Privacy Mitigating Algorithms for Medical Action Recognition

by

Leiyu Xie

A doctoral thesis submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy (PhD), from Newcastle University.

June 2023

Intelligent Sensing and Communication (ISC) Research Group,
School of Engineering,
Newcastle University,
Newcastle Upon Tyne, UK, NE1 7RU.

CERTIFICATE OF ORIGINALITY

I at this moment certify that the work which is being presented in the thesis entitled '**Privacy Mitigating Algorithms for Medical Action Recognition**', in fulfilment of the requirements for the award of **the degree of Doctor of Philosophy (PhD)** is an authentic record of my work carried out during a period from October 2019 to August 2023 under the supervision of Dr Syed Mohsen Naqvi, School of Engineering, Newcastle University. The matter presented in this thesis has not been submitted for the award of any other degree elsewhere.

..... (Signed)

..... (candidate)

I dedicate this thesis to my loving parents.

Abstract

Video-based human action recognition has become a crucial component in recent years for many applications, such as human machine interaction, video surveillance and healthcare-related systems. The primary task for human action recognition is to analyse the human behavior based on the given action data. However, different from the general action recognition, medical action recognition is more challenging due to the data limitation, privacy protection and noisy annotations issues. In this thesis, in order to improve the medical action recognition performance and the robustness of system by addressing the aforementioned issue, a variety of enhanced approaches are proposed.

The first contribution aims to focus on human multiple fall events classification using a deep neural network framework by reducing the redundant information and presenting a two-stage framework. The proposed redundant reducing theory is developed to remove the unimportant part, including the redundant empty frames from the video and the redundant body parts from the processed privacy-mitigated human skeleton data. In addition, the proposed two-stage framework is designed for addressing the imbalanced data issue from the data limitation. To improve the classification performance, the gating parameter is utilized along with the proposed structure.

The second contribution relates to address the noisy annotation issue for multiple fall events classification, since the quality of the annotations plays a key role in the data-driven methods. The proposed noisy annotation managing system includes two parts: cascaded noisy annotation purification and noisy annotation learning framework, which is called JoCoT. The purification theory is based on the principle of the joint distribution probability density function to identify and prune the incorrect annotations. JoCoT is proposed for

fully exploiting the potential of the noisy instances with a trinity network. The small loss theory is utilized for selecting the clean instances. Moreover, both the co-regularization and contrastive learning with joint loss function are applied for enhancing the performance.

The third contribution focuses on extracting the novel direction-level features by the proposed signal image generation (SIG) to further protect the privacy information, which could assist the position-level feature to improve the performance by investigating their complementary benefits in different stages for medical action recognition. A one-shot learning framework is developed to address the medical data limitation issue, and together with the cross-attention mechanism (CsA) is used to reduce the misclassification bias for the similar medical action issue. Moreover, dynamic time warping (DTW) module is proposed to minimize the temporal mismatching issue between the instances from the same category, thereby improving the performance.

The proposed contributions are evaluated on the UP-Fall, NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD benchmark datasets, which are widely used for medical action recognition. Detailed evaluations on the benchmarks, along with the comparisons with the recent state-of-the-art methods, confirm the effectiveness of the proposed approaches on medical action recognition.

Contents

Statement of Originality

The contributions of this thesis are mainly to focus on improving the performance and robustness of medical action recognition. The following international journal and conference papers verify the novelty of the contributions.

In Chapter 3, a redundant information reduction strategy is proposed to remove the inessential information in the data processing stage to achieve better multiple fall events classification performance. Then a two-stage framework based on the deep neural network is designed to address imbalanced data issue and further improve the performance. Additionally, a gating method is exploited empirically for controlling the filtering ability in the initial stage for better discriminating ability. These research outputs have been presented in:

- **L. Xie**, Y. Yang, Z. Fu, and S. M. Naqvi, 'Skeleton-based Fall Events Classification with Data Fusion', in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), 2021.*

- **L. Xie**, Y. Sun, J. A. Chambers, and S. M. Naqvi, 'Two-stage Fall Events Classification with Human Skeleton Data', in *arXiv preprint arXiv:2208.12027, 2022.*

In Chapter 4, a novel cascaded learning framework is proposed to purify the corrupted dataset for addressing the noisy annotation issue. The noisy annotations are pruned by four different methods, which are based on the principle of joint distribution probability density function. Moreover, learning with the noisy label algorithm is proposed to fully exploit the potential of noisy instances and enhance the robustness of the proposed framework. The small-loss theory is applied for the clean instances selection, along with the Kullback-Leibler divergence is utilized for avoiding the networks converging and extending the effective training process. The outputs of these solutions are presented in:

- **L. Xie**, Y. Sun, J. A. Chambers, and S. M. Naqvi, 'Privacy Preserving Multiclass Fall Classification based on Cascaded Learning and Noisy Labels Handling', in *IEEE International Conference on Information Fusion (FUSION), 2022.*

- **L. Xie**, Y. Sun, and S. M. Naqvi, 'Learning with Noisy Labels for Human Fall Events Classification: Joint Cooperative Training with Trinity Networks', in *ACM Transactions on Computing for Healthcare (under review).*

In Chapter 5, a multiple-level fusion within a novel one-shot learning framework is proposed to address the data limitation for medical action recognition. Both the direction-level and position-level features are extracted and transformed by the proposed signal-level image transformation (SIG) method for further mitigating the privacy information. A cross-attention mechanism is developed to address the similar action issue, together with the dynamic time warping (DTW) module for aligning the temporal information between the instances. Moreover, feature-level and decision-level fusion approaches are proposed to further enhance performance by exploiting the complementary benefits among different features. The contributions of this chapter are presented in:

- **L. Xie**, Y. Yang, Z. Fu, and S. M. Naqvi, 'One-shot Medical Action Recognition with a Cross-Attention Mechanism and Dynamic Time Warping', in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.*

- **L. Xie**, Y. Yang, Z. Fu, and S. M. Naqvi, 'MF-OSMAR: Multiple-level Fusion of One-Shot Learning for Privacy Preserved Skeletal Human Medical Action Recognition', in *IEEE Transactions on Multimedia (under review)*

Acknowledgements

First of all, I would like to show my strongest gratitude to my supervisor, Dr Mohsen Naqvi for his responsible and continuous support during my PhD study, for his generous advice, kind enlightening and thoughtful concern throughout the past four years, especially during the unprecedented pandemic. I have learned and benefited intensely from his professional guidance and meticulous attitude for all the related research and the writing up of this thesis. This thesis would never be accomplished without his constant and patient encouragement. I feel a lifelong sense of pride to be one of his research students. I do wish I could have the opportunity to work with him again in the future.

I am also tremendously thankful to Professor Jonathon Chambers, Dr Zeyu Fu and Dr Yang Sun, who assisted me to enter my research field and kept giving me suggestions during my PhD research study period. I have gained valuable knowledge from them.

I would also like to express my appreciation to my friends and colleagues: Federico Angelini, Yang Xian, Jiawei Yan, Yuxing Yang, Yi Li, Shuanglin Li and Yichun Li, for your professional assistance and the understanding you provided in daily life during the past four years in the United Kingdom.

Finally but most importantly, I would like to express my deepest love to my family, especially to my beloved parents. I am extremely thank you for raising me up, as well as providing me with a growing-up environment filled with love, respect, support and understanding without any reservations.

Leiyu Xie

August, 2023

List of Acronyms

DNN	Deep Neural Network
CNN	Convolutional Neural Network
GNN	Graph Neural Network
RNN	Recurrent Neural Network
KL	Kullback-Leibler Divergence
RF	Random Forest
SVM	Support Vector Machine
MLP	Multiple Layer Perceptron
AdaBoost	Adaptive Boosting
SIG	Signal-level Image Generation
CsA	Cross Attention Mechanism
DTW	Dynamic Time Warping
FP	False Positive
FN	False Negative

List of Symbols

$\neq$	Not equal to
$\geq$	Greater than or equal to
$\sum$	Summation
$\in$	Belongs to
$\forall$	For all elements
ln	Logarithmic function with Euler number
log	Logarithmic function
exp	Exponential function
$\times$	Multiplication
∇	Gradient
$\subseteq$	Subset
$\sqrt{\cdot}$	Square root
arccos	Inverse cosine function
$\cap$	Intersection

$(\cdot)^{\mathbf{T}}$	Transpose operator
p	Probability distribution
f	State transition function
max	Maximum value
min	Minimum value
argmin	Argument of the minimum
argmax	Argument of the maximum
$\|\cdot\|$	Paradigm function
$\|\cdot\|_F$	Frobenius normalization
$\|\cdot\|$	Cardinality of a set

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Motivation

Due to the decreasing of birth rate and the increasing in the average lifespan in the past 80 years, the aging population has been a worldwide problem [12]. According to the report from the World Health Organization (WHO), the aging population (aged 65 or over) has reached 20% of the population worldwide and will reach 28% by 2050, which is approximately 1.5 billion [13, 14]. Generally, deterioration with age leads to cognitive, physical and sensory functionalities reduction [15,16]. At least 35% of aging people may suffer one time or more medical actions per year and these medical actions may have various consequences for the aging population, the healthcare systems and society [14]. With the increasing of aging population, human medical action recognition has played an important role in recent years for many applications. The major task of human medical action recognition is to accurately analyze and classify medical actions based on the given action sequences [17]. Since the video camera has become ubiquitous with the development of industrial technologies, video data analytics plays a crucial role in the intelligent medical action recognition area [18, 19]. However, privacy protection is one of the most challenging issues in the medical area, otherwise will lead to various invasions to the people, especially to the aging population which lacks distinguishing abilities, such as identity theft, legal liability, financial fraud and doctor-patient trust issues. The ultimate objective of it is to automatically discover,

analyze and understand medical actions according to the visual and privacy-protected information from the video cameras.

Human medical action recognition which represents an essential and fundamental step has enormously advanced in many applications, such as nursing homes, hospitals and assisted living homes. Example applications of medical action recognition are illustrated in Figure 1.1. The major interest of medical action recognition is to facilitate the autonomous intelligent system to obtain better comprehension and analyzing ability for human-involved events which are related to medical actions, such as vomiting, falling down, headache, and staggering. Many researchers have been seeking various reliable action recognition approaches which enable the intelligent system to acquire better human action features from the sequences, therefore facilitating comprehensive human action interpretation [20, 21]. However, there still exist many challenging issues which need to be further addressed. These problems are particularly caused by privacy protection and dataset limitation, such as imbalanced data, limited data, noisy annotations, redundant information, temporal mismatching and similar action. To address these challenging issues, there are various approaches that have been proposed for this task area. This thesis focuses on utilizing the video-based data for addressing the aforementioned issues of medical action recognition tasks by only replying on the privacy-mitigated approaches.

Amongst the medical actions, with the rapid development of the aging population, falling down has surpassed cardiovascular diseases and cancer becoming the primary reason for death and health effects in the aging population during these years [22, 23]. With the significant advancements in perception sensors, recent human fall detection approaches have benefited remarkably in either wearable or non-wearable sensors. The wearable-based approaches using wearable sensors, such as accelerometers and gyroscopes to detect fall events, but present challenges in terms of high hardware costs, battery exhaustion, and potential data loss in tracking due to the resistance

or forgetfulness of the aging population to consistently wear these devices. The non-wearable sensor methods use such as the environment sensors or the WiFi signal to detect fall events, which are sensitive to the environment and more difficult to distinguish the falls under the multiple subject scenarios. The majority of previous research using non-wearable sensors primarily focused on the binary fall detection task, which involved determining if the fall events had taken place or not. However, different falling events can cause varying levels of physical harm to aging population. For instance, the extent of injuries incurred from falling onto a chair is extremely different from those caused by falls on stairs, which are particularly leading to head injuries for aging population. Therefore, it is imperative to classify specific fall events to account for the unique nature and severity of injuries associated with different falling events.

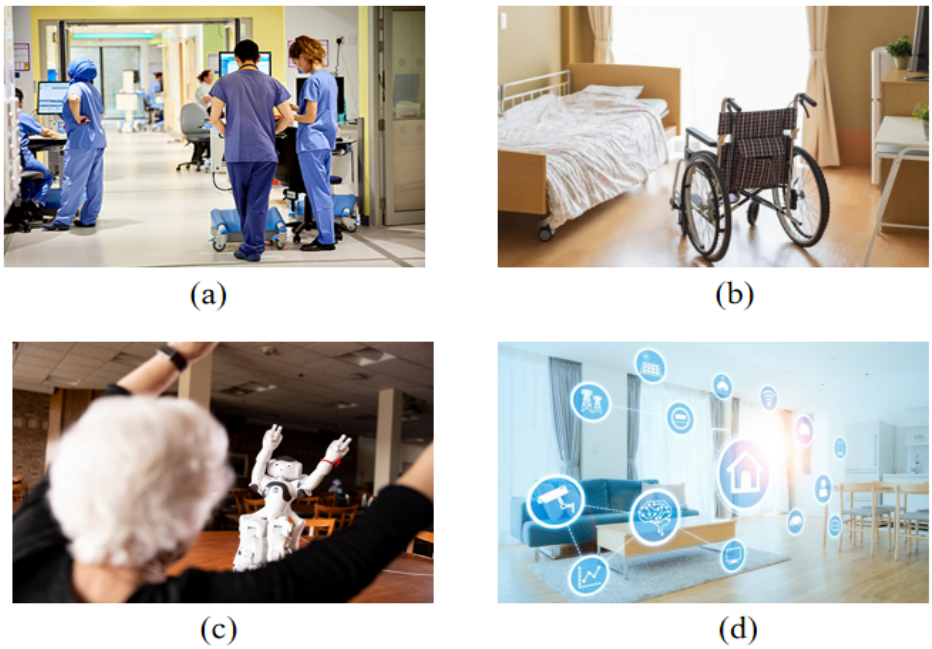


Figure 1.1: Example application scenes of medical action recognition: (a) Hospital. (b) Nursing home. (c) Assisted living for aging population. (d) Smart home with AI.

The rapidly developed deep learning approach is essentially the data-driven approach [24, 25]. The performance of the model not only depends on the quality of

the data but also on the quality of the data annotations. Due to the aforementioned issue, there exist noisy label issues for medical action recognition because the occurrence of most medical actions is much less than normal human actions. Moreover, the anonymizing or de-identifying operation for protecting medical privacy information may lead to the loss of crucial information for medical actions, which may result in noisy labels. The occurrence of noisy label issues primarily stems from biases in determining the same categories of actions, negligence in judging similar data, and the potential for careless errors during data annotation. Improving label quality can be achieved through labeling by experts, iterative labeling, and the use of standardized annotation methods. Noisy label issue commonly arises the datasets that are manually or semi-automatically labelled. Since the noisy labels have a significant negative impact on the performance, the model may not be able to demonstrate promising performance. It is necessary to provide suitable learning with noisy label and dataset purification algorithms.

Owing to the low occurrence and privacy protection reasons for medical action data, in addition to the aforementioned noisy label issue, the problem of data and annotation lacking issue is also existing in this area. More recent learning with limited labelled data methods are developed based on the one-shot learning framework to mitigate this issue, which indicates learning novel categories with single labeled data. However, applying the conventional one-shot learning framework directly is inadequate to achieve the desired medical action recognition performance, since there exist several limitations which need particular attention to further improve its capability in medical action recognition applications. Firstly, the performance of the entire framework is heavily dependent on the preprocessing stage since the high-quality training instances could provide more informative features to the models and lead to a more explainable artificial intelligence system. Extracting more informative features from the limited raw action sequences is beneficial for better interpreting the relationship

between different body parts of the subjects. The framework is also required to be able to clearly distinguish similar medical actions such as vomiting and abdominal pain as it may confuse the model and decrease the classification performance. Moreover, the model should include a mechanism to determine the aforementioned temporal mismatching issue, thereby synchronising the temporal information of the action sequences. Ultimately, in the context of medical action recognition, the association between different extracted features should be well exploited to improve the efficacy of the framework.

1.2 Aims and Objectives

This thesis aims to fully exploit the approaches based on deep learning techniques to obtain better promising performance for privacy-preserved medical action recognition. The main goal is to address the aforementioned limitations existing in the medical action recognition area by improving several principal components. The particular objectives are:

- Objective 1: Improving the fall events classification performance by applying the proposed redundant information reduction method.

In Chapter 3, a redundant information reduction method which processes the useless information from the raw action sequences is developed for multiple fall events classification, which aims to avoid the impact of redundant information and decrease the computational cost.

- Objective 2: Improving the fall events classification performance by applying the two-stage framework selection and filtering the incorrect classification.

In Chapter 3, the establishment of enhanced fall events classification reliability is divided into two stages: Identifying and filtering out the normal actions from the raw

imbalanced dataset by utilizing the parameter-controlled gating approach and specific multiple fall events classification. This is beneficial to be less prone to incorrect action recognition and the imbalanced data issue.

- Objective 3: Improving the quality of dataset annotations by applying noisy annotation cleaning learning and cascaded learning pipeline.

In Chapter 4, noisy label detection and pruning algorithms are used to retrieve accurate and qualitative annotations for achieving promising performance. The extracted human skeleton data is utilized as the training data for handling the dynamic lighting conditions.

- Objective 4: Enhancing the robust ability the improving the performance by applying the proposed learning with noisy label algorithm.

In Chapter 4, in order to fully exploits the potential effectiveness of noisy instances, a joint cooperative training method using the trinity networks is used to mine the non-corrupted annotations and decrease the negative impact of the noisy labels for human fall events classification.

- Objective 5: Elevating the learning with limited data ability by performing one-shot learning for medical action recognition.

In Chapter 5, to enhance the one-shot learning framework, the dynamic time warp mechanism and the cross-attention mechanism are jointly used for mitigating the aforementioned temporal mismatching and similar medical action issues, respectively. Moreover, a novel action data generation approach is applied for better privacy protection.

- Objective 6: Improving the one-shot learning system to be further accurate in medical action recognition by extracting further feature information and applying multiple fusion approaches at diverse levels.

In Chapter 5, a novel human action feature is extracted from the initial sequences to represent the direction-level information. Moreover, a multiple-level fusion approach is developed to exploit the potential complementary relations between different features for further improving performance at both feature-level and decision-level.

1.3 Thesis outline

The remainder of this thesis is structured as follows:

Chapter 2, in general, gives a relevant literature review of privacy-mitigating medical action recognition algorithms, and also explains the background preliminaries that are helpful to derive and evaluate the proposed techniques in the thesis. The existing challenges associated with medical action recognition are first discussed from four perspectives. Moreover, related developed research algorithms are allocated by carrying through the major components of the proposed system, where the limitations of these approaches are given. The primary components for medical action recognition in this thesis are also provided, including the deep neural network, peer network and ProtoNet. After that, in order to compare with the state-of-the-art approaches, both benchmark datasets and evaluation metrics are given detailed for the medical action recognition system.

The major technical contributions of this thesis are divided into the next three chapters. Chapter 3 aims to address the first and second objectives and provides contributions to improve the model performance by filtering the redundant information and reducing the computational cost via the proposed framework. This chapter is mostly based on the works in [26] and [27]. Chapter 4 illustrated the proposed algorithms and pertains to the third objective via the proposed dataset purification method, which indicates removing the corrupted annotations in the early preprocessing stage of the entire fall events classification framework. In order to fully exploit the potential of the corrupted instances, this chapter also aims to deal with

the fourth objective via the proposed learning with the noisy label algorithm for multiple fall events classification. The technical parts in this chapter were previously presented in [28] and [29]. Chapter 5 continuously deals with data issues which are the fifth and sixth objectives by introducing a novel one-shot learning framework and extracting various human action features in the early data processing stage. Furthermore, a multiple-level feature fusion approach is developed to further improve medical action recognition by constructing the complementary relations between different extracted features from the limited raw data. This contribution was partly published in [30] and [31]. In the final chapter, the contribution chapters are summarized and a discussion of the future work is provided.

Chapter 2

RELEVANT LITERATURE REVIEW AND PRELIMINARIES

2.1 Introduction

Due to the distinctive attributes of the medical action recognition task as mentioned before, it requires a more accurate and robust performance with privacy protection methods for the real-world environments. For this purpose, many researchers have been investigating related algorithms to fulfil these requirements in recent years. The summary of the recent progression for medical action recognition, which is mostly relevant to the proposed approaches in this thesis is provided in this chapter. In fact, the improved medical action recognition approaches are from diverse dimensions such as model upgrading and data pre-processing, therefore it is difficult to classify them in a universal criterion. Therefore, this chapter introduces the corresponding existing algorithms along with the limitations of the current medical action recognition frameworks. The main challenges of medical action recognition are first introduced, which are primarily divided into four aspects: multiple fall classification, privacy protection, data limitations and noisy annotation issues. Following by the recently relevant developed methods for these limitations. After that, the evaluation instances from four different benchmark datasets together with the applied evaluation metrics in this thesis are presented for comparing the proposed approaches with the other

state-of-the-art methods.

2.2 The challenges of medical action recognition

Different from normal action recognition, medical action recognition becomes more challenging, especially with the multiple fall classification, privacy protection, data limitations and noisy annotations issues. In the following subsections, relevant solutions to these challenges are presented and discussed.

2.2.1 Multiple fall classification

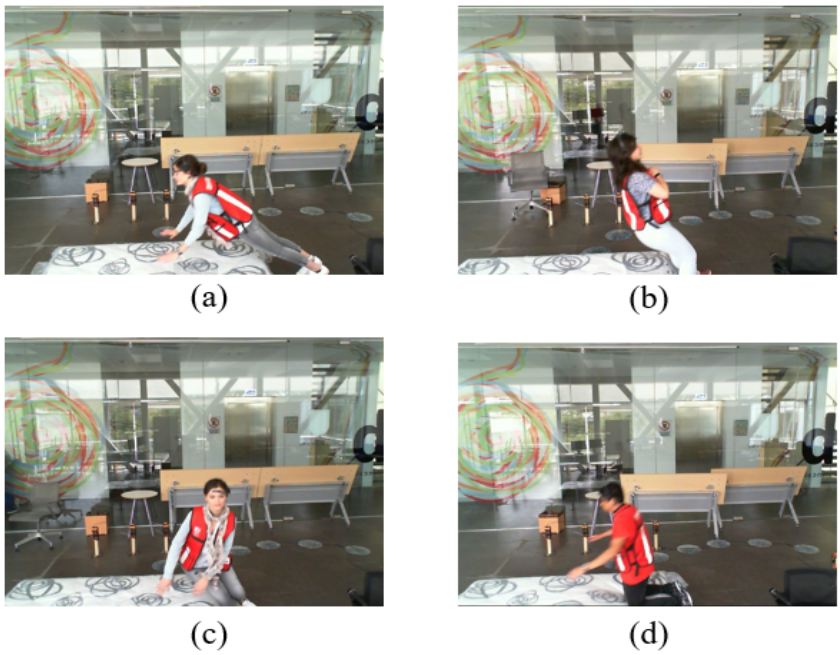


Figure 2.1: Example visual illustrations of multiple fall classification from the UP-Fall dataset: (a) Hand falling. (b) Backwards falling. (c) Sideways falling. (d) Knee falling.

According to the conventional fall detection research works, the approaches in this area are primarily divided into two categories: wearable sensor-based and video-based. The wearable sensor-based methods aim to detect fall events by analyzing the significant physiological variations of the human body, which can be captured

by wearable sensors, such as accelerometers [32–34], gyroscopes [35–37] and pressure sensors [38–40]. For example, a single gyroscope with three different identification measures was proposed in [41], which are angular velocity, angular acceleration and the body angle variations of the targets. The empirical thresholds were set for discriminating the normal actions and fall actions. A sensor fusion method which used both accelerometers and gyroscopes was established in [42], which is more accurate and robust than using single sensors. However, due to the major target of fall detection being aging population, wearable sensor-based methods are suffering from several perspectives. Firstly, aging population often exhibits resistance to adopting such intrusive wearable equipment due to subjective discomfort [43, 44]. Secondly, cognitive impairments-related disease or memory decline among the aging population can lead to objective forgetting to equip the device [45, 46]. Thirdly, long-term usage of the device without recharging may result in failure to capture the body variation information [47]. Moreover, these wearable devices could potentially exacerbate the injuries to the aging population, such as fractures.

The video-based fall detection approach is the most widely applied in recent decades due to the extreme developments of computer vision and deep learning techniques, which normally utilized the information captured from static RGB cameras to distinguish normal actions and fall actions [12, 48, 49]. Different from the wearable sensor-based method, most static RGB cameras are non-intrusive and wired thereby preventing the aforementioned limitations in wearable sensor-based methods [13]. However, these vision-based approaches may suffer from the dark scene and dynamic illumination [50]. Moreover, in contrast to wearable sensor-based approaches, traditional vision-based research predominantly focused on binary fall detection tasks, which involve determining whether a fall event has occurred or not [13]. There was a marked scarcity of attention dedicated to the distinct fall events classification tasks. Moreover, the majority of public fall events benchmark datasets are designed with

the lack of types of falls or amount of the subjects [51–53], with only the public UP-Fall [1] dataset containing the sufficiently large scale and multimodal collection of multiple fall event categories, which is presented in Figure 2.1.

The injuries incurred by the aging population due to various types of fall incidents exhibit substantial variability. For instance, falling onto a chair might lead to minimal injuries, while falling down from a staircase may result in hindbrain injuries, thereby endangering the overall safety and well-being of the aging population. Consequently, according to the aforementioned factors, the fall detection research field is in dire need of vision-based methods that can comprehensively categorize various fall event categories.

2.2.2 Privacy protection

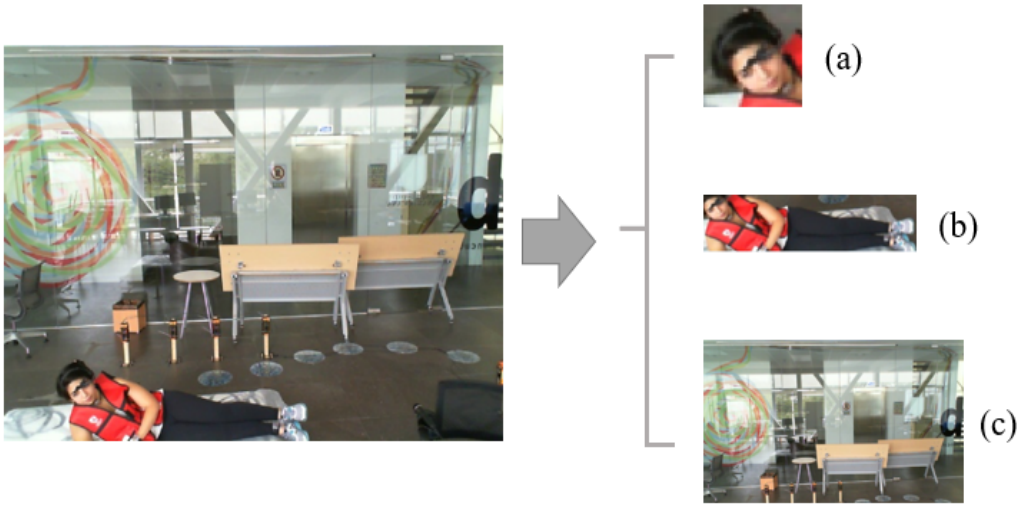


Figure 2.2: Medical action recognition illustration of privacy protection issue: the personal information of the target needs to be protected from leaking. (a) Facial information. (b) Dressing information. (c) Background information.

In medical-related areas, privacy protection is one of the most challenging tasks. Medical data contains sensitive health, personal and even familial information, the leakage of which could cause disruptions in their lives and financial concerns [54, 55].

An example illustration of private information leakage is shown in Figure 2.2. As for video-based medical action recognition, which is a medical-related task, it is important to handle action sequences for preserving the privacy of targets. With the increasing extent of privacy processing, more data information may lose and result in the model performance decreasing [56]. The objective of this task is to achieve a balance between maximizing model accuracy and preserving privacy such as facial information, dressing information and background information. The video-based privacy-preserving approaches for medical action recognition are primarily divided into two categories: low-resolution processing and training with human skeleton sequences.

The utilization of low-resolution techniques aims to preserve visual privacy features while remaining the model performance by reducing the resolution of the raw images [57–59]. The approach proposed in [58], aims to involve the down-sampling of the original videos to obtain extremely low-resolution frames, along with utilizing specialized models to enhance the practicality. Another approach proposed in [59] involves the initial identification of the most critical visual privacy features through user surveys, following the resolution reduction processing on the raw data to ensure privacy protection. In [60], the researchers applied both spatial and temporal resolution reduction on the raw video sequences to preserve privacy information, and the experimental outcomes are thoroughly analyzed. However, while these methods effectively ensure privacy protection, the usage of low-resolution techniques significantly decreases the model performance [60–62], contradicting the requirements of the medical field which necessitates accurate performance.

Utilizing human skeleton sequences offers a preserving privacy solution by retaining positional information of the landmarks while removing all other raw image details [63–66],. This approach enables privacy protection while maintaining model performance [65]. Moreover, skeletal data can mitigate the impact of dynamic illumi-

nation and darkness and reduce computational costs. In [67], a directed acyclic graph data paradigm was proposed for processing the skeleton sequences to enhance model performance. In [68], a dynamic network was designed that simultaneously learns both temporal and spatial information from skeleton data, thereby enhancing model representation and generalization. Based on the foundations of [68], [63] extracted skeleton features at different levels and fused them to enhance model robustness and performance. However, these skeleton sequences still reveal the position and landmarks information upon observation. In this thesis, both Chapter 3 and Chapter 4 utilize the extracted skeleton data for model training. To further enhance privacy protection and address the aforementioned issue, Chapter 5 introduces a novel signal-level image generation (SIG) method for skeleton representation, rendering training data impervious to leaking the positions of landmarks or motion information upon observation.

2.2.3 Data limitations

Medical action detection datasets exhibit several types of data limitations, are primarily consisted of two following categories: the limited quantity of medical action instances and the medical annotation lacking. Addressing these two issues within deep learning-based medical tasks presents significant challenges, as the quality of the dataset constitutes the most important factor influencing the performance [69–71].

