Expansion-Squeeze-Excitation Fusion Network for Elderly Activity Recognition

Xiangbo Shu, Jiawen Yang, Rui Yan, and Yan Song

Abstract—This work focuses on the task of elderly activity recognition, which is a challenging task due to the existence of individual actions and human-object interactions in elderly activities. Thus, we attempt to effectively aggregate the discriminative information of actions and interactions from both RGB videos and skeleton sequences by attentively fusing multi-modal features. Recently, some nonlinear multi-modal fusion approaches are proposed by utilizing nonlinear attention mechanism that is extended from Squeeze-and-Excitation Networks (SENet). Inspired by this, we propose a novel Expansion-Squeeze-Excitation Fusion Network (ESE-FN) to effectively address the problem of elderly activity recognition, which learns modal and channel-wise Expansion-Squeeze-Excitation (ESE) attentions for attentively fusing the multi-modal features in the modal and channel-wise ways. Specifically, ESE-FN firstly implements the modal-wise fusion with the Modal-wise ESE Attention (M-ESEA) to aggregate discriminative information in modal-wise way, and then implements the channel-wise fusion with the Channel-wise ESE Attention (C-ESEA) to aggregate the multi-channel discriminative information in channel-wise way (referring to Figure 1). Furthermore, we design a new Multi-modal Loss (ML) to keep the consistency between the single-modal features and the fused multi-modal features by adding the penalty of difference between the minimum prediction losses on single modalities and the prediction loss on the fused modality. Finally, we conduct experiments on a largest-scale elderly activity dataset, i.e., ETRI-Activity3D (including 110,000+ videos, and 50+ categories), to demonstrate that the proposed ESE-FN achieves the best accuracy compared with the state-of-the-art methods. In addition, more extensive experimental results show that the proposed ESE-FN is also comparable to the other methods in terms of normal action recognition task.

Index Terms—Elderly activity recognition, Activity recognition, Fusion network, Multi-modal fusion.

I. INTRODUCTION

With the increasing population of the society, the aging of the population is becoming more and more serious. Coupled with the absence of young people from home all year round, the proportion of empty-nesters in many countries is increasing sharply. Since the physical quality and movement ability of the elderly begin to decline, some kinds of dangers are inclined to occur in life. In such circumstances, some institutions are trying to provide intelligent monitoring for the daily activities of the elderly by employing some advanced technologies in the fields of artificial intelligence and computer vision in recent years. As the key to intelligent monitoring, elderly activity recognition is attracting increased attentions.

By reviewing the normal action recognition methods, we can divide these methods into two categories based on the types of input data, i.e., RGB-based action recognition approach [1]–[8], and skeleton-based action recognition approach [9]–[15]. For these two types of action recognition methods, various deep neural networks, as the currently main models, have shown remarkable ability to model human actions, such as Convolution Neural Network (CNN) [14], Recurrent Neural Network (RNN) [15], Convolutional 3D (C3D) [16], Long short-term memory (LSTM) [17]. Graph Convolutional Networks (GCN) [18], and so on. Generally, the RGB-based action recognition approach mainly gets motion information of actions from RGB videos, which is disturbed by the background information to some extent. The skeleton-based action recognition approach faces a challenge for recognizing the actions with similar postures. Thus, a natural way is to jointly model motion information from both RGB videos and skeleton sequences [19], [20]. One impressive solution of these methods is to build two-branch deep neural networks, that firstly learn multi-modal features from RGB and skeleton modalities, and then fuse them.

Compared with normal action recognition, elderly activity recognition is a more challenging task due to the existence of individual actions and human-object interactions in elderly activities, where many human-object interactions are local; the amplitudes of many elderly activities are unapparent; and the movements of some elderly activities are particularly similar. For example, Figure 2 shows two groups of the typical elderly activities, i.e., “Blowing hair” vs “Combing hair”, and “Making call” vs “playing mobile”. For the comparison between “Blowing hair” and “Combing hair”, the motion trajectory and amplitude are very similar in spatiotemporal space, except for the subtle clues of different objects in hand, i.e., comb and blower. For the comparison between “Making call” and “Playing mobile”, most motions and the interacting object...
Playing mobile
Making call
Blowing hair
Combing hair
(a) Comparison between Blowing hair and Combing hair.
(b) Comparison between Making call and Playing mobile.

