised learning is also applied in several real-world tasks, such as in facial landmark analysis [28,29] and digital healthcare [30].

2.2.2. Supervised Methods

ST-GCN [6] is a spatial–temporal method to tackle human action recognition with GCN, which laid a foundation for subsequent studies. Moreover, 2s-AGCN [31] was proposed to address the limitations of ST-GCN, including its ability to only deal with first-order information. Instead, bone information including length and direction was incorporated in addition to the joint information. Additionally, a data-driven approach was used to parameterize both the global graph and individual graph to extend the flexibility of the model. MS-G3D [7] was a feature extractor by fusing a disentangled multi-scale aggregation scheme and a unified spatial–temporal graph convolution (G3D) operator, and dilated convolutions [32] were applied to multi-scale aggregation, which effectively controlled the complexity of the network architecture. MST-GCN [33] is a multi-scale spatial graph convolution (MS-GC) module combined with a multi-scale temporal graph convolution (MT-GC) module that models distance joints relations and long-range temporal information. In addition, both MS-G3D and MST-GCN considerably increased temporal receptive fields, but they were adopted in different ways. MS-G3D utilized paralleled 3×1 kernel sizes combined with dilated windows, while MST-GCN used a single block of a hierarchical architecture.

Conventional GCN-based methods often have high computational complexity. To reduce the computational burden of GCN, Shift-GCN [34], which was inspired by shift convolution [35], exploited the lightweight shift graph operations to provide flexible receptive fields for both spatial graphs and temporal graphs. However, the proposed ConMLP does not contain any GCN structure. While it achieves state-of-the-art of performance, the computational complexity is substantially reduced, even compared with Shift-GCN.

2.3. Multi-Layer Perceptron

2.3.1. Advances in MLP Architecture

Multi-layer perceptron (MLP)-based models have recently received a lot of attention. MLP-Mixer [36] showed that even without using CNN and self-attention, the model based on MLP could also achieve excellent performance in image classification. MLP-Mixer relied solely on an MLP that was repeatedly implemented in the spatial domain or feature channels, as well as matrix multiplication, data scaling, and nonlinear layers. This was much more efficient in terms of computational complexity.

ResMLP [37] adopted a similar approach to MLP-Mixer, using two MLPs acting on different directions of image patches. Its strength was that ResMLP could rapidly train a high-performing model on a smaller ImageNet branch. gMLP [38] introduced a spatial gating unit into the model without using self-attention. It evaluated performance on both computer vision and natural language processing. Compared with MLP-Mixer, it reduced parameters by 66% while improving the performance by 3%. These models were similar in structure but differ in block design details.

Without blindly emphasizing the importance of MLP architecture, RepMLP [39] and CycleMLP [40] were attempts to combine MLP with CNNs or self-attention mechanisms.

2.3.2. MLP's Resurgence

There are many reasons for the remarkable performance of MLPs, including the increase in computing power and datasets improvements. Moreover, modern MLP models have many commonalities in their implementation. For example, most of them applied Gelu [41] as the activation function [36–38,40], and layer normalization [42] was exploited instead of batch normalization [36,37]. Additionally, adding skip connections [43] was a common method while extending more layers [36,37].

Our paper follows these practices, using Gelu and layer normalization. The skip connection will be investigated in future work. In Section 4.6, the role played by the number of MLP hidden layers is also explored.

3. ConMLP Framework Design

Assume that an input skeleton sequence can be represented as $x = (x_1, x_2, \ldots, x_T)$ containing T consecutive skeleton frames, where $x_i \in \mathbb{R}^{S \times J \times 3}$ contains coordinates of S subjects with J different 3D body joints. The training dataset $\Phi = \left\{x^i\right\}_{i=1}^{N}$ consists of N skeleton sequences from different actions, and their views and subjects may be different. Each skeleton sequence x^i corresponds to a label y_i, where $y_i \in \{a_1, a_2, \ldots, a_c\}$, a_i denotes i^{th}

action class, and c is the number of classes. The goal is to learn a set of valid representations by x^i without introducing any labels.