On one hand, due to rarely occurrence of medical actions compared to normal human actions, two challenges are raised. Firstly, this leads to an imbalance of quantities issue within the dataset, where the number of normal action sequences greatly surpasses that of medical action sequences. The direct consequence of data imbalance is to exhibit overfitting during the training process [72–74], subsequently diminishing model performance and generalization ability. Some methods tackled the imbalanced issue by employing focal loss [75], aiming to enhance model performance by atten-

uating attention towards numerous sample categories and augmenting focus on the few sample categories. In Chapter 3 of this thesis, a two-stage framework is proposed to mitigate the imbalanced dataset issue. Secondly, the limited quantity of medical action samples avoids applying the conventional deep-learning-based framework for training. Recent research works have focused on exploiting one-shot or few-shot frameworks to mitigate this challenge [76–79]. These frameworks were trained with the numerous action categories to acquire prior knowledge, followed by fine-tuning their parameters using the few sample categories. A detailed explanation of the well-known one-shot learning frameworks will be presented in the following section, which is utilized as the backbone of the third contribution of this thesis. This issue also serves as one of the motivations for the approach proposed in Chapter 5 of this thesis.

On the other hand, the second challenge is caused by the annotation lacking, stemming from the dual concerns of privacy protection and the expensive costs associated with annotations. This issue serves as the second motivation for the approach proposed in Chapter 5 of this thesis. Over the past few years, various training approaches have been employed to accomplish this issue. As aforementioned, the one-shot learning framework can tackle this issue by training with a small set of few sample categories (medical actions) which are correctly annotated. Furthermore, self-supervised learning frameworks were investigated to tackle the annotation-lacking issue [80–82]. By obtaining information from the entire dataset and then generating pseudo-labels, these approaches facilitated the model through the training stage without numerous annotations. However, the efficacy of these approaches is constrained by the quality of the generated pseudo-labels, potentially decreasing the performance [82]. The semi-supervised learning frameworks were embraced for addressing this issue by training the model with a subset of annotated instances and then generating pseudo-labels for unlabeled instances [83–85]. Similar to self-supervised learning approaches, the efficacy of semi-supervised learning heavily depends on the quality of the generated

pseudo-labels. Moreover, the aforementioned data imbalance issue and the noisy annotation which will be discussed in the following subsection could derive extremely negative impacts on the performance of semi-supervised learning frameworks due to the characteristics biases. In this thesis, a one-shot learning framework will be utilized to address the aforementioned data limitation issue.

2.2.4 Noisy annotations

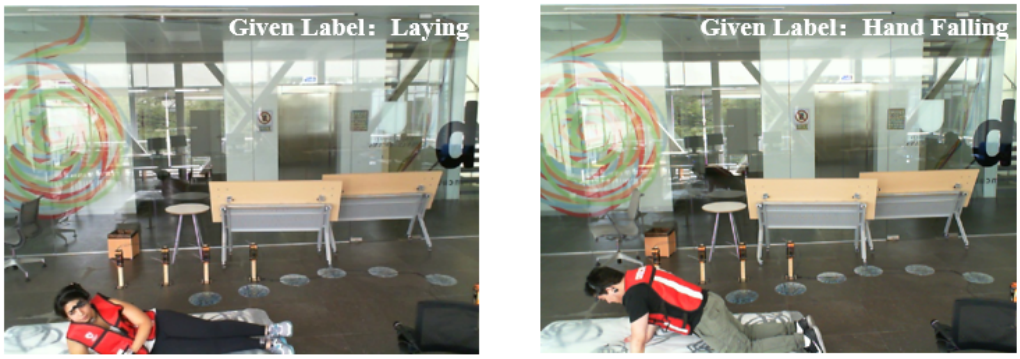


Figure 2.3: Example visual illustration of noisy annotations from the UP-Fall dataset: (a) Given label: laying, Correct Label: sideways falling. (b) Given label: hand falling, Correct Label: knee falling.

In data-driven deep learning methods, the quality of the annotations plays a fundamental role in the performance of the frameworks [69]. The low quality of annotations results in incorrect training direction which can deteriorate the classification performance of the neural models. The fundamental origins of noisy annotations in datasets can be attributed to the following primary points: firstly, the heavy time cost and the heavy labor cost associated with the manual annotation method facilitate some researchers to apply fully or semi-automated annotation methods, consequently resulting in bring the noisy annotations into the dataset during the inaccurate models [86]. Secondly, some researchers employ multiple annotators for collaborative annotation work, which can introduce noise due to reasons such as work fatigue, time constraints

and disparities in the understanding of identical data categories [70]. Finally, the imbalanced data issue can lead annotators to identify the incorrect annotations to the infrequently data categories due to the rarity of the instances. The example visual illustration of noisy annotations is shown in Figure 2.3.

The noisy annotations bring substantial negative impacts on the model performance. A long-term training process can lead the model to update parameters towards the incorrect direction and overfit to noisy instances [71]. Furthermore, medical-related research works necessitate promising results due to their direct relevance to the life security of patients [71]. Hence, it becomes paramount to tackle the noisy annotations issues for medical action recognition. To address this issue, researchers have exploited various methods for noisy annotation handling [5–7, 87]. In recent years, the DNN model has been proven to learn the simple (clean) data at the beginning of the training process and then gradually trend to adapt the hard (noisy) data with the increasing of the training epochs [88]. In addition, the clean data also have been proven to have a smaller loss value than the noisy data [89], which inspired the researchers to investigate this small-loss theory as the foundation for discriminating the clean instances and noisy instances during the training process.

Since the DNN models have an extensive capacity for fitting to noisy instances, therefore the updating strategy plays a crucial role in robust learning with noisy label methods [69, 90]. The disagreement updating strategy was proposed in [87] to enrich the effective training process rather than overfitting to the noisy instances, which trains two DNN models simultaneously and attempts to update the parameters by utilizing the instances only if the predictions from the two models are different. Han et al. [5] attempted to improve the noisy label learning framework by introducing the cross-update theory, which indicates training two DNN models simultaneously by applying the useful data with small loss values to the peer network for the parameter updating. However, the two DNN models may easily tend to converge with the in-

creasing training epochs by employing this co-training algorithm [5]. In order to keep the DNN models diverged to keep learning the knowledge from the useful instances and to further improve the performance, it is needed to address the aforementioned converge issue by investigating the novel updating strategy. In this thesis, both a data purification method based on the joint probabilities matrix and a noisy label learning framework are proposed for addressing the data quality issue.

2.3 Primary components for medical action recognition

2.3.1 Deep neural network

The deep neural network (DNN) models have been exploited for addressing many practical video processing tasks such as human action recognition, human tracking and abnormal human behavior detection. A DNN model primarily comprises three distinct layers: the input layer, hidden layers, and the output layer, which contain numerous neurons for delivering and processing the feature information.

The DNN model aims to minimize the loss value between the training targets and the output predictions for updating the parameters by utilizing the weights and biases. The gradient descent strategy is applied to decrease the weights and the biases for minimizing the loss value to fit the dataset and obtain the model with the best performance.

Moreover, multiple DNN variation models are proposed for specific tasks, such as convolutional neural networks (CNN) for processing the image data [91, 92], graph neural networks (GNN) for processing the graph data [93, 94] and recurrent neural networks (RNN) for handling the data with time series information. In this thesis, DNN models are utilized as the backbone in both Chapter 3 and Chapter 4. Additionally, one of the most extensively employed CNN architectures is employed in Chapter 5 for feature learning.

2.3.2 Peer networks for learning with noisy annotations

For the proposed learning with noisy labels algorithm in Chapter 4, the peer networks are utilized as the backbone model for mining the clean instances. The peer network consists of two vanilla DNNs with identical structures, and each mini-batch of data is simultaneously fed into these two DNN models for training. Following the small-loss criteria, instances with small loss values are considered clean samples. The loss values of the data from each mini-batch are ranked after each training epoch. Based on the proportion of noise, it is determined how many clean samples exist within a mini-batch. Initially, the training aims to maintain a higher proportion of clean samples and gradually decrease this proportion as epochs increase. The DNN model is inclined to learn from simple clean samples in the initial stage of the training process before tackling more challenging noisy samples. The peer network models have the ability to filter out the noisy instances at the beginning of the training since they have not yet received the impacts from the noisy instances. The peer networks will gradually overfit the noisy data with the training epochs increasing and thereby decrease the discriminatory capability between the clean instances and noisy instances.

The cross-update and agreement strategies are also employed for enhancing the noisy-tolerant ability in the proposed learning with noisy labels framework. Intuitively, different classifiers can generate distinct decision boundaries leading to different learning abilities. Additionally, the random initialization of the two DNN models results in divergent initial parameters, thereby enhancing the disparate learning capabilities of peer models. The models update the parameters by using the loss values of the selected clean samples from their peer models at the end of each training epoch, rather than individually updating the parameter by using the loss values of the entire mini-batch instances. Through this peer-review update strategy, the two DNN models can adaptively refine training errors originating from the peer network, even if where the selected instances transmitted from the peer model are not entirely clean.

2.3.3 The prototypical network for medical action recognition

Definition of one-shot learning

The objective of one-shot learning is different from conventional supervised learning. Conventional supervised learning aims to learn the features from the training set during the training process and generalize to the testing set to acquire promising performance. However, one-shot learning aims to obtain the distinguishing capability to learn from the limited samples, which is called 'learn to learn'.

To be precise, the samples in the training set and the testing set are independent of conventional supervised learning, but the categories are the same. Nevertheless, both the samples and categories of the instances utilized for testing are independent of instances from the training set in one-shot learning. As a result, one-shot learning presents more difficulty compared to conventional supervised learning. There are two definitions which are widely used in one-shot learning: the support set and the query set. The support set is utilized for fine-tuning the model by learning with limited instances with novel classes after the pre-training stage. The labels of the instances in the support set will be provided to the model for further information learning. The support set is often described using the notation 'n-way-k-shot', where 'n' denotes the number of novel classes in the support set and 'k' is the number of samples per class. i.e. '5-way-1-shot' indicates the existence of five novel classes in the support set, with each class having one instance. The query set is employed to evaluate the model after the meta-training stage based on the support set. The experimental results on the query set instances are utilized as the conclusive performance of the framework.

The preliminaries of the prototypical network

In Chapter 5 of this thesis, ProtoNet [9] is selected as the one-shot learning framework. For each class, it calculates the mean vector of the embedded instances from the support set as the prototype for that class. The model is tested on the query

set by calculating a specific variation of Euclidean distance between the instances from the query set and the acquired prototypes of each class, which is called regular Bregman divergence [95]. Bregman divergence represents a form of distance between two points in a specific space. When these points are subject to arbitrary probability distributions, the averaged point is invariably the one with the minimized average distance to these points in the specific space, which is used for calculating the distance between the prototype instance and the query instance. The illustration of ProtoNet is provided in Figure 2.4.

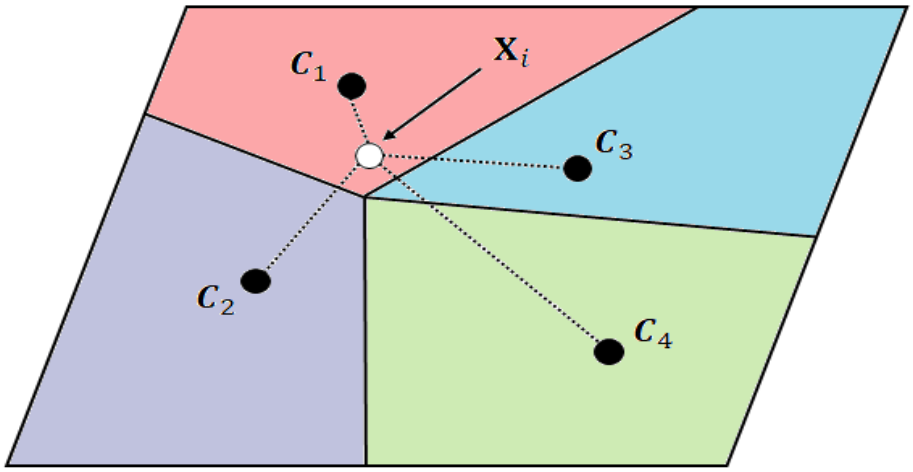


Figure 2.4: The illustration of prototypical networks in one-shot learning scenarios. The one-shot prototypes $\mathbf{c}_k$ are calculated as the anchors for determining the distances with the query actions $\mathbf{x}_i$, thereby discriminating the similarities between the prototypes and query actions.

In the one-shot learning task, a small support set of N labeled instances $I = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_i \in \mathbb{R}^D$ indicates the instance features with D dimensions and $y_i \in \{1, \ldots, K\}$ is the ground truth labels. I_k indicates the instance set with class k.

The M dimensional prototype of each class $\mathbf{c}_k \in \mathbb{R}^M$ is obtained after the ProtoNet

computing. The mean vectors of the embedded instances are presented as:

$$\mathbf{c}_k = \frac{1}{|I_k|} \sum_{(\mathbf{x}_i, y_i) \in I_k} f_\phi \left(\mathbf{x}_i \right) \tag{2.3.1}$$

where f_ϕ is the encoder function which embedding the features $\mathbb{R}^D \rightarrow \mathbb{R}^M$ with learnable parameter ϕ. The Bregman divergence is utilized for evaluating the similarity between the instances, which is formulated as follows:

$$\mathrm{dis}_\varphi \left(\mathbf{z}, \mathbf{z}' \right) = \varphi(\mathbf{z}) - \varphi \left(\mathbf{z}' \right) - \left(\mathbf{z} - \mathbf{z}' \right)^{\mathbf{T}} \nabla \varphi \left(\mathbf{z}' \right) \tag{2.3.2}$$

where φ is the differentiable and $\mathbf{z}, \mathbf{z}'$ denotes the points to be calculated for distances. ∇ is the gradient function and $\mathbf{T}$ is the transpose operation. ProtoNet then calculates the distance between query instance $\mathbf{x}$ and each class prototype and calculates the classification probability distribution by softmax:

$$p_{\boldsymbol{\phi}}(y = k \mid \mathbf{x}) = \frac{\exp \left(- \mathrm{dis} \left(f_{\boldsymbol{\phi}}(\mathbf{x}), \mathbf{c}_k \right) \right)}{\sum_{k'} \exp \left(- \mathrm{dis} \left(f_\phi(\mathbf{x}), \mathbf{c}_{k'} \right) \right)} \tag{2.3.3}$$

where $p_{\boldsymbol{\phi}}(y = k \mid \mathbf{x})$ represents the probability of instance $\mathbf{x}$ from the query set belongs to class k. $\mathrm{dis}(\cdot)$ is the distance function as shown in equation 2.3.2 and $\exp(\cdot)$ is the exponential function.

$$J(\phi) = - \log p_\phi(y = k \mid \mathbf{x}) \tag{2.3.4}$$

The learning process of Protonet is to minimize the value of $J(\phi)$ by utilizing the stochastic gradient descent (SGD). This is equivalent to maximising $p_{\boldsymbol{\phi}}(y = k \mid \mathbf{x})$ value to obtain the predicted classes.

2.4 Evaluation Datasets

To demonstrate the effectiveness and applicability of the proposed medical action recognition methods, it is essential to evaluate the experimental performance of these proposed methods. In this section, several widely used and public medical action recognition benchmark datasets are explained, which are applied to demonstrate the efficacy of the proposed methods. The image example illustrations from the following benchmark datasets are provided in both Figure 2.5 and Figure 2.6.

2.4.1 UP-Fall Dataset

UP-Fall dataset is a large-scale fall detection dataset which provides multimodal sensor data, including the accelerators, EEG, infrared and RGB cameras. These data comprise raw and action features from 17 healthy subjects from 12 different human actions, which includes falling using hands, falling using knees, falling sidewards, falling backwards and falling sitting in an empty chair. Each of them has three trials. Moreover, the dataset provides two different experimental use cases for research. The RGB images are utilized for the experiments and fed into Alphapose [8] for the skeletal data extraction to remove the noisy information and protect the privacy information. The amount of skeleton data groups is 220,660 after applying the pre-processing approach, which will be shown detailed in Chapter 3.

2.4.2 NTU RGB+D 60

NTU RGB+D 60 is a 3D large-scale human action dataset which provides skeleton data sequences. Each action is captured by 3 Kinect V2 cameras at the same height but with different horizontal angles, which are $-45°$, $0°$ and $45°$. This dataset provides RGB images, depth map sequences, 3D skeleton data and infrared (IR) videos for each sample sequence. For privacy protection considerations, skeleton sequences data are selected as the training data in this thesis. The skeleton data sequences consist of

56,880 instances for 60 types of human actions. These sequences are recorded from 40 different subjects with 17 different scene conditions, each subject provides 25 pose landmarks. The age range of the subjects is from 10 to 35 and the actions are labelled from A001 to A060.

2.4.3 NTU RGB+D 120

NTU RGB+D 120 is an extended version of NTU RGB+D 60. It contains 120 types of human actions recorded from 106 subjects in 155 different scene conditions and each subject also provides 25 pose landmarks. There are 114,480 skeleton sequences including daily, mutual and medical-related actions. This dataset is the most widely used benchmark for action recognition in recent years, which has diversity in camera views, environmental conditions and human subjects. The extended actions are labelled from A061 to A120.

2.4.4 PKU-MMD

PKU-MMD dataset is a 3D large-scale dataset which contains 1,076 long human action sequences in 51 human action classes, which include daily actions, interactive actions and medical-related actions. This dataset is recorded from 66 subjects in 3 different camera views and the ages of the subjects are from 18 to 40. There are over 20,000 instances provided with multi-modality data, including RGB, depth, infrared radiation and skeleton sequences. The dataset is captured by the Kinect V2 cameras and the skeleton data is chosen for training. The skeleton data consists of 3-dimensional locations of 25 human landmarks for detected targets in scenes.

2.5 Evaluation Metrics

This section describes the four common evaluation metrics to evaluate the action recognition performance of the related medical action recognition approaches. One

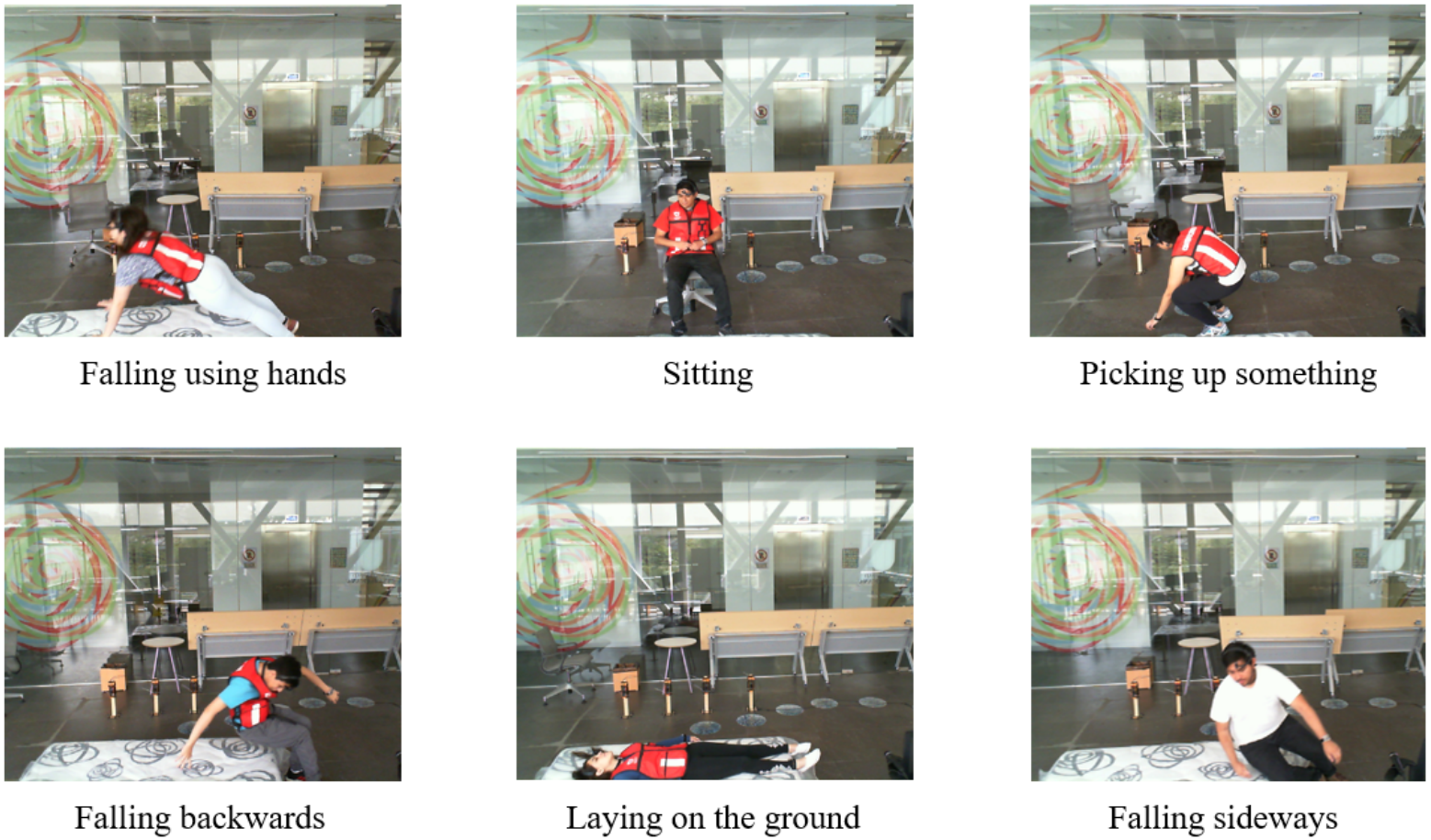


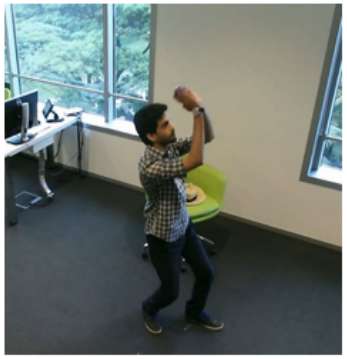
Figure 2.5: Example image frames from the UP-Fall dataset [1], which is used in Chapter 3 and Chapter 4 for evaluating the proposed methods.

NTU 60
A043 Falling

NTU 120
A061 Put on headphone

PKU-MMD
A011 Falling

NTU 60
A051 Kicking



NTU 120
A063 Shoot at basket

PKU-MMD
A030 Reading

Figure 2.6: Example image frames from the NTU RGB+D 60 [2], NTU RGB+D 120 [3] and PKU-MMD [4] benchmark datasets, which are applied in Chapter 5 for evaluating the proposed methods. Detailed annotations are given along with the action descriptions.

term is precision which is widely used for fall detection retrieval to evaluate the accuracy of positive predictions made from the framework. The second is recall which determines the proportion of actual positive cases that can be correctly identified by the model. The third is the F1 score which is one of the widely used tools to give the individual and brief performance and it is relevant for the imbalanced dataset. The last is Top-1 accuracy which plays the most important role in the action recognition field, which will be applied to compute the benchmark performance of the proposed method, so as to accomplish a fair comparison with the other state-of-the-art medical action recognition methods in the ranking board.

2.5.1 Precision

The precision score is to measure the average medical action classification errors between the predicted action categories and the ground truth, which is formulated as:

$$Precision = \frac{TP}{TP + FP} \tag{2.5.1}$$

Precision measures the performance of the proposed algorithm between the predicted positive and the true positive instances. It is mainly applied in Chapter 4 to evaluate the proposed noisy label learning method, which is used for inspecting the rate of the predicted clean instances to the ground truth clean instances.

2.5.2 Recall

The recall score measures the proportion of actual positive cases that were correctly identified by the model, the definition of recall is:

$$Recall = \frac{TP}{TP + FN} \tag{2.5.2}$$

The model with a high recall score indicates that it has a strong ability to correctly identify positive instances. Recall is one of the crucial evaluation metrics in medical-related areas, which will be utilized in Chapter 3 to determine the model selection.

2.5.3 F1 Score

The F1 score as the evaluation metric, is designed to measure the performance of the fall events classification framework between the human daily normal actions and different fall events, and it is calculated as:

$$\begin{aligned} F_1 \text{ score } &= 2 \cdot \frac{\text{Precision } \cdot \text{ Recall}}{\text{Precision } + \text{ Recall}} \\ &= 2 \cdot \frac{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \\ &= \frac{2TP}{2TP + FP + FN} \end{aligned} \tag{2.5.3}$$

where the range of $F1$ score is $[0, 1]$, higher value indicates the proposed framework has better performance. The scores are reported in percentage formats. Since the occurrence of medical actions is not as frequent as human normal actions, which will lead to data imbalanced issues. F1 score combines both precision and recall, which is suitable for evaluating the imbalanced dataset and remains effective in the presence of imbalanced data and missing label issues. Due to the fact that the F1 score provides a comprehensive and intuitive manner to analyze the action classification ability of the proposed framework, it is applied in Chapter 3 as the main evaluation metric for addressing Objective 1 and Objective 2.

Combining precision and recall, it is suitable for unbalanced datasets and remains effective in the presence of unbalanced data or missing labels by calculating the harmonic mean of both recall and precision. The F1 score provides a comprehensive and intuitive way to evaluate the classification ability of a model.

2.5.4 Top-1 Accuracy

The top-1 score as accuracy is often considered the most important evaluation metric for appraising the medical action recognition performance. It refers to the percentage of times that model correctly predicts the highest-probability class for the specific input sequence, and it is calculated as:

$$Top-1 = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.5.4}$$

It is mainly comprised of the following two error categories: the false positive and the false negative. A classification or mismatched error is likely to happen between similar human medical actions, such as neck pain and headache. In order to achieve promising Top-1 accuracy, these errors are expected to be as few as possible. This Top-1 accuracy is the most widely-used evaluation metric for comparing the proposed algorithms with the benchmarks in Chapter 4 and Chapter 5. The scores are illustrated in the percentage format for the evaluations.

2.6 Summary

In this chapter, the challenges of medical action recognition were first introduced along with the existing approaches as well as the limitations of these approaches were investigated. Then, the three primary components of the proposed approaches in this thesis were presented: deep neural network for multiple fall events classification, peer network for noisy label issue and ProtoNet for data limitation issue. This thesis aims to address the aforementioned challenges for medical action recognition tasks. Four benchmark datasets have also been presented and the evaluation metrics were provided to evaluate the proposed approaches. In the next chapter, the redundant information reduction theory and the two-stage framework for the specific multiple fall classification task will be first developed to improve the performance.

Chapter 3

PRIVACY MITIGATING HUMAN FALL EVENTS CLASSIFICATION USING DATA FUSION AND CASCADED LEARNING

3.1 Introduction

In human action recognition, due to the vulnerability of the aging population, it is important for the framework to be accurately and promptly aware of the occurrence of fall events since it has been proven that the time taken to be detected after the fall events for the aging population is positively correlated with the severity of the injuries they suffer.

Conventional fall detection methods are empirically divided into two main categories, which are wearable sensor-based and video-based. Due to the memory decline of the aging population, the lack of willingness to wear the devices, and the power consumption of wearable devices, there are numerous limitations and uncertainties associated with wearable sensor-based approaches and may not be effective and sustainable in dealing with fall detection in real-world scenarios. Alternatively, the human action recognition methods in data-driven mechanisms based on the video sequence

have been widely developed to analyse human behavior with promising model performance. A substantial proportion of the established methods aims to apply the human skeleton sequence to avoid the impact of the dynamic illumination and preserved the privacy information. Moreover, the computational cost for calculating the skeleton sequence is much less than the video sequence. However, previously established video-based algorithms have primarily focused on addressing the occurrence of fall events, which is the binary detection task. Due to the relatively fragile physical condition of aging population, different fall events lead to distinct physical and emotional injuries. For instance, the risk posed by falling backwards from stairs is significantly more critical than falling into a chair. On the other hand, exploring the redundant information in the human skeleton sequences could provide better performance and robustness in fall events classification with less computation time.

In order to address the aforementioned limitations of the existing methods, this chapter presents a deep neural network-based (DNN) framework for fall events classification in video. The proposed system primarily exploits two concepts: redundant information reduction and a two-stage cascaded learning framework for fall events classification. The redundant information reduction technique aims to remove the skeleton parts which are relatively low weights for the fall classification by considering improving the performance and robustness in the data processing stage. A skeleton feature extractor is applied to distillate the skeleton information and following the skeleton data preprocessing stage to remove the undesired subjects and redundant empty frames. After that, the redundant information reduction technique is applied and four classic clustering classifiers are used for verifying the efficiency of the proposed method, which are Random Forest (RF), Support Vector Machine (SVM), Multiple Layer Percepton (MLP) and Adaptive Boosting (AdaBoost). In order to mitigate the imbalanced data issue and further improve the performance of the fall events classification on the framework level, a two-stage multiple fall events clas-

sification framework based on DNN is proposed for addressing these issues with the extracted skeleton features. In the initial stage, the model focuses on discriminating the normal actions and fall actions based on the re-annotated binary labels. In the conclusive stage, the DNN is applied for the multiple fall events classification based on prior knowledge from the initial stage. Moreover, the gating parameter is utilized along with the proposed structure to further boost the classification performance, which would refine the controlling of the initial stage for better discriminating ability.

1. A novel DNN-based strategy is developed for multiple fall events classification on video data.

2. The redundant data issue is mitigated on the data processing level with reduced computational cost and the privacy information is protected by utilizing the extracted skeleton features.

3. A two-stage learning framework for human fall events classification based on the extracted skeleton features is proposed to address the imbalanced data issue on the framework level and further improve the performance.

This chapter addresses the first objective and second objectives of the thesis, which match the deep learning-based fall events classification using the redundant information reduction method published in [26], and the DNN-based fall events classification using the proposed two-stage learning to improve the performance which presented in [27]. Section 3.2 introduces the proposed fall events classification algorithms in detail as four main components, including raw RGB image processing, skeleton feature extraction, redundant information reduction and the proposed two-stage learning framework. Experimental results are shown in Section 3.3 and the conclusion and discussions are reported in Section 3.4. The overview of the proposed fall events classification system is shown in Figure 3.1.

