electronics

MDPI

Article

Interpretable Multi-Channel Capsule Network for Human Motion Recognition

Peizhang Li, Qing Fei *, Zhen Chen and Xiangdong Liu

School of Automation, Beijing Institute of Technology, Beijing 100081, China; lpz@bit.edu.cn (P.L.); chenzhen76@bit.edu.cn (Z.C.); xdliu@bit.edu.cn (X.L.)
* Correspondence: feiqing@bit.edu.cn

Abstract: Recently, capsule networks have emerged as a novel neural network architecture for human motion recognition owing to their enhanced interpretability compared to traditional deep learning networks. However, the characteristic features of human motion are often distributed across distinct spatial dimensions and existing capsule networks struggle to independently extract and combine features across multiple spatial dimensions. In this paper, we propose a new multi-channel capsule network architecture that extracts feature capsules in different spatial dimensions, generates a multi-channel capsule chain with independent routing within each channel, and culminates in the aggregation of information from capsules in different channels to activate categories. The proposed structure endows the network with the capability to independently cluster interpretable features within different channels; aggregates features across channels during classification, thereby enhancing classification accuracy and robustness; and also presents the potential for mining interpretable primitives within individual channels. Experimental comparisons with several existing capsule network structures demonstrate the superior performance of the proposed architecture. Furthermore, in contrast to previous studies that vaguely discussed the interpretability of capsule networks, we include additional visual experiments that illustrate the interpretability of the proposed network structure in practical scenarios.

Keywords: human motion recognition; capsule networks; multi-channel routing; interpretability

check for updates

Citation: Li, P.; Fei, Q.; Chen, Z.; Liu, X. Interpretable Multi-Channel Capsule Network for Human Motion Recognition. *Electronics* **2023**, *12*, 4313. https://doi.org/10.3390/electronics12204313

Academic Editors: Richard Jiang and Angela Lombardi

Received: 22 September 2023
Revised: 7 October 2023
Accepted: 16 October 2023
Published: 18 October 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Human motion recognition in videos aims to identify the actions or behaviors of one or more individuals from video sequences where the occurrences of these actions or behaviors can be characterized as multi-articular coordinated movements. During the execution of a specific motion by the human body, such as waving, boxing, handclapping, running, walking, and jogging, multiple articulations often aggregate, disperse, or undergo periodic variations according to specific patterns. Consequently, actions are typically represented across consecutive video frames rather than in isolation within a single frame [1,2]. Therefore, the effective spatiotemporal representation of multi-articular motion information is a critical aspect of human action recognition.

In recent years, numerous studies (e.g., [3,4]) have been dedicated to representing motion using visual features extracted from video frames and optical flow fields, yielding significant advancements. However, these methods necessitate manual feature selection. Neural network techniques possess the capability to automatically extract and learn effective representations of multi-articular motion in both temporal and spatial dimensions. Empirical evidence demonstrates that features learned from neural network techniques surpass manually selected features [5–9]. Nevertheless, end-to-end neural network architectures are often referred to as black-box architectures due to their lack of transparency in intermediate layers. This opacity diminishes the interpretability of neural networks,

rendering it challenging for researchers to comprehend the feature extraction and classification mechanisms within motion recognition networks. Consequently, this impedes further network optimization and guidance for motion analysis.

Henceforth, there arises a necessity to construct innovative neural network architectures based on interpretable modules to supplant conventional networks. Capsule networks, as introduced in [10–12], have been employed in image classification tasks. Capsules constitute a collection of neurons designed to emulate distinct entities or components of entities. These capsules within the network undergo protocol routing algorithms, enabling the capsule networks to establish relationships from parts to wholes, thereby affording the potential to unveil the enigmatic aspects of deep learning. Benefiting from the interpretability conferred by capsule networks, they are gradually displacing traditional neural networks in numerous domains characterized by elevated demands for both security and interpretability [13–16]. In reference [13], the authors introduced an interpretable deep learning framework utilizing capsule networks for feature selection to identify the genomic profiles of different subcellular types, rendering decision-making processes transparent through the analysis of weight parameters among distinct genomes. In reference [14], the authors applied capsule networks to brain tumor classification tasks, revealing that their utilization not only distinguishes between tumor types but also demonstrates significant correlations between the features employed by the network and manually curated features. In reference [15], the authors proposed a self-explanatory breast diagnosis capsules network (BDCN) embedded with semantic embeddings, achieving local interpretability in diagnostic models. The literature [16] introduced an interpretable capsule network termed iCapsNets, which exhibits both local and global interpretability in the context of text classification.

In addition to the scenarios outlined above that require heightened interpretability, capsule networks are also being actively deployed in the realm of human motion recognition [17–21]. However, these endeavors predominantly employ capsule networks to process spatiotemporal data associated with human motion without affording comprehensive explication of the capsules' interpretability. Simultaneously, these undertakings neglect the consideration of the importance and distinctiveness of features pertaining to different channels (spatial coordinate axes) during the feature extraction process in relation to their role in determining motion categories. Evidently, in certain activities such as waving and boxing, joint trajectory features along one particular directional axis wield significantly more influence in determining the motion type than joint trajectory features along other directional axes. Thus, the dissection and aggregation of features from disparate channels offer a more precise means of classifying motion. In summary, it is hoped that a novel capsule network architecture can be proposed, one that possesses the capability to independently extract features from different channels and amalgamate these features in an interpretable manner to represent motion categories.