This paper adopts a widely used evaluation metric [10,18]. First, data augmentations are applied to the input mini-batch data. Two copies of mini-batch data enhanced by augmentations are forward propagated through the base encoder network. Then, the representations of the output are further trained by a projection head, which is not used in the inference stage. Next, the contrastive loss is calculated based on the output of the projection network. Finally, a linear classifier is trained on top of them, while the representations are frozen (Figure 1).

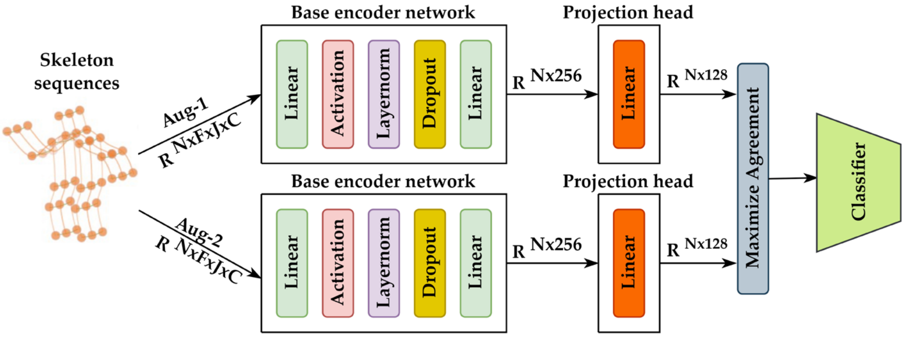

Figure 1. An overlook of the ConMLP framework. N denotes the mini-batch size; F denotes the number of frames in a skeleton sequence; J denotes the number of joints; C denotes the 3D coordinates.

3.1. Data Augmentation Schemes

For self-supervised contrastive learning, data augmentation plays a crucial role. Positive samples of the same class as the anchor are generated through data augmentation, and the selection of an augmentation scheme can directly affect the quality of representation learning [18]. For each piece of the input skeleton sequence, i.e., each action, two sets of augmented data are generated, each representing a varied view of the skeleton sequence.

Different from the augmentation schemes for images, skeleton-based data augmentation mainly includes shear, reverse, rotation, joint mask, Gaussian blur, Gaussian noise, channel mask, etc. Shear and reverse are adopted in this paper for optimal performance [44,45].

Shear transformation is one of the spatial linear transformations. In the context of an image, it can be interpreted as stretching either side of a rectangle, eventually turning it into a parallelogram. However, for skeleton data, each joint is projected in a predefined direction, so that a skeleton frame is projected to a certain viewpoint. A generic 3D shear transformation matrix can be used to generate sheared sequences for skeleton data. The shear transformation matrix can be defined as:

$$S = \begin{pmatrix} 1 & S_X^Y & S_X^Z \\ S_Y^X & 1 & S_Y^Z \\ S_Z^X & S_Z^Y & 1 \end{pmatrix}$$

where $S_X^Y, S_X^Z, S_Y^X, S_Y^Z, S_Z^X, S_Z^Y \in [-1, 1]$ are randomly generated shear factors, which control the amplitude of transformation from one dimension to another. The joints coordinate of the original skeleton sequence can all be transformed by this matrix.

Reverse refers to reversing the view of the temporal order. For NTU RGB+D datasets with (N, F, J, C) structures, the reverse is exploited for the second dimension, i.e., frames. One skeleton sequence has a 50% chance of being reversed. Nevertheless, the reverse is one of the best-proven augmentations for skeleton sequences.

3.2. Base Encoder Network

The base encoder network is to transform input data into representation vectors. The two groups of augmented data are forward propagated through the base encoder network, with a pair of representation vectors being generated.

In this paper, a naive MLP with 256 hidden layers is employed as the base encoder network. Following a common MLP design principle, the base encoder consists of linear layers followed by activation, normalization, dropout, and linear layers (Figure 1). Gelu [41] is applied for activation. For normalization, layernorm [42] is used for training stability instead of applying batch normalization. As a convention, dropout is used to avoid overfitting.

Moreover, a ResNet50 is also applied as the base encoder network for comparison purposes, which is explained in detail in the experimental sections.

3.3. Projection Head

The purpose of using the projection head is to map the representations generated by the base encoder network to a 128-dimensional latent space, which serves as the basis for classification. Usually, the projection head can be a single linear layer network or an MLP with only one hidden layer. In this research, to further reduce the computational complexity and reflect the advantage of the proposed framework, the linear layer network is used as the projection head by default. The option of using an MLP with one hidden layer will be considered in future work.