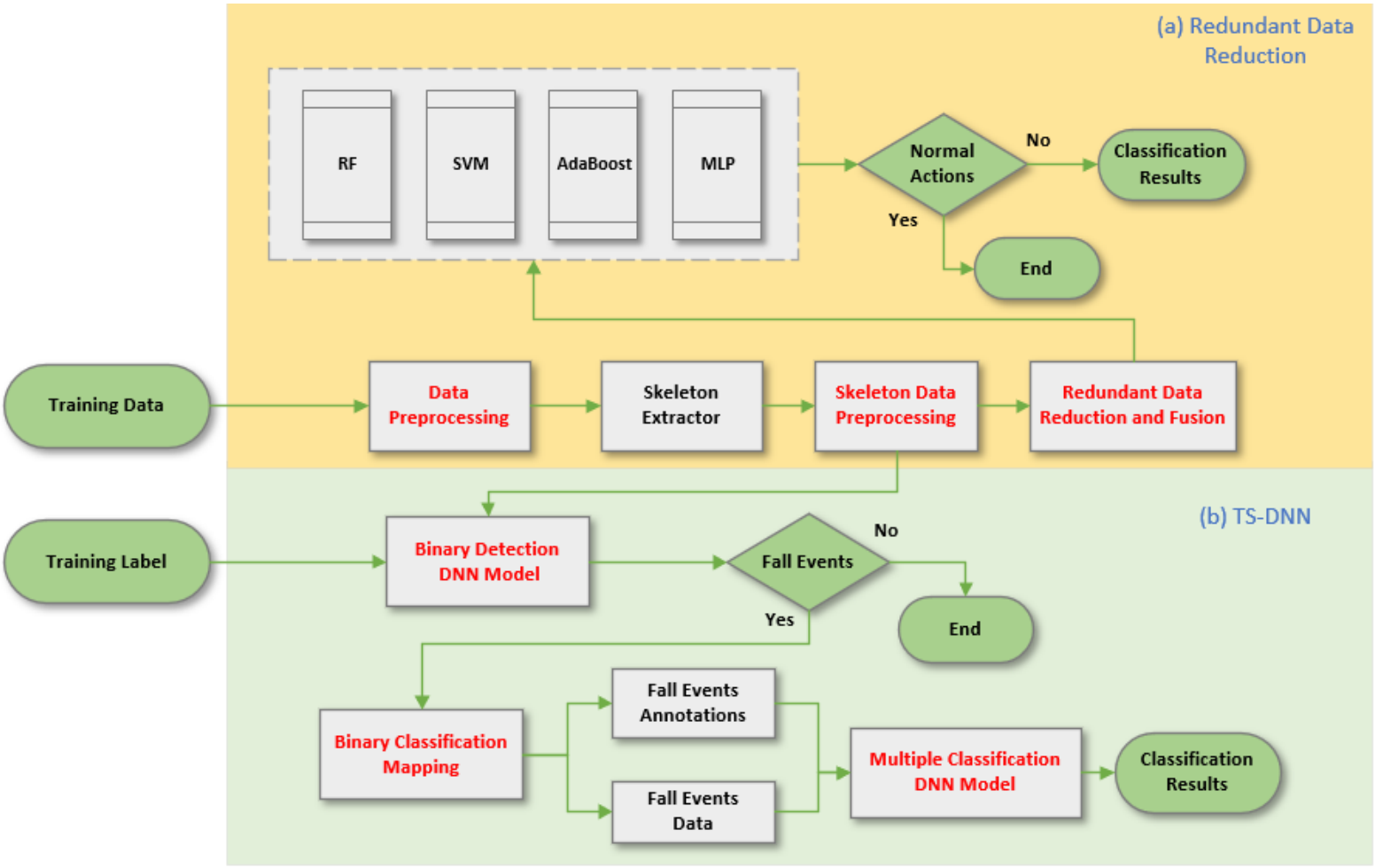


Figure 3.1: The overall block diagram for this chapter. (a) Part 1 of this chapter aims to reduce the redundant information from the raw action sequences and protect privacy. Four classical clusters are shown in dotted lines as the classifiers. (b) Part 2 of this chapter proposed a two-stage DNN-based framework to address the imbalanced data issue. The illustration of the DNN model is provided in section 3.2.2. The main contribution is highlighted in red.

3.2 Proposed Method

3.2.1 Data Processing

Data Preprocessing

Figure 3.2: An illustration of dataset recording environment for the UP-Fall dataset.

In this chapter, the experiments are implemented on the widely-used UP-Fall dataset. This dataset provides various daily normal actions and different fall events in video sequences. The UP-Fall contains various kinds of normal activities, such as walking, jumping and sitting. Figure 3.2 shows the recording environment of the UP-Fall dataset. The UP-Fall dataset contains two perspectives of video data from two Kinect cameras, which are the front view and the side view, respectively. The video data from the side view camera (CAM1) is selected as the data to the framework for the initial pre-processing. According to the description of the UP-Fall dataset, part of the action sequences are missing, in order to match the sequences and annotations, the synchronization operation is applied between the instances and the annotations. The annotations are transformed into the numpy file for the model training. The instances are sorted by the subject ID, trial ID and timestamp ID. The instances from the same subject, same actions and same trial are packaged into the same folder for sending to the skeleton extractor in the next step.

Skeleton Extractor

Since the video sequences exist dynamic illumination and leakage the privacy information, such as dress information, facial information and background information. Moreover, the computational cost for the video data is relatively higher than applying the skeleton sequences. To tackle these issues, AlphaPose is introduced for extracting the human skeleton features from the video data, which is open-sourced by Shanghai Jiaotong University (https://www.mvig.org/research/alphapose.html). The model was pre-trained on the COCO detection dataset and achieved 89.2% accuracy as the motion estimation performance. Readers are referred to [8] for further details of the model instructions. A pose graph which consists of 17 keypoints of the human body is generated by applying the image data to the model. Both the confidence scores c and the coordinates of the values are included in the set of the keypoints, which are denoted as $(\mathrm{x}, \mathrm{y}, \mathrm{c})$. Therefore, the model estimates a total of $3 \times 17 = 51$ sets of characteristics for each target in every frame. In this way, the initial RGB image sequences are converted into the skeleton sequences and will be utilized for model training after the following major target selection step.

Skeleton Data Preprocessing

(a) (b) (c)

Figure 3.3: The illustrations of the instance with a redundant target from the UP-Fall dataset, a pedestrian could be observed walking in the corridor. The data with different settings are conducted as follows: (a) The initial RGB image data with a redundant target. (b) The skeleton rendered data from AlphaPose. (c) The redundant data with only skeleton information before the major target selection step.

The UP-Fall dataset recordings were conducted within a controlled laboratory setting featuring glass walls and an adjacent corridor. With the meticulous data checking, two distinct reasons for interference with the recorded targets were identified during the data collection process: the first pertains to subject shadows resulting from reflections on the glass walls and the second involves the presence of individuals walking across the corridor during dataset recording. Subsequent to the initial data processing step, where raw image data is transformed into human skeleton data, there arise instances in which the output files contain multiple skeletons within a single frame. This situation leads to the inadvertent inclusion of undesired and redundant targets in the skeleton data. To tackle this issue systematically, it becomes imperative to perform a targeted selection of the desired targets while effectively eliminating redundant ones from the processed skeleton data. To be precise, the confidence scores of the targeted subjects are utilized for filtering out the undesired subjects. The confidence scores are ranked and the target with the highest scores is defined as the desired subject. This rigorous selection process ensures that the output files solely consist of the intended targets, thereby mitigating interference arising from subject reflection and corridor passersby. Through meticulous data selection and refinement, the objective is to obtain a refined and reliable human skeleton dataset for subsequent analysis and research within the context of the UP-Fall dataset, recorded in the laboratory environment. Moreover, when the detected subjects are walking out of the field of view of the camera, the dataset will contain empty frames. These empty frames are redundant for the training stage and will affect the performance of the proposed framework. Therefore, these empty frames will be removed from the dataset in advance to reduce the impact on the proposed framework performance. Figure 3.3 illustrates an example with a redundant target, which includes the original RGB image, the rendered image from AlphaPose, and the skeleton-only image.

Redundant Data Reduction and Fusion

In the acquired skeleton data, a total of 17 keypoints are extracted, including five facial keypoints and twelve body keypoints. However, the facial keypoints demonstrate comparatively not as informatics as the body keypoints for the fall events. To be precise, the five facial keypoints will be fused to identify the proposed optimal fusion approach. The detailed redundant data reduction is $\mathcal{F}_r(\mathrm{x}_r, \mathrm{y}_r, \mathrm{c}_r)$ formulated as follows:

$$\mathcal{F}_r(\mathrm{x}_r, \mathrm{y}_r, \mathrm{c}_r) = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} (\mathrm{x}_i, \mathrm{y}_i, \mathrm{c}_i) \tag{3.2.1}$$

where $\mathcal{M}$ represents the number of redundant keypoints to be fused, x_i and y_i denote the x-axis and y-axis coordinates of each keypoint, respectively. c_i represents the confidence score associated with each keypoint.

By employing this redundant data reduction technique, the resulting facial keypoints retain critical fall-related information while exhibiting reduced computation time compared to the original facial keypoints. Furthermore, the performance of fall event classifications is evaluated under four occlusion scenarios in the following experiment section, which are right arm, left arm, right leg, and left leg occlusion, to analyze the impact of missing data due to occlusion and privacy protection measures. The proposed framework holds significant practical implications for fall detection under occlusion and privacy protection scenarios with enhancements of 3% to 9%, offering valuable insights into mitigating challenges faced in real-world fall detection applications.

3.2.2 Proposed DNN

Since the size of the skeleton data extracted from the video image is much less than the raw image data, the skeleton data may have a trade-off between privacy protection

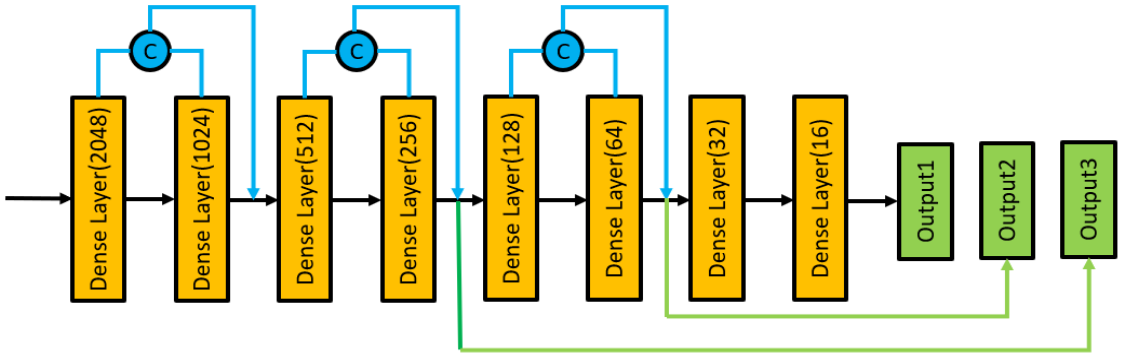


Figure 3.4: The proposed DNN architecture for mitigating information loss in the proposed data purification method.

and promising performance. To overcome the possible information loss, a new DNN model is proposed as shown in Figure 3.4. It has 8 dense layers and 3 concatenation operations between the layers in order to reuse the skeleton information. Meanwhile, 3 sub-outputs from different layers are extracted to generate the final weighted output, which is defined as the inner-ensemble. The sub-outputs from different layers have different sensitivity and precision for different activities. The weighted loss for the three outputs is utilized for updating the parameters of the proposed vanilla DNN model, the loss function is formulated as follows:

$$\mathcal{L} = -\sum_{i=1} \omega_i(\widetilde{y}_i \cdot \log(p_i^*)) \tag{3.2.2}$$

where ω_i represents the loss weight of the sub-output i. And $\widetilde{y}_i$ and p_i^* indicate the target and prediction for the sample in the ith sub-output, respectively. The final loss is the weighted combination of three sub-losses, therefore the number of sub-loss is 3, and the weight ratio of the sub-output is 1:1:2.

3.2.3 Cascaded Learning

Initial Stage: Binary Fall Classification

The aforementioned imbalanced data issue may lead negative impact on the classification performance of fall events. In order to tackle this problem, the proposed framework focuses on identifying the normal and abnormal actions at the initial stage and trends to the multiple fall events classification at the conclusive stage by proposing a vanilla DNN model. The objective of the initial stage is to filter out the normal actions from the entire dataset and thereby address the imbalanced data problem. Due to the fall events may lead to serious health and life risks for the aging population, it is crucial for the model trained in the initial stage to be capable of detecting most fall events and achieving a high recall measure. The recall measure comparisons between RF and the proposed vanilla DNN model for fall detection ability are presented in Table 3.1.

Table 3.1: Comparison between the DNN and the RF [11] at the initial stage for model selection by using recall measure.

Method	No. of Falls Detected / Ground Truth	Recall
DNN	1788 / 1803	0.96
RF [11]	1605 / 1803	0.86

According to Table 3.1, it can be observed that there are a total of 1803 ground truth instances. Due to the requirements of the medical-related task, it is essential to detect as many true positive instances as possible and minimize false negatives. The recall represents the proportion of true positive instances correctly predicted among all actual positive instances. Therefore, recall is utilized as the evaluation metric for the best model selection. The proposed DNN model achieved a recall of 96%, which is 10% higher than the recall of the pre-trained RF model on the UP-Fall dataset. This performance indicates that the proposed DNN model is more suitable for the binary classification task of distinguishing fall events from normal activities in the

initial stage. Since it is a binary classification during the initial stage, the binary cross entropy is used as the loss function, which is shown in equation 3.2.3. In order to overcome the information loss in the network layers, multiple weighted outputs are utilized from different layers as sub-outputs to reuse the information.

$$\mathcal{L}_a = -\sum_i \omega_i (y_i^g \cdot \log(p_i^p) + (1 - y_i^g) \cdot \log(1 - p_i^p)) \tag{3.2.3}$$

where ω_i represents the loss weight of ith sub-output. y_i^g and p_i^p respectively indicate the ground truth and instances prediction from the ith sub-output. $\mathcal{L}_a$ represents the loss function for the initial model. The weight ratio of the sub-outputs is 1:1:2. Moreover, the sigmoid activation function is used as the activation function of the output layers in the proposed DNN model. Since the efficacy of the proposed approach significantly depends on the performance achieved during the initial stage, particularly focusing on minimizing errors during the training process in the initial stage. To mitigate classification errors, two thresholds denoted as ξ_{fp} and ξ_{fn} are introduced to handle the occurrences of false positives and false negatives, respectively.

Conclusive Stage: Multiple Fall Events Classification

Upon completing the training of the DNN model in the initial stage, a subset of training data exclusively containing fall events is obtained. Subsequently, the second DNN model is trained using this refined subset, wherein the labels represent distinct categories of different fall events. The purpose of the second model in the conclusive stage is to address the classification of multiple fall events. The definition of loss function from the proposed conclusive stage is as follows:

$$\mathcal{L}_b = -\sum_i \omega_i (y_i^g \cdot \log(p_i^p)) \tag{3.2.4}$$

Since the classification task in the conclusive stage involves multiple fall events,

the sparse categorical cross-entropy loss function $\mathcal{L}_b$ is employed for classifying the multiple fall events, which is shown in equation 3.2.4. Moreover, weighted sub-outputs are utilized, and the final loss function in the conclusive stage consists of several sub-output losses with a weight ratio of 1:1:2. Furthermore, the sigmoid activation functions are employed in the output layers.

During the testing stage, the instances are initially evaluated by the trained binary classification model in the initial stage. The instances are then fed into the multi-class fall classification model in the conclusive stage to determine the specific fall event category if it is identified as positive in the initial stage. The proposed two-stage learning approach for multiple fall events classification is outlined in Algorithm 1 as follows:

Algorithm 1: Two-stage learning fall events classification

Input : Training data D, original labels L, binary labels L_{bin}, binary model epoch T_a^{max}, multi-class model epoch T_b^{max}

Initialize Binary model M_a,
Multi-class classification model M_b
for $T_a \leftarrow 1, 2, 3, ..., T_a^{max}$ **do**
  Based on D and L_{bin},
  Train M_a by using equation (1) ;
end
After M_a is trained:
Obtain $Q_{bin} \leftarrow M_a, D$
  // Generate binary classification map ;
Obtain $D_{multi} \leftarrow D, Q_{bin}$
  // Generate multi-class falls data ;
Obtain $L_{multi} \leftarrow L, Q_{bin}$
  // Generate multi-class falls labels ;
for $T_b \leftarrow 1, 2, 3, ..., T_b^{max}$ **do**
  Based on D_{multi} and L_{multi},
  Train M_b by using equation (2) ;
end

Output: Trained binary classification model M_a;
Trained multi-class classification model M_b

3.3 Experiments

3.3.1 Datasets

The UP-Fall dataset comprises 2 Kinect cameras capturing front-view and side-view images of the subject. The raw data from the side view camera is selected for training the model in this chapter and Chapter 4, in which the resolution of the images is 640×480 in PNG format. After the redundant empty frame reduction processing, a total of 220,660 frames were utilized for the human skeleton feature extraction. The dataset is divided into two parts for the experiments: 70% (154,462 frames) for training and 30% (66,198 frames) for testing.

The AlphaPose algorithm is utilized to extract the skeleton features with 17 keypoints of the human body, encompassing 5 facial keypoints and 12 body keypoints. The proposed redundant data reduction theory primarily concentrates on the facial keypoints. To compare with the baseline results, the five facial keypoints are considered as redundant data for reduction. The proposed simple and efficient method significantly improves both results and calculation time, which are demonstrated in the following sections. Additionally, an ablation study experiment is carried out on the left and right arms (shoulder, wrist, elbow) as well as the left and right legs (hip, knee, ankle) of the human body. This is done to examine the impact of occlusion on the experimental outcomes for various body parts. To address occlusion scenarios in this research, each classification method employs a 10-fold cross-validation approach. This ensures a robust experimental validation basis for the theory of redundant data reduction.

3.3.2 Parameter Settings

The UP-Fall dataset encompasses 5 types of falls and 7 types of normal human activities. Adam is selected as the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and the epsilon

is set as $1e^{-8}$. The initial learning rate is set as 0.0001 and the batch size is set as 1024. Moreover, 300 instances from the training set are selected as the validation set for the best model selection in the proposed two-stage cascaded learning framework. The training and testing evaluations were conducted on a workstation equipped with a CPU i7-9700k and a GPU Nvidia GTX 1660Ti with 6GB of RAM.

3.3.3 Experimental Results for Redundant Data Reduction

Table 3.2 presents the F1 score performance comparisons, which use RF, SVM, MLP and AdaBoost as the classifiers for evaluation. The detailed parameter settings are set as default as in [11]. Better performance for each label is highlighted in bold. For the detailed description of each label, Label-1 denotes forward falling using hands, Label-2 represents forward falling using knees, Label-3 indicates backwards falling, Label-4 is sideways falling and Label-5 is falling to an empty chair.

Table 3.2: Comparison using F1 score (%) based on four classifiers between the baseline and the proposed method.

Methods	Label-1	Label-2	Label-3	Label-4	Label-5
RF_Baseline [11]	76.1	73.9	66.1	79.2	84.5
RF_Proposed	**85.2**	**82.6**	**77.6**	**86.2**	**84.5**
SVM_Baseline	79.0	77.0	73.1	82.1	83.3
SVM_Proposed	**84.0**	**81.3**	**76.9**	**86.3**	**86.5**
MLP_Baseline	**65.1**	53.1	51.6	51.5	55.8
MLP_Proposed	64.2	**54.6**	**54.1**	**56.3**	**59.1**
AdaBoost_Baseline	74.8	**73.9**	70.0	**80.5**	78.5
AdaBoost_Proposed	**76.9**	72.1	**70.1**	79.6	**79.1**

It can be observed from Table 3.2 that the proposed method using RF and SVM shows higher improved performance compared to the other two classification methods, with enhancements of 3% to 9%, respectively. The classification improvements for MLP and AdaBoost generally range from 1% to 3%. Additionally, it is noteworthy that the performance of MLP for Label-1 exhibits a relatively small reduction, while the performance of Label-2 and Label-4 also slightly decreases in AdaBoost. Fur-

thermore, both Label-3 and Label-5 demonstrate relatively significant improvements for all classification methods after applying the proposed redundant data reduction method. The reason for these performance improvements is that during the different fall events, most of the physical variations of the human body are mainly localized on the trunk and limbs of the human body, rather than the facial landmarks. In this case, the facial landmarks information is relatively similar for different fall events which are redundant or adverse for the fall events classification performance. By utilizing the proposed method, the differences between the different fall events are more distinct for the model training and reduce the misclassification issue. It needs to be mentioned that different from the other 3 classifiers, AdaBoost achieves no significant improvements after applying the proposed method. The reason for this is that AdaBoost is one of the integrated learning methods and is sensitive to noisy annotations, which motivates the dataset purification method in the next chapter.

Table 3.3: Ablation study precision performance (%) under four body parts occlusion scenes with four clustering classifiers for fall classification.

Methods	Left Arm	Right Arm	Left Leg	Right Leg	Baseline
RF	92.8	92.4	92.5	93.1	93.7
SVM	84.2	84.7	84.2	85.2	87.9
MLP	72.5	71.6	72.2	72.8	72.9
AdaBoost	84.8	85.3	85.1	85.9	86.0

The experimental results confirm the efficacy of the proposed framework for enhancing the performance of multiple fall events classification. In addition, to further investigate whether the proposed theory for other body parts could affect the fall events classification results, the remaining skeleton features are manually into four parts: left arm, right arm, left leg, and right leg. Ablation study experiments are further conducted on these four parts under occlusion conditions compared with the baseline to verify if these body parts are also redundant for the fall events classification. Table 3.3 examines the impact on precision performance when different parts of the human body are occluded. Precision is selected as the performance metric as it

indicates the proportion of true positive instances that are from the predicted positive instances. The performance of the baseline is provided as the comparison group.

From Table 3.3, it is evident that the precision result of SVM decreases by approximately 2%-3% in each occlusion situation. Meanwhile, the performance of RF, MLP, and AdaBoost remains tiny decreasing, respectively. This observation can be attributed to the symmetry of the body structure, where missing a certain part of the data does not have a definitive increase in the performance but leads to a slight decrease. It could be seen that different the RF, MLP and AdaBoost, SVM achieves a slight decrease compared with the baseline. The reason for this is SVM is sensitive to the missing data, especially to the large-scale dataset (landmark occluded indicates parts of the raw data are missed). This finding is significant for further fall detection research involving missing data due to occlusion or privacy protection considerations.

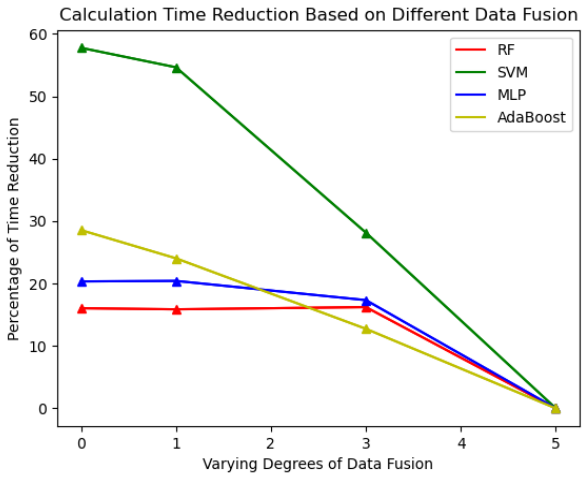


Figure 3.5: Time reduction by using different degrees of redundant data reduction, x-axis denotes the remaining processed redundant keypoints and y-axis denotes the percentage of the time reduction.

Figure 3.5 illustrates the computation time of different classifiers achieved after the proposed redundant information fusion. The x-axis represents the number of reduced redundant keypoints, while the y-axis indicates the percentage of time reduction. It can be observed that the proposed method significantly reduces the computational

time with the reduction level increasing. Especially SVM, compared to the result with baseline, the calculation time is reduced by more than 50%. And the calculation time percentages of the other classifiers compared with the raw data have been reduced by approximately 15%-30%, respectively.

3.3.4 Experimental Results for Two-stage Framework

Table 3.1 has demonstrated that the DNN outperforms RF in terms of the initial stage for selecting the fall events from the entire dataset. Therefore, in this section, the results of the proposed two-stage learning for the multiple fall events classification are presented. Furthermore, the performance of four classifiers used in are illustrated in Figure 3.6, which clearly shows that RF is the best classifier in terms of performance. In terms of this, RF performance is selected as the comparison group for verifying the efficacy of the proposed two-stage learning framework.

Table 3.4: The F1 score comparisons of multiple fall events classification on the UP-Fall dataset among the single RF, single DNN and the proposed TS-DNN.

Methods	HF	KF	BF	SF	SDF
Single RF [11]	**0.88**	0.85	0.82	0.87	0.87
Single DNN	*0.85*	*0.83*	*0.83*	*0.88*	*0.88*
TS-DNN (ξ_{fp},ξ_{fn}=0)	*0.84*	***0.85***	***0.85***	***0.89***	***0.89***

Table 3.5: The computational time cost comparisons between the proposed TS-DNN and the four selected clustering classifiers.

Methods	RF	SVM	MLP	AdaBoost	*TS-DNN*
Time cost (s) ↓	135.4	2480.7	242.5	115.5	**23.1**

Table 3.4 presents a performance comparison between the baseline approaches and the proposed two-stage DNN (TS-DNN) model. The results indicate that the single DNN achieves superior classification performance for backwards falling, sideways falling and falling to a chair. The overall multiple events classification performance is further improved with the proposed TS-DNN. It is noteworthy that no thresholds

Figure 3.6: The F1-score performance of multiple fall events classification by using the UP-Fall dataset. The RF achieves the best performance in all five fall events classification performance compared with other classification methods.

are incorporated into the TS-DNN, as shown in Table 3.6 (where $\xi_{fp} = 0$, $\xi_{fn} = 0$). Moreover, since the fall classification system requires high accuracy and fast response time, the computational time cost comparisons between the proposed TS-DNN and the four clustering classifiers are provided in Table 3.5. It could be seen that the computational time cost for the proposed TS-DNN is much faster than the other baselines, which indicates that our model is better than the other classifiers.

Table 3.6: The performance comparison with the single RF in F1 score evaluation metric for UP-Fall dataset multiple fall events classification by using different settings.

Methods	HF	KF	BF	SF	SDF
Single RF [11]	**0.88**	0.85	0.82	0.87	0.87
TS-DNN (ξ_{fp},ξ_{fn}=0.3)	*0.72*	*0.76*	*0.63*	*0.74*	*0.82*
TS-DNN (ξ_{fp},ξ_{fn}=0.1)	0.84	0.86	0.84	0.87	0.88
TS-DNN (ξ_{fp},ξ_{fn}=0.05)	*0.85*	*0.86*	*0.85*	*0.88*	*0.89*
TS-DNN (ξ_{fp},ξ_{fn}=0.03,0.02)	*0.85*	***0.86***	***0.85***	***0.89***	***0.89***
TS-DNN (ξ_{fp},ξ_{fn}=0)	*0.84*	*0.85*	*0.85*	*0.89*	*0.89*

Nevertheless, the hand-falling (HF) performance in the TS-DNN is lower compared

to that of the single model. This discrepancy may be attributed to misclassifications by the binary fall classification model in the initial stage. Consequently, two thresholds (ξ_{fp} and ξ_{fn}) will be introduced to re-classify the misclassified fall and no-fall categories in the initial stage, to enhance the performance in the conclusive stage.

Based on Table 3.6, the TS-DNN with the proposed gating parameters ($\xi_{fp} = 0.03$ and $\xi_{fn} = 0.02$) exhibits improved performance compared to the TS-DNN without thresholds and the single RF. The proposed TS-DNN with gating parameters achieves better performance for both falling with hands and falling with knees, which refines the human action discriminating ability in the initial stage. RF achieves the best performance in hand falling (HF) at approximately 88% but the proposed TS-DNN is 85%, the assumption of this is that hand falling is similar to knee falling, which will lead to confusion to the model for the action classification. For the backward falling, sideway falling and sit-down falling, the proposed TS-DNN achieves 2-3% improvements compared with the RF method. By considering the F1 score, the proposed two-stage framework demonstrates the highest accuracy in multiple fall events classification. The introduction of the DNN-based binary classifier in the initial stage eliminates the data imbalanced problem and leads to better classification performance of the multi-class classifier in the conclusive stage. Additionally, the gating parameters in the initial stage contribute to improving the final classification performance in the conclusive stage.

3.4 Summary

In summary, this chapter achieves the contribution to improving the human fall events classification and enhancing the robustness by proposing two components: redundant data reduction and the two-stage learning framework. For both of these two parts, the initial RGN data were first sent for the skeleton feature extraction to preserve privacy, thereby avoiding the impact from the dynamic illumination and reducing

the computing time. The redundant empty frames and redundant targets were then removed according to the confidence scores. The redundant data reduction technique was exploited to remove the unnecessary keypoints among the skeleton sequences. In order to avoid the imbalanced issue which typically exists in the fall events dataset, the proposed two-stage learning framework was designed for distinguishing the fall events from the entire dataset in the initial stage and trend to classify different fall events in the conclusive stage. Additionally, the gating parameter with the proposed structure was exploited for further improving the classification performance by controlling the discriminating ability in the initial stage. The redundant data reduction evaluation in Section 3.3.3 showed the contributions of the proposed method, as well as demonstrated the ablation study for occlusion handling of the other four body parts. Evaluations on the UP-Fall dataset were further provided in Section 3.3.4, which confirms the proposed method achieved improved performance compared to other approaches.

Although the proposed fall events classification method has achieved improved performance, the quality of the annotations may arise a negative impact on the classification performance with the data-driven algorithms in the real-world environment, since the UP-Fall dataset is manually annotated by multiple annotators. The next chapter focuses on increasing the quality of the annotation by purifying the corrupted dataset, as well as exploiting the potential of the corrupted instances by the robust framework for better fall events classification performance.