To this end, in this paper, we propose a novel interpretable multi-channel capsule network (MCCN) architecture which employs a multi-channel trajectory transformation graph formed from multi-channel spatial coordinates over multiple temporal intervals to classify motion. Within the MCCN framework, feature capsules from distinct channels are stacked to form capsule chains where capsules within each chain independently undergo routing without mutual interference. Ultimately, MCCN computes the activation probabilities of individual top-level capsule chains and employs them to denote the probabilities of different categories. We conducted experiments on the KTH dataset and compared the algorithm's effectiveness with existing methods. Moreover, we employed visualization techniques to illustrate the multi-channel routing process and expounded on the interpretability of MCCN in the context of motion recognition. Specifically, the main contributions of this paper can be summarized in the following three points:

(1) A multi-channel capsule network architecture was proposed. This structure allows for the independent routing of feature capsules within each channel and aggregates

the routing results from all channels, endowing the network with the capability for multi-channel feature extraction;

(2) The interpretability of capsule networks in motion recognition tasks was expounded upon. A visualization method was introduced to depict the activation states of capsules, elucidating the sparse interpretability of capsule networks during motion recognition;

(3) Validation of the proposed architecture on the KTH dataset was conducted. Experimental results demonstrate that the introduced network structure exhibits superior accuracy and robustness when compared to traditional capsule networks and single-channel capsule networks.

The organizational structure of this paper is as follows: Section 2 presents MCCN's core elements including the design of multi-channel capsule chains for complex spatiotemporal trajectories, collaborative dynamic routing, and the formulation of the loss function. Section 3 presents the utilized dataset and experimental network architecture, followed by an analysis of the experimental results and interpretability. Finally, Section 4 draws a conclusion based on the entire paper.

2. Materials and Methods

Figure 1 illustrates the pipeline of our interpretable multi-channel capsule network. Multiple-channel spatial coordinates over a continuous time span are utilized to construct a multi-channel trajectory transformation graph. This graph is subsequently directed into the MCCN pipeline for the purpose of extracting convolutional features. These features, extracted from multiple channels, are amalgamated into the lower-level capsule chains. It is essential to note that capsules within distinct lower-level chains autonomously and collaboratively undergo dynamic routing to the uppermost level, thus forming upper-level capsule chains. The activation of these upper-level capsule chains corresponds to various categories of motion.

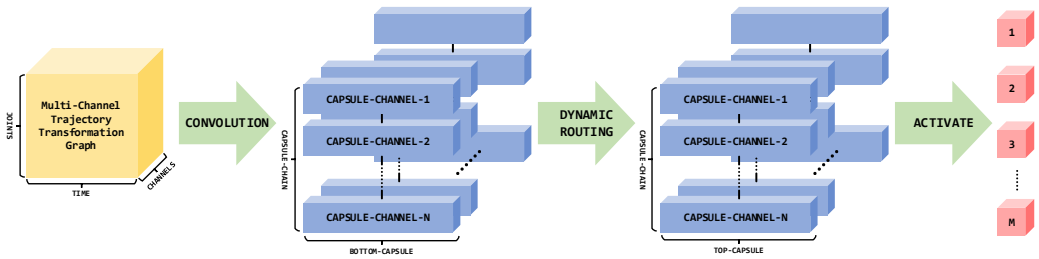


Figure 1. Pipeline of the multi-channel capsule network.

2.1. Multi-Channel Trajectory Characteristic Capsule Chain

The multi-channel trajectory transformation graph is composed of spatial coordinates over continuous time intervals across multiple joints, each represented along different coordinate axes. Each coordinate transformation graph along an axis is represented using a matrix wherein the rows signify the joints and the columns denote the coordinate values of the respective joints over continuous time intervals along that axis. The aggregation of these coordinate transformation graphs, referred to as coordinate graphs, along various axes (distinct channels) culminates in the formation of the multi-channel trajectory transformation graph.

The multi-channel trajectory feature capsule chain is divided into two distinct capsule chains: the lower-level capsule chain and the upper-level capsule chain. Each of these capsule chains comprises capsules linked together, representing instantiation parameters extracted from different channels. In a specific context, multiple-channel convolution is applied to the multi-channel trajectory transformation graph, giving rise to multi-channel local feature detectors. Subsequently, convolution is employed to generate multi-channel lower-level feature capsules. These lower-level feature capsules from different channels are concatenated to form the lower-level capsule chain. It is important to note that capsules

within the lower-level capsule chain do not share coupling coefficients and visual invariant matrices [10]. Following the mapping through visual invariant matrices, the lower-level capsule chain progressively converges toward the upper-level capsule chain via collaborative dynamic routing. The upper-level capsule chain serves the purpose of representing instantiation parameters for various human motion categories.

2.2. Cooperative Dynamic Routing Structure

The capsule chains are denoted as Ξ_i,

$$\Xi_i = stack\left\{s_i^1,\ s_i^2, \ldots, s_i^N\right\}, \tag{1}$$

where i denotes the number of capsule chains within that layer. The operation $stack\ \{\cdot\}$ signifies the row-wise vector stacking process to construct a matrix, with N representing the total number of channels. Furthermore, Ξ_i^k designates the k-th component within a specific capsule. Here, the parameter k ranges from 1 to N.

$\hat{\Xi}_{j|i}$ represents the prediction matrix of the lower-level i capsule chain with respect to the upper-level j capsule chain. It is derived by applying a chain transformation of visual invariant matrices W to the individual capsule components within the lower-level capsule chain,

$$\hat{\Xi}_{j|i} = stack\left\{W_{ij}^1 s_i^1,\ W_{ij}^2 s_i^2, \ldots, W_{ij}^N s_i^N\right\}. \tag{2}$$