Fig. 2. Examples of elderly activities. The elderly activities, “Blowing hair” vs “Combing hair”, as well as “Making call” vs “Playing mobile” shows most of similar motions in spatiotemporal space, where only the interacting objects (e.g., comb and blower are marked by yellow circle) are different or only few local movements (right-hand movements are marked by red circle) are distinguishable.

(e.g., mobile) are the same as each other, except for local hand movement in some frames. Therefore, how to capture and fuse the discriminative information in RGB and skeleton modalities is crucial for modeling elderly activities.

Multi-modal data or features fusion across different modalities has always been a hot topic for years [21]–[26]. Since multi-modal features often contain irrelevant (modal-specific) information, the straightforward feature-level or score-level fusion will degrade the discriminative ability of the fused features. The key to successful fusion is how to reinforce the discriminative information while suppressing the irrelevant information among multi-modal features. To this end, some works [27]–[29] propose to utilize the linear attention to selectively fuse multi-modal features. For example, Hori et al. [28] propose a multi-modal attention model that selectively fuses multi-modal features based on different attention factors. Witnessing the impressive classification performance of Squeeze-and-Excitation Networks (SENet) [30] on ImageNet, some researchers [31]–[34] extend linear fusion to the nonlinear fusion via Squeeze-and-Excitation (SE). For example, He et al. [31] propose an improved temporal Xception network to integrate multi-modal information by combining both the early and later fusion of multiple modalities. Kuang et al. [32] utilize a multi-layer fusion network to fuse the features from different layers, which fuses the low, mid, and high-level information from different modalities in a unified framework.

In this work, we consider designing a couple of Squeeze-and-Excitation based networks as nonlinear attention mechanisms for effectively aggregating the motion information in video and skeleton modalities by additionally bringing in expansion, called Expansion-Squeeze-Excitation (ESE) in this work. First, we design a new Modal-fusion Net (M-Net) consisting of several convolutional layers, a global average pooling layer, and several fully connected layers to capture
the local and global modal-wise dependencies among features via a modal-wise ESE, as shown in Figure 1. Second, we design a new Channel-fusion Net (C-Net) consisting of a convolutional layer, a global average pooling layer, and several fully connected layers to capture the channel-wise dependencies among features via a channel-wise ESE. By regarding M-Net and C-Net as the modal and channel-wise nonlinearity mechanisms, we build a modal-fusion module and a channel-fusion module to fuse multi-modal features with Modal-wise Expansion-Squeeze-Excitation Attention (M-ESEA) and Channel-wise Expansion-Squeeze-Excitation Attention (C-ESEA). The former can be seen as a rough fusion way, while the latter can be seen as a fine-grained fusion way.

Formally, we propose a novel Expansion-Squeeze-Excitation Fusion Network (ESE-FN) to jointly adopt the modal and channel-wise fusions with the corresponding ESE attentions for capturing modal and channel-wise dependencies among features. The overview framework of ESE-FN is shown in Figure 3, which mainly consists of four parts, i.e., feature extractor module, modal-fusion module, channel-fusion module, and multi-modal loss. First, ESE-FN randomly samples RGB frames in different video clips, which are fed into the backbone to extract the RGB features. Likewise, the skeleton features can be also extracted from sampled skeleton sequences. Second, the concatenated RGB and skeleton features are fed into the modal-fusion module for attentively fusing features in modal-wise way, where M-Net aims to learn the Modal-wise Expansion-Squeeze-Excitation Attention (M-ESEA). Third, the output from the modal-fusion module is further fed into the channel-fusion module for attentively fusing features in channel-wise way, where C-Net aims to learn the Channel-wise Expansion-Squeeze-Excitation Attention (C-ESEA). Finally, we utilize three types of features, i.e., RGB features, skeleton features, fused multi-modal features, to construct three sub-losses, which are integrated into a new Multi-modal Loss (ML) for optimizing the whole network. Here, compared with traditional recognition loss, ML additionally brings in the penalty of difference between the minimum prediction losses on single modalities and the prediction loss on the fused modality, which can keep the consistency between the single-modal features and the fused multi-modal features.

Overall, the main contributions of this work can be summarized as follows.