Chapter 4

PRIVACY MITIGATING DATA PURIFICATION AND JOINT COOPERATIVE TRAINING WITH NOISY LABELS FOR HUMAN FALL EVENTS CLASSIFICATION

4.1 Introduction

This chapter is mainly aimed at addressing the noisy annotation issue for the fall events classification. As aforementioned in the previous chapters, different types of fall events will lead to various injuries in human bodies. However, conventional video-based fall detection only focuses on whether the fall event has occurred but ignores different types of fall events, which may cause different injuries to elderly people [12, 47]. Moreover, DNN models have exhibited impressive performance in recent years [96, 97]. Their success depends on high-quality labels, and a massive amount of data [91]. However, obtaining high-quality data annotation is expensive and time-consuming because the labels of the dataset are all manually annotated. Therefore, some annotators choose to use the non-manufactured, semi-manufactured or online

survey methods to improve the data annotation efficiency [98–100], which leads to incorrect annotations due to the differences in cognitive definitions of annotators and model performance. Therefore, a robust fall classification system which could address the noisy annotations and improve the classification performance is needed for better healthcare.

Handling the corrupted instances plays a key element in the model performance. Existing work in [101] has attempted to perform a two-stage measurement filtering theory to address the noisy annotation issue. This noisy instance filtering approach typically comprises the confidence score ranking and the probabilistic thresholds to estimate noise, which can be efficiently applied to the fall events classification task. However, all the benchmarks applied for this method are large-scale image datasets for classification tasks. Due to privacy protection requirements for the medical action task, the human skeleton is the training data for feeding into the model rather than the image data, which has much lower information density compared to image data. On the other hand, recent noisy label learning approaches [5–7] shows the small loss algorithm has been demonstrated to discriminate the noisy instance during the training stage rather than utilizing the two-stage pipeline. However, the direction of the model update mainly depends on the selected clean instances from the single peer network module, which is not robust enough due to the overfitting problem, especially with the deep noise rate.

This chapter aims to mitigate the noisy annotations issue on the extracted skeleton data from the UP-Fall dataset by presenting a noise managing system which is mainly divided into two parts: a cascaded noisy dataset purification method and a noisy label learning framework with trinity networks (JoCoT). The proposed cascaded noisy dataset purification algorithm falls into the aforementioned confident learning to filter out the corrupted instances by three steps, which are instance counting, confidence score ranking and noisy data pruning. Four different pruning methods are

provided for the noisy instance cleaning. They are based on the principle of a joint distribution probability density function and focus on label quality by characterizing and identifying noisy labels. JoCoT is proposed to fully exploit the potential of the noisy instances and enhance the robustness of the framework, especially with deep noise rates. Specifically, it trains a trinity network that includes two teacher modules and one student module. The consensus outputs of the two teacher modules are fed into the student module to guide the clean instances mining. The peer network is used with the proposed structure for selecting clean instances during the training process. Moreover, both the co-regularized and contrastive learning with joint loss function are applied for keeping the peer networks converged, which is able to enhance the model performance. The main contributions of this chapter are listed as follows:

1. The data purification algorithm with four different pruning methods is applied to human skeleton data from the UP-Fall dataset to clean the corrupted annotations.

2. Noisy label learning with trinity networks (JoCoT) is developed to improve the robustness and the performance of human fall events classification with different noise categories and rates via the consensus-based noisy instances selection method.

3. Empirical results demonstrate the efficiency of the proposed noise managing system is superior to many state-of-the-art approaches. Additionally, sufficient verification experiments with different estimated noises are conducted for further analysis and discussion.

In this chapter, since the applied UP-Fall dataset is manually annotated and exists the noisy label issue. Therefore, the dataset purification algorithm which is called clean lab is introduced to remove the corrupted instances and improve the fall events classification performance. In addition, to fully exploit the knowledge

potential of the noisy instances rather than pruning them, a learning with noisy labels algorithm called JoCoT is developed for human fall events classification and verified with different noise distributions and rates. This chapter targets to fulfill the third and fourth objectives of this thesis, which are the enhanced data purification for human fall events classification based on the vanilla DNN model presented in [28], and the enhanced learning with noisy label algorithm using privacy-preserved skeletal data for human fall events classification which is submitted to ACM Transaction on Computing for Health [29]. The rest of this chapter is organised as follows: Section 4.2 demonstrates the corrupted dataset annotation purification method. Section 4.3 describes the proposed training process with noisy labels. Section 4.4 performs the experimental results, analysis and discussion. Furthermore, the summary of this chapter is presented in Section 4.5.

4.2 Formulation of Data Purification

4.2.1 Overview

The overview of the proposed dataset purification theory for fall events classification is shown in Figure 4.1, which primarily consists of two steps: counting and pruning with ranking. Firstly, both the entire corrupted dataset and corrupted annotations are fed into the vanilla-DNN model to generate the prediction coarse $p_f(x)$ for each instance, the structure of the vanilla-DNN model is set as the same as the model applied in section 3.2.2.

The determined possibility threshold τ_f is generated by $p_f(x)$ for determining the ground truth labels, which is the expected self-confidence score for each category. After that, the confidence joint counting matrix between the predicted labels and true labels is generated and recalibrated for counting the noisy annotation by proposing the confidence joint probability distribution matrix, which is precisely formulated in

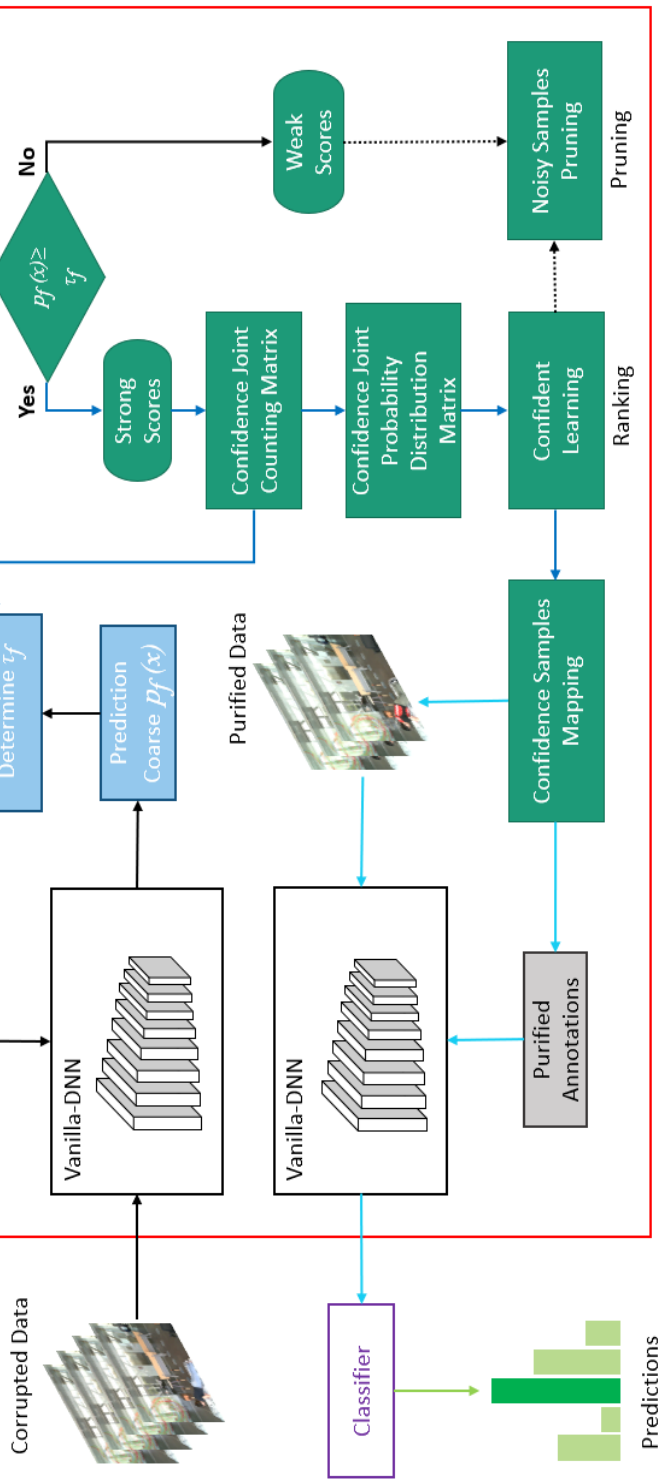


Figure 4.1: Overview of the proposed cascaded data purification for human fall events classification. The same vanilla-DNN model as illustrated in section 3.2.2 is utilized as the backbone in the training stage. The design in the black dashed line is indicated to remove the noisy instances and the main contribution parts are annotated in the red box.

equation 4.2.4. Then four different noisy annotation ranking methods are utilized for dataset cleaning. The data with weak scores identified as noisy samples are pruned. On the other hand, the purified data and annotations are mapped from the confidence map. Finally, the clean dataset is fed into the same vanilla DNN model again for retraining, then obtain the conclusive classification performance for human multiple fall events.

4.2.2 Confident Learning for Noisy Label Pruning

Assume corrupted training dataset $\mathcal{D} = \{\boldsymbol{x}, \widetilde{y}\}_M^N \in (\mathbb{R}, \{1, 2, ..., M\})^N$, which denotes the dataset contains N samples in M categories with noisy label $\widetilde{y}$ for samples $\boldsymbol{x}$, along with the correct labels are given as y^*. The possibility threshold is defined as:

$$\tau_f = \frac{1}{|\mathcal{D}_{\widetilde{y}=f}|} \sum_{\boldsymbol{x} \in \mathcal{D}_{\widetilde{y}=f}} \widehat{p}_f(\boldsymbol{x}) \tag{4.2.1}$$

where τ_f is the possibility threshold for all samples labelled as $\widetilde{y} = f$. If the predicted probability has $\widehat{p}_f(\boldsymbol{x}) < \tau_f$, then it will be suspected as a wrong annotation. The estimated dataset $\widehat{\mathcal{D}}_{\widetilde{y}=d,y^*=f}$ is defined as follows:

$$\widehat{\mathcal{D}}_{\widetilde{y}=d,y^*=f} = \left\{ \boldsymbol{x} \in \mathcal{D}_{\widetilde{y}=d} : \widehat{p}_f(\widetilde{y} = f; \boldsymbol{x}) \geq \tau_f \right\} \tag{4.2.2}$$

$$C_{\widetilde{y},y^*}[d][f] := \left| \widehat{\mathcal{D}}_{\widetilde{y}=d,y^*=f} \right| \tag{4.2.3}$$

where each cell of the unnormalized confident joint counting matrix $C_{\widetilde{y},y^*}[d][f]$ is the number of samples with the noisy label is d but correct label is f. $|\mathcal{D}_{\widetilde{y}=d}|$ denotes the sample amounts with $\widetilde{y} = d$.

$$\widehat{J}_{\widetilde{y}=d,y^*=f} = \frac{\frac{C_{\widetilde{y}=d,y^*=f}}{\sum_{f\in[M]} C_{\widetilde{y}=d,y^*=f}} \cdot |\mathcal{D}_{\widetilde{y}=d}|}{\sum\limits_{d\in[M],f\in[M]} (\frac{C_{\widetilde{y}=d,y^*=f}}{\sum_{f\in[m]} C_{\widetilde{y}=d,y^*=f}} \cdot |\mathcal{D}_{\widetilde{y}=d}|)} \tag{4.2.4}$$

where the instance amount of $C_{\widetilde{y},y^*}$ is not equal to the true instance amount due to the limitation from τ_f, it needs to be refined by $\mathcal{D}_{\widetilde{y}=d}$. After the joint confident probability distribution matrix $\widehat{J}_{\widetilde{y}=d,y^*=f}$ is obtained. The noisy labels are pruned by the following methods:

Confusion: Estimate the i-th noisy labels as $\widetilde{y}_i \neq \text{argmax}_{f\in[M]}\widehat{p}_f(\widetilde{y}=f;\boldsymbol{x}_i)$, $\forall \boldsymbol{x}_i \in \mathcal{D}$, which indicates to be selected by off-diagonal elements of confusion matrix.

PBC: For each class $d \in [M]$, select $N \cdot \sum_{f\in[M]:f\neq d}(\widehat{J}_{\widetilde{y}=d,y^*=f}[d])$ samples for filtering with lowest confidence are identified as noisy labels.

PBNR: $N \cdot \widehat{J}_{\widetilde{y}=d,y^*=f}$ samples in off-diagonal are selected and has $\boldsymbol{x} \in \mathcal{D}_{\widetilde{y}=d}$.

C+NR: Align the previous PBC and PBNR with the operation 'and'.

4.3 Noisy Labels Learning with Trinity Networks

Table 4.1: Comparisons between the other algorithms and the proposed JoCoT.

	Co-teaching [5]	Co-teaching+ [6]	JoCoR [7]	*JoCoT*
Small Loss	✓	✓	✓	✓
Cross Update	✓	✓	✗	✓
Joint Training	✗	✗	✓	✓
Agreement	✗	✗	✓	✓
Disagreement	✗	✓	✗	✗
Consensus	✗	✗	✗	✓

In this section, detailed explanations of the proposed JoCoT are presented. As shown in Figure 4.2, both teacher modules instruct the student module to mine the reliable and clean instances from the corrupted dataset in the training stage. Each mini-batch of the corrupted instances are simultaneously fed into the teacher modules and the prediction indexes for the noisy annotations are obtained. According to the predictions from the teacher modules, a consensus-based data selection strategy is

applied for choosing reliable and clean data, then fed the selected data into the student module for training. The consensus algorithm is updated in both teacher modules after each iteration. The comparisons of state-of-the-art approaches with the proposed JoCoT are listed in Table 4.1.

4.3.1 Preliminaries

In order to verify the proposed algorithm JoCoT, the multi-class dataset is defined as $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, N indicates the number of instances in the dataset. For the data training, $\boldsymbol{x}_i$ denotes the tensor quantity of i-th instance and the corrupted annotation $y_i \in \{1, 2, ..., M\}$, M indicates the classes of human activities. Both of the teacher modules have the peer networks, which are denoted as $F(\boldsymbol{x}, \boldsymbol{\Theta_1})$, $F(\boldsymbol{x}, \boldsymbol{\Theta_2})$ and $G(\boldsymbol{x}, \boldsymbol{\Phi_1})$, $G(\boldsymbol{x}, \boldsymbol{\Phi_2})$, respectively. For the instance $\boldsymbol{x}$, $\forall \boldsymbol{x} \in \mathcal{D}_n$, $\mathcal{D}_n$ is the mini-batch dataset which is fetched from $\mathcal{D}$. Moreover, the prediction probabilities of the instances $\boldsymbol{x}_i$ from the two teacher modules are denoted as $\boldsymbol{p}_1 = \left[p_1^1, p_1^2, \ldots, p_1^M\right]$, $\boldsymbol{p}_2 = \left[p_2^1, p_2^2, \ldots, p_2^M\right]$ and $\boldsymbol{q}_1 = \left[q_1^1, q_1^2, \ldots, q_1^M\right]$, $\boldsymbol{q}_2 = \left[q_2^1, q_2^2, \ldots, q_2^M\right]$, which could also be considered as the "softmax" layer outputs from the network parameters $\boldsymbol{\Theta_1}$, $\boldsymbol{\Theta_2}$, $\boldsymbol{\Phi_1}$, $\boldsymbol{\Phi_2}$, respectively.

For the proposed approach, both teacher modules could individually predict the annotations but train the network to update the parameters simultaneously with the peer paradigm. The cross-entropy loss $\mathcal{L}_1$ which applied in the teacher module is as follows:

$$\mathcal{L}_1(\boldsymbol{x}_i, y_i) = \mathcal{L}_s(\boldsymbol{x}_i, y_i) = \mathcal{L}_{s_1}(\boldsymbol{x}_i, y_i) + \mathcal{L}_{s_2}(\boldsymbol{x}_i, y_i) \tag{4.3.1}$$

where $\mathcal{L}_{s_1}$ and $\mathcal{L}_{s2}$ indicate the sub-losses from the teacher module. The classifier function f maps the feature $\boldsymbol{x}_i$ to the label space and is defined as $f(\cdot) : \mathcal{X} \to \mathbb{R}^M$, $\boldsymbol{e}_{y_i}$ is a one-hot vector which equals to 1 if the predicted label is the same as y_i, otherwise equals to 0. For the second teacher module, there exists the joint-training

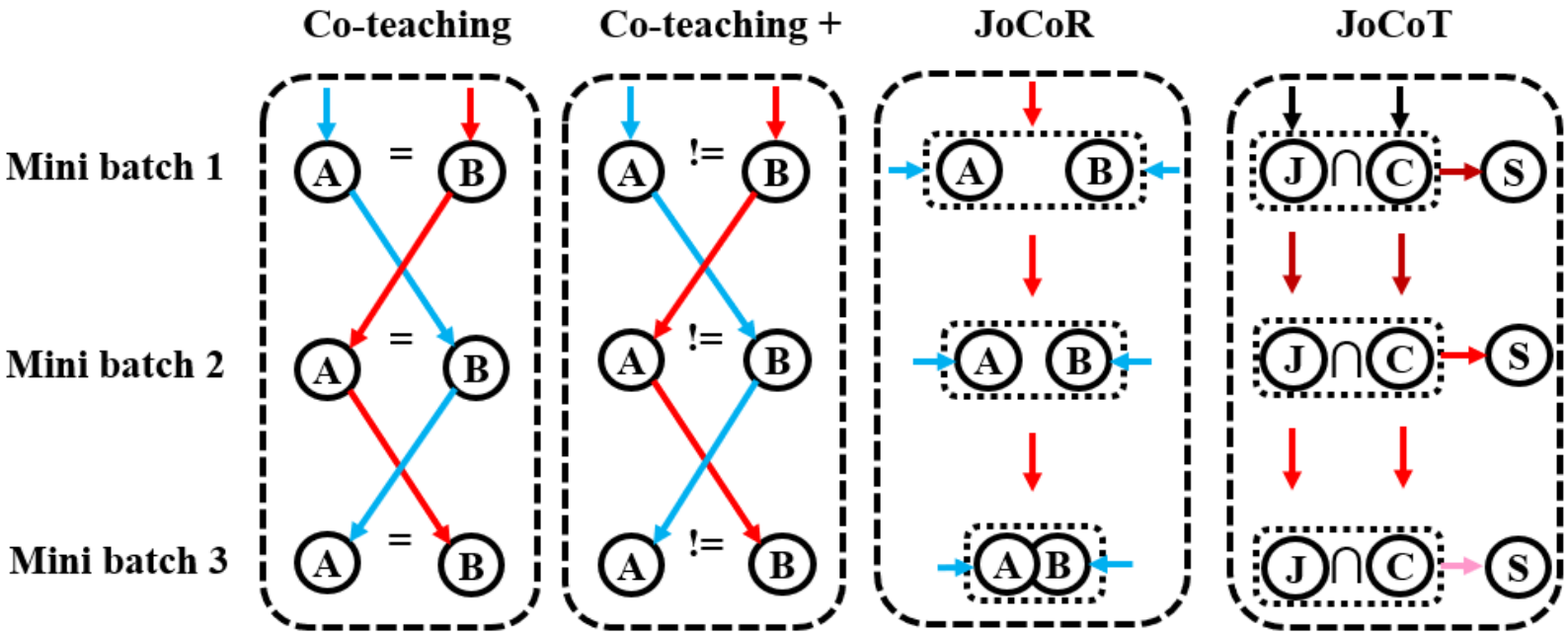


Figure 4.2: The comparisons of the noisy label learning algorithms between Co-teaching [5], Co-teaching+ [6], JoCoR [7] and the proposed JoCoT. The selected errors are from the prediction errors, which are predicted by the peer networks A and B in Co-teaching, Co-teaching+ and JoCoR. For the JoCoT, since there exist both Co-teaching and JoCoR modules inside, therefore the errors are assumed from those two modules which are J and C. First panel: Co-teaching maintains two networks (A&B). The parameters of the two networks are cross-updated with the agreement (=). Second panel: Co-teaching+ also maintains two networks (A&B). Cross-updated is also applied in the peer networks but using disagreement (!=). Third panel: JoCoR maintains two networks (A&B) and uses the joint loss function which contains both the contrastive loss and classification loss to make the predictions closer to each other. Fourth panel: JoCoT contains two teacher modules (J&C) which denote the JoCoR and Co-teaching respectively, and apply the consensus ($\cap$) of the modules predictions to train the student module (S) to make the predictions of the corrupted data closer to the ground truth labels. (Best viewed in colored version)

stage. The precise loss function $\mathcal{L}_2$ for $\mathbf{x}_i$ is defined as followed:

$$\mathcal{L}_2(\boldsymbol{x}_i, y_i) = (1-\lambda)\mathcal{L}_1\left(\boldsymbol{x}_i, y_i\right) + \lambda\mathcal{L}_c(\boldsymbol{x}_i) \tag{4.3.2}$$

where $\mathcal{L}_s$ and $\mathcal{L}_c$ denote the supervised loss and contrastive loss functions in the joint loss function. Moreover, λ is the weight parameter. The range of λ is between 0.05 and 0.95, it is empirically selected for different noise rates. The detailed explanations of the $\mathcal{L}_s$ and $\mathcal{L}_c$ are presented in the following subsections.

4.3.2 Classification Loss

Since the algorithm is proposed for addressing the multi-classification noisy label learning, the cross-entropy loss is chosen as the supervised classification loss function for all the peer networks from the teacher modules. It is widely used in multi-class classification tasks for calculating the loss value between the annotations and the predictions. Since the small-loss selection could help the network filter the noisy labels, the cross-update theory [5] is applied to reduce the annotation errors. The complete supervised loss function $\mathcal{L}_s$ is defined as follows:

$$\begin{aligned}\mathcal{L}_s(\boldsymbol{x}_i, y_i) &= \mathcal{L}_{s1}(\boldsymbol{x}_i, y_i) + \mathcal{L}_{s2}(\boldsymbol{x}_i, y_i) \\ &= -\textstyle\sum_{i=1}^{N}\sum_{m=1}^{M} y_i \log(q_1^m(\boldsymbol{x}_i)) \\ &- \textstyle\sum_{i=1}^{N}\sum_{m=1}^{M} y_i \log(q_2^m(\boldsymbol{x}_i))\end{aligned} \tag{4.3.3}$$

where $q_1^m(\boldsymbol{x}_i)$ and $q_2^m(\boldsymbol{x}_i)$ denote the predicted probabilities of i-th instance from the peer networks for m-th label.

4.3.3 Contrastive Loss

According to the previous work [5], the DNN model will tend to learn the simple and clean instances at the beginning of the training process. The peer networks will reach

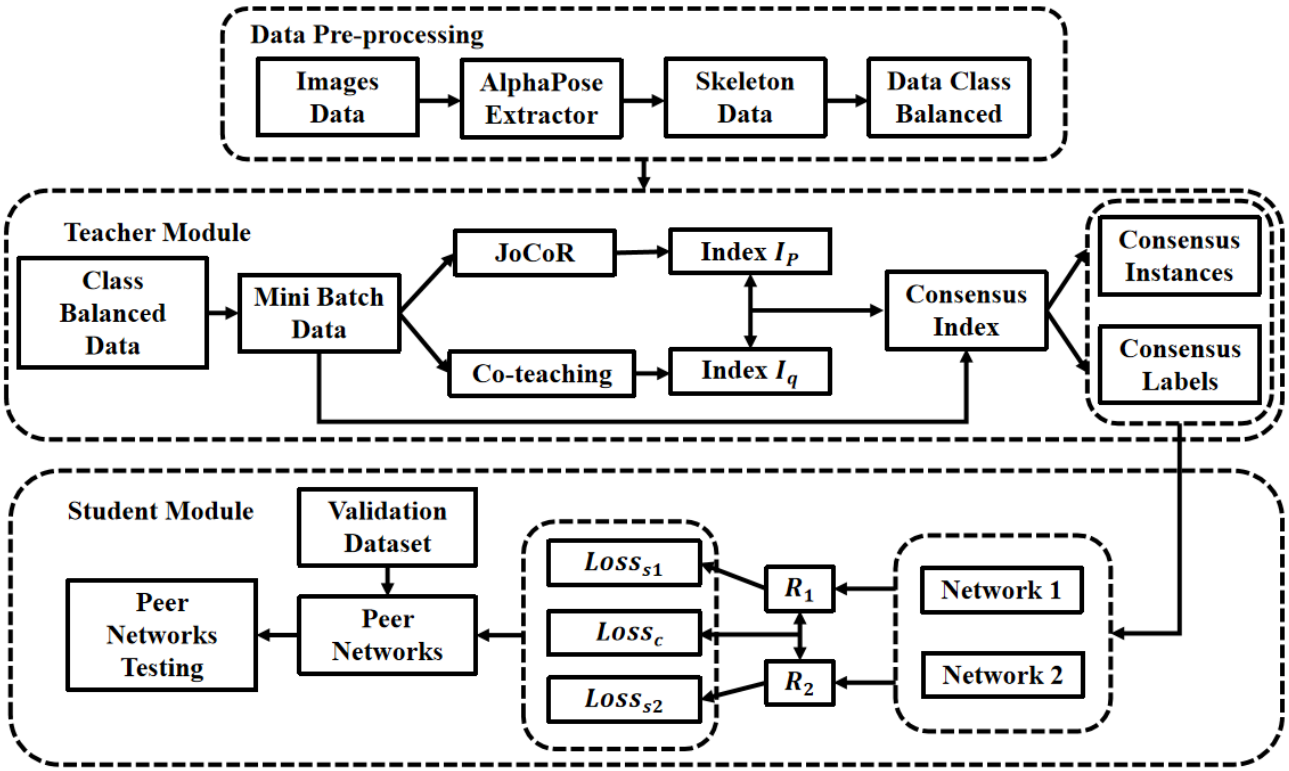


Figure 4.3: The schematic of the proposed JoCoT. It is divided into three modules, data pre-processing, teacher module and student module. Firstly, the original RGB image data will be fed into the AlphaPose extractor [8] to obtain the human skeleton data, each sample contains 17 body landmarks. Since the original dataset is imbalanced, the skeleton data is re-scaled to make sure each class of the activities contains the same number of samples. Secondly, the skeleton data will be fed into the teacher modules, which have JoCoR and Co-teaching for predicting the corrupted instances and applying the proposed consensus method to obtain the consensus 'clean' instances and labels. Finally, those instances and labels will be fed into the student module for the network parameters updating. The joint loss function is applied between the predictions R_1 and R_2 from the peer networks to mine the noisy instances, which is precisely described in equation 4.3.2. The validation dataset is also supported for guiding the network updating towards the correct direction.

a consensus on most of the data but not on the corrupted data. In order to guide the models to find more clean and reliable data, thereby achieving better generalization ability and performance. Co-regularization is constructed as the contrastive term in the loss function for the second teacher module, which could help the classifiers to maximize the agreement. One of the specific forms of Jensen-Shannon (JS), Kullback-Leibler (KL) divergence, is utilized to calculate the loss value between the predictions from the peer networks. The symmetric KL divergence based contrastive loss function $\mathcal{L}_c$ is specifically defined as follows:

$$\mathcal{L}_c = D_{KL}(\boldsymbol{q}_1 \,||\, \boldsymbol{q}_2) + D_{KL}(\boldsymbol{q}_2 \,||\, \boldsymbol{q}_1) \tag{4.3.4}$$

The detailed definition of sub terms $D_{KL}(\boldsymbol{q}_1 \,||\, \boldsymbol{q}_2)$ and $D_{KL}(\boldsymbol{q}_2 \,||\, \boldsymbol{q}_1)$ in $\mathcal{L}_c$ are as:

$$D_{KL}(\boldsymbol{q}_1 \,||\, \boldsymbol{q}_2) = \sum_{i=1}^{N}\sum_{m=1}^{M} q_1^m(\boldsymbol{x}_i) \log \frac{q_1^m(\boldsymbol{x}_i)}{q_2^m(\boldsymbol{x}_i)} \tag{4.3.5}$$

$$D_{KL}(\boldsymbol{q}_2 \,||\, \boldsymbol{q}_1) = \sum_{i=1}^{N}\sum_{m=1}^{M} q_2^m(\boldsymbol{x}_i) \log \frac{q_2^m(\boldsymbol{x}_i)}{q_1^m(\boldsymbol{x}_i)} \tag{4.3.6}$$

where $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ indicate the prediction probabilities of the peer networks. Hence, the negative memorization effects of noisy labels could be mitigated, and the classification performance could be further improved.

4.3.4 Small Loss Selection

The peer networks are more likely to reach a consensus on most clean instances but will have different predictions on the noisy data. According to the previous work focusing on the LNL algorithms [5, 102], the instances with small loss are more likely to be clean. Therefore, the network model will be improved if the training data are with small-loss values. With the same decaying settings in JoCoR and Co-teaching, a parameter factor $R(T_k)$ is defined to determine the proportion of the instances that

should be selected related to the noise rate τ. After each iteration, the small loss values will be sorted for selecting the mini-batch data. As mentioned in Chapter 2, the DNN model trends to learn the easy and clean instances at the beginning of the training process, but overfit to the noisy instances with the prolong of the training process, i.e., it has the best noisy instances filtering out ability at the beginning of the training process. Therefore, in order to maximally keep the clean instances, $R(T_k)$ will be gradually reduced to $1-\tau$ with the epochs increasing, i.e., $R(T_k)$ is the largest at the beginning and then gradually decrease. Hence, clean instances could be kept and noisy instances will be dropped before the network overfits to the noisy instances. The calculation of the remember rate $R(T_k)$ is as follows.

$$R(T_k) = 1 - \min\left\{\frac{T_k}{T}\tau, \tau\right\} \tag{4.3.7}$$

where T_k and T indicate current epochs and total decay epoch numbers in the training stage. It could be observed that the $R(T_k)$ will gradually decrease with the epochs increasing to keep the network trained with 'clean' instances during the current training process. The detailed definitions of the instances selection algorithm in teacher modules are as follows:

$$\bar{\mathcal{D}}_p = \text{argmin}_{\mathcal{D}'_n:|\mathcal{D}'_n|\geq R(t)|\mathcal{D}_n|}\mathcal{L}_1(\mathcal{D}'_n) \tag{4.3.8}$$

$$\bar{\mathcal{D}}_q = \text{argmin}_{\mathcal{D}'_n:|\mathcal{D}'_n|\geq R(t)|\mathcal{D}_n|}\mathcal{L}_2(\mathcal{D}'_n) \tag{4.3.9}$$

where $\bar{\mathcal{D}}_p$ and $\bar{\mathcal{D}}_q$ indicate the small loss selection instances from the sorted mini-batch data $\mathcal{D}'_n$ in the teacher modules. In this case, the influences from the noisy data will be reduced.