The coupling coefficients c_{ij}^k are employed to aggregate all capsules from the k channels within the lower-level capsule chains to the upper-level capsule chains. This aggregation process is executed using a "squashing" function to compress the capsule vectors,

$$s_j^k = squashing\left(\sum_i c_{ij}^k \hat{\Xi}_{j|i}^k\right). \tag{3}$$

In order to enhance the nonlinearity of the compression function, a slight modification was introduced to the "squashing" function proposed in reference [10],

$$squashing\ (x) = \frac{\|x\|^2}{0.5 + \|x\|^2} \frac{x}{\|x\|}. \tag{4}$$

The capsules within the top layer, each belonging to the j channel, are stacked together to form the j capsule chain,

$$\Xi_j = stack\left\{s_j^1,\ s_j^2, \ldots, s_j^N\right\}. \tag{5}$$

The coupling coefficients c_{ij}^k for all capsules within the k channels of the same layer sum up to 1. These coupling coefficients are computed from the initial logits b_{ij}^k using the "softmax" function,

$$c_{ij}^k = \frac{exp\left(b_{ij}^k\right)}{\sum_l exp\left(b_{il}^k\right)}. \tag{6}$$

b_{ij}^k is initialized as a zero vector and it is iteratively updated through the iterations to achieve consistency between the higher-level capsule vectors and the prediction vectors [10],

$$b_{ij}^k := b_{ij}^k + s_j^k \cdot \hat{\Xi}_{j|i}^k. \tag{7}$$

Differing from the conventional capsule network architecture, the top-level instantiation parameters are no longer represented by individual capsules but rather by capsule chains. Therefore, it is no longer appropriate to rely solely on the vector's magnitude to represent probabilities. Instead, an additional cooperative activation probability denoted as a_j is introduced to account for the collaborative influence of top-level capsule chains

across different channels. Furthermore, these cooperative activation probabilities will also serve to represent category probabilities and reconstruction masks. The calculation method for a_j is as follows:

$$a_j = \frac{tr\left(\left(\Xi_j\right)^T\left(\Xi_j\right)\right)}{1 + tr\left(\left(\Xi_j\right)^T\left(\Xi_j\right)\right)}, \tag{8}$$

where $tr(\cdot)$ denotes the trace of a matrix (the sum of its eigenvalues) and T represents the transpose operation.

Details about the cooperative dynamic routing are listed in Algorithm 1.

Algorithm 1 Cooperative dynamic routing algorithm

Input: s_i^k. // Lower-level capsule
Output: a_j. // Cooperative activation probability.
procedure Routing (s_i^k)
  stacking the low-level capsules into capsule chain: $\Xi_i \leftarrow stack\left\{s_i^1, s_i^2, \ldots, s_i^k\right\}$
  for all capsule i in layer m, capsule j in layer $(m+1)$ and channel k:
    $b_{ij}^k \leftarrow 0$
    randomly initialized visual invariant matrices W_{ij}^k
  for r iterations **do**
    for all capsule i in layer m and channel k: $c_{ij}^k \leftarrow softmax\left(b_{ij}^k\right)$
    for all capsule i in layer m and channel k: $\hat{\Xi}_{j|i}^k \leftarrow W_{ij}^k s_i^k$
    for all capsule j in layer $m+1$ and channel k:
      $s_j^k \leftarrow squashing\left(\sum_i c_{ij}^k \hat{\Xi}_{j|i}^k\right)$
    for all capsule i in layer m, capsule j in layer $(m+1)$ and channel k:
      $b_{ij}^k \leftarrow b_{ij}^k + s_j^k \cdot \hat{\Xi}_{j|i}^k$
  for all capsule j in layer $(l+1)$ and channel k: $\Xi_j \leftarrow stack\left\{s_j^1, s_j^2, \ldots, s_j^k\right\}$
  for all top-level capsule chain j: $a_j \leftarrow \frac{tr\left(\left(\Xi_j\right)^T\left(\Xi_j\right)\right)}{1+tr\left(\left(\Xi_j\right)^T\left(\Xi_j\right)\right)}$
  return a_j
end procedure

Utilizing the obtained collaborative activation probabilities a_j, a loss function is devised for optimizing network parameters.

2.3. Loss Function

The cooperative activation probability is introduced into the marginal loss function,

$$L_Margin_j = T_j max\left(0,\ m^+ - a_j\right)^2 + \lambda\left(1 - T_j\right) max\left(0,\ a_j - m^-\right)^2, \tag{9}$$

where L_Margin_j represents the marginal loss for the j top-level capsule (i.e., the j category) and the T_j is considered only when the j capsule is activated (indicating the activation of the j category). The parameters m^+, m^-, and λ serve as hyperparameters for the classification network.

Similarly, the reconstruction loss function $L_Reconstruct_j$ is introduced as a part of the loss function to encourage the encoding of instantiation parameters of multi-channel trajectory transformation graphs by capsules. A multi-channel fully connected network with ReLU activation functions is selected as the reconstruction network. During training, the reconstruction target capsules (only activating the j capsule) are stimulated by applying masks to capsules other than those related to the category. The objective is to minimize the Euclidean distance between the input multi-channel trajectory transformation graph and the output layer. The reconstruction loss is assigned a relatively low weight ε to reduce its

impact on the overall loss function, ensuring it plays a minor role within the complete loss function. The complete loss function is given by

$$Loss = L_Margin_j + \varepsilon \times L_Reconstruct_j. \quad (10)$$

3. Experiment and Results

In this section, we apply MCCN and other methods separately to process a generated multi-channel human joint trajectory dataset. This evaluation aims to assess the performance and interpretability of MCCN in the context of analyzing human motion data in comparison to alternative methods.

3.1. Experimental Setup

3.1.1. Dataset

In our experiments, to obtain multi-channel human motion trajectories, we utilized the human motion dataset created in [22] known as "The KTH Dataset", as depicted in Figure 2. This database contains six types of human motions (walking, jogging, running, boxing, hand waving, and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3, and indoors s4, as illustrated below. All videos were taken over homogeneous backgrounds with a static camera with a 25 fps frame rate.