3.4. Contrastive Loss Function

The contrastive loss function for self-supervised learning used in this paper is the Normalized Temperature-scaled Cross-Entropy Loss (NT-Xent Loss) [18] (Equation (1)):

$$L^{self} = -\sum_{i \in I} \log \frac{\exp(z_i \bullet z_{j(i)} / \tau)}{\sum\limits_{a \in A(i)} \exp(z_i \bullet z_a / \tau)} \tag{1}$$

where $i \in I \equiv \{1, 2, \ldots, 2N\}$ is the index of one arbitrary augmented sample, which is called the anchor. N refers to the size of mini-batch. $z_l = PROJ(ENC(\widetilde{x}_l)) \in \mathbb{R}^{D_p}$ is the 128-dimensional latent space obtained from the augmented sample $\widetilde{x}_l$, which is learned by the base encoder network followed by the projection network. $z_{j(i)}$ refers to the augmented sample generated by the sample with index i. *j(i)* is called the positive, and the remaining $2(N-1)$ indices are called the negatives. The symbol $\bullet$ denotes the inner product. $\tau \in \mathbb{R}^+$ is a temperature parameter. $A(i) \equiv I \backslash \{i\}$ represents the index value of all values of set I in $2N$ except i.

Triplet loss [16] is a special case of the loss defined in SupCon [14], with only one positive and one negative sample. The margin can be expressed as two times of temperature in the loss function defined in the supervised contrastive learning. For N-pair [17], although the loss does not define a temperature parameter, it can be considered as a case of SupCon loss with a certain transformation. As for SimCLR [18], when the all-positives set in the $2N$ samples is restricted to contain only a view of the same source action as that of the anchor, the self-supervised contrastive loss can be expressed in a form of supervised contrastive learning [14]. The above conclusions can be mathematically proved [14].

Therefore, the loss function definition of SupCon can not only be a self-supervised learning loss but also be used as supervised contrastive learning with human-annotated labels. When the supervised contrastive loss function is employed, the augmented sample and samples with the same label in a mini-batch are all simultaneously used as positives. Moreover, the gradient calculation of SupCon has the intrinsic ability to mine hard positives and negatives, which ensures the maximum utilization of hard samples [14].

The supervised contrastive learning loss function adopted in this paper is a unity of the above contrastive losses (Equation (2)) [14]:

$$L^{\text{sup}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \bullet z_p / \tau)}{\sum\limits_{a \in A(i)} \exp(z_i \bullet z_a / \tau)} \tag{2}$$

where $P(i) \equiv \{p \in A(i) : \widetilde{y}_p = \widetilde{y}_i\}$ corresponds to the set of indices of all positives in $A(i)$ except i. $|P(i)|$ is its cardinality.

3.5. Classifier

The classifier adopts a linear classification network with one single, fully connected layer followed by a cross-entropy as the loss function.

3.6. The Superiority of FLOPs

The default base encoder network used in this paper is an MLP with 256 hidden layers. The FLOPs of the MLP network are the summation of the FLOPs of the encoder and the head. The fully connected layer in the encoder contributes most of the portion, while the computation of the head is very low. Meanwhile, using the ResNet50 as the base encoder network is an option in the experimental sections for comparison purposes. Because the ResNet50 network has relatively larger FLOPs, only the FLOPs of the encoder are considered, while the FLOPs of the head can be ignored. To reflect the strength of ConMLP in computational complexity, the FLOPs of typical GCN models are also listed as a reference.

Table 1 shows an evaluation of Shift-GCN [34], ST-GCN [6], 2s AS-GCN [46], 2s AGCN [31], 2s AGC-LSTM [47], and 4s DGNN [48]. Moreover, prefix 2s refers to a two-stream fusion strategy by using "joint stream" and "bone stream", and 4s refers to the addition of "joint motion stream" and "bone motion stream" [34]. The GCN-based models generally require large FLOPs, which are several orders of magnitude larger than ConMLP. Although the computational complexity of ConMLP is relatively low, the performance of ConMLP can be comparable to that of GCN-based state-of-the-art models.

Table 1. Computational complexity comparison between ConMLP- and GCN-based models.