4.3.5 Consensus-Based Data Selection

The framework of the proposed JoCoT contains three modules, including two teacher modules and one student module. Intuitively, the instances jointly selected by the two teacher modules are more reliable and clean-confident than those selected by a single teacher module. The teacher modules in the framework not only stabilize the selection process but also eliminate the noisy data. To be precise, the two teacher modules will be trained in parallel, and the peer networks will select the clean instances in each mini-batch. A reasonable method is designed to select reliable clean instances, which uses the consensus-based decision from the predictions of the teacher modules. Finally, the reliable clean instances will be fed to the student module for guiding the parameter updating.

The framework of JoCoT is illustrated in Figure 4.3. The teacher modules which contain the peer networks are trained at the same time. After that, the consensus decision between the predictions from the teacher modules will guide the student module to classify the noisy data. A mini-batch of the input instances will be separately and simultaneously trained into the teacher modules. Each pair of networks in each teacher module will obtain a prediction index for the input data. Therefore, there will be four output indexes from the teacher modules. The consensus decision contains two steps which are called inner consensus and outer consensus. Inner consensus indicates that the consensus decision between the peer networks outputs from the same teacher module and outer consensus indicates the consensus decision between different teacher modules. The consensus-based data selection will be repeated in each data iteration for the teacher modules until the training stage is finished. After that, the consensus-based clean instances will be fed into the student module for model training. The detailed schematic of the proposed consensus method is shown in Figure 4.4. This multi-step consensus data selection strategy could guide the networks to utilize more robust and reliable clean instances for parameter updating. The original

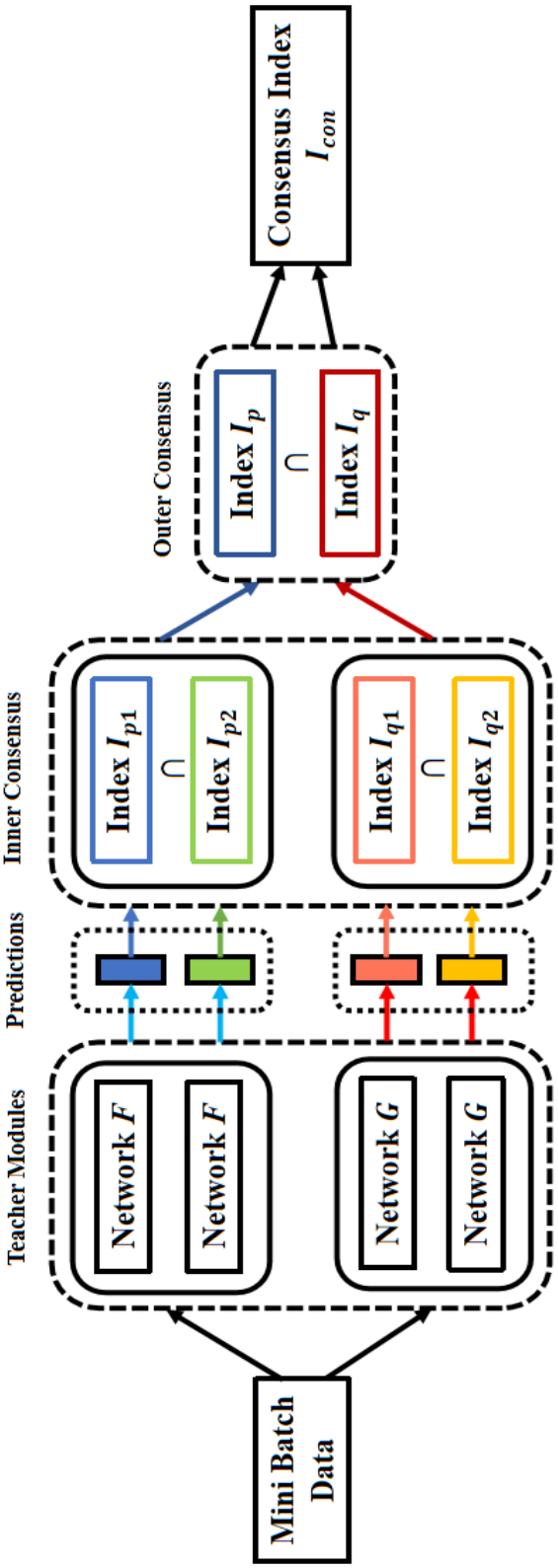


Figure 4.4: The schematic of the proposed consensus method. Networks F indicate the peer networks of JoCoR [7] and Networks G indicate the peer networks of Co-teaching [5]. According to the four different colors predictions of the networks, the inner consensus will be obtained. Finally, the outer consensus data for both of the inner consensus indexes, I_p for JoCoR and I_q for Co-teaching will be obtained as I_{con}. This schematic will be repeated in each iteration until the end of the teacher modules training. (Best viewed in colored version)

data will be used as the testing set for both teacher modules and student module. The testing sets for the three modules are the same, and the student module outputs are reported as the final results.

$$\begin{aligned} I_{con} &= \left(I_{\boldsymbol{p}_1} \cap I_{\boldsymbol{p}_2}\right) \cap \left(I_{\boldsymbol{q}_1} \cap I_{\boldsymbol{q}_2}\right) \\ &= I_{\boldsymbol{p}} \cap I_{\boldsymbol{q}} \end{aligned} \tag{4.3.10}$$

where $I_{\boldsymbol{p}_1}, I_{\boldsymbol{p}_2}$ and $I_{\boldsymbol{q}_1}, I_{\boldsymbol{q}_2}$ indicate the output indexes from the teacher modules. I_{con} demonstrates each mini-batch final consensus-based selected data index. The calculation order of the equation will firstly be the inner brackets union (inner consensus) and then outside the brackets union (outside consensus). Afterwards, the selected instances and labels corresponding to the I_{con} will be fed into the student module for training. The detailed process of JoCoT is shown in Algorithm 2.

Algorithm 2: JoCoT Teacher Modules Training

input : Epoch $\mathcal{T}_{max}$, learning rate λ, iteration $\mathcal{N}_{max}$, sub networks F, G with randomly initialization parameters $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2$ and $\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2$;

for $t = 1, 2, 3, ..., \mathcal{T}_{max}$ **do**
  Shuffle training set $\mathcal{D}$;
  for $n = 1, 2, 3, ..., \mathcal{N}_{max}$ **do**
    Fetch mini-batch $\mathcal{D}_n$ from $\mathcal{D}$;
    $\boldsymbol{p}_1, \boldsymbol{p}_2 = F(\boldsymbol{x}, \boldsymbol{\Theta_1}), F(\boldsymbol{x}, \boldsymbol{\Theta_2})\ \forall \boldsymbol{x} \in \mathcal{D}_n$;
    Calculate loss value $\mathcal{L}_1$ using $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$;
    $\boldsymbol{q}_1, \boldsymbol{q}_2 = G(\boldsymbol{x}, \boldsymbol{\Phi_1}), G(\boldsymbol{x}, \boldsymbol{\Phi_2})\ \forall \boldsymbol{x} \in \mathcal{D}_n$;
    Calculate loss value $\mathcal{L}_2$ using $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$;
    Obtain the small-loss selections $\bar{\mathcal{D}}_p$ and $\bar{\mathcal{D}}_q$ by equation 4.3.5 and 4.3.6 from $\mathcal{D}_n$;
    Update the peer network parameters $\boldsymbol{\Theta_1}, \boldsymbol{\Theta_2}$ and $\boldsymbol{\Phi_1}, \boldsymbol{\Phi_2}$;
    Obtain the inner consensused data index sets $\boldsymbol{I}_p$, $\boldsymbol{I}_q$;
  end
  Calculate $\boldsymbol{I}_p \cap \boldsymbol{I}_q$ by equation 4.3.10;
  Obtain the outer consensused data index set $\boldsymbol{I}_{con}$;
  Joint both $\boldsymbol{I}_{con}$ and mini-batch data $\mathcal{D}_n$ for the consensused-based clean instance set $\hat{\mathcal{D}}_n$;
  Update $R(t)$ by equation 4.3.7
end

output: Consensused-based clean instance set $\hat{\mathcal{D}}_n$

4.4 Experiments and Performance Analysis

4.4.1 Dataset for Data Purification

Table 4.2: Abbreviations of five fall events in the proposed method

HF	Hands forward Falling
KF	Knees forward Falling
BF	Backward Falling
SF	Sideways Falling
SDF	Sit Down Falling

In the UP-Fall dataset, there are 5 fall events and 6 normal daily activities in both 2 perspective cameras. The image data from a sideway camera is selected as the training data, which is named as CAM1. The work in [11] is the baseline, several clustering classifiers were utilized to perform the fall detection performance. To avoid any confusion, the dataset is divided into 12 classes with one more class called unknown activity. By using Alphapose [8], the human skeleton will be obtained which contains 17 joint points. Each feature point contains 3 dimensions which are joint scores and 2-D coordinates. Therefore, 51 attributes are used as features for each image for the model to classify the five fall events. Table 4.2 shows the details of the five fall actions in the UP-Fall dataset. It is highlighted that different from the other fall events, the subjects are facing the camera when they are recording the sideways fall. In the other 4 fall events, subjects are side-way in the camera field of view.

4.4.2 Data Purification hypermeter settings

Figure 4.5: An example of forward falling using hands. (a) is the original RGB image, (b) is the skeleton data extracted from AlphaPose [8].

The ratio of the number of falls and no-fall in the data set is approximately 3:97. After the preprocessing step, in total there are 220,660 groups of skeleton data, 154,462 in training and 66,198 testing sets, respectively. Besides, parts of the training set are set as the validation set in order to prevent the overfitting issue. The batch size was set as 128 and the initial learning rate is 0.0001. The overall training epoch was set as 300 and the random seed is 42. Adam was selected as the optimizer and the monument was set as 0.9. Figure 4.5 shows the skeleton data example which is extracted by using AlphaPose. The experiments are conducted on a workstation with 4 GeForce GTX 1080Ti GPUs, and 16GB of RAM. The proposed framework is implemented based on Keras.

4.4.3 Data Purification Results

It can be seen in Figure 4.6 that among the four classification methods, RF has the highest performance but the DNN also achieves competitive performance. According to Table 4.3, MLP has the lowest inference time with the lowest classification performance. Meanwhile, the proposed CD-DNN has the second shorter inference time and achieves competitive performance compared with the best classification performance of all falls events.

The RF earns first place in all events classification, however, from Table 4.3, the inference time of the proposed CD-DNN (1.23 seconds) is shorter than RF (2.13 seconds), which has 73.17 % improvement in the inference time. Therefore, the RF- and DNN-based fall detection algorithms have trade-off between the classification performance and its inference time. Since fall detection always requires low inference time, the proposed CD-DNN will be the better choice.

Table 4.3: Inference time of the proposed method and other methods by using cleaned data.

Methods	MLP	*CD-DNN*	RF	KNN	SVM
Time (s)	0.34	*1.23*	2.13	146.87	157.64

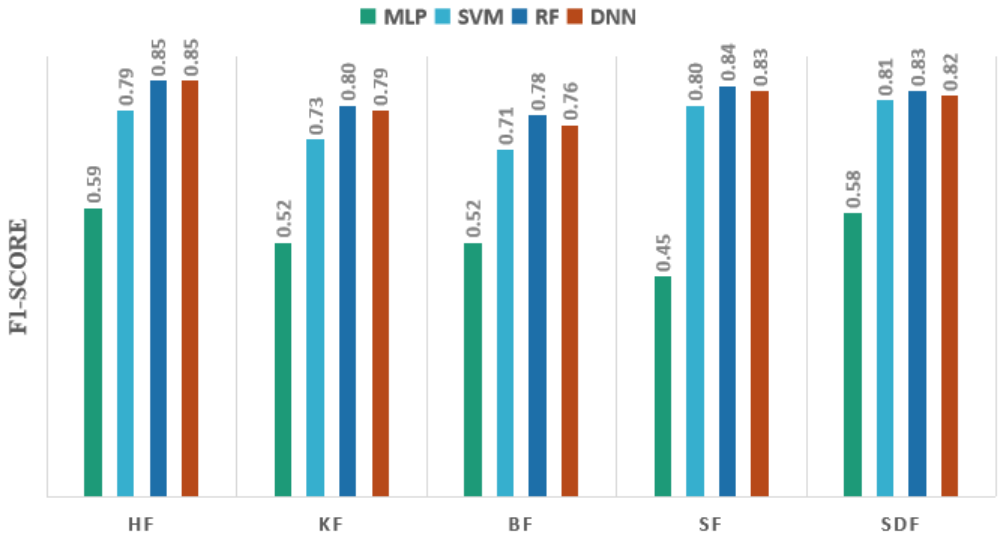


Figure 4.6: The F1 score comparisons between the clustering classifiers and the DNN on the UP-Fall dataset. The fall events abbreviations are shown in Table 4.2.

Figure 4.7: Examples of different types of noisy labels.

In Table 4.4, the comparison results between the baseline method [11] and the proposed methods. It can be observed that after using the proposed noisy instance pruning methods, almost all the fall events classification performance has improved. These results confirm the importance of confident learning in noisy instance pruning methods when using skeleton data. Meanwhile, within the noisy instance pruning methods, the *Confusion* method can achieve the best performance. After comparing the results between *RF-Confusion* and *DNN-Confusion* in Table 4.4, *RF-Confusion* earns the best performance in 2 events (HF and KF) while *DNN-Confusion* gives the best performance in 3 events (BF, SF and SDF).

Table 4.4: Classification performance using F1-score with proposed cascaded data purification methods.

Methods	HF	KF	BF	SF	SDF
RF [11]	0.85	0.80	0.78	0.84	0.83
RF-PBC	0.83	0.81	0.78	0.83	0.83
RF-C+NR	0.84	0.82	0.80	0.86	0.86
RF-PBNR	0.86	0.83	0.79	0.87	0.84
RF-Confusion	**0.88**	**0.85**	**0.82**	**0.87**	**0.87**
DNN	0.85	0.79	0.76	0.83	0.82
CD-DNN-PBC	0.83	0.82	0.79	0.84	0.84
CD-DNN-C+NR	0.86	0.80	0.79	0.84	0.83
CD-DNN-PBNR	0.82	0.81	0.81	0.87	0.84
CD-DNN-Confusion	**0.85**	**0.83**	**0.83**	**0.88**	**0.88**

The visualization results are provided in Figure 4.7, which shows examples of the noisy labels found by confident learning in the dataset. According to the observation, in terms of fall events, in Figure 4.7 (a), the true label is falling by using hands, but the given label is falling by using knees. In Figure 4.7 (b), the true label is falling by using knees but the given label is falling by using hands. Moreover, in daily activities, Figure 4.7 (c) shows the case where the given label is laying but the true label is side-way falling.

The results confirm that different from previous work, where confident learning was applied at the image level, the noisy label issue with skeleton data can also

be addressed. This is beneficial for classification models trained with accurate supervision. Based on classification performance and computational cost, the proposed CD-DNN appears to be the best choice for fall event classification using the extracted privacy-preserving skeleton data.

4.4.4 Dataset for Learning with Noisy Labels

Table 4.5: Comparison between the original and re-scaled training set and detailed description for UP-Fall dataset.

Activity ID	Original	Re-scaled	Description
1	935	940	Falling using hands
2	922	958	Falling using knees
3	1,073	966	Falling backwards
4	902	942	Falling sideways
5	1,157	975	Falling to a chair
6	28,326	962	Walking
7	39,762	964	Standing
8	33,241	967	Sitting
9	1,162	955	Picking up objects
10	16,305	965	Jumping
11	29,682	961	Laying
12	995	965	Unknown

The custom dataset for JoCoT is selected from the UP-Fall dataset. Since the proportion of positive and negative samples in the UP-Fall dataset is imbalanced, their ratio is approximately 3:97. Therefore, the dataset is rescaled by randomly selecting 1,200 samples from the original dataset in each activity. In total, there are 14,400 groups of skeleton data. the dataset consists of three main components, training set, testing set, and validation set, and the ratio is 8:1:1. Validation set was applied to select the model with the best classification performance. Table 4.5 shows the details of the UP-Fall dataset, including the description and the numbers of each activity. The dataset has five fall events and seven normal human indoor activities. The number of fall events is much smaller than the normal activities in the original training set. In order to simulate the data corruption situation in the real world,

noisy labels were added automatically to the dataset by using the noise-generating matrix which will be introduced in the section 4.4.6. The imbalanced dataset problem significantly affects the actual noise rate far away from the noise rate settings and leads to a biased experimental performance. To be precise, the significant disparity in the quantity of data samples among different categories can lead to an imbalance in the generated noise label types, consequently affecting experimental outcomes. For instance, considering a noise label ratio of 0.1 in the case of pairflip noise with the original dataset, when generating from falling to a chair to walking, there are approximately 116 generated noisy instances. However, when generating from walking to standing, the number of generated noisy instances is around 2830, showcasing a substantial discrepancy that can significantly impact experimental results. Therefore, re-scaling the dataset is required for addressing the noise generation problem.

4.4.5 Learning with Noisy Labels Parameter Settings

The vanilla DNN network architectures are applied in both two teacher modules and one student module. The experimental settings in both student and teacher modules are the same. ReLU was used as the activation function. For the Adam optimizer used in all the experiments, similar to the settings in the baseline, the momentum is set as 0.9. Due to the skeleton data containing less information than the RGB image data, the initial learning rates were set as 0.0001. The batch size was set to 128. The epochs were set as 300 for mining the clean instances in the consensus data selection stage. With the same learning rate decay point as in both teacher modules, the learning rate decayed gradually and linearly to zero from 80 to 300 epochs. The clean validation set guided the network training in the correct direction and prevented the over-fitting issue from noisy label instances. Different λ were chosen to achieve the best performance under different noise rates and types. All the experiments were conducted on a workstation with 4 GeForce GTX 1080Ti GPUs and 16GB of RAM.

4.4.6 Learning with Noisy Labels Results

Noise Types

To estimate the real-world noisy dataset, the dataset needs to be corrupted by using the noise-generating matrix $\mathcal{W}$. There are several noise types, e.g., pairflip and symmetric. Details of them will be introduced below, $\mathcal{R}$ denotes the noise rate and M denotes the number of activities in the dataset:

(i) Pairflip flipping, a noise type which flips the ground truth to another specified activity among the entire dataset. The noisy matrix for pairflip is shown below:

$$\mathcal{W} = \begin{bmatrix} 1-\mathcal{R} & \mathcal{R} & 0 & \cdots & 0 \\ 0 & 1-\mathcal{R} & \mathcal{R} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & 1-\mathcal{R} & \mathcal{R} \\ \mathcal{R} & 0 & \cdots & 0 & 1-\mathcal{R} \end{bmatrix} \tag{4.4.1}$$

(ii) Symmetry flipping, which indicates that the noisy label is uniformly distributed over all labels except the true label, with an equal probability distribution. The detailed noise matrix explanation is shown below:

$$\mathcal{W} = \begin{bmatrix} 1-\mathcal{R} & \frac{\mathcal{R}}{M-1} & \frac{\mathcal{R}}{M-1} & \cdots & \frac{\mathcal{R}}{M-1} \\ \frac{\mathcal{R}}{M-1} & 1-\mathcal{R} & \frac{\mathcal{R}}{M-1} & \cdots & \frac{\mathcal{R}}{M-1} \\ \vdots & \frac{\mathcal{R}}{M-1} & 1-\mathcal{R} & & \vdots \\ \vdots & \vdots & & \ddots & \frac{\mathcal{R}}{M-1} \\ \frac{\mathcal{R}}{M-1} & \frac{\mathcal{R}}{M-1} & \cdots & \frac{\mathcal{R}}{M-1} & 1-\mathcal{R} \end{bmatrix} \tag{4.4.2}$$

Pairflip analysis

Flipping-Rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Co-teaching [5]	87.16	85.16	76.51	75.10	47.88	32.84	13.45	8.50
Co-teaching+ [6]	86.75	85.57	81.85	61.79	46.57	20.83	14.22	11.41
JoCoR [7]	88.18	86.12	83.52	78.71	44.32	25.63	12.03	11.04
JoCoT	**89.15**	**86.54**	**84.30**	**79.30**	**50.14**	**37.25**	**20.78**	**12.25**

Table 4.6: Average test accuracy (%) of Pairflip noise with different noise rate on UP-Fall.

Flipping Level	Average	LR-Avg	HR-Avg
Co-teaching [5]	53.33	74.36	18.26
Co-teaching+ [6]	51.12	72.51	15.49
JoCoR [7]	53.69	76.17	16.23
JoCoT	**57.46**	**77.89**	**23.43**

Table 4.7: Average test accuracy (%) of Pairflip noise with different noise rate levels on UP-Fall.

To verify JoCoT performance at different levels of noise rate, pairflip noise from 0.1 to 0.8 are generated into the dataset for the proposed JoCoT for cleaning. Different noise levels are introduced in the experiments to simulate the corrupted dataset in the real-world environment.

As shown in Table 4.6, the test accuracy of Co-teaching and Co-teaching+ are not stable, Co-teaching+ shows better performance than Co-teaching with 30% noisy data, the accuracy reaches 81.85%. However, it drops significantly to 61.79% when the noise rate increases to 40%. Co-teaching achieves better performance than Co-teaching+ at 40% noisy data, which is 75.10% and without a significant drop but a steady decrease. JoCoT still achieves 79.30% test accuracy with the same noise rate setting, which is 40% of pairflip noisy labels. In a similar situation between JoCoR and Co-teaching+, it could be observed that the performance of JoCoR is better than Co-teaching+ at the low noise rate levels but trends to worse than Co-teaching+ when the rates are achieved 50% and 70%, which achieves 44.32% and 12.03%, respectively. It verifies that the robustness of the baseline approaches is not stable enough for

responding to different noise rate settings. Although JoCoR achieves the best average performance in different levels of noise rate among the baseline approaches, it could not gain all the best performance with all the noise rate settings. Different from those baselines, the proposed JoCoT achieves the best performance of the others among all the noise rate levels. Besides these, four algorithms have performance dropping when the noise rate reaches 50%, and the accuracy decreases by at least 25%. It is assumed that the most important reason is the density distribution of the pairflip noisy instances is imbalanced among all the activity categories.

According to the definition of pairflip in equation 4.4.1, each of the activities only exists one specified noisy activity. Thus the density distribution of the pairflip noisy instances is imbalanced. Besides this, the noisy labels ratio of the corrupted dataset reaches a threshold (50%), which indicates the noisy instances become more than half of the original clean dataset. The impact of the weight of the clean instances on the model is lower than the noisy instances, even though the DNN model first learns the clean and simple instances. Those factors can exacerbate the negative influence of the noisy samples, which leads to the incorrect parameter updating direction and the performance dropping. According to Table 4.7, the average test accuracy of JoCoT is better than the other algorithms. It also presents the average testing accuracy in low noise rates (LR-Avg from 0.1 to 0.5 cases) and the average accuracy in high noise rates (HR-Avg from 0.6 to 0.8 cases). It could be observed that the performance improvement at HR-Avg is more significant than in the LR-Avg noise rate. The robust noisy label training method could mine valuable and clean instances even if the noise rate is relatively high. The student could distillate more reliable knowledge from the peer teacher rather than a single teacher. These confirm that the JoCoT has better noise tolerance and robustness to address the high-level noise rate issue for pairflip settings.

Results on Symmetric

Flipping-Rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Co-teaching [5]	87.73	86.29	85.33	83.47	81.11	78.70	69.70	51.79
Co-teaching+ [6]	85.77	84.65	84.31	80.16	77.24	73.56	68.35	46.69
JoCoR [7]	89.37	87.82	86.07	83.36	80.96	79.08	70.87	51.68
JoCoT	**89.44**	**88.10**	**86.32**	**84.49**	**82.55**	**80.63**	**73.71**	**57.35**

Table 4.8: Average test accuracy (%) of Symmetry noise with different noise rate on UP-Fall.

Flipping Level	Average	LR-Avg	HR-Avg
Co-teaching [5]	78.02	84.79	66.73
Co-teaching+ [6]	75.09	82.43	62.87
JoCoR [7]	78.65	85.52	67.21
JoCoT	**80.32**	**86.18**	**70.56**

Table 4.9: Average test accuracy (%) of Symmetry noise with different noise rate levels on UP-Fall.

Similar to the discussion in pairflip section, Table 4.8 shows the performance of different algorithms with different symmetric noise rates. Among all the noise rate levels, it could be easily found that Co-teaching+ achieved the worst performance. The average performance for Co-teaching+ is 75.09%. According to Table 4.9, the LR-Avg and HR-Avg performances are much worse than the other algorithms, which are 82.43% and 62.87%, respectively. Regarding Co-teaching and JoCoR, JoCoR achieves better performance than Co-teaching from 10% to 30% noise rate. However, slightly lower than Co-teaching when the rates in 40%, 50% and 80%. This same situation happened in symmetric noise, verifying that the baseline algorithms are not robust enough for the skeleton data-based noisy label learning on the UP-Fall dataset. Overall, the average JoCoR performance is 78.65%, almost the same as the average performance of Co-teaching, which reaches 78.02%. According to Table 4.8, JoCoT also achieves the best performance for symmetric noise among all the noise rate levels. When the noise rate achieves 80%, JoCoT can also maintain the test accuracy at 57.35% which indicates more than half of the data could be classified correctly.

This verifies that the proposed JoCoT could effectively and robustly mine the clean instance under different noise rates. Moreover, according to Table 4.9, the average accuracy of JoCoT reaches 80.32%, which is almost 5% higher than Co-teaching+. The same in the pairflip section, the improvement of JoCoT obtained in HR-Avg is more than in the LR-Avg. With the increase of the noise rate, the improvements that the proposed JoCoT obtains are improving. This further verified the noise-tolerant ability of JoCoT even under a high-level noise rate, more reliable and robust clean instances could be mined based on the consensus-based clean instance selection algorithm.

In contrast to the pairflip performance change in Table 4.6, there is no performance dropping for symmetric in Table 4.8. The performance of all the algorithms decreased steadily with the increase in the noise rate. This is due to the noisy instances being distributed uniformly in the dataset according to the definition of symmetric noise in equation 4.4.2. Therefore, different from the pairflip noise, the distribution of the noisy data is balanced. This could help to prevent the performance from dropping. JoCoT achieves the best performance among all the noise rate settings with symmetric noises, which justifies its clean instance mining and noise tolerant ability are better than the other approaches.

Pairflip analysis

Table 4.10 indicates the precision performance for the noisy data with pairflip noise in the last training epoch. Since part of the original dataset is applied as the testing set, label precision experiments are conducted on the training set. According to Table 4.10, it could be significantly observed that the proposed JoCoT outperforms all the other baseline algorithms. The same with the test accuracy in Table 4.6, Co-teaching+ could find the least noisy instances on the average of all the noise rates. Comparing Co-teaching with Co-teaching+ in Table 4.10, it could be observed that

the precision of Co-teaching outperforms Co-teaching+ in most of the noise rates but not in 10% and 50%. This may be due to the disagreement applied in Co-teaching+ was to select the incorrect instances for mitigating the peer networks diverged, this may guide the parameter updating towards the erroneous direction. When the rate achieves 50%, there exists a significant drop in all the approaches, as aforementioned, this is due to the noisy instances becoming over half of the dataset and misleading the training process.

Flipping-Rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Co-teaching [5]	71.75	75.77	82.20	74.44	50.94	47.39	59.75	76.27
Co-teaching+ [6]	73.03	74.51	78.33	69.17	52.70	44.81	57.55	71.41
JoCoR [7]	73.06	78.44	80.18	78.87	51.82	48.96	58.29	75.58
JoCoT	**81.26**	**85.35**	**84.78**	**87.01**	**60.99**	**61.46**	**74.40**	**85.64**

Table 4.10: The precision of noisy data (%) for Pairflip noise with different noise rates on UP-Fall.

Flipping Level	Average	LR-Avg	HR-Avg
Co-teaching [5]	67.31	71.02	61.14
Co-teaching+ [6]	65.19	69.55	57.92
JoCoR [7]	68.15	72.48	60.94
JoCoT	**77.61**	**79.97**	**70.62**

Table 4.11: Average Noisy label precision (%) of Pairflip noise with different noise rate levels on UP-Fall.

In Table 4.11, the average precision of Co-teaching+ reaches 65.19% and 67.31% for Co-teaching. Regarding JoCoR, the same situation happens. It could obtain 68.15% on average, which is higher than Co-teaching, but the precision at 30%, 70% and 80% noise rates are lower than Co-teaching. Since JoCoR applied designed loss function between the peer networks predictions, it is assumed that the reason for this is due to the skeleton data containing no such enough information for maintaining the JoCoR performance at some noise levels. This leads the peer networks from JoCoR to learn the incorrect information and mislead the direction of the parameter updating. No matter whether in the LR-Avg or HR-Avg, JoCoT always has significant

improvement for finding the noisy instances and retains the clean instances for information mining. According to Table 4.11, JoCoT has a more significant improvement of precision at a high rate of noise than at a low rate, which reaches around 13% improvement than Co-teaching+.

Moreover, as shown in Tables 4.10 and 4.11, the precision of JoCoT achieves 77.61% for the average precision, which achieves around 10% improvement over JoCoR. Both the performances in LR-Avg and HR-Avg significantly outperformed all the other baseline algorithms, which confirms the robustness and noisy data mining ability of the proposed JoCoT in all the pairflip noise levels.