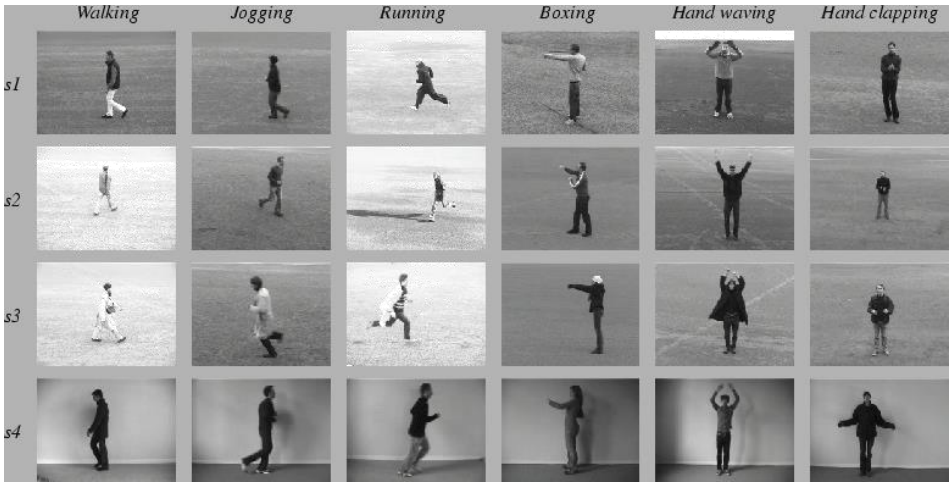


Figure 2. The KTH dataset.

To acquire the multi-joint trajectories during human motion, we followed the approach outlined in [23,24] which is based on the regional multi person pose estimation framework. This framework allowed us to extract spatial coordinates of key joints and subsequently map their horizontal and vertical coordinates to the range of 0–255. We used these mapped coordinates to construct a multi-channel trajectory transformation graph, as described in [25]. The pipeline for constructing the transformation graph is illustrated as shown in Figure 3.

Finally, the acquired multi-channel trajectory transformation graph was categorized into six distinct classes based on the nature of the observed motions. Subsequently, a partitioning was performed to allocate 85% of the data for the training dataset and 15% for the testing dataset, adhering to an 85–15 ratio.

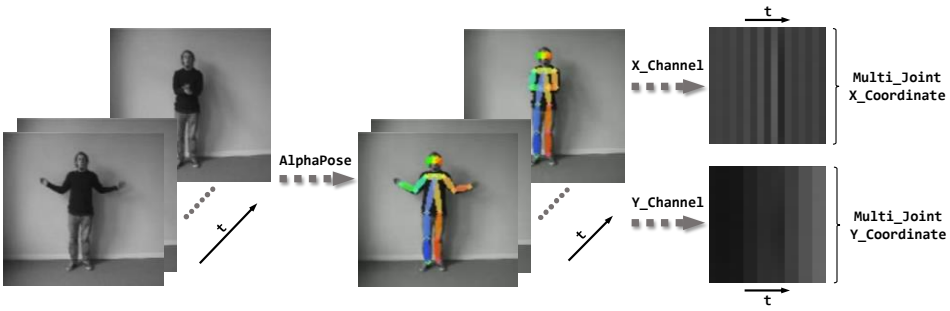


Figure 3. The pipeline for constructing a multi-channel trajectory transformation graph.

3.1.2. Baseline

To validate the effectiveness of MCCN, we established four distinct neural networks as control experiments.

(1) ResNet18: The ResNet18 is a deep convolutional neural network introduced by He et al. in 2015 [5];
(2) Modified capsule network with dual-channel convolution (mCapsNet): mCapsNet is a modified version of the original capsule network (CapsNet). It incorporates dual-channel convolutional layers to enable the network to accept dual-channel image inputs for feature extraction. Importantly, this modification retains the structure of the capsule and routing layers while adapting the reconstruction part by modifying the output neuron count to match the number of pixels in the dual-channel images;
(3) X-Channel capsule network (X-CapsNet): X-CapsNet is designed to perform motion classification using only the motion trajectory coordinates from the X channel as the dataset. It extracts and generates feature capsules in X dimensions and conducts routing based on these capsules;
(4) Y-Channel capsule network (Y-CapsNet): Y-CapsNet is designed for motion classification using only the motion trajectory coordinates from the Y channel as the dataset. It extracts and generates feature capsules in the Y dimension and conducts routing based on these capsules.

As a control experiment, the number of lower-level feature capsules used for routing in mCapsNet, X-CapsNet, and Y-CapsNet is set to be the same as MCCN, with each network having 64 lower-level feature capsules.

3.1.3. Architectures

The MCCN architecture comprises several fundamental components meticulously designed to harness the dual-channel features effectively. This network structure is illustrated in Figure 4 and can be succinctly described as follows:

(1) Initial Dual-Channel Feature Extraction Layers;
    - MCCN initiates with convolutional layers equipped with batch normalization (BN) and rectified linear unit (ReLU) non-linear activation functions;
    - Each convolutional layer employs a 3×3 kernel, operates with a stride of 1, ingests input from the dual channels, and produces an output with 128 channels.
(2) Dual-Channel Lower-Level Capsule Layer;
    - Subsequent to the convolutional layers, the architecture incorporates dual-channel lower-level capsule layers;
    - These lower-level capsules are derived through convolutional operations using a 9×9 kernel, a stride of 2, 128 input channels, and 8 output channels;
    - Each capsule within this layer exhibits a length of 16, resulting in a total of $4 \times 4 \times 4$ capsules per channel;
    - Capsules at corresponding positions within the two channels are aggregated to form capsule chains, each with a length of 2.

(3) Dual-Channel Top-Level Capsule Layer;

- The top-level capsules are established through routing procedures from the lower-level capsules and undergo four routing iterations;
- Each channel consists of 6 capsules, each having a length of 32;
- Capsules at corresponding positions within the dual channels are integrated to form capsule chains, each possessing a length of 2.