- We deeply explore the characteristics of elderly activities, and propose a novel Expansion-Squeeze-Excitation Fusion Network (ESE-FN) to attentively fuse multi-modal features in the modal and channel-wise ways for effectively addressing the problem of elderly activity recognition.
- To well capture the discriminative information of multi-modal features, we design a flexible Modal-fusion Net (M-Net) and Channel-fusion Net (C-Net) to learn Modal-wise Expansion-Squeeze-Excitation Attention (M-ESEA) and Channel-wise Expansion-Squeeze-Excitation Attention (C-ESEA) for capturing the modal and channel-wise dependencies among features, respectively.
- To keep the consistency between the single-modal features and the fused multi-modal features, we design a new Multi-modal Loss (ML) to additionally measure the difference between the minimum prediction loss on single modalities and the prediction loss on the fused modality.
- By conducting extensive experiments on both elderly activity recognition and normal action recognition tasks, we illustrate that the proposed ESE-FN method achieves the SOTA performance compared with the other competitive methods.

The rest of this paper is organized in the following. Section II surveys some works related to RGB-based action recognition, skeleton-based action recognition, and multi-modal fusion. Section III introduces the proposed method for elderly activity recognition in details. Section IV presents results...
and analysis of experiments, followed by the conclusions in Section V.

II. RELATED WORK

We survey some works related to RGB-based action recognition, skeleton-based action recognition, and multi-modal fusion.

A. RGB-based Action Recognition

Most existing RGB-based Action Recognition methods can be categorized into two classes. The first class is based on a two-stream network that usually uses RGB and optical flow to model spatial and temporal information, respectively. Karen et al. [35] proposed a two-stream network to model spatial and temporal information in RGB and optical flow frames for the first time. Subsequently, Wang et al. [36] proposed a Temporal Segment Network (TSN) to model long-range temporal action. Crasto et al. [37] proposed a distillation network that uses optical flow to distill RGB data, which can make the RGB-based model learn the temporal information. It is well known that optical flow is computed by using the RGB frames, which is time-consuming and would bring in a bottleneck. The second class is based on a series of 3D convolutional networks, such as, C3D [16], I3D [7], T3D [38], Res3D [39], and so on, which are extended from 2D networks in spatiotemporal dimension. Due to the computation consumption of general 3D convolutional networks, Qiu et al. [40] proposed a Pseudo-3D residual network (P3D) that decomposes the convolutions into separate 2D spatial and 1D temporal filters. Moreover, Feichtenhofer et al. [3] proposed to use two different frame rates to accelerate train 3D networks.

The above-mentioned methods are proposed to address the problem of normal action recognition. Compared with the normal action recognition task, the elderly activity recognition task requires more discriminative information to identify some subtle individual actions and human-object interactions in elderly activities. Due to the lack of temporal information on RGB data, the network implementing on RGB data cannot accurately describe the actions with obvious temporal information. Although some temporal information can be captured from optical flow data, the calculation of optical flow is too time-consuming.

B. Skeleton-based Action Recognition

So far, there are many action recognition works based on skeleton [9], [10], [11], [13], [41], called skeleton-based action recognition methods. The main target of skeleton-based action recognition is to learn temporal and spatial information from skeleton sequences. In the early stages, researchers utilized various deep neural networks, e.g., Recurrent Neural Network (RNN), Convolution Neural Network (CNN), and Long Short-Term Memory (LSTM), to model skeleton motions for capturing temporal and spatial information. For example, Zhang et al. [9] proposed an adaptive model designed by CNN and RNN to solve the problem of view difference during the skeleton data collection and action shooting. Banerjee et al. [10] used CNN to extract different complementary motion information from skeleton sequences, e.g., distance and angle between joint points, to better model the temporal information of skeleton data. Recently, researchers used Graph Convolution Network (GCN) to model temporal and spatial information from skeleton sequences by treating skeletons, joints, and bonds as graphs, nodes, and edges, respectively [11], [12], [42]. For example, Yan et al. [11] proposed a Spatial-Temporal Graph Convolutional Network (ST-GAN) that employs graph convolution to aggregate the joint features in the spatial dimension. Liu et al. [42] proposed a multi-stream graph convolutional network to avoid the missing of structural information in the training phase. Cheng et al. [12] proposed a Shift Graph Convolutional Network (Shift-GCN) for solving the problem of inflexible acceptance domain of graph convolution network on both spatial and temporal dimensions.