Symmetric analysis

Flipping-Rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Co-teaching [5]	79.31	85.73	89.42	90.61	90.20	91.08	92.50	92.08
Co-teaching+ [6]	77.71	81.51	82.88	84.10	83.44	84.32	84.96	84.09
JoCoR [7]	82.50	85.99	90.28	91.50	90.13	91.44	92.18	90.72
JoCoT	**88.72**	**91.22**	**93.46**	**94.24**	**93.24**	**94.64**	**95.30**	**95.97**

Table 4.12: The precision of noisy data (%) of Symmetry noise with different noise rates on UP-Fall.

Flipping Level	Average	LR-Avg	HR-Avg
Co-teaching [5]	88.87	87.05	91.89
Co-teaching+ [6]	82.88	81.93	84.46
JoCoR [7]	89.34	88.08	91.44
JoCoT	**93.35**	**92.18**	**95.30**

Table 4.13: Average Noisy label precision (%) of Symmetric noise with different noise rate levels on UP-Fall.

Regarding JoCoR and Co-teaching, JoCoR could achieve higher precision than Co-teaching on average. However, as shown in Table 4.12, the JoCoR could not achieve better precision than Co-teaching among all the noise rates. For example, when the noise rate increases to 70% and 80%, Co-teaching reaches 92.50% and 92.08% respectively, but JoCoR achieves 92.18% and 90.72% respectively, which are

lower than Co-teaching. As opposed to this, the proposed JoCoT could consistently achieve the best performance. This confirms that the JoCoT could always find most of the noisy instances and verify its robustness.

Tables 4.12 and 4.13 show the precision of the algorithms with different noise rates of symmetric noise. Co-teaching+ shows the lowest precision among the noisy label learning methods under different noise rates. The average precision is 82.88%. This may be due to the incorrect clean instances selection of disagreement strategy applied in Co-teaching+ since the skeleton contains much less information than the information the RGB image contains. The average precision of JoCoT reaches 93.35%. JoCoR has only 89.34% for the averaged precision. Both LR-Avg and HR-Avg of JoCoT have nearly 4% precision improvements than JoCoR. Each noise rate setting with JoCoT could achieve at least around 3% improvement over JoCoR. In order to simultaneously analyze the performance in both Table 4.8 and Table 4.12, it could be observed that even if the noisy data precision of JoCoT achieves almost 95% with 60%-80% symmetric noisy instances, which means over 95% noisy data could be found, the accuracy of JoCoT still decreases from over 80% to 57.35%. This is because the number of noisy instances increases in percentage with the increasing of the noise rate. As aforementioned in the dataset setting section, there are 11,520 samples in the training set after the re-scaling operation. Even if the precision of JoCoT achieves around 96% at 80% noise rate. There still exist $11520 \times (1 - 95.97\%) \times 80\% \approx 371$ noisy instances in the dataset, which JoCoT does not select in the corrupted dataset. The noisy instances which are not found by JoCoT with a 10% noise rate, which is approximately $11520 \times (1 - 88.72\%) \times 10\% \approx 130$ in the dataset. It is much smaller than the number of noisy instances at 80% noise rate. The more noisy instances that could not be found, the more significant the performance dropping occurs in the test accuracy. The results in Table 4.8, Table 4.12 and Table 4.13 further confirm the noise tolerant and clean instances mining ability of JoCoT.

Lastly, it is noteworthy to emphasize that the proposed JoCoT framework is almost algorithm-driven. The primary procedure of it for mining clean instances is not heavily reliant on GPUs, unlike certain other deep learning methods. Consequently, JoCoT can be trained not only on GPUs but also on workstations with only CPUs. The training time for each experimental set is approximately 1 hour. Furthermore, the computational power and model size required for JoCoT has also been optimized. The proposed JoCoT requires 13.76M bytes for the model parameters and around 8 seconds for each epoch. Since the JoCoT contains both teacher and student modules, its computational cost is relatively large than the baseline approaches.

4.5 Summary

This chapter aims to tackle the noisy annotation issue for human fall events classification by proposing the following two components: noisy annotations purification and noisy label learning framework, which is call JoCoT. The noisy annotations purification strategy which contains four different pruning methods was utilized for cleaning the corrupted annotations from the dataset. The noisy label learning with trinity networks was exploited to improve the robustness of the framework by applying the corrupted dataset, which could fully exploit the potential of the noisy annotations. The small selection, agreements and cross-update algorithms were utilized for selecting the clean samples in the training stage. Additionally, contrastive learning was applied for improving performance by extending the training process. Finally, the consensus-based decisions were fed into the model for conclusive decisions. The data purification evaluation in Section 4.4.3 showed the contributions of different purification approaches. Evaluations on the widely-used UP-Fall dataset were further demonstrated in Section 4.4.6 to verify the proposed noisy label learning algorithm compared to the other state-of-the-art methods. Furthermore, the analytical experiments with different noisy types were presented to confirm the efficiency of the

proposed JoCoT algorithm.

Overall this chapter has demonstrated the effectiveness of the proposed methods for addressing the noisy label issues both in dataset purification and learning with noisy labels perspectives. However, the data and annotation-lacking issue caused by the privacy issue may be still limited for medical action recognition tasks. In order to address this issue comprehensively and extend its applicability rather than be limited to fall events classification tasks, the proposed approach aims to generalize it for the medical action recognition task with the same privacy issue in the following chapter. Therefore, the next chapter focuses on handling the data and annotation-lacking issue, as well as improving the performance from the feature fusion perspective with limited annotated data by performing a multiple-level fusion of a one-shot learning framework.

Chapter 5

MULTIPLE LEVEL FUSION OF ONE-SHOT LEARNING WITH PRIVACY MITIGATING DATA FOR MEDICAL ACTION RECOGNITION

5.1 Introduction

Previous chapters contributed to improving the fall events classification performance by reducing the redundant information and addressing the noisy label issues for fall

Figure 5.1: The illustrations between the general actions and medical actions.

events classification. However, the data and label lacking as a challenging issue due to the characterisation of the medical data presented in this thesis has not been effectively addressed. Moreover, in order to broaden the range of the application rather than focusing on fall events, this chapter aims to address the data and annotation lacking issues for medical action recognition. Figure 5.1 shows some examples of general actions and medical actions. Different from the general action recognition task, medical action recognition has much fewer samples. There are several reasons leading to this issue. Firstly, similar to other medical applications, privacy protection is one of the most challenging topics for medical action recognition [28, 30]. Secondly, medical actions such as coughing, headache, or falling occur much less frequently than normal human actions in daily life [58]. Some of these medical actions may not even occur within a month. As a result, the number of samples available for analysis is significantly smaller compared to normal behaviors. Furthermore, there are still challenging issues that hinder performance, such as temporal mismatching and similar actions.

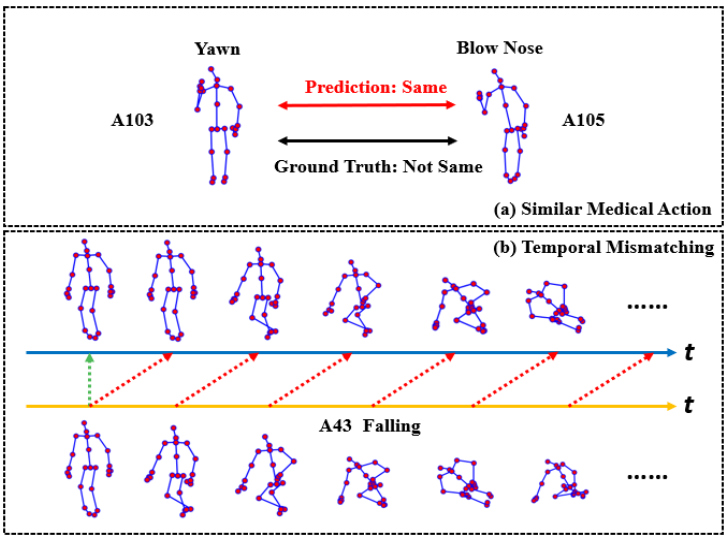


Figure 5.2: The illustration of similar medical actions and temporal mismatching issues, which are the two primary limitations existing in the conventional one-shot learning methods.

With the development of one-shot learning, this learning-with-limited-samples

framework has contributed attractive success to medical action recognition with human skeleton sequences, which could effectively preserve personal information and represent the pose by human landmarks [103, 104]. One-shot learning indicates to fed the model with a single instance for each category after prior knowledge trained [105, 106]. There are two primary limitations in this learning framework that need to be improved as mentioned before: temporal mismatching and similar actions, which are illustrated in Figure 5.2. Firstly, since the same labelled action sequences may have different action timing and temporal lengths. Precisely, variations in the speed at which subjects demonstrate the same action can result in different lengths of action sequences [107]. This temporal mismatching issue will decrease the model performance for the framework when matching the sequences from the support and query sets [108]. Secondly, there exist similar medical actions in the real-life environment, such as headache and neck pain, taking on the glasses and taking off the glasses. Similar actions may confuse the model and lead the incorrect training directions since they have different importance for their landmarks [109]. This may also degrade the experimental performance drastically when matching the support and query set because the unimportant landmarks can have a noise-like effect during the matching process [110].

Furthermore, the fusion of data features has been verified to enhance the recognition performance and robustness in recent years, it also improves the reliability and the explainability of the artificial intelligence system [63, 111–114]. However, the previous work has not taken into consideration the information regarding the changes in the angles of the human skeleton extracted from raw skeleton sequences and the skeleton data used in previous studies still retains visible residual landmarks information regarding privacy concerns. Most previous works for skeletal data action recognition simply applied the feature vectors only consisting of the coordinates of the 3D human landmarks, which could be considered as the position-level features of the skeleton

data. However, the direction-level features from the raw skeleton data, which represent the features of angles between the skeletal bones and coordinate system are not extracted. Precisely, the angles of the bones are complementary with the position of the joint features, which is more informative in interpreting the direction information in the temporal dimension for action recognition. Figure 5.3 demonstrates the motivation of the proposed method which combines both the position and direction-level information for medical action recognition.

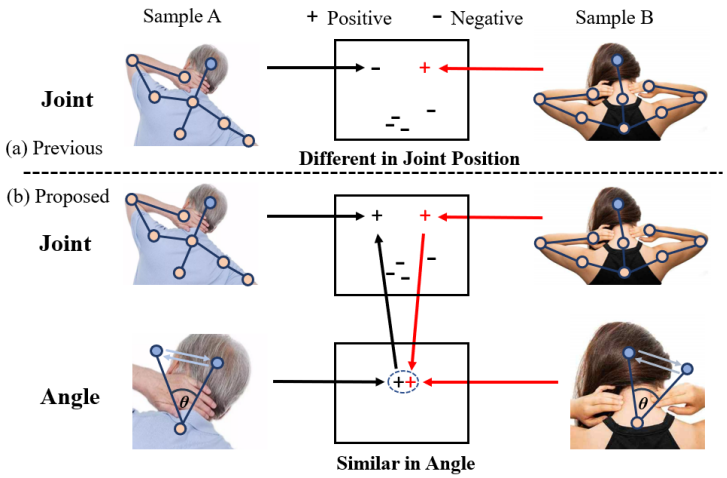


Figure 5.3: Neck pain actions in joint and angle formats. (a) The previous approaches using joint features predict them as different actions. (b) The proposed angle features and fusion method enhance the recognition performance and consider the samples as the same actions.

In this work, a one-shot learning framework is proposed for aiming to deal with the aforementioned issues with a cross-attention mechanism and dynamic time warping via a multi-level fusion approach for medical action recognition. The overview of the proposed medical action recognition framework is illustrated in Figure 5.4 and Figure 5.8. The proposed approach consists of two main components: one-shot learning and multi-level fusion. For one-shot learning, there are three modules, which are signal image generation (SIG), cross-attention (CsA) and dynamic time warping (DTW). In the SIG stage, more informative features are obtained from the limited given data. The skeleton sequences will be first extracted from the raw data. After that, joint

and angle-transformed signal-level images are distilled from the skeleton sequences to further mitigate privacy leakage issues. Moreover, the pixel values from the transformed images are normalized in the pre-processing stage to avoid performance issues caused by significant variations in the temporal dimension. The CsA module aims to guide the model to prioritize the more crucial landmarks from the data, which helps the model discriminate similar actions and reduce the misclassification situation. The DTW module is designed to assist the model in aligning the temporal information of two instances during the matching stage, which helps alleviate the performance degradation caused by mismatched action timings. Multi-level fusion is designed with joint and angle features at both feature and decision levels. The intuition of this design is to enable the maximum strengths from the two different features that exhibit a complementary relationship and also mitigate the sensitivity of the misclassification occurring in the original single feature rule. NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD Benchmark evaluations are provided to verify the improved performance over the other state-of-the-art one-shot learning methods. To summarize, the main contributions of the proposed work are listed as follows:

1. A novel feature transformation method that distilled the skeleton sequences for both human joints and angles is proposed for further preserving privacy and improving the recognition performance.

2. A novel medical action recognition approach is proposed by conducting a feature cooperative training method within the one-shot learning framework, along with a multiple-level fusion to further improve the framework performance.

3. Both the cross-attention and the dynamic time warping modules are applied to address similar medical actions and temporal mismatching issues to enhance the training process.

4. Experimental results on NTU RGB+D 60 [2], NTU RGB+D 120 [3] and PKU-

MMD [4] are provided to confirm improved performance better than other state-of-the-art one-shot learning action recognition methods.

This chapter aims to address the fifth and sixth objectives of this thesis, which relate to the one-shot learning framework based on medical action recognition including cross-attention and dynamic time warping modules published in IEEE ICASSP 2023 [30], both the feature transformation for human skeleton sequences and the multiple level fusion theory are exploited by submitting to IEEE Transactions on Multimedia [31].

The rest of this chapter is presented as follows: Section 5.2 presents the details of each module in the proposed one-shot learning framework, which includes the signal image transformation, cross-attention and dynamic time warping modules. Section 5.3 explains the multiple-level fusion process, which is realized by applying the transformed features on both the feature level and decision level. Then the training objectives for the proposed one-shot learning framework including the prediction distribution and loss function are formulated in Section 5.4. Extensive experimental results on the benchmarks and discussions are presented in Section 5.5 to verify the effectiveness of the proposed one-shot learning framework with the multiple-level fusion approach for medical action recognition. The conclusions of this chapter are summarized in Section 5.6.

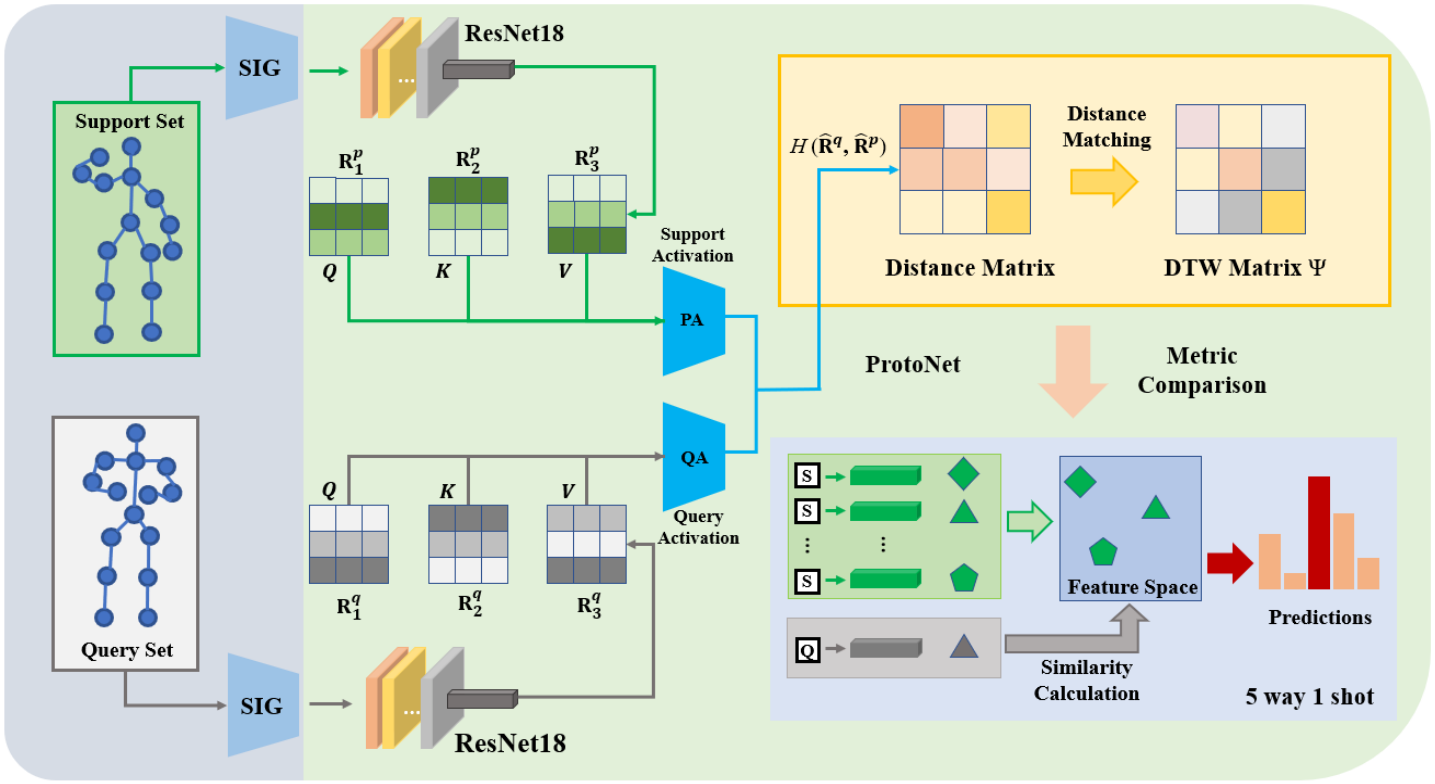


Figure 5.4: The illustration of the proposed one-shot learning framework which contains the SIG, CsA and DTW modules. The SIG module first transforms the input skeleton sequences into signal-level images before being fed into the ResNet18 encoder for feature extraction. The encoded features from both support and query sets are fused via the CsA module for metric learning in the ProtoNet framework [9]. DTW module is exploited to address the temporal information mismatching issue which could be obtained via Equations 5.2.6 and 5.2.7. The vectors from the support and query set are finally mapped to the feature space for similarity calculation and to obtain conclusive results.

5.2 The Proposed One-shot Learning Framework

5.2.1 Overview

The overview of the proposed one-shot learning framework is illustrated in Figure 5.4. The foundation of one-shot learning was introduced in detail in Chapter 2, Section 2.3.3. Firstly, for both the support set and the query set, raw human skeleton sequences are transformed into signal-level images. Then the instances are fed into the ResNet18 followed by the cross-attention for enhancing the discriminative capability in similar medical action recognition. The resulting cooperative outputs from the cross-attention module are further processed through the proposed dynamic time warping module to mitigate the temporal information mismatching issue. Finally, ProtoNet [9] is applied for calculating the similarity between the instances from the support set and the query set for conclusive predictions according to the outputs from the feature space.

5.2.2 Preliminaries

To address the limitations of the one-shot medical action recognition abovementioned, a ProtoNet-based [9] novel one-shot learning framework is proposed to train with multiple level fusion by signal level images. The details of ProtoNet are provided in section 2.3.3 of Chapter 2. The proposed algorithm will be demonstrated on the dataset $\mathcal{D} = \{(\boldsymbol{S}_i, y_i)\}_{i=1}^{\mathcal{N}}$ which includes $\mathcal{N}$ skeleton sequences $\boldsymbol{S}_1, ...\boldsymbol{S}_{\mathcal{N}}$, with given labels $y_i \in \{1, \ldots, M\}$. Two different human action features, $\mathbf{F}_j$ and $\mathbf{F}_a$ are extracted from $\boldsymbol{S}_i$ with signal-level image representation, The proposed model aims to train the extracted features $\mathbf{F}_j$ and $\mathbf{F}_a$ from dataset $\mathcal{D}$ to get the joint feature representation $\mathbf{x}_j = f_\delta(\mathbf{F}_j)$ and the angle feature representation $\mathbf{x}_a = f_\delta(\mathbf{F}_a)$. And utilizes $\mathbf{x}_j$ and $\mathbf{x}_a$ with the multiple-level fusion approaches by calculating the distance based on the proposed metric learning framework for human medical action recognition.

5.2.3 Signal Images Transformation

Joints Transformation

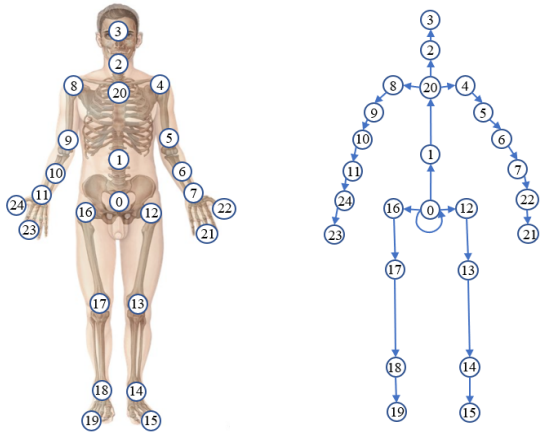


Figure 5.5: The left sketch illustrates the joint labels for each body part from NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD. The right sketch shows the bone labels which are extracted for the angle features, the direction of arrows indicates the bone directions.

Based on the human physical structure, the human bones are manually designed from the landmarks in the raw skeleton sequences. The customised joint connections for the human bone information from NTU RGB+D and PKU-MMD datasets are shown in Figure 5.5. Different from most of the available skeleton-based approaches, the transformed signal level images will be applied as the training data, the same as in [30]. The raw skeleton sequence $\boldsymbol{S}$ from NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD datasets denote as $\boldsymbol{S} = \mathbb{R}^{X \times T \times V}$, where X denotes the number of joints in each skeleton, T denotes the temporal length of the sequence and V denotes the 3D coordinates position of each joint. Since the pixels of RGB image has three colour channels, $\boldsymbol{S}$ could be transferred into image representation as $\mathbf{F}_j \in \{0, 1, ..., 255\}^{H \times W \times 3}$. Hence, the total number of joints X and the temporal length T are transformed into the image height of H and the image width W, respectively.

To balance the impacts of the pixel values on the model performance, a normalization operation is applied to the transformed images. The overall transformation

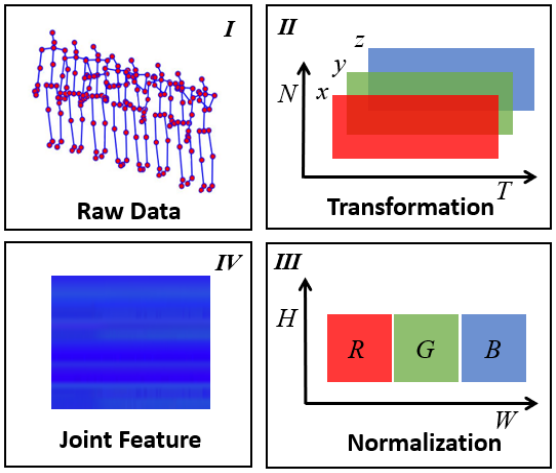


Figure 5.6: The illustration of joint feature image transformation.

process of the joint feature is illustrated in Figure 5.6.

Angles Transformation

In order to obtain different types of human action features from the raw skeleton sequences. The angles of the human bones are extracted as the other action features from the raw skeleton sequences.

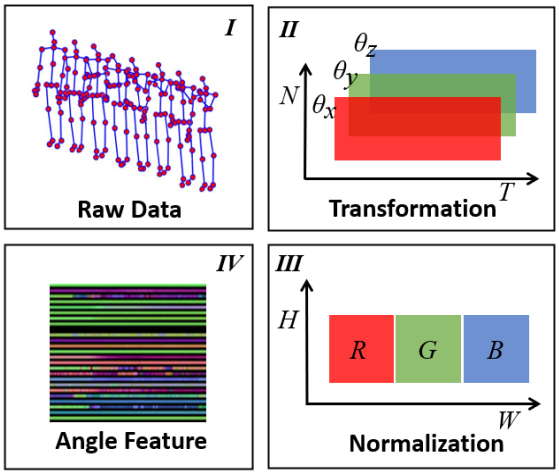


Figure 5.7: The illustration of angle feature images transformation.

Since the provided data is in 3D coordinates, in order to facilitate the acquisition of angle information, three normal vectors are established based on a coordinate system. By calculating the angle information between each bone with the three normal vectors,

respectively in the time dimension and performing image transformation operations, human angle features are complementary to the joint information for recognition accuracy. The angles with the three coordinate system planes are formulated as follows:

$$\theta_x = \arccos \frac{|\mathbf{v}_x \cdot \mathbf{v}_b|}{|\mathbf{v}_x| \cdot |\mathbf{v}_b|} \tag{5.2.1}$$

$$\theta_y = \arccos \frac{|\mathbf{v}_y \cdot \mathbf{v}_b|}{|\mathbf{v}_y| \cdot |\mathbf{v}_b|} \tag{5.2.2}$$

$$\theta_z = \arccos \frac{|\mathbf{v}_z \cdot \mathbf{v}_b|}{|\mathbf{v}_z| \cdot |\mathbf{v}_b|} \tag{5.2.3}$$

where $\mathbf{v}_b$ denotes the bone vector and $\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z$ are the normal vectors from the three dimension, respectively. The same as the joint image features, the angle information will be also transferred into $\mathbf{F}_a \in \{0, 1, ..., 255\}^{H \times W \times 3}$. $\theta_x, \theta_y, \theta_z$ will be calculated as the values of three colour channels for the pixels after the normalization. Similar to the joint feature, temporal length T and the number of joints X are transferred into the height W and the width H of the SIG images, respectively. The illustration of the transformation process for angle SIG images is shown in Figure 5.7.

5.2.4 Cross Attention Mechanism

After transferring the skeleton sequence into the signal-level representation, a cross-attention module between the support set and query set is exploited in the proposed framework. In previous cross-attention approaches, typically only one of the two modules involved in the computation was focused on. This is due to the aim of cross-attention is to utilize information from one module in another to enhance the model performance. To extend the capabilities of the cross-attention module, it is supposed to focus on more modules and capture a more comprehensive set of associative information. This mechanism could decrease the difficulties in discriminating against similar human actions. This is because the spatial relationship between the

different parts of the human body plays an important role in distinguishing similar actions. For example, coughing is similar to neck pain, the difference between them is that the head joint of coughing is more important than the neck joint for the neck pain action. Furthermore, the representation of the support set should adaptively change the importance of each joint according to the representation of the query set or vice versa. The transformed representations of query set $\hat{\mathbf{R}}^q$ and support set $\hat{\mathbf{R}}^p$ are formulated as follows:

$$\hat{\mathbf{R}}^q = \mathrm{S}\left(\frac{\mathbf{M}_1^q\mathbf{R}^q \cdot [\mathbf{M}_2^q\mathbf{R}^p]^{\mathbf{T}}}{\sqrt{d}}\right)\mathbf{M}_3^q\mathbf{R}^q \tag{5.2.4}$$

$$\hat{\mathbf{R}}^p = \mathrm{S}\left(\frac{\mathbf{M}_1^p\mathbf{R}^p \cdot [\mathbf{M}_2^p\mathbf{R}^q]^{\mathbf{T}}}{\sqrt{d}}\right)\mathbf{M}_3^p\mathbf{R}^p \tag{5.2.5}$$

where the representation of the query set and support set could be formulated as $\mathbf{R}^q$ and $\mathbf{R}^p$, respectively. And $\mathrm{S}(\cdot)$ denotes the Softmax function which is applied for calculating the weights of different human body parts. $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ are the transformation matrices, which contain the trainable parameters. $\mathbf{T}$ denotes the transpose matrix operation and d indicates the latent dimension of the skeleton joints.

$$H\left(\hat{\mathbf{R}}^q, \hat{\mathbf{R}}^p\right) = \left\|\hat{\mathbf{R}}^q - \hat{\mathbf{R}}^p\right\|_F \tag{5.2.6}$$

where H denotes the distance between the two samples from the support and query set, respectively. This will be applied in prototypical networks [9] for classifying different human actions. And $\|\cdot\|_F$ indicates the Frobenius normalization which is a kind of specific L_2 regularization between the matrices. In equations 5.2.4 and 5.2.5, $\mathbf{M}_2^q\mathbf{R}^p$ and $\mathbf{M}_2^p\mathbf{R}^q$ allow the model to interact with the feature information from different perspectives. The model can improve its accuracy in recognizing and classifying similar actions by incorporating its attention mechanism to further pay attention to the important joints through the interactive information between the

support and query sets. This approach aims to reduce the learning inductive bias and minimize the occurrence of misclassifications of similar actions. By applying higher attention weights to the more crucial parts of the human body, the model enhances its ability to distinguish between similar actions and ultimately enhances its overall accuracy.

5.2.5 Dynamic Time Warping

There exist several factors (e.g. different experimental subjects, speed, duration of the recording and action timing) that result in the temporal information mismatching issue between the support set and query set actions. For the instance in support set $P = \left\{\hat{\mathbf{R}}_1^p, \hat{\mathbf{R}}_2^p, \cdots \hat{\mathbf{R}}_m^p\right\}$ and the query set $Q = \left\{\hat{\mathbf{R}}_1^q, \hat{\mathbf{R}}_2^q, \cdots \hat{\mathbf{R}}_m^q\right\}$, m is the length of the resized signal image. The mismatching issue will directly affect the Euclidean distance calculation between $\hat{\mathbf{R}}_i^p$ and $\hat{\mathbf{R}}_j^q$ and decrease the classification performance. To tackle this issue, the most popular temporal information alignment approach is exploited, which is the dynamic time warping approach from [108] to address this issue.

$$\Psi(i,j) = E(i,j) + \min\{\Psi(i-\zeta, j-\zeta), \Psi(i-\zeta, j), \Psi(i, j-\zeta)\} \tag{5.2.7}$$

where $\Psi(i,j)$ indicates the cumulative distance between the i-th frame from the query set and the j-th frame from the support set. ζ is the time mismatch hyperparameter. Each element in *E(i, j)* is generated according to equation 5.2.6. In practice, each pair of instances is correspondingly associated with each other to compute a correlation distance response map. The instances could be aligned by measuring the similarity between them since DTW allows for the non-linear mapping of the temporal dimension. The $\hat{P}$ and $\hat{Q}$ will be updated from support set P and the query set Q after the

DTW module, respectively.