(4) Reconstruction Layers.

- The reconstruction segment comprises three fully connected layers designed to accommodate the dual channels;
- Activation probabilities are employed to preserve the capsule chain with the highest probability, thereby masking other capsule chains;
- The three fully connected layers are equipped with neuron counts of 512, 1024, and 289, respectively.

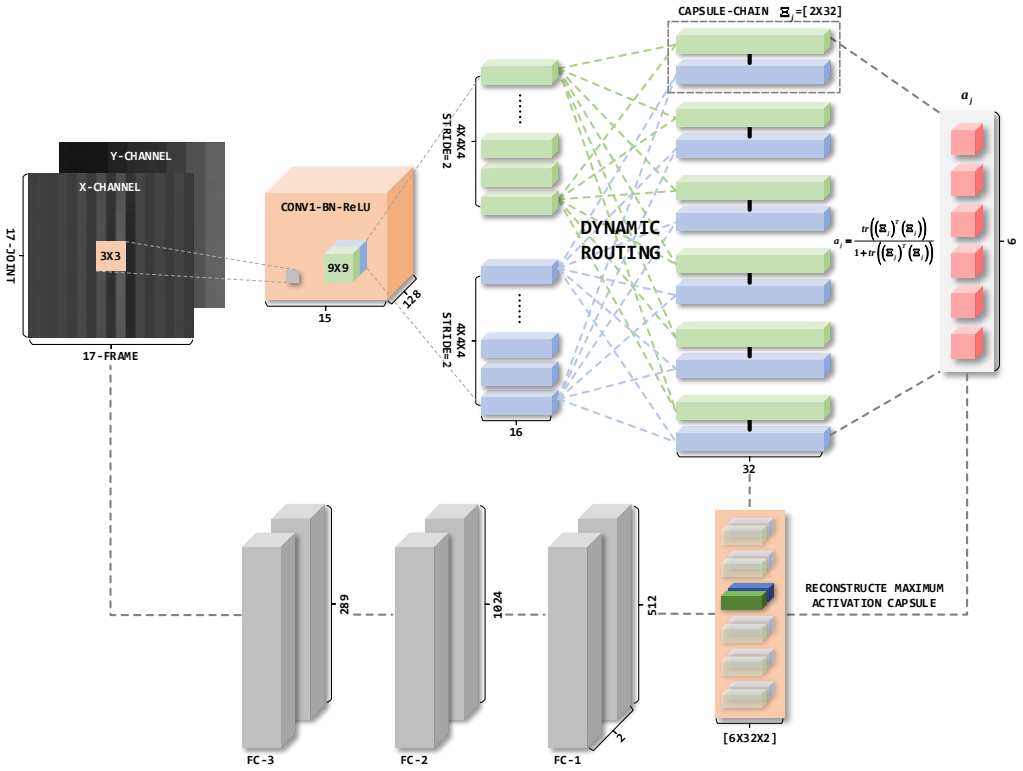


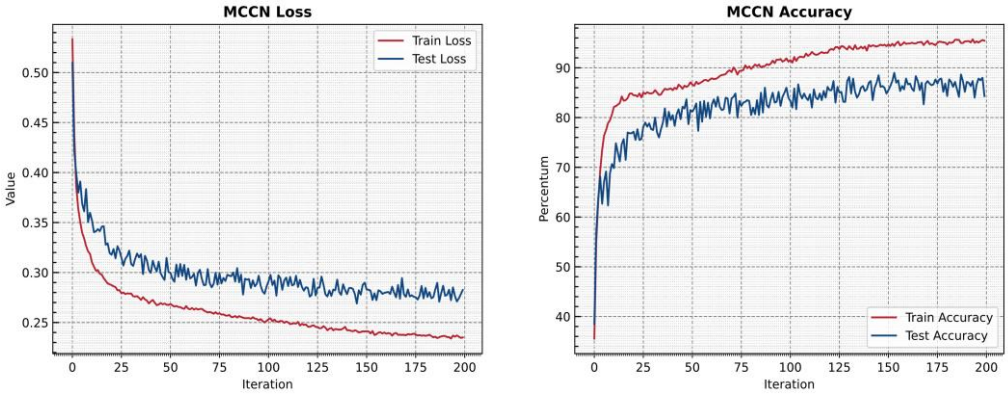
Figure 4. The network structure of MCCN.

The intricate details of the network structure and its parameters are meticulously illustrated in Figure 4.

3.2. Results Analysis

The results for MCCN on both the training and testing datasets, including loss values and accuracy, are presented in Figure 5. MCCN has demonstrated excellent performance on both the training and testing datasets. Additionally, to assess the reconstruction quality, visualizations of real trajectory graphs (GroundTruth) and trajectory graphs generated by MCCN's reconstruction (reconstruction) within a single batch are displayed in Figure 6.

Upon examination of Figure 6, it is evident that utilizing the capsule chain with the highest activation probability allows for the effective reconstruction of trajectory graphs that closely resemble the ground truth data, albeit with some minor loss of fine features. This observation underscores the proficiency of the top-level capsules in encoding essential human motion trajectory features.



(**a**) (**b**)

Figure 5. The performance of MCCN: (**a**) the loss value of MCCN on the training and testing datasets and (**b**) the accuracy of MCCN on the training and testing datasets.