Models	FLOPs (G)
4s DGNN [48]	126.80
2s AGC-LSTM [47]	54.40
2s AGCN [31]	35.80
2s AS-GCN [46]	27.00
ST-GCN [6]	16.20
Shift-GCN [34]	2.50
ConMLP *	1.30
ConMLP	0.05

Note that all the FLOPs are for processing one action sample. * Denotes using ResNet50 as the base encoder network.

The computational complexity of MLP- and ResNet50-based networks are all calculated by an openly available FLOPs computing framework [49]. The FLOPs of the GCN-based model are from the analysis with Shift-GCN [34]. All the FLOPs in this section are based on training one single action sample.

4. Experiments

4.1. Datasets

4.1.1. NTU RGB+D Datasets

NTU RGB+D [50] contains 56,880 action clips in 60 classes, which were simultaneously captured by 3 camera views. The recommendation metric protocols are Cross-View (X-View) and Cross-Subject (X-Sub). NTU RGB+D 120 [51] is an expansion of NTU RGB+D.

It contains 114,480 clips in 120 classes. Similarly, Cross-Setup (X-Set) and Cross-Subject (X-Sub) are recommended as evaluation protocols.

4.1.2. Data Extraction and Pre-Processing

In this research, the raw data were divided into a training set and a test set. According to the official NTU dataset statement, there are some incomplete data in the datasets, which needed to be removed accordingly. The training and test sets used in this paper were as follows: For NTU RGB+D, there were 37,646 and 18,932 clips for X-View, while there were 40,091 and 16,487 clips for X-Sub. For NTU RGB+D 120, there were 54,468 and 59,477 clips for X-Set, while there were 63,026 and 50,919 clips for X-Sub, respectively.

4.2. Default Metrics

MLP was applied as the base encoder network, which was optimized by SGD with a learning rate of 0.001 and weight decay of 0.0005. Additionally, a linear project head was used to transform the representations to a 128-dimensional latent space. The same definition of supervised contrastive learning [14] was chosen as the loss function, with a temperature of 0.07. A mini-batch size of 512 for training was randomly generated for 5000 epochs. The learning rate was decayed with a cosine schedule without restarts [52].

4.3. Comparison with the State of the Art

ConMLP was evaluated on two datasets, NTU RGB+D [50] and NTU RGB+D 120 [51]. Compared with other self-supervised approaches, ConMLP achieved the state of the art on several views of the datasets. It achieved the highest recognition rate on the X-View of NTU RGB+D, with reduced performance on all three other views. The results of the other methods are similar to this. Overall, the performances of these methods on NTU RGB+D 120 are generally not as good as those on NTU RGB+D due to the increased number of classes (Table 2).

Table 2. Top 1 accuracy comparison with self-supervised methods.

Models	**NTU RGB+D**		**NTU RGB+D 120**	
	X-View (%)	**X-Sub (%)**	**X-Set (%)**	**X-Sub (%)**
4s-MG-AL [20]	68.0	64.7	49.5	46.2
CRRL [21]	73.8	67.6	57.0	56.2
Tanfous et al. [22]	76.3	67.0	59.1	61.5
SKT [23]	77.1	72.6	64.3	62.6
GLTA-GCN [24]	81.2	61.2	51.1	49.1
MCAE-MP [27]	82.4	51.9	46.1	42.3
SRCL [19]	82.5	76.7	67.5	67.1
Thoker et al. [26]	85.2	76.3	67.9	67.1
Wang et al. [25]	85.8	80.2	85.5	84.2
VaRe [8]	96.7	92.0	89.4	87.6
ConMLP *	92.2	64.8	62.5	73.9
ConMLP	96.9	75.4	77.2	76.4

* Denotes using ResNet50 as the base encoder network.

On the other hand, in terms of supervised learning methods, the performance of ConMLP is also very outstanding. Similar to the results of self-supervised learning, the accuracies on certain views achieve the state of the art (Table 3).

Table 3. Top 1 accuracy comparison with supervised methods.