5.3 Multiple Level Feature Fusion

5.3.1 Multiple Feature Fusion

In this context, multiple-feature fusion actually enhances the representation of the human action features. It refers to merging both the joint and angle signal-level images at an early stage of the proposed framework. It involves combining the complementary feature information and increasing the correlation between the joint and angle before any further processing stage begins. The commonly used position-level information of the skeleton data only contains 2D or 3D coordinates of the joints. Nevertheless, the angles of the human bones, which are regarded as direction-level information are naturally complementary to the position-level information. Typically, the movement differences between the angles and the directions are more informative and discriminative for action recognition.

After the signal image transformation, joint and angle images are merged as one single image representation, respectively. However, the temporal dimensions of each raw skeleton sequence are different, to ensure consistent signal image size after transformation, the feature images are resized with customised image resolutions. Early fusion can benefit from cooperative learning and feature sharing across both the position and direction-level information, potentially leading to improved performance. Moreover, the computational complexity could be simplified and straightforward to implement and interpret due to the increasing feature dimensions. A detailed illustration of early fusion is shown in Figure 5.8 (a).

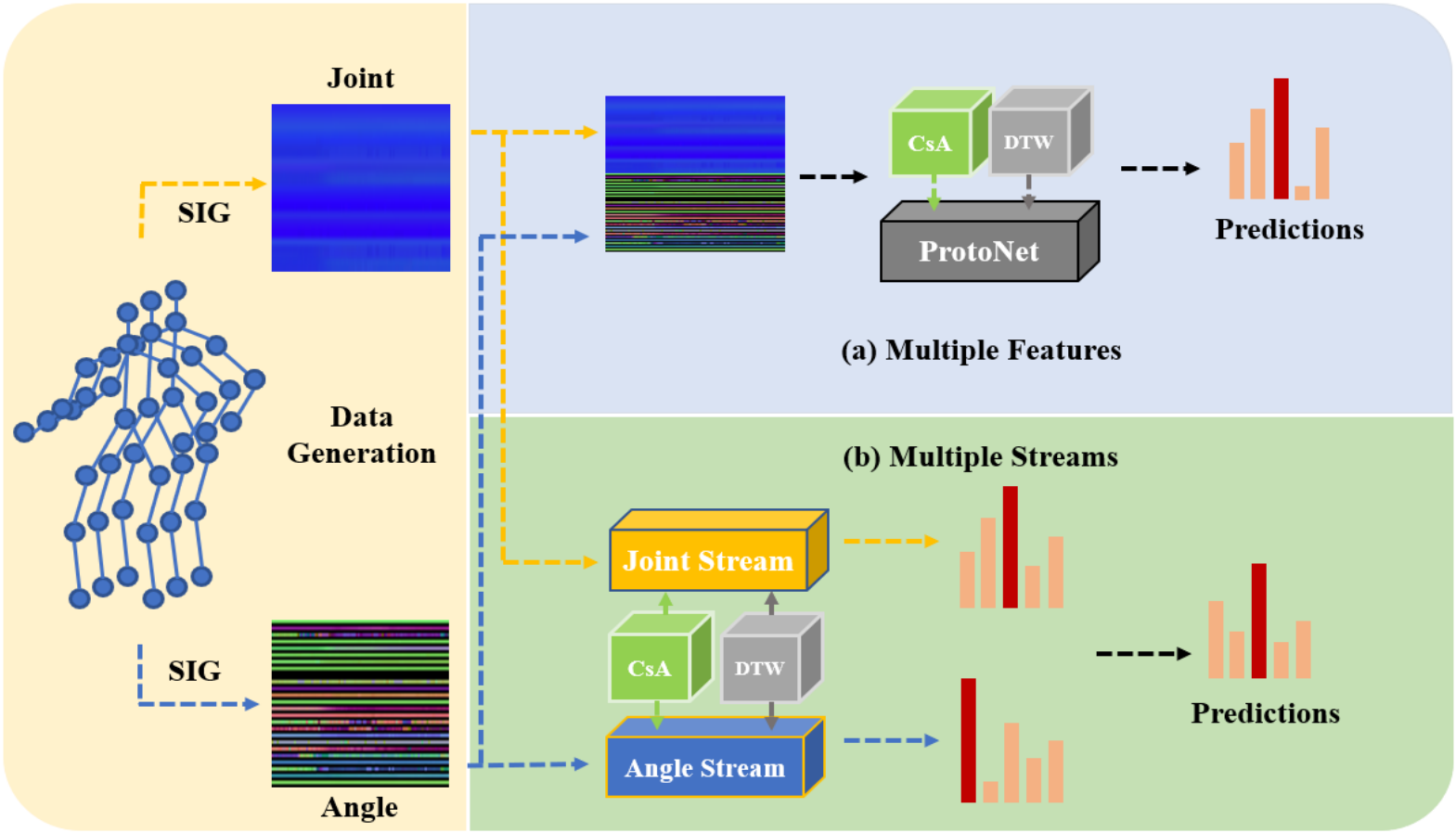


Figure 5.8: The overall illustration of proposed multiple-level fusion method. Raw skeleton sequences will be first transformed into signal-level images by the SIG method and then fed to different fusion levels for medical action recognition. (a) Multiple feature fusion aims to concatenate joint and angle features as the final feature and fed it into the proposed one-shot learning framework for medical action recognition. (b) Both the proposed CsA and DTW modules will be employed in the multiple streams. The probability scores from the multiple streams will be calculated for the final predictions (Best view this in the color version).

5.3.2 Multiple Stream Fusion

Since the multiple feature fusion merges joint and angle information at the beginning of the proposed framework. The fused features may be sensitive to noise and outliers if it includes potentially noisy information in the individual feature. Furthermore, due to the various complexity of the human action dataset, late fusion is needed for action recognition. Without the correlations between the joint and angle information, the prediction probabilities value will be weight averaged and obtained for the fusion recognition task. The pipeline of late fusion is demonstrated in Figure 5.8 (b) for further explanation. Both the joint and angle features are trained in a multiple stream network model, and the late fusion matching prediction is computed as follows:

$$\mathcal{P}_m = \mathcal{P}_j + \alpha \mathcal{P}_a \tag{5.3.1}$$

where $\mathcal{P}_j$ and $\mathcal{P}_a$ are the joint prediction values and angle prediction values from the sub-streams, respectively. α indicates the importance of weight and is set as 1 as the default. A detailed value analysis of the weight settings will be discussed in the experiment session.

5.4 Training Objectives

For the support set, there are N classes with M labelled support actions in each class. As the definition in [9], each prototype indicates the mean vector of the support embedding points which belong to its categories. The prototypical representation of each instance could be formulated as follows:

$$C_m = \frac{1}{M} \sum_{\left(\mathcal{A}_i^p, y_i^p\right)} f_\phi\left(\mathcal{A}_i^p\right) \times \mathbb{F}\left(y_i^p = m\right) \tag{5.4.1}$$

where $\mathcal{A}_i^p$ is the actions from the support set and $\mathbb{F}$ is the indicator function. f_ϕ

denotes the action encoder for the learnable parameter ϕ and y_i^p denotes the ground truth label for the i-th instance from the support set. The prediction distribution of the actions from the query set is defined as follows:

$$p_\phi\left(y = m \mid \mathcal{A}^q\right) = \frac{\exp\left(-\operatorname{dis}\left(f_\phi\left(\mathcal{A}^q\right), C_m\right)\right)}{\sum_{m'} \exp\left(-\operatorname{dis}\left(f_\phi\left(\mathcal{A}^q\right), C_{m'}\right)\right)} \tag{5.4.2}$$

where $\mathcal{A}^q$ denotes the query actions. And exp is the exponential function and $\operatorname{dis}(\cdot)$ is the distance function. $p_\phi\left(y = m \mid \mathcal{A}^q\right)$ represents the probability of instance $\mathcal{A}^q$ from the query set belongs to class m. The model aims to maximise $p_{\boldsymbol{\phi}}(y = m \mid \mathcal{A}^q)$ value to obtain the predicted classes, which means minimising the distance between the query instances and prototype. The matching loss $\mathcal{L}_{\mathrm{m}}$ is derived as shown in equations 5.4.3 and 5.4.4.

$$\mathcal{L}_d = -\frac{1}{B \times r} \sum_{b}^{B} \sum_{i}^{r} \|\mathbf{U}_{bi}\| \tag{5.4.3}$$

$$\mathcal{L}_{\mathrm{m}} = -\frac{1}{N^q} \sum_{i}^{N^q} \log p_\phi\left(\hat{y}_i = y_i \mid \mathcal{A}_i^q\right) + \lambda \mathcal{L}_{\mathrm{d}} \tag{5.4.4}$$

where $\mathcal{L}_d$ is the disentanglement loss function for decreasing the linear dependence between the skeleton key points. B denotes the batch size, b is the b-th instance from the batch and r is the length of the resized image. $\mathbf{U}_{bi} \in \mathbb{R}^{X \times d}$ denotes the i-th second action position of b-th updated image presentation. Model training proceeds to minimise the negative log probability of the ground truth y_i with the optimizer. N^q indicates the number of query actions and $\hat{y}_i$ is predicted label of the i-th instance. $\|\cdot\|$ indicates the paradigm function for regularization and λ is the weight hyperparameter of $\mathcal{L}_d$ for obtain the optimal model performance.

5.5 Experiments

5.5.1 Datasets

This section briefly introduces the three selected public and available datasets for the experiments. All three datasets are large-scale datasets and provide the human skeleton sequence data. These datasets are distinctively and roughly divided into training (80%), validation (10%) and testing (10%) sets. The instances in the testing sets are entirely unseen during the training stage.

PKU-MMD

PKU-MMD dataset is a 3D large-scale dataset which contains 1076 long skeleton sequences in 51 action classes. It is recorded from 66 subjects in 3 different camera views. There are over 20,000 instances provided with multi-modality data, including RGB, depth, infrared radiation and skeleton sequences. Five medical-related actions are selected as the testing set which contains, falling, backache, heart pain, headache and neck pain (A11, A42, A43, A44, A45). There are 5 randomly selected classes applied as the validation set and 41 classes of actions for the training set.

NTU RGB+D 60

NTU RGB+D 60 is a 3D large-scale human action dataset which provides skeleton data sequences. The skeleton data sequences consist of 56,880 instances for 60 types of human actions. These sequences are recorded from 40 different subjects with 17 different scene conditions, each subject provides 25 pose landmarks. For evaluation of medical action recognition, 6 actions with medical conditions are considered for testing, they are cough, falling down, headache, chest pain, back pain and neck pain (A41, A43, A44, A45, A46, A47).

NTU RGB+D 120

Similar to NTU RGB+D 60, it is an extended version of NTU RGB+D 60. It contains 120 types of human actions recorded from 106 subjects in 155 different scene conditions and each subject also provides 25 pose landmarks. There are over 114,000 skeleton sequences including daily, mutual and medical-related actions. For medical action recognition, 12 actions with medical conditions are selected as testing set, they are cough, staggering, falling, headache, chest pain, back pain, neck pain, vomiting, fan self, yawn, stretch oneself and blow nose (A41, A42, A43, A44, A45, A46, A47, A48, A49, A103, A104, A105). There are 12 classes selected as the validation set and the rest of the actions are set as the training set which contains 96 classes.

5.5.2 Implementation Details

For all the datasets, ResNet18 [91] is selected as the backbone to encode the proposed image-level pose feature, which is the most widely used for image-processing tasks. Adam optimizer [115] is applied to the experiments with an initial learning rate of 0.001 with a decay of 0.5. The random seed is chosen as 7 and ProtoNet [9] is selected for classifying the action by calculating the distances between the instances. The parameter of DTW is set to default as 1. Furthermore, all the experiments are conducted under the Ubuntu system by using Pytorch deep learning framework, on a workstation with 4 GeForce GTX 1080Ti GPUs.

5.5.3 Performance Analysis

In this section, the performance analysis is presented to evaluate the effectiveness of the proposed approach, including the ablation study of different proposed modules as well as the experiments with different parameter settings. The experiments are conducted on the widely-used and well-known action recognition datasets for this purpose, which are NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD datasets.

Ablation Study

Table 5.1: The 5-way-1-shot Top-1 accuracy (%) comparisons with SOTA methods on NTU RGB+D 120, NTU RGB+D 60 and PKU-MMD for medical action recognition.

Approaches	NTU-120	NTU-60	PKU-MMD
Skeleton-DML [104]	39.3	45.1	34.8
SL-DML [116]	41.8	33.6	36.0
PAMMAR [30]	67.9	56.5	57.3
Angle	64.3	56.1	68.4
Joint	67.9	56.5	57.3
MF-OSMAR (MF)	**68.6**	**57.3**	**72.0**
MF-OSMAR (MS)	**68.6**	**59.2**	**68.6**

Table 5.2: The 5-way-1-shot accuracy (%) of the proposed MF-OSMAR for medical action recognition with different resolutions on NTU RGB+D 120 dataset.

Features	Size 192		Size 160	
	DTW	CsA+DTW	DTW	CsA+DTW
Joint [30]	64.6	67.9	64.2	66.0
Angle	63.0	64.3	62.6	62.7
MF-OSMAR (MF)	**67.8**	**68.6**	**66.6**	**67.0**
MF-OSMAR (MS)	**67.1**	**68.6**	**66.2**	**68.3**

Table 5.3: The 5-way-1-shot accuracy (%) of the proposed MF-OSMAR for medical action recognition with different signal image resolutions on NTU RGB+D 60 dataset.

Features	Size 192		Size 160	
	DTW	CsA+DTW	DTW	CsA+DTW
Joint [30]	52.9	56.5	52.8	54.1
Angle	52.1	56.1	52.1	55.2
MF-OSMAR (MF)	**55.0**	**57.3**	**53.5**	**55.6**
MF-OSMAR (MS)	**58.2**	**59.2**	**56.5**	**58.0**

Table 5.4: The 5-way-1-shot accuracy (%) of the proposed MF-OSMAR for medical action recognition with different signal image resolutions on the PKU-MMD dataset.

Features	Size 192		Size 160	
	DTW	CsA+DTW	DTW	CsA+DTW
Joint [30]	56.4	57.3	55.9	56.7
Angle	65.2	68.4	64.2	65.0
MF-OSMAR (MF)	**66.6**	**72.0**	**66.3**	**66.8**
MF-OSMAR (MS)	**65.4**	**68.6**	**65.2**	**65.3**

In order to investigate the contributions of different proposed components of the proposed approaches, the ablation study is performed from two perspectives. Firstly, the experiments with only DTW module or both DTW and CsA modules are conducted individually after the raw skeleton sequences are processed by the proposed SIG transformation operation. Next, the impacts of the multiple features (MF) and multiple streams (MS) are analyzed with the overall proposed framework including SIG, CsA and DTW. Furthermore, ablation studies for specific medical actions by joints and angles are provided. Finally, the fusion hyperparameter ablation study is conducted to analyze the impact of the fusion weights on the late fusion performance.

Table 5.1 illustrates the comparisons with the SOTA approaches on medical action recognition. Table 5.2, 5.3 and 5.4 report the detailed evaluations on NTU RGB+D 120, NTU RGB+D 60 and PKU-MMD for the proposed multiple-level fusion approach, respectively. Compared to PAMMAR [30], the proposed method here effectively reduces incorrect recognition times and improves NTU RGB+D 120 by 0.7%, NTU RGB+D 60 by 2.7% and PKU-MMD by 14.7%. This is because NTU RGB+D 120 is a massive scale dataset, which is much more complex and challenging than the other two datasets. Comparison between DTW and DTW+CsA verifies the cross-attention module correctly directs the model to focus on the important parts, even with the early fusion. Moreover, the model is employed with different image resolutions to further verify the contributions of each proposed module. The smaller resolution of the transformed image leads to a slight decrease in the model performance, which is due to the information loss during the image resize operation.

Analysis for Specific Classes

To investigate the performance of each specific action, quantitative ablation study experiments are conducted for each specific medical action on NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD datasets of the proposed one-shot learning approach in

Table 5.5: Ablation Study on the specific classes on NTU RGB+D 60 (NTU 60), NTU RGB+D 120 (NTU 120) and PKU-MMD datasets for 5-way-1-shot medical action recognition with Top 1 Accuracy (%). The tag w/o CsA indicates the model only contains the DTW module and w/ CsA indicates the model contains both DTW and CsA modules. The experimental results are obtained with the transformed signal images with 192×192 resolutions.

#	Medical Action	Datasets	Joint		Angle		Joint+Angle (MF)		Joint+Angle (MS)	
			w/o CsA	w/ CsA	w/o CsA	w/ CsA	w/o CsA	w/ CsA	w/o CsA	w/ CsA
A41	Cough	NTU 60	33.2	42.5	31.4	43.1	34.2	40.2	43.1	**45.8**
A43	Falling down	NTU 60	97.7	98.2	95.0	95.8	**99.4**	97.8	98.6	99.1
A44	Headache	NTU 60	62.5	57.4	55.3	57.1	57.7	57.7	63.4	**63.5**
A45	Chest pain	NTU 60	32.9	45.9	32.4	44.9	37.3	47.5	46.0	**48.1**
A46	Back pain	NTU 60	49.8	53.6	48.0	54.7	59.1	**60.2**	53.1	56.9
A47	Neck pain	NTU 60	42.9	42.8	41.3	41.6	43.1	43.0	42.1	**45.2**
A41	Cough	NTU120	36.5	37.5	42.9	43.0	43.0	43.5	45.0	**45.6**
A42	Staggering	NTU 120	88.4	92.2	91.0	93.1	90.3	93.0	93.0	**95.7**
A43	Falling down	NTU 120	89.2	95.7	89.0	94.8	98.6	98.9	95.0	**99.5**
A44	Headache	NTU 120	61.8	63.1	64.1	66.5	**70.2**	66.1	66.8	65.7
A45	Chest pain	NTU 120	48.9	50.4	49.4	50.4	51.9	**54.8**	53.9	54.1
A46	Back pain	NTU 120	62.6	65.1	61.2	63.4	64.7	**68.8**	65.7	67.8
A47	Neck pain	NTU 120	45.3	54.9	46.3	49.8	53.7	53.7	48.0	**55.9**
A48	Vomiting	NTU 120	65.3	63.6	70.9	72.3	60.8	**73.2**	72.5	73.0
A49	Fan self	NTU 120	62.4	73.2	53.9	63.9	68.4	**75.2**	62.2	72.6
A103	Yawn	NTU 120	60.5	68.1	67.1	67.4	**75.4**	69.3	69.8	74.4
A104	Stretch oneself	NTU 120	72.5	78.9	66.6	69.3	**82.3**	74.2	72.4	79.0
A105	Blow nose	NTU 120	58.2	59.8	59.5	56.0	**64.5**	61.7	63.6	60.4
A11	Falling	PKU-MMD	98.5	99.6	96.6	97.5	93.8	97.7	99.4	**99.8**
A42	Backache	PKU-MMD	49.6	49.6	60.0	**75.3**	72.1	72.2	63.8	74.7
A43	Heart pain	PKU-MMD	42.8	43.4	65.7	65.9	56.0	**67.7**	60.1	61.9
A44	Headache	PKU-MMD	47.0	50.1	59.1	60.8	60.9	**69.7**	56.1	58.1
A45	Neck pain	PKU-MMD	44.5	46.6	46.3	45.6	45.8	**51.0**	47.8	48.4

this section.

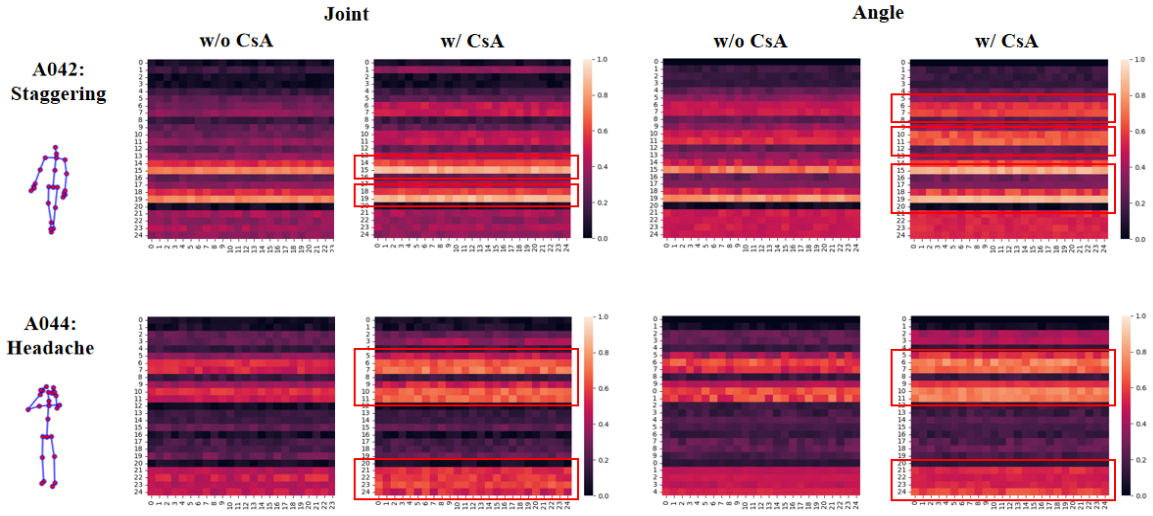


Figure 5.9: The heatmap visualization for 5-way-1-shot medical action recognition implemented by the proposed MF-OSMAR on A042 (staggering) and A044 (headache) from NTU RGB+D 120 dataset without (w/o) CsA and with (w/) CsA modules. The predicted important body parts are highlighted in the red boxes. (Best viewed in the color version)

Table 5.5 reports the Top 1 accuracy (%) performance on the medical action for the three datasets. In Table 5.5, w/o CsA indicates the framework without CsA module and w/ CsA indicates the framework with CsA module is employed. To compare the performance of joint and angle without CsA module in Table 5.5, for some specific medical actions, such as vomiting and headache in NTU RGB+D 120, backache and heart pain from PKU-MMD dataset, the performance of the single angle feature relatively outperforms single joint feature performance. This is due to the fact that the changing of the angles of the human physical body is more significant in these medical action sequences, which contributes informative knowledge to the angle features for model training. It could be easily observed that most of the best performances for each specific medical action are under the proposed multiple-stream or multiple-feature fusion method, which verifies that the joint information benefits the complementary relation for enhancing the medical action recognition performance. The visualization heatmap comparisons between A042 staggering and A044 headache are illustrated

in Figure 5.9, which demonstrates the query activation attention weights when the instances from the query and support set are the same. According to Table 5.5, the staggering and headache medical actions from the NTU 120 RGB+D dataset perform better results from angle features rather than joint features. For the A042 staggering, both the arms and legs of the subjects keep significant movements in the sequences. It could be observed that the model correctly focuses on the important body parts after the CsA module, which are labelled as 15, 19, 14, 18, 7, 11, 6 and 10. For the A044 headache, the subjects keep holding their heads by using their arms. The heatmaps for A044 demonstrate higher weights after getting through the CsA module for the arms and hands, which are labelled as 6, 10, 11, 7, 21, 22, 23 and 24. Furthermore, the weights of the head are slightly improved in the angle features after the CsA module which corresponds to the minor movements of the head, which are labelled as 2 and 3. Detailed illustrations of the landmarks label are shown in Figure 5.5. Comparing the heatmaps from the joint and angle w/o CsA for A042, it could be observed that landmarks 6, 7, 10 and 11 have more important weights in angle than joints, the same situation happens in A044, in which the color of these parts is brighter than the joint. The reason is that the direction-level features are more informative than the position-level features in interpreting A042 actions. the visualisation results demonstrate the CsA module guides the network to focus on the correct and important parts for each specific medical action, which is beneficial for distinguishing similar actions and improving performance. The experimental results illustrated in the tables verify the complementary relation between the position and direction-level features for further enhancing performance.

Overall, the proposed data fusion approaches achieve the best performance among almost all the medical actions compared to the single feature performance in all the datasets. This verifies that the joint and angle features are complementary in the spatial and temporal dimensions. Joints provide the position of the landmarks and

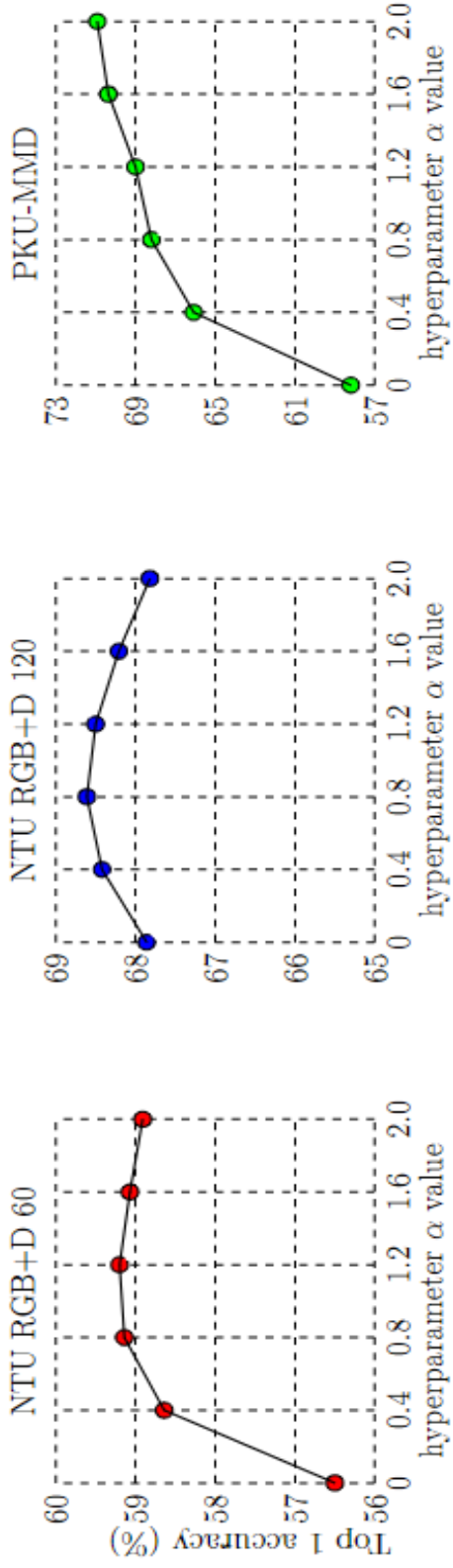


Figure 5.10: Comparisons of Top 1 accuracy (%) with different α hyperparameter settings on NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD datasets. The ablation study is computed under the full model, which contains both the CsA and DTW modules. The resolution sizes are set as 192×192.

Figure 5.11: The UMAP [10] visualization for 5-way-1-shot medical action recognition implemented by proposed MF-OSMAR on NTU-RGB+D 120 dataset with different proposed features, which are joint features(right), angle features (middle) and joint+angle features (left). The distributions of the medical actions in different colors demonstrate the classification performance. (Best viewed in the color version)

angles provide the bending magnitude of the subject bodies. Both the changing rates of these two features are also supplied in the temporal dimension. Compared to PAMMAR [30], the proposed fusion method achieves 6.1% and 6.6% improved performance for headache and back pain on NTU RGB+D 60, respectively. Similar to NTU RGB+D 120 dataset, the proposed data fusion improves vomiting by 9.6% and 7.3%, respectively. According to the results from the three datasets, particularly in PKU-MMD, falling achieved over 99% accuracy. This is because both the joint and angle of falling have noticeable changes based on the position and bending angles. There exist significant differences between the first frame and the last frame for falling. The experimental evaluation above imply that the proposed multiple-scale fusion method could considerably strengthen the one-shot learning framework performance for medical action recognition. Moreover, the UMAP [10] visualization results of medical actions from the NTU RGB+D 120 dataset are shown in Figure 5.11. The uniform manifold approximation and projection (UMAP) demonstrates the clustering relationship by dimension reduction operation of the features. The larger distance from the other clustering demonstrates higher performance. For example, as shown in Figure 5.11, falling down, staggering and stretch oneself actions are more centralized and relatively have higher performance in Table 5.5. According to Table 5.5, the performance of cough and neck pain are always relatively lower than the other medical action for the proposed one-shot learning framework. As a suggestion for future work, exploiting the feature relation between joint and angle on both spatial and temporal dimensions, which is useful for capturing the long-term latency of the trajectories of important body parts could further improve the proposed fusion method performance.

Analysis for Parameters

The experiments on the datasets are also conducted to analyze the impacts of different critical parameter settings. The proposed approach is first evaluated on different

transformed image resolutions, as illustrated in Table 5.6. Top 1 accuracy (%) is employed to investigate the relative influences on the performance. With the increase of resolutions from 32 to 192, the performance of the proposed approach is largely improved from 58.9% to 67.9%. This is because there exists a positive correlation between the image resolution and the performance. Since the images contain more information with larger image resolutions.

Table 5.6: The 5-way-1-shot accuracy (%) of the proposed method using joint features on different signal image resolutions for medical actions on NTU RGB+D 120 dataset.

Resolutions	**Baseline** [104]	**DTW**	**CsA+DTW**
32×32	33.4	56.0	**58.9**
64×64	33.7	60.7	**62.0**
96×96	37.6	59.5	**64.7**
144×144	40.7	63.5	**65.4**
160×160	42.0	64.2	**66.0**
192×192	41.8	64.6	**67.9**

Furthermore, the ablation study performance on the fusing hyperparameter α is analyzed for late fusion, which controls the fusing weight of each classifier with 6 different settings, the experimental results of different importance are shown in Figure 5.10. Since both the joint and angle contain different action features in different dimensions. In order to determine the best weight value settings for different datasets, an appropriate value is experimentally determined for α from equation 5.3.1 to manage the importance of the weight between the joint features and angle features. To this end, a set of pilot tests $\alpha = \{0, 0.4, 0.8, 1.2, 1.6, 2.0\}$ is conducted on NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD datasets. All the performance of the three datasets increases when the value of α changes from 0 to 0.4. The best important weight selections for NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD are 1.2, 0.8 and 2.0, respectively. The above experimental results demonstrate that the joint and the angle feature are complementary for enhancing the medical action recognition performance. One thing that needs to be noted is that the largest image

resolutions with the full proposed model are used and keep them fixed during this ablation study experiment.