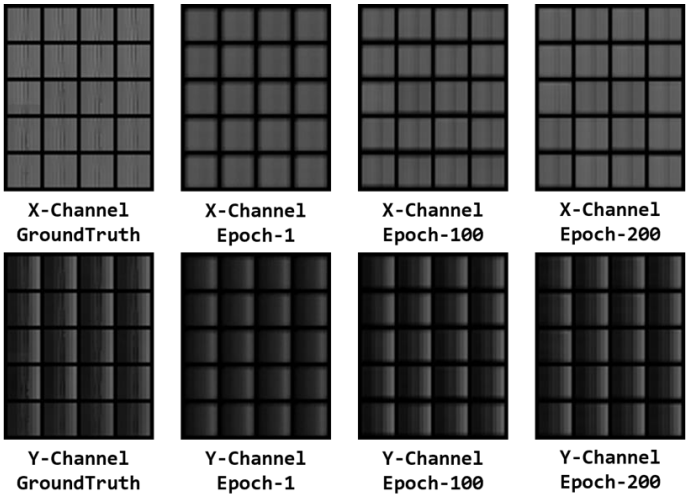
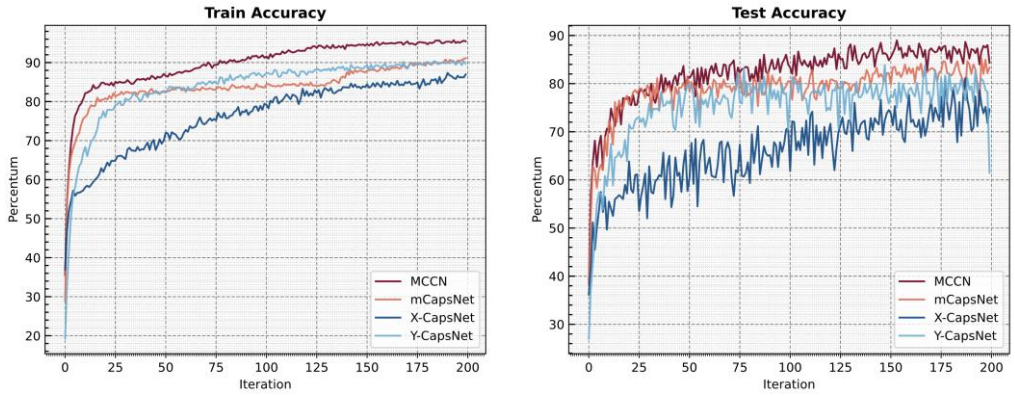


Figure 6. GroundTruth trajectory graph versus MCCN reconstructed trajectory graph.

Experimental trials were conducted on the dataset employing MCCN, CNN (ResNet18), mCapsNet, X-CapsNet, and Y-CapsNet and the findings are presented in Figure 7 and Table 1, as depicted.



(**a**) (**b**)

Figure 7. The performance of MCCN with other methods: (**a**) the accuracy of MCCN and other methods on the training datasets and (**b**) the accuracy of MCCN and other methods on the testing datasets.

Table 1. The performance comparison between MCCN and other methods.

Methods	#Param. (M)	Test Accuracy (%)	Visual Interpretability
ResNet18 (baseline)	11.1785	86.91	No
MCCN (Ours)	3.0380	**87.60**	**Yes**
mCapsNet	1.7464	82.26	No
X-CapsNet	1.4496	74.54	No
Y-CapsNet	1.4496	79.39	No

The results indicate that the performance of MCCN on the test dataset is slightly superior to that of ResNet18, demonstrating a proficient capability in accomplishing classification tasks while exhibiting fewer model parameters and a certain degree of interpretability. The visual interpretability of MCCN will be elaborated upon in the following section. The mCapsNet, through the incorporation of dual-channel features in its convolutional layers, exhibits superior classification performance on both the training and test datasets when compared to X-CapsNet and Y-CapsNet. A scrutiny of the performance of X-CapsNet and Y-CapsNet on the training and test datasets reveals that, concerning motion classification, the features from the Y-channel prove more effective than those from the X-channel (on the training dataset, Y-CapsNet even approaches the performance of mCapsNet). However, the substantial fluctuations observed in the test dataset for X-CapsNet and Y-CapsNet to some extent suggest the presence of overfitting in the case of single-channel features. In summary, focusing solely on the analysis of human motion trajectories through single-channel processing leads to the loss of numerous crucial features, resulting in inadequate classification performance and oscillations in performance on the test dataset. The exclusive utilization of convolutional layers to fuse dual-channel features, to some extent, enhances the performance on the test dataset (reducing oscillations); nevertheless, accuracy levels still fall short of meeting classification requirements.

3.3. Visual Interpretability

The neural activity within active capsules represents various attributes of specific entities present in the images, including a wide range of instantiated parameters [10]. The coupling coefficients utilized by lower-level capsules when routed to the top-level capsules reflect the importance of feature entities represented by different lower-level capsules to the category entities represented by the top-level capsules during the process of multi-channel collaborative routing. To better visualize the connectivity between lower-level capsules and top-level capsules during multi-channel routing, coupling coefficient matrix heat maps (CCMHMs) are constructed using the coupling coefficients between lower-level capsules and top-level capsules. For the sake of clarity in presentation, the CCMHMs aggregate the cooperative coefficients of multiple-channel capsules through arithmetic averaging, simplifying the visualization of routing relationships. (Similarly, it is also possible to independently construct CCMHMs for each channel to visualize the capsule activation states of different channels). The horizontal axis of the heat map represents the lower-level capsules (64 in total) while the vertical axis represents the top-level capsules (representing six motion categories). Each cell in the heat map represents the routing coupling coefficient from a lower-level capsule, after channel combination, to a top-level capsule, with lighter colors indicating larger coefficient values. It is worth noting that, for better visualization, the vertical scales of the heat maps are not entirely uniform. Figure 8a–f sequentially depicts the CCMHMs for the six types of motion, with each subfigure showing three CCMHMs corresponding to Epoch 5, 50, and 100.