Models	NTU RGB+D		NTU RGB+D 120	
	X-View (%)	**X-Sub (%)**	**X-Set (%)**	**X-Sub (%)**
ST-GCN [6]	88.3	81.5	/	/
AS-GCN [46]	94.2	86.8	/	/
AGC-LSTM [47]	95.0	89.2	/	/
2s-AGCN [31]	95.1	88.5	/	/
DGNN [48]	96.1	89.9	/	/
MS-G3D [7]	96.2	91.5	88.4	86.9
Shift-GCN [34]	96.5	90.7	87.6	85.9
MST-GCN [33]	96.6	91.5	88.8	87.5
STF [9]	96.9	92.5	89.9	88.9
ConMLP *	93.0	73.9	75.7	87.5
ConMLP	96.9	75.1	87.2	77.8

* Denotes using ResNet50 as the base encoder network.

To verify whether the MLP-based encoder network was more advantageous under the lower FLOPs demand, a ResNet50-based encoder network with a nonlinear neural network as the project head was also evaluated under the same evaluation metrics. The experimental results demonstrated that the MLP-based encoder network with 256 hidden layers has higher overall performance (Tables 2 and 3).

4.4. The Case without Contrastive Learning

To validate the superiority of the contrastive learning method, a model with no contrastive learning loss function setups but a cross-entropy loss function, was evaluated for comparison purposes. Each individual skeleton sequence was fed directly into the base encoder network without any data augmentations. The representations from the project head were classified by the cross-entropy loss function instead of the contrastive loss function. Overall, the accuracies of the model with contrastive learning loss function were better. Moreover, the results obtained by self-supervised contrastive learning even surpassed the case in which supervised cross-entropy loss was deployed (Table 4).

Table 4. Top 1 accuracy comparison with cross-entropy loss.

	Base Encoder Network	NTU RGB+D		NTU RGB+D 120	
		X-View (%)	**X-Sub (%)**	**X-Set (%)**	**X-Sub (%)**
SL with cross-entropy loss	ResNet50	91.4	64.4	75.4	75.3
	MLP	94.5	63.2	76.2	86.5
SSL with contrastive loss	ResNet50	92.2	64.8	62.5	73.9
	MLP	96.9	75.4	77.2	76.4
SL with contrastive loss	ResNet50	93.0	73.9	75.7	87.5
	MLP	96.9	75.1	87.2	77.8

SL and SSL denote supervised learning and self-supervised learning, respectively.

4.5. Determine the Optimal Hyperparameters

As for the default settings, a learning rate of 0.001, temperature of 0.07, and epochs of 5000 were applied. The above optimal combination of hyperparameters is based on a series of experiments.

The learning rate was checked for values of 0.0001, 0.001, 0.005, 0.01, and 0.1. The bigger the learning rate, the worse the performance. However, when the learning rate was adjusted to 0.0001, the recognition accuracy dropped dramatically. The best accuracy was achieved when the learning rate was 0.001 (Table 5).

Table 5. Top 1 accuracy from various learning rates.

Dataset	Learning Rate				
	0.0001	**0.001**	**0.005**	**0.01**	**0.1**
NTU RGB+D X-View (%)	28.1	96.9	95.3	90.6	88.3

Temperature is considered a critical factor for contrastive learning. The lower the temperature, the greater the contribution to penalizing the hard cases. However, contrastive loss concentrates a few nearest samples when a very low temperature is applied, which can seriously degenerate the performance [53]. Additionally, five values were evaluated: 0.01, 0.04, 0.07, 0.1, and 0.5. The best recognition accuracy was achieved at the temperature of 0.07 (Table 6). This is also consistent with the default setting in SupCon [14]. These results demonstrate that an extremely low temperature can cause an obvious decrease in recognition accuracy.

Table 6. Top 1 accuracy from various temperatures.

Dataset	Temperature				
	0.01	**0.04**	**0.07**	**0.10**	**0.50**
NTU RGB+D X-View (%)	89.8	95.3	96.9	95.3	92.2

The number of epochs explored was high. In the cases in which the epochs of 1000, 2000, 3000, 4000 and 5000 were measured, our results show that the larger the epoch, the better the recognition performance (Table 7). In addition, there was no significant performance improvement after taking the epoch value above 5000.

Table 7. Top 1 accuracy from various epochs.

Dataset	Epoch				
	1000	**2000**	**3000**	**4000**	**5000**
NTU RGB+D X-View (%)	85.2	93.8	96.1	96.1	96.9

Moreover, an ablation study for cosine schedule [52] was adopted, and no significant influence was observed.