5.5.4 Benchmark Evaluations

The proposed framework is evaluated on the NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD datasets. Quantitative experimental results compared with recent state-of-the-art approaches are shown in Table 5.7. These state-of-the-art approaches include: Attention Network [117], Fully Connected [117], Average Pooling [118], APSR [3], TCN [119], SL-DML [116], Skeleton-DML [104], SMAM-Net [78], ALCA-GCN [120], PAMMER [30], PartProtoNet [77], CTR-GCN-KP [79] and MMTS [121]. In order to conduct a fair comparison with the state-of-the-art approaches, the dataset partitioning is consistent with the SOTA methods in this part.

Table 5.7: The 5-way-1-shot human action recognition accuracy (%) comparisons with SOTA methods on NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD with general dataset partitioning. † indicates the SIG method is applied.

Approaches	**NTU-60**	**NTU-120**	**PKU-MMD**
Attention Network [117]	-	41.0	-
Fully Connected [117]	60.9	42.1	56.4
Average Pooling [118]	59.8	42.9	58.1
APSR [3]	-	45.3	-
TCN [119]	64.8	46.5	56.1
CTR-GCN-KP [79]	-	68.1	-
† SL-DML [116]	71.4	50.9	67.0
ALCA-GCN [120]	-	57.6	-
PartProtoNet [77]	-	65.6	-
† Skeleton-DML [104]	71.8	54.2	68.6
SMAM-Net [78]	73.6	56.4	70.4
† PAMMER [30]	69.9	58.3	78.5
MMTS [121]	-	69.8	-
† *MF-OSMAR*	**76.3**	**76.0**	**82.6**

As could be observed from Table 5.7, the proposed method (MF-OSMAR) reports state-of-the-art accuracy compared with the other recent state-of-the-art one-shot

learning approaches for human action recognition on NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD datasets with general dataset partitioning, which indicates the proposed method is capable to provide more valuable action feature information to the model. It is worth noting that the full proposed model, which contains both CsA and DTW modules is applied for comparison with the state-of-the-art methods in this part of the experiments by the 192×192 resolutions.

5.5.5 Failure Case

According to Table 5.5, the proposed method performs better in recognizing some of the medical actions rather than the baseline approach [30], such as "Falling down" and "Staggering", demonstrating that the proposed method of enhancing the complementary relation features in distinguishing different medical actions. However, the specific experimental results of the action "Cough" are relatively not promising, which are around 45%. This demonstrates the proposed method is insensitive in capturing the medical actions with the tiny range of movements. Presumably, this is because the pixel values of the angle features for these medical actions are relatively limited. Moreover, since the backbone is chosen to use ResNet18, the kernel size of it is fixed when taking the instances, the feature values may extend to disappear after several layers compared to the other actions. Therefore, the discriminative capability of MF-OSMAR based on barely on the actions with limited movements is insufficient for some challenging cases.

5.6 Summary

In summary, this chapter contributed to addressing the data and label-lacking issue to improve the medical action recognition performance by proposing a one-shot learning framework with a multiple-level fusion theory. By performing the one-shot learning framework including cross-attention and dynamic time warping modules, the simi-

lar medical action misclassify and temporal information mismatching issues are both effectively mitigated. The signal image transformation approach was proposed to alleviate the privacy leakage issue by transforming the human skeleton sequence into the RGB image. By employing the multiple-level fusion method, the medical action features were enriched and the misclassified errors were alleviated, thereby maximizing the complementary benefits between the joint features and the angle features for various similar medical actions. In Section 5.5.3, the detailed performance analysis including the extensive ablation study and visualization results was conducted to exploit the contributions from different proposed components. Evaluations on the NTU RGB+D 60, NTU RGB+D 120 and PKU-MMD with general partitioning in Section 5.5.4 were provided to verify the effectiveness of the proposed approach compared with other state-of-the-art methods.

Furthermore, the failure case of medical action recognition was given and discussed in Section 5.5.5. In future work, techniques for extracting more robustness and deeper feature together with occlusion-aware medical action recognition will be exploited for better real-world scenarios generalization ability.

Chapter 6

CONCLUSIONS AND FUTURE WORK

Privacy mitigating algorithms provide a variety of solutions for addressing the existing issues in medical action recognition tasks, including data limitations, noisy annotations and privacy protection. This thesis focuses on the multiple human fall events classification task rather than the conventional video-based binary fall detection task in the first and second contribution chapters since fall events play a key role in the reasons for aging population death and different fall events result in varying degrees of injury. The proposed algorithms are then generalized to medical action recognition for a border application field in the third contribution chapter. Overall, this thesis has effectively accomplished the six objectives aforementioned in Chapter 1, by developing three different methods which are described in detail in the three contribution chapters to address the challenging issues for medical action recognition. Extensive experimental evaluations on various medical action recognition-related benchmark datasets verify the effectiveness of the proposed algorithms, along with detailed investigation in both quantitative and qualitative ways were provided to compare with the recent state-of-the-art methods.

In the final chapter, the contributions of this thesis will be summarized along with the limitations in Section 6.1. Section 6.2 will provide future research plans that could further improve the robustness and performance of medical action recognition.

6.1 Conclusions

In Chapter 3, a new redundant data reduction theory was first developed in the data processing stage. The main functionality of this method was to enhance the data quality by removing unimportant information including the redundant empty frames and the redundant body parts related to the fall events. The main advantage is that this method tackles redundant information in the data processing stage, therefore it is plugin into either the data-driven methods or the model-driven methods. For the RF classification method, the proposed redundant data reduction method achieved 7% to 11% improvements for different fall events. Furthermore, to deal with the imbalanced data issue for multiple fall events, a two-stage deep neural network-based framework was designed to efficiently filter the normal human activities in the initial stage and focused on multiple fall events classification in the conclusive stage. The gating parameter settings were made to control the data filtering ability in the initial stage. The proposed methods achieved approximately 2% to 3% improvements in multiple fall events classification on the extracted privacy-mitigated skeleton data from the UP-Fall dataset. Evaluations on the UP-Fall benchmark dataset further verified the effectiveness of the proposed methods compared with other state-of-the-art methods. Both of the detailed experimental results analysis and discussion were presented in Chapter 3.

Chapter 4 was particularly dedicated to demonstrating the paramount importance of annotation quality to data-driven methods, and thus proposed a novel noisy annotations managing system. There are two primary components included in the proposed system: noisy annotation purification and noisy label learning. For the proposed noisy annotation purification method, the entire corrupted dataset was fed into the model for obtaining the initial prediction coarse. Followed by generating the confidence joint counting matrix and counting the noisy annotation from the raw corrupted dataset with the refined confidence joint probability distribution matrix.

After that, four noisy annotation pruning methods were proposed for purifying the corrupted dataset. The proposed noisy label learning algorithm is developed with a trinity network to fully exploit the potential of the noisy instance, which includes two teacher modules and one student module. The peer network structures were utilized for mining the clean instances. The small loss algorithm was used for distinguishing the noisy annotations since the clean data performs relatively small loss values at the initial training process. The co-regularization algorithm was utilized as a contrastive term in the proposed Kullback-Leibler loss function for maximizing the clean data agreement. Moreover, cross update approach was used for exchanging the perceived clean instance during the training process between the peer networks. Furthermore, the consensus-based data selection theory was developed to enhance the robustness of the proposed system by allocating the clean data agreement from the teacher modules to guide the student module training. Evaluations on the UP-Fall dataset confirmed the efficacy of the proposed purification method in Chapter 4, which achieved 3% to 7% improvements for different fall events, respectively. Moreover, experiments with different estimated noise types and noise rates on the UP-Fall dataset were also provided. The proposed JoCoT achieved approximately 4% and 2% improvements for the average of all rate settings for the pairflip and symmetric noise, respectively. For the deep corruption, JoCoT achieved 5% and 3% improvements, respectively. Improved experimental results demonstrated the superiority of the proposed noisy label learning algorithm in discriminating the noisy instances and promoting classification performance in different noisy situations.

In Chapter 5, the entire medical action recognition process was typically established from a multiple-level fusion perspective. The first contribution targeted at addressing the issues of data limitation by proposing a one-shot learning framework with both the cross-attention mechanism and dynamic time warping module. Since similar medical action may lead to misclassification and decreased performance, the

cross-attention mechanism aims to guide the model more concerned with the important landmarks for each medical action. By aligning the temporal information between the instances, the dynamic time warping module effectively addressed the temporal mismatching issue. In order to further mitigate the privacy leakage issue, both the position and direction-level features were distilled and transformed into the proposed signal-level images, since they were able to better protect the privacy information compared with the features from the unprocessed raw human skeleton sequences. Then, the proposed multiple-level fusion was developed to investigate the complementary benefits between the direction-level and position-level features as the second contribution, such a multiple-level fusion approach was performed promising to remedy the shortage of the usage of single features. Evaluations on the NTU-60, NTU-120 and PKU-MMD benchmark datasets confirmed the advantageous performance of the proposed method compared with the recent state-of-the-art one-shot learning medical action recognition methods in both visualization and statistics perspectives. Moreover, extensive ablation study experiments confirmed the efficacy of each proposed module. Furthermore, for the benchmark evaluation with the general dataset partitioning, the proposed MF-OSMAR achieved 2.7%, 6.2% and 4.1% improvements for NTU-60, NTU-120 and PKU-MMD benchmark datasets, respectively.

6.2 Future Work

Although the proposed approaches have achieved improved and promising performance in the medical action recognition research area, there are still several limitations that can be addressed for further comprehensive this study. This section aims to provide some prospective research directions that may insight any researchers planning to dedicate to this research field.

In this thesis, human fall events classification as one of the computer vision tasks has been exploited to achieve improved performance by removing redundant informa-

tion and addressing the imbalanced data issue. However, the applied dataset is mostly containing a single target from a single camera perspective, this technique can obtain a better generalization ability to the real-world environment which may contain multiple targets with fall events by combining the person re-identification and multiple human tracking algorithms. It is also possible that other kinds of approaches such as semantic segmentation, multiple-view understanding and multimodal learning, can be available for better classification of human fall events.

Even though this thesis has addressed the noisy annotation issue by pruning the noisy instance and learning with the noisy instance. It should be noted that performing a noisy annotation correction theory during the training process can also achieve promising performance. This theory requires an effective noisy instance selection mechanism that can accurately filter the incorrect annotation, along with a powerful annotation correction ability either based on the designed network architecture or joint probability density distribution matrix. In this way, the corrupted dataset can be purified during the training process by taking advantage of cleaning the dataset without removing instances and training with correcting. Moreover, estimating the density distribution matrix of noisy annotations from the corrupted dataset is also a good option to particularly deal with real-world noise rather than manually generated noise.

In the medical action recognition research area, addressing the widespread occlusion and frequent occlusion is still a challenging issue, and its quality directly affects the medical action recognition performance. Consequently, various future research towards to tackle this issue. Firstly, the position and posture of the occluded target during the occlusions can be estimated by historical information. Recent possible solutions based on deep learning methods to accomplish this issue include Self-Supervised Learning and Transformers. Secondly, the target with frequent occlusion can be beneficial by incorporating the person re-identification algorithm, thereby bet-

ter recognising the desired target during the occlusion.

Due to the distinctive attributes of medical action recognition, such as fall events holds significant lethality for aging population. Therefore, an accurate medical action prediction system is required rather than mere recognition after their occurrence. This system requires an effective human medical action forecasting mechanism that could accurately predict the next movement of the target based on the historical trajectories, which guide the system to predict the next occurring events according to past events. Understanding human motion can be applied by utilizing such as Federated Learning, Spatial-Temporal Networks, Self-Supervised Learning and Transformers. To this end, more accurate human medical action trajectories can be predicted according to the aforementioned deep learning methods but the conventional linear prediction algorithms can be prevented or improved.

References

[1] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, "Up-fall detection dataset: A multimodal approach," *Sensors*, vol. 19, no. 9, p. 1988, 2019.

[2] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1010–1019.

[3] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 42, no. 10, pp. 2684–2701, 2019.

[4] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," *arXiv preprint arXiv:1703.07475*, 2017.

[5] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *arXiv preprint arXiv:1804.06872*, 2018.

[6] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 7164–7173.

[7] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 726–13 735.

[8] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *ICCV*, 2017.

[9] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[10] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[11] 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. 33 532–33 542, 2021.

[12] L. Ren and Y. Peng, "Research of fall detection and fall prevention technologies: A systematic review," *IEEE Access*, vol. 7, pp. 77 702–77 722, 2019.

[13] X. Wang, J. Ellul, and G. Azzopardi, "Elderly fall detection systems: A literature survey," *Frontiers in Robotics and AI*, vol. 7, pp. 71–94, 2020.

[14] S. Usmani, A. Saboor, M. Haris, M. A. Khan, and H. Park, "Latest research trends in fall detection and prevention using machine learning: A systematic review," *Sensors*, vol. 21, no. 15, p. 5134, 2021.

[15] J. S. McPhee, D. P. French, D. Jackson, J. Nazroo, N. Pendleton, and H. Degens, "Physical activity in older age: perspectives for healthy ageing and frailty," *Biogerontology*, vol. 17, pp. 567–580, 2016.

[16] C. Qiu, G. Johansson, F. Zhu, M. Kivipelto, and B. Winblad, "Prevention of cognitive decline in old age—varying effects of interventions in different populations," *Annals of Translational Medicine*, vol. 7, no. Suppl 3, 2019.

[17] J. Yin, J. Han, C. Wang, B. Zhang, and X. Zeng, "A skeleton-based action recognition system for medical condition detection," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2019.

[18] G. Yao, T. Lei, and J. Zhong, "A review of convolutional-neural-network-based action recognition," *Pattern Recognition Letters*, vol. 118, pp. 14–22, 2019.

[19] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, pp. 2259–2322, 2021.

[20] T. Özyer, D. S. Ak, and R. Alhajj, "Human action recognition approaches with video datasets—a survey," *Knowledge-Based Systems*, vol. 222, p. 106995, 2021.

[21] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, 2019.

[22] T. Xu, Y. Zhou, and J. Zhu, "New advances and challenges of fall detection systems: A survey," *Applied Sciences*, vol. 8, no. 3, pp. 418–428, 2018.

[23] Z. Zhang, C. Conly, and V. Athitsos, "A survey on vision-based fall detection," in *Proceedings of the 8th ACM international conference on PErvasive technologies related to assistive environments*, 2015, pp. 1–7.

[24] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A survey on active deep learning: from model driven to data driven," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–34, 2022.

[25] Q. Sun and Z. Ge, "A survey on deep learning for data-driven soft sensors," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 5853–5866, 2021.

[26] L. Xie, Y. Yang, F. Zeyu, and S. M. Naqvi, "Skeleton-based fall events classification with data fusion," in *2021 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2021, pp. 1–6.

[27] L. Xie, Y. Sun, J. A. Chambers, and S. M. Naqvi, "Two-stage fall events classification with human skeleton data," *arXiv preprint arXiv:2208.12027*, 2022.

[28] L. Xie, Y. Sun, J. Chambers, and S. M. Naqvi, "Privacy preserving multi-class fall classification based on cascaded learning and noisy labels handling," in *2022 25th International Conference on Information Fusion (FUSION)*. IEEE, 2022, pp. 1–6.

[29] L. Xie, Y. Sun, and S. M. Naqvi, "Learning with noisy labels for human fall events classification: joint cooperative training with trinity networks." 2022.

[30] L. Xie, Y. Yang, Z. Fu, and S. M. Naqvi, "One-shot medical action recognition with a cross-attention mechanism and dynamic time warping," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[31] L. Xie, Y. Yang, Z. Fu, and M. Naqvi, "Mf-osmar: Multiple level fusion of one-shot learning for privacy preserved skeletal human medical action recognition," 2023.

[32] B. Aguiar, T. Rocha, J. Silva, and I. Sousa, "Accelerometer-based fall detection for smartphones," in *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2014, pp. 1–6.

[33] F. A. S. F. de Sousa, C. Escriba, E. G. A. Bravo, V. Brossa, J.-Y. Fourniols, and C. Rossi, "Wearable pre-impact fall detection system based on 3d accelerometer and subject's height," *IEEE Sensors Journal*, vol. 22, no. 2, pp. 1738–1745, 2021.

[34] L. Chen, R. Li, H. Zhang, L. Tian, and N. Chen, "Intelligent fall detection method based on accelerometer data from a wrist-worn smart watch," *Measurement*, vol. 140, pp. 215–226, 2019.

[35] A. Jefiza, E. Pramunanto, H. Boedinoegroho, and M. H. Purnomo, "Fall detection based on accelerometer and gyroscope using back propagation," in *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. IEEE, 2017, pp. 1–6.

[36] Q. T. Huynh, U. D. Nguyen, L. B. Irazabal, N. Ghassemian, B. Q. Tran *et al.*, "Optimization of an accelerometer and gyroscope-based fall detection algorithm," *Journal of Sensors*, vol. 2015, 2015.

[37] Q. T. Huynh, U. D. Nguyen, S. V. Tran, A. Nabili, and B. Q. Tran, "Fall detection system using combination accelerometer and gyroscope," in *Proc. of the Second Int. l Conf. on Advances in Electronic Devices and Circuits (EDC 2013)*, 2013.

[38] P. Youngkong and W. Panpanyatep, "A novel double pressure sensors-based monitoring and alarming system for fall detection," in *2021 Second International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP)*. IEEE, 2021, pp. 1–5.

[39] C. M. Lee, J. Park, S. Park, and C. H. Kim, "Fall-detection algorithm using plantar pressure and acceleration data," *International Journal of Precision Engineering and Manufacturing*, vol. 21, pp. 725–737, 2020.

[40] H.-W. Tzeng, M.-Y. Chen, and J.-Y. Chen, "Design of fall detection system with floor pressure and infrared image," in *2010 International Conference on system science and engineering*. IEEE, 2010, pp. 131–135.

[41] A. K. Bourke and G. M. Lyons, "A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor," *Medical engineering & physics*, vol. 30, no. 1, pp. 84–90, 2008.

[42] X. Li, L. Nie, H. Xu, and X. Wang, "Collaborative fall detection using smart phone and kinect," *Mobile Networks and Applications*, vol. 23, pp. 775–788, 2018.

[43] A. Singh, S. U. Rehman, S. Yongchareon, and P. H. J. Chong, "Sensor technologies for fall detection systems: A review," *IEEE Sensors Journal*, vol. 20, no. 13, pp. 6889–6919, 2020.

[44] A. H. M. Yusoff, S. M. Salleh, and M. O. Tokhi, "Towards understanding on the development of wearable fall detection: an experimental approach," *Health and Technology*, vol. 12, no. 2, pp. 345–358, 2022.

[45] S. N. Patel, A. Lathigara, V. Y. Mehta, and Y. Kumar, "A survey on vision-based elders fall detection using deep learning models," in *Futuristic Trends in Networks and Computing Technologies*. Springer, 2022, pp. 447–465.

[46] M. J. A. Nahian, M. H. Raju, Z. Tasnim, M. Mahmud, M. A. R. Ahad, and M. S. Kaiser, "Contactless fall detection for the elderly," *Contactless Human Activity Analysis*, pp. 203–235, 2021.

[47] A. Ramachandran and A. Karuppiah, "A survey on recent advances in wearable fall detection systems," *BioMed research international*, vol. 2020, 2020.

[48] A. Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, "Vision-based fall detection with convolutional neural networks," *Wireless communications and mobile computing*, vol. 2017, 2017.

[49] R. Espinosa, H. Ponce, S. Gutiérrez, L. Martínez-Villaseñor, J. Brieva, and E. Moya-Albor, "A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the up-fall detection dataset," *Computers in biology and medicine*, vol. 115, p. 103520, 2019.

[50] C. A. U. Hassan, F. K. Karim, A. Abbas, J. Iqbal, H. Elmannai, S. Hussain, S. S. Ullah, and M. S. Khan, "A cost-effective fall-detection framework for the elderly using sensor-based technologies," *Sustainability*, vol. 15, no. 5, p. 3982, 2023.

[51] X. Ma, H. Wang, B. Xue, M. Zhou, B. Ji, and Y. Li, "Depth-based human fall detection via shape features and improved extreme learning machine," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1915–1922, 2014.

[52] Z. Zhang, C. Conly, and V. Athitsos, "Evaluating depth-based computer vision methods for fall detection under occlusions," in *Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II 10*. Springer, 2014, pp. 196–207.

[53] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 3, pp. 489–501, 2014.

[54] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature medicine*, vol. 25, no. 1, pp. 37–43, 2019.

[55] Y. Huang, Y. J. Li, and Z. Cai, "Security and privacy in metaverse: A comprehensive survey," *Big Data Mining and Analytics*, vol. 6, no. 2, pp. 234–247, 2023.

[56] V. Srivastav, A. Gangi, and N. Padoy, "Human pose estimation on privacy-preserving low-resolution depth images," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2019, pp. 583–591.

[57] M. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[58] E. Chou, M. Tan, C. Zou, M. Guo, A. Haque, A. Milstein, and L. Fei-Fei, "Privacy-preserving action recognition for smart hospitals using low-resolution depth images," *arXiv preprint arXiv:1811.09950*, 2018.

[59] Y. Wang, Z. Cheng, X. Yi, Y. Kong, X. Wang, X. Xu, Y. Yan, C. Yu, S. Patel, and Y. Shi, "Modeling the trade-off of privacy preservation and activity recognition on low-resolution images," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–15.

[60] J. Dai, J. Wu, B. Saghafi, J. Konrad, and P. Ishwar, "Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 68–76.

[61] M. U. Kim, H. Lee, H. J. Yang, and M. S. Ryoo, "Privacy-preserving robot vision with anonymized faces by extreme low resolution," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 462–467.

[62] N. Miyazaki, K. Tsuji, M. Zheng, M. Nakashima, Y. Matsuda, and E. Segawa, "Privacy-conscious human detection using low-resolution video," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 326–330.

[63] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 026–12 035.

[64] C. Plizzari, M. Cannici, and M. Matteucci, "Spatial temporal transformer network for skeleton-based action recognition," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*. Springer, 2021, pp. 694–701.

[65] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," *arXiv preprint arXiv:2002.05907*, 2020.

[66] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Computer Vision and Image Understanding*, vol. 208, p. 103219, 2021.

[67] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7912–7921.

[68] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[69] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[70] G. Algan and I. Ulusoy, "Image classification with deep learning in the presence of noisy labels: A survey," *Knowledge-Based Systems*, vol. 215, p. 106771, 2021.

[71] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical image analysis*, vol. 65, p. 101759, 2020.

[72] Z. Li, K. Kamnitsas, and B. Glocker, "Analyzing overfitting under class imbalance in neural networks for image segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 3, pp. 1065–1077, 2020.

[73] S. Park, J. Lim, Y. Jeon, and J. Y. Choi, "Influence-balanced loss for imbalanced visual classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 735–744.

[74] M. Saqlain, Q. Abbas, and J. Y. Lee, "A deep convolutional neural network for wafer defect identification on an imbalanced dataset in semiconductor manufacturing processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 3, pp. 436–444, 2020.

[75] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[76] L. Wang, J. Liu, and P. Koniusz, "3d skeleton-based few-shot action recognition with jeanie is not so na\" ive," *arXiv preprint arXiv:2112.12668*, 2021.

[77] T. Chen, D. Zhou, J. Wang, S. Wang, Q. He, C. Hu, E. Ding, Y. Guan, and X. He, "Part-aware prototypical graph network for one-shot skeleton-based action recognition," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023, pp. 1–8.

[78] Z. Li, X. Gong, R. Song, P. Duan, J. Liu, and W. Zhang, "SMAM: Self and mutual adaptive matching for skeleton-based few-shot action recognition," *IEEE Transactions on Image Processing (TIP)*, vol. 32, pp. 392–402, 2022.

[79] X. Wang, X. Xu, and Y. Mu, "Neural koopman pooling: Control-inspired temporal dynamics encoding for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10 597–10 607.

[80] Y. Hua, W. Wu, C. Zheng, A. Lu, M. Liu, C. Chen, and S. Wu, "Part aware contrastive learning for self-supervised action recognition," *arXiv preprint arXiv:2305.00666*, 2023.

[81] F. Askari, R. Jiang, Z. Li, J. Niu, Y. Shi, and J. J. Clark, "Self-supervised video interaction classification using image representation of skeleton data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5228–5237.

[82] V. Rani, S. T. Nabi, M. Kumar, A. Mittal, and K. Kumar, "Self-supervised learning: A succinct review," *Archives of Computational Methods in Engineering*, vol. 30, no. 4, pp. 2761–2775, 2023.

[83] Z. Xing, Q. Dai, H. Hu, J. Chen, Z. Wu, and Y.-G. Jiang, "Svformer: Semi-supervised video transformer for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 816–18 826.

[84] B. Xu and X. Shu, "Pyramid self-attention polymerization learning for semi-supervised skeleton-based action recognition," *arXiv preprint arXiv:2302.02327*, 2023.

[85] Z. Tu, J. Zhang, H. Li, Y. Chen, and J. Yuan, "Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition," *IEEE Transactions on Multimedia*, 2022.

[86] B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama, "A survey of label-noise representation learning: Past, present and future," *arXiv preprint arXiv:2011.04406*, 2020.

[87] E. Malach and S. Shalev-Shwartz, "Decoupling" when to update" from" how to update"," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[88] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 233–242.

[89] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International Conference on Machine Learning, ICML*. PMLR, 2018.

[90] F. R. Cordeiro and G. Carneiro, "A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations?" in *2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*. IEEE, 2020, pp. 9–16.

[91] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[92] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[93] Z. Hu, Y. Dong, K. Wang, K.-W. Chang, and Y. Sun, "Gpt-gnn: Generative pre-training of graph neural networks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1857–1867.

[94] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, "Qa-gnn: Reasoning with language models and knowledge graphs for question answering," *arXiv preprint arXiv:2104.06378*, 2021.

[95] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, and J. Lafferty, "Clustering with bregman divergences." *Journal of machine learning research*, vol. 6, no. 10, 2005.

[96] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi, "Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2277–2291, 2019.

[97] Z. Fu, X. Lai, and S. M. Naqvi, "Enhanced detection reliability for human tracking based video analytics," in *22th International Conference on Information Fusion (FUSION)*, 2019.

[98] Y. Yan, R. Rosales, G. Fung, R. Subramanian, and J. Dy, "Learning from multiple annotators with varying expertise," *Machine learning*, vol. 95, no. 3, pp. 291–327, 2014.

[99] X. Yu, T. Liu, M. Gong, and D. Tao, "Learning with biased complementary labels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[100] A. Blum, A. Kalai, and H. Wasserman, "Noise-tolerant learning, the parity problem, and the statistical query model," *Journal of the ACM (JACM)*, vol. 50, no. 4, pp. 506–519, 2003.

[101] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021.

[102] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 447–461, 2015.

[103] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, "Delving deep into one-shot skeleton-based action recognition with diverse occlusions," *IEEE Transactions on Multimedia (TMM)*, 2023.

[104] R. Memmesheimer, S. Häring, N. Theisen, and D. Paulus, "Skeleton-DML: Deep metric learning for skeleton-based one-shot action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[105] Y. Zou, Y. Shi, D. Shi, Y. Wang, Y. Liang, and Y. Tian, "Adaptation-oriented feature projection for one-shot action recognition," *IEEE Transactions on Multimedia (TMM)*, vol. 22, no. 12, pp. 3166–3179, 2020.

[106] Y. Han, T. Zhuo, P. Zhang, W. Huang, Y. Zha, Y. Zhang, and M. Kankanhalli, "One-shot video graph generation for explainable action reasoning," *Neurocomputing*, vol. 488, pp. 212–225, 2022.

[107] D. D. Ram, U. Muthukumaran, and N. S. Fatima, "Enhanced human action recognition with ensembled dtw loss function in cnn lstm architecture," in *Proceedings of Third International Conference on Sustainable Expert Systems: ICSES 2022*. Springer, 2023, pp. 491–508.

[108] N. Ma, H. Zhang, X. Li, S. Zhou, Z. Zhang, J. Wen, H. Li, J. Gu, and J. Bu, "Learning spatial-preserved skeleton representations for few-shot action recognition," in *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, 2022.

[109] D. Ahn, S. Kim, H. Hong, and B. C. Ko, "STAR-Transformer: A spatio-temporal cross attention transformer for human action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3330–3339.

[110] T.-K. Kang, G.-H. Lee, K.-M. Jin, and S.-W. Lee, "Action-aware masking network with group-based attention for temporal action localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6058–6067.

[111] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 9532–9545, 2020.

[112] F. Meng, H. Liu, Y. Liang, J. Tu, and M. Liu, "Sample fusion network: An end-to-end data augmentation network for skeleton-based human action recognition," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 11, pp. 5281–5295, 2019.

[113] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi, "Multi-level cooperative fusion of gm-phd filters for online multiple human tracking," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2277–2291, 2019.

[114] F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, "2d pose-based real-time human action recognition with occlusion-handling," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1433–1446, 2019.

[115] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[116] R. Memmesheimer, N. Theisen, and D. Paulus, "Signal level deep metric learning for multimodal one-shot action recognition," *arXiv preprint arXiv:2004.11085*, 2020.

[117] J. Liu, G. Wang, P. Hu, L. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[118] J. Liu, G. Wang, L. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2017.

[119] A. S., "One-shot action recognition towards novel assistive therapies," *arXiv preprint arXiv:2102.08997*, 2021.

[120] A. Zhu, Q. Ke, M. Gong, and J. Bailey, "Adaptive local-component-aware graph convolutional network for one-shot skeleton-based action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[121] J. Lee, M. Sim, and H.-J. Choi, "MMTS: Multimodal teacher-student learning for one-shot human action recognition," in *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2023, pp. 235–242.