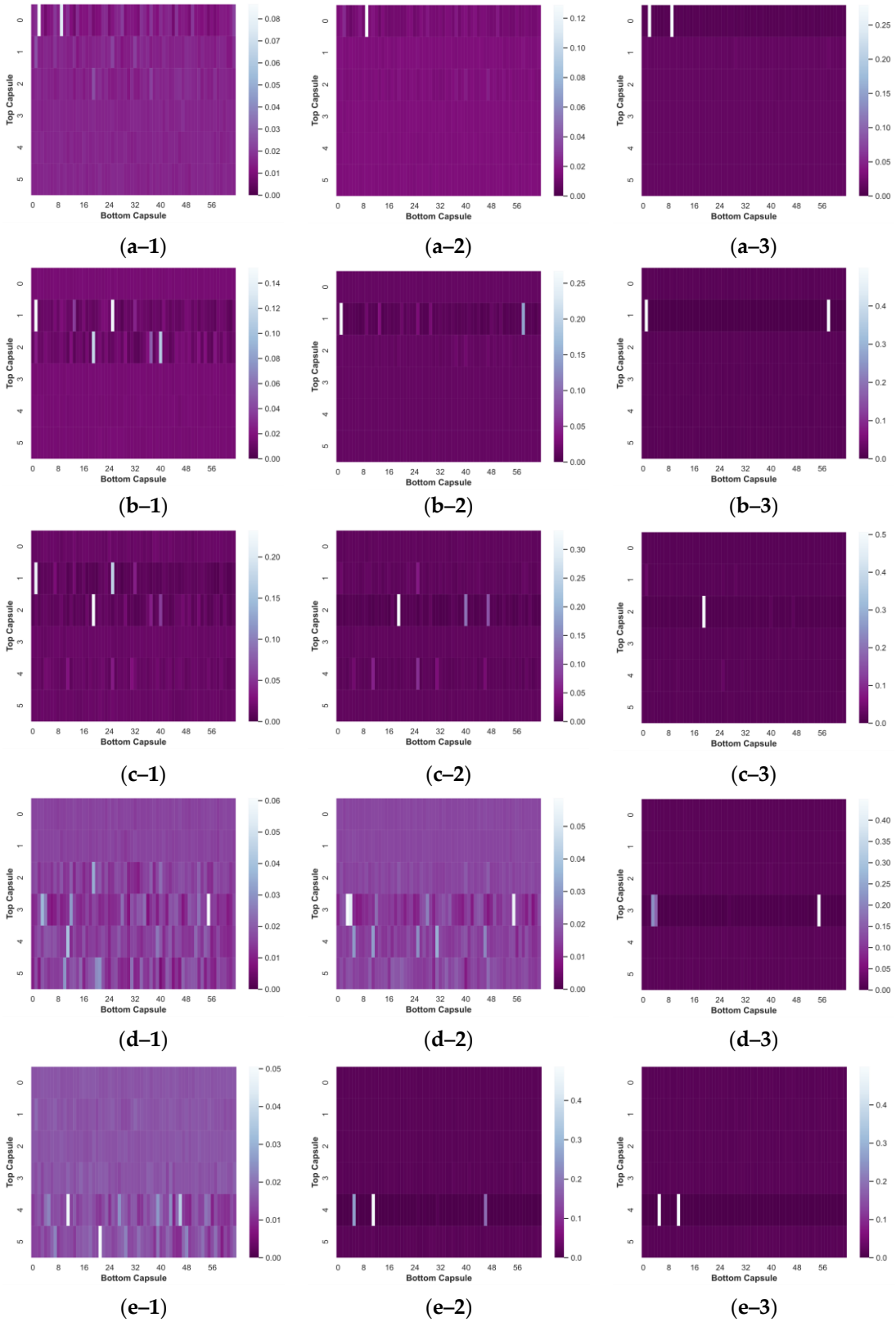
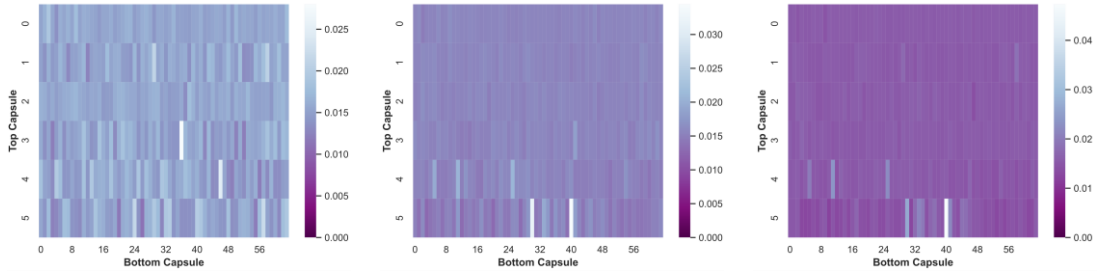


Figure 8. *Cont.*



(**f–1**) (**f–2**) (**f–3**)

Figure 8. The coupling coefficient matrix heat maps of MCCNs is a figure. The subfigures (**a**–**f**) represent the CCMHMs for six categories of motion data input: waving, boxing, handclapping, running, walking, and jogging. The numbers 1–3 following each subfigure denote the iteration epochs, specifically, the 5th, 50th, and 100th epochs, respectively.

From Figure 8, it is evident that during the early iterations, the activation of top-level capsules appears to be relatively chaotic. However, as the iterations progress towards Epoch 200, in each category of motion's CCMHMs, the top-level capsules that are activated correspond to the capsules representing that particular category (for instance, in the case of the first motion category as shown in Figure 8a–3, the larger activated capsules are consistently located at Top Capsule-0). This outcome effectively demonstrates the classification robustness of MCCN, highlighting its ability to robustly categorize different types of motion.