4.6. Number of Hidden Layers Is Critical

Another reason contributing to the excellent results of the MLP architecture is the number of hidden layers. An MLP with 256 hidden layers was employed as the base encoder network by default in this paper.

To verify the reason why the MLP architecture outperforms its ResNet50 counterpart, the number of hidden layers of the MLP was gradually decreased from 256 to 128, 64, 32, and 16. However, as the number of hidden layers decreased, the performance decreased dramatically (Table 8). This also confirms that more network layers lead to more useful features being learned.

Table 8. Top 1 accuracy from various numbers of hidden layers.

Dataset	Number of Hidden Layers				
	16	**32**	**64**	**128**	**256**
NTU RGB+D X-View (%)	13.3	17.2	71.1	93.0	96.9

As for deeper network architecture, skip connections [43] can be set [36,37,40]. This will be explored in future work.

4.7. Computational Complexity and Numbers of Parameter

In this section, the computational complexity and the number of parameters of both the MLP-based model and the ResNet50-based model are compared and analyzed. The FLOPs and the number of parameters listed in Table 9 are the values when processing one action sample. Because the model needs to be utilized in training and inference, it should be calculated twice, whereas the classifier is only utilized at inference time.

Although the parameters of the ResNet50 model are not much larger than those of the MLP model, the difference in computational cost is several orders of magnitude. Therefore, models using MLP as the base network have considerable advantages in terms of computational resource consumption. Nevertheless, the recognition accuracy of the MLP-based contrastive learning framework still achieves state-of-the-art performance.

Table 9. Computational complexity and parameters comparison.

		FLOPs (M)	**Parameters (M)**
MLP-based	Model	23.24 × 2	11.62 × 2
	Classifier	0.06	0.03
	Total	46.54	23.27
ResNet50-based	Model	655.32 × 2	23.67 × 2
	Classifier	0.50	0.25
	Total	1311.14	47.59

Note that all the FLOPs are for processing one action sample.

Moreover, compared with the GCN-based architecture, even Shift-GCN [34], which is known for its low computational complexity, in addition to ConMLP, still has advantages. Shift-GCN achieves an accuracy of 95.1% on NTU RGB+D; yet it claimed as many as 2.5G FLOPs. In contrast, ConMLP's FLOPs were only 46.54M, but with an accuracy of 96.9%.

5. Conclusions

The aim of this research is to recognize human actions through skeleton data. The goal is to control the consumption of computing resources and reduce the computational complexity of the model as much as possible while keeping the high recognition accuracy.

ConMLP, which is proposed in this paper, is a simple learning framework based on a naive MLP architecture. In this paper, the performance of the model on NTU RGB+D and NTU RGB+D 120 datasets while using self-supervised contrastive learning is mainly analyzed. Additionally, the loss function is extended to the case of supervised learning. The corresponding model performances are analyzed in comparison with both self-supervised learning and supervised learning approaches.

Moreover, the ResNet50 is used as an alternative option to the base encoder network to make a comparison with the MLP-based network. On the premise of obtaining state-of-the-art recognition accuracies, the advantages of ConMLP, such as low computational complexity and a small number of parameters, are highlighted.