It can be found that skeleton-based action recognition methods mainly design an effective model to capture temporal and spatial information from skeleton sequences, which is difficult to recognize human-object actions, e.g., blowing hair, combing hair, etc.

C. Multi-modal Fusion

Many works [19], [20], [24], [43]–[46] consider aggregating information from multi-modal data for various image and video content analysis tasks, such as video action recognition, video anomaly detection, and localization, and so on. For example, Liu et al. [19] and Dan et al. [20] used the skeleton data and RGB frames for action recognition. Chen et al. [45] used depth maps and skeleton data for human action recognition. Li et al. [47] used RGB frames and Flows for video anomaly detection and localization. On the one hand, different modality data have different distributions. On the other hand, multi-modal data often contain irrelevant (modal-specific) information. Thus, how to aggregate information from multi-modal features is the main problem. Multi-modal data or features fusion across different modalities is always a hot topic, some classical methods [27]–[29] were proposed to utilize the linear embedding or attention mechanism to fuse multi-modal features. For example, Hori et al. [28] propose a multi-modal attention model that selectively fuses multi-modal features based on learned attention. With the advancement of Convolutional Neural Networks (CNN), some nonlinear fusion approaches of multi-modal features are proposed by designing various convolutional networks as the attention mechanism [31]–[34], [48]. For example, Kuang et al. [32] proposed a multi-modal fusion network based on CNN for face anti-spoofing detection. Fooladgar et al. [33] used an efficient attention method based on CNN architecture to fuse RGB data and depth maps in channel-wise way. Su et al. [49] used the soft attention method to interact multi-modal data in channel-wise way. To fully fuse the multi-modal features, we consider interacting information of multi-modal features not only in channel-wise way but also in modal-wise way. In this work, we design a flexible Expansion-Squeeze-Excitation (ESE) in Modal-fusion Net (M-Net) and Channel-fusion Net (C-Net) to learn modal and channel-wise nonlinear attention.
for capturing the modal and channel-wise dependencies among features, respectively.

III. METHODOLOGY

In this section, we mainly introduce Expansion-Squeeze-Excitation Fusion Network (ESE-FN). Specifically, we first revisit the SENet as a warm-up, and then introduce the framework of ESE-FN in details, including Modal-fusion Net (M-Net), Channel-fusion Net (C-Net), and Multi-modal Loss (ML).

A. Revisiting SENet

Squeeze-and-Excitation Networks (SENet) [30] has shown remarkable performance in the ImageNet database by explicitly capturing the channel-wise dependencies between feature maps. Here, the channel-wise dependencies are quantified via the nonlinear attention corresponding to each channel. The nonlinear attention is obtained by Squeeze-and-Excitation (SE). SENet has been proven that learning channel-wise nonlinear attention can improve the discriminative ability of features.

In SENet, Squeeze-and-Excitation (SE) can be divided into two steps: squeeze and excitation in turn, which are used for obtaining channel-wise representation and channel-wise nonlinear attention. Specifically, assuming $C$ feature maps, denoted by $X = \{X_c\}_{c=1}^{C}$, for one feature map $X_c$ in the $c$-th channel, the squeeze operation is performed as follows,

$$
Y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j),
$$

where $H$ and $W$ denote the height and width of the feature map, respectively. $Y_c$ is the result of global average pooling for the $c$-th channel. In the squeeze step, all feature maps are transformed to a channel-wise representation $Y = [Y_1, Y_2, \cdots, Y_c, \cdots, Y_C]^T$, which is a one-dimensional vector. In the excitation step, the channel-wise representation $Y$ is fed into multi fully-connected layers to obtain the channel-wise attention $W$, as follows,

$$
W = \sigma(F_1(\text{Relu}(F_2(Y)))),
$$

where $F_\ast$ denotes a Fully Connected (FC) layer in this work, and $\sigma$ is a sigmoid function. Finally, the original feature maps $X$ can be updated to $\hat{X}$ by the following equation

$$
\hat{X} = X \otimes W,
$$

where $\otimes$ denotes channel-wise multiplication between $X$ and $W$. Compared with the original feature maps $X$, the updated feature map $\hat{X}$ has enhanced the discriminative information, while suppressing the useless or irrelevant modal-specific information to some extent.