The interpretability of MCCN is demonstrated by the sparsity exhibited when lower-level capsules cluster into top-level capsules, resulting in each category of motion being "sparse", meaning that when a lower-level capsule significantly contributes as a feature (with a substantial coupling coefficient) to represent one category of top-level capsule, it will not serve as a prominent feature for representing other categories of top-level capsules. In other words, a category of top-level capsules can be collectively represented by multiple lower-level capsules but a single lower-level capsule will only serve as a prominent feature for one top-level capsule, akin to a "capsule tree". This sparsity ensures that each category of motion possesses its own set of "distinctive feature capsules" where these "distinctive feature capsules" correspond one-to-one with motion categories, thereby imparting interpretability to the entire classification network.

While the semantics represented by capsules may not necessarily align with human-understandable semantics, the sparsity of these semantic capsules allows individuals to label the "distinctive feature capsules" for each category based on prior knowledge. On the other hand, neural networks possess superior feature extraction capabilities compared to humans. Relying solely on human-understandable semantics would compromise the network's ability to fit the data. This is precisely why this paper emphasizes the importance of neural networks having interpretability rather than human-comprehensibility.

4. Conclusions and Future Works

In this paper, we introduced a method for composing capsule chains in a multi-channel manner and routing and activating them in capsule-based models which enhances the integration of salient features from various channels. Systematic experiments were conducted to compare our multi-channel capsule network with an existing capsule network and a single-channel capsule network. The experiments have validated the competitive performance of our method in terms of accuracy and generalization. Additionally, we visualized the routing process of the multi-channel capsule network and demonstrated its interpretability in the context of human motion recognition through sparsity analysis.

As a prospect for future work, exploring the matching relationships between feature capsule chains and joint decompositions is an intriguing avenue as it has been shown that each type of motion is composed of "dedicated" capsule chains. Such matching

relationships could potentially enhance our understanding of neural network decisions and provide insights into guiding motion analysis.

Author Contributions: Conceptualization, P.L., X.L. and Z.C.; methodology, P.L. and Q.F.; software, P.L.; validation, Q.F.; writing—original draft preparation, P.L.; writing—review and editing, X.L., Q.F. and Z.C.; visualization, P.L. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Technology Research and Demonstration of National Scientific Training Base Construction of China (No.2018YFF0300800).

Data Availability Statement: The KTH dataset used in this paper is available for open access at https://www.csc.kth.se/cvap/actions/ (accessed on 1 May 2023). The source code and related support data can be found at https://github.com/teabun0805/MCCN.git (accessed on 7 October 2023). For additional support or inquiries, please feel free to contact the corresponding author.

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

References

1. Shih, H.-C. A survey of content-aware video analysis for sports. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1212–1231. [CrossRef]
2. Shi, Y.; Tian, Y.; Wang, Y.; Huang, T. Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Trans. Multimed.* **2017**, *19*, 1510–1520. [CrossRef]
3. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
4. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 3–6 December 2013; pp. 3551–3558.
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
6. Wu, Z.; Wang, X.; Jiang, Y.G.; Ye, H.; Xue, X. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 461–470.
7. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
8. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]
9. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
10. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3859–3869.
11. Patrick, M.K.; Adekoya, A.F.; Mighty, A.A.; Edward, B.Y. Capsule networks—A survey. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1295–1310.
12. Haq, M.U.; Sethi, M.A.J.; Rehman, A.U. Capsule Network with Its Limitation, Modification, and Applications—A Survey. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 891–921. [CrossRef]
13. Wang, L.; Nie, R.; Yu, Z.; Xin, R.; Zheng, C.; Zhang, Z.; Cai, J. An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell RNA-sequencing data. *Nat. Mach. Intell.* **2020**, *2*, 693–703. [CrossRef]
14. Afshar, P.; Plataniotis, K.N.; Mohammadi, A. Capsule networks' interpretability for brain tumor classification via radiomics analyses. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3816–3820.
15. Chen, D.; Zhong, K.; He, J. BDCN: Semantic Embedding Self-Explanatory Breast Diagnostic Capsules Network. In Proceedings of the China National Conference on Chinese Computational Linguistics, Hohhot, China, 13–15 August 2021; pp. 419–433.
16. Wang, Z. iCapsNets: Towards interpretable capsule networks for text classification. *arXiv* **2020**, arXiv:2006.00075.
17. Duarte, K.; Rawat, Y.; Shah, M. VideoCapsuleNet: A simplified network for action detection. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 2–8 December 2018; pp. 7610–7619.
18. Zheng, X.; Liang, X.; Wu, B.; Wang, J.; Guo, Y.; Zhang, X.; Ma, Y. A Multi-scale Interaction Motion Network for Action Recognition Based on Capsule Network. In Proceedings of the 2023 SIAM International Conference on Data Mining, Minneapolis, MN, USA, 27–29 April 2023; pp. 505–513.

19. Voillemin, T.; Wannous, H.; Vandeborre, J.P. 2d deep video capsule network with temporal shift for action recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 3513–3519.
20. Ha, M.H.; Chen, O.T.C. Deep neural networks using capsule networks and skeleton-based attentions for action recognition. *IEEE Access* **2021**, *9*, 6164–6178. [CrossRef]
21. Yu, Y.; Tian, N.; Chen, X.; Li, Y. Skeleton capsule net: An efficient network for action recognition. In Proceedings of the 2018 International Conference on Virtual Reality and Visualization (ICVRV), Qingdao, China, 22–24 October 2018; pp. 74–77.
22. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 23–26 August 2004; pp. 32–36.
23. Fang, H.S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Lu, C. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 7157–7173. [CrossRef] [PubMed]
24. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2334–2343.
25. Li, P.; Fei, Q.; Chen, Z.; Yao, X.; Zhang, Y. Characteristic Behavior of Human Multi-Joint Spatial Trajectory in Slalom Skiing. *J. Adv. Comput. Intell. Intell. Inform.* **2022**, *26*, 801–807. [CrossRef]

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.