Author Contributions: Conceptualization, C.D. and Z.X.; methodology, C.D.; software, C.D.; validation, C.D. and Y.W.; formal analysis, C.D. and Y.W.; investigation, C.D. and Y.W.; resources, C.D., Y.W. and Z.X.; data curation, C.D.; writing—original draft preparation, C.D.; writing—review and editing, C.D., Z.X. and M.C.; visualization, C.D.; supervision, Z.X. and M.C.; project administration, Y.L. and J.F.; funding acquisition, J.F. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chinese National Natural Science Foundation grant number 62071378.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in NTURGB-D at https://doi.org/10.1109/CVPR.2016.115 and https://doi.org/10.1109/TPAMI.2019.2916873, reference number [50,51].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lemieux, N.; Noumeir, R. A Hierarchical Learning Approach for Human Action Recognition. *Sensors* **2020**, *20*, 4946. [CrossRef] [PubMed]
2. Shotton, J.; Sharp, T.; Fitzgibbon, A.; Blake, A.; Cook, M.; Kipman, A.; Finocchio, M.; Moore, R. Real-Time Human Pose Recognition in Parts from Single Depth Images. *Commun. ACM* **2013**, *56*, 116–124. [CrossRef]
3. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A New Representation of Skeleton Sequences for 3d Action Recognition. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017.
4. Lee, G.C.; Loo, C.K. On the Post Hoc Explainability of Optimized Self-Organizing Reservoir Network for Action Recognition. *Sensors* **2022**, *22*, 1905. [CrossRef] [PubMed]
5. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data. In Proceedings of the 16th IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017.
6. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, New Orleans, LA, USA, 2–7 February 2018.
7. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020.
8. Pan, Q.; Zhao, Z.; Xie, X.; Li, J.; Cao, Y.; Shi, G. View-Normalized and Subject-Independent Skeleton Generation for Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, 1. [CrossRef]
9. Ke, L.; Peng, K.C.; Lyu, S. Towards to-a-T Spatio-Temporal Focus for Skeleton-Based Action Recognition. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 1131–1139. [CrossRef]
10. Zhang, R.; Isola, P.; Efros, A.A. Colorful Image Colorization. In Proceedings of the 14th European Conference on Computer Vision, ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 649–666.
11. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised Representation Learning by Predicting Image Rotations. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
12. Noroozi, M.; Favaro, P. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In Proceedings of the 14th European Conference on Computer Vision, ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 69–84.
13. Kolesnikov, A.; Zhai, X.; Beyer, L. Revisiting Self-Supervised Visual Representation Learning. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 15–20 June 2019.
14. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. In Proceedings of the 34th Conference on Neural Information Processing Systems, NeurIPS 2020, Online, 6–12 December 2020.
15. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014.
16. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015.
17. Sohn, K. Improved Deep Metric Learning with Multi-Class N-Pair Loss Objective. In Proceedings of the 30th Annual Conference on Neural Information Processing Systems, NIPS 2016, Barcelona, Spain, 5–10 December 2016.
18. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual, 12–18 July 2020.
19. Zhang, W.; Hou, Y.; Zhang, H. Unsupervised Skeleton-Based Action Representation Learning Via Relation Consistency Pursuit. *Neural Comput. Appl.* **2022**, *34*, 20327–20339. [CrossRef]
20. Yang, Y.; Liu, G.; Gao, X. Motion Guided Attention Learning for Self-Supervised 3d Human Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8623–8634. [CrossRef]
21. Wang, P.; Wen, J.; Si, C.; Qian, Y.; Wang, L. Contrast-Reconstruction Representation Learning for Self-Supervised Skeleton-Based Action Recognition. *IEEE Trans. Image Process.* **2022**, *31*, 6224–6238. [CrossRef] [PubMed]
22. Tanfous, A.B.; Zerroug, A.; Linsley, D.; Serre, T. How and What to Learn: Taxonomizing Self-Supervised Learning for 3d Action Recognition. In Proceedings of the 22nd IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, 4–8 January 2022.
23. Zhang, H.; Hou, Y.; Zhang, W. Skeletal Twins: Unsupervised Skeleton-Based Action Representation Learning. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, 11–15 July 2022.

24. Qiu, H.; Wu, Y.; Duan, M.; Jin, C. Glta-Gcn: Global-Local Temporal Attention Graph Convolutional Network for Unsupervised Skeleton-Based Action Recognition. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, 11–15 July 2022.
25. Wang, H.; Yu, B.; Xia, K.; Li, J.; Zuo, X. Skeleton Edge Motion Networks for Human Action Recognition. *Neurocomputing* **2021**, *423*, 1–12. [CrossRef]
26. Thoker, F.M.; Doughty, H.; Snoek, C.G.M. Skeleton-Contrastive 3d Action Representation Learning. In Proceedings of the 29th ACM International Conference on Multimedia, MM 2021, Virtual, 20–24 October 2021.
27. Xu, Z.; Shen, X.; Wong, Y.; Kankanhalli, M.S. Unsupervised Motion Representation Learning with Capsule Autoencoders. In Proceedings of the 35th Conference on Neural Information Processing Systems, NeurIPS 2021, Virtual, 6–14 December 2021.
28. Zhu, C.; Li, X.; Li, J.; Dai, S.; Tong, W. Multi-Sourced Knowledge Integration for Robust Self-Supervised Facial Landmark Tracking. *IEEE Trans. Multimed.* **2022**, 1–13. [CrossRef]
29. Dong, X.; Yu, S.I.; Weng, X.; Wei, S.E.; Yang, Y.; Sheikh, Y. Supervision-by-Registration: An Unsupervised Approach to Improve the Precision of Facial Landmark Detectors. In Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–23 June 2018.
30. Krishnan, R.; Rajpurkar, P.; Topol, E.J. Self-Supervised Learning in Medicine and Healthcare. *Nat. Biomed. Eng.* **2022**, *6*, 1346–1352. [CrossRef] [PubMed]
31. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 15–20 June 2019.
32. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016.
33. Chen, Z.; Li, S.; Yang, B.; Li, Q.; Liu, H. Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual, 7–12 February 2021.
34. Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; Lu, H. Skeleton-Based Action Recognition with Shift Graph Convolutional Network. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020.
35. Wu, B.; Wan, A.; Yue, X.; Jin, P.; Zhao, S.; Golmant, N.; Gholaminejad, A.; Gonzalez, J.; Keutzer, K. Shift: A Zero Flop, Zero Parameter Alternative to Spatial Convolutions. In Proceedings of the 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–23 June 2018.
36. Tolstikhin, I.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. Mlp-Mixer: An All-Mlp Architecture for Vision. In Proceedings of the 35th Conference on Neural Information Processing Systems, NeurIPS 2021, Virtual, 6–14 December 2021.
37. Touvron, H.; Bojanowski, P.; Caron, M.; Cord, M.; El-Nouby, A.; Grave, E.; Izacard, G.; Joulin, A.; Synnaeve, G.; Verbeek, J.; et al. Resmlp: Feedforward Networks for Image Classification with Data-Efficient Training. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, 1–9. [CrossRef] [PubMed]
38. Liu, H.; Dai, Z.; So, D.R.; Le, Q.V. Pay Attention to Mlps. In Proceedings of the 35th Conference on Neural Information Processing Systems, NeurIPS 2021, Virtual, 6–14 December 2021.
39. Ding, X.; Zhang, X.; Han, J.; Ding, G. Repmlp: Re-Parameterizing Convolutions into Fully-Connected Layers for Image Recognition. *arXiv* **2021**, arXiv:2105.01883.
40. Chen, S.; Xie, E.; Ge, C.; Chen, R.; Liang, D.; Luo, P. Cyclemlp: A Mlp-Like Architecture for Dense Prediction. *arXiv* **2021**, arXiv:2107.10224.
41. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (Gelus). *arXiv* **2016**, arXiv:1606.08415.
42. Lei Ba, J.; Ryan Kiros, J.; Geoffrey Hinton, E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016.
44. Rao, H.; Xu, S.; Hu, X.; Cheng, J.; Hu, B. Augmented Skeleton Based Contrastive Action Learning with Momentum Lstm for Unsupervised Action Recognition. *Inf. Sci.* **2021**, *569*, 90–109. [CrossRef]
45. Budisteanu, E.A.; Mocanu, I.G. Combining Supervised and Unsupervised Learning Algorithms for Human Activity Recognition. *Sensors* **2021**, *21*, 6309. [CrossRef] [PubMed]
46. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 15–20 June 2019.
47. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An Attention Enhanced Graph Convolutional Lstm Network for Skeleton-Based Action Recognition. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 15–20 June 2019.
48. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-Based Action Recognition with Directed Graph Neural Networks. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 15–20 June 2019.
49. GitHub. GitHub-Sovrasov/Flops-Counter.Pytorch: Flops Counter for Convolutional Networks in Pytorch Framework. Available online: https://github.com/sovrasov/flops-counter.pytorch (accessed on 3 November 2022).

50. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu Rgb+D: A Large Scale Dataset for 3d Human Activity Analysis. In Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016.
51. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. Ntu Rgb+D 120: A Large-Scale Benchmark for 3d Human Activity Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2684–2701. [CrossRef] [PubMed]
52. Loshchilov, I.; Hutter, F. Sgdr: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
53. Wang, F.; Liu, H. Understanding the Behaviour of Contrastive Loss. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021, Nashville, TN, USA, 20–25 June 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.