Through the above analysis, we can find that SENet uses the global average pooling to interact spatial information. Since the global average pooling is implemented from the global view, the learned nonlinear attention is significantly affected by noise (e.g., a large area of background). In this work, to well capture the local and global spatial information of features, we consider interacting spatial information from the local and global views.

B. Overview of ESE-FN

For the task of elderly activity recognition, we attempt to aggregate the discriminative information from both RGB videos and skeleton sequences by attentively fusing multi-modal features. Thus, multi-modal features fusion across multiple modalities is key to modeling the motion among RGB videos and skeleton sequences. Based on Squeeze-and-Excitation in SENet, we design a new Modal-fusion Net (M-Net) and a new Channel-fusion Net (C-Net) to learn the Modal-wise Expansion-Squeeze-Excitation Attention (M-ESEA) and Channel-wise Expansion-Squeeze-Excitation Attention (C-ESEA) for capturing the modal and channel-wise dependencies among features, respectively. By integrating M-Net and C-Net as the main components, a novel Expansion-Squeeze-Excitation Fusion Network (ESE-FN) is proposed to attentively fuse the multi-modal features with M-ESEA and C-ESEA. The framework of ESE-FN is shown in Figure 3, which mainly consists of four parts, i.e., feature extractor module, modal-fusion module, channel-fusion module, and multi-modal loss.

Denoted $\{v_{ri}|i=1,2,3,\ldots,N\}$ as an video set with size $N$, we first split each video into $T$ clips and random sample a frame from each clip. Then, we get the frame set $\{r_i|i=1,2,3,\ldots,T\}$ as one input video $r$, which are fed into the RGB backbone (e.g., ResNeXt101 [50]) to extract the RGB feature $f_r \in R^{d_r \times 1}$ ($d_r$ denotes the feature dimension in this paper), as follows,

$$
f_r = \text{RGB}_\text{Backbone}(r).
$$

Likewise, for video $r$, the corresponding skeleton sequence $s$ is obtained and fed into the skeleton backbone (e.g., Shif-GCN [12]) to extract the skeleton feature $f_s \in R^{d_s \times 1}$,

$$
f_s = \text{Skeleton}_\text{Backbone}(s).
$$

Generally, the multi-modal features, i.e., $f_r$ and $f_s$, have different sizes and cannot be directly concatenated. Thus, we use two MLPs to unify the size of $f_r$ and $f_s$, and then concatenate them, as follows,

$$
f = \text{Concat}(H_r(f_r), H_s(f_s)),
$$

where $f \in R^{d \times n}$, $n$ is the number of modality, $H_r$ and $H_s$ are MLP, and Concat($\cdot$) is a modal-wise concatenation operator.

Subsequently, we transpose the feature $f$ and feed it into M-Net for modal-wise fusion, as follows,

$$
h_m = \text{M-Net}(\text{Trans}(f)),
$$

where Trans($\cdot$) is a transpose operator. Likewise, $h_m$ is transposed and fed into C-Net for channel-wise fusion, as follows,

$$
h_{mc} = \text{C-Net}(\text{Trans}(h_m)),
$$

We sum up feature $h_{mc}$ in modal-wise way and get the fused multi-modal feature $f_{rs}$.

Finally, RGB feature $f_r$, skeleton feature $f_s$ and fused multi-modal feature $f_{rs}$ are fed into a new Multi-modal Loss (ML) for optimizing all parameters of ESE-FN, as follows,

$$
\Theta = \text{argmin}(\mathcal{L}(F_r(f_r), F_s(f_s), F_{rs}(f_{rs}))),
$$
where Θ denotes the parameter set of the proposed ESE-FN. The following sections introduce M-Net, C-Net, Multi-modal Loss in details.

C. Modal-fusion Net (M-Net)

The detailed configuration of Modal-fusion Net (M-Net) is shown in Figure 4. M-Net can attentively aggregate the local and global spatial discriminative information of multi-modal features via M-ESEA in modal-wise way. Specifically, we firstly extend the transposed feature \( f \in \mathbb{R}^{m \times d} \) to \( f \in \mathbb{R}^{m \times d \times 1} \) as the input of M-Net. Similar to Squeeze-and-Excitation in SENet, we divide the implementation of M-Net into three steps, i.e., modal-wise expansion, modal-wise squeeze, and modal-wise excitation in turn, which are detailed in the following.

Modal-wise expansion step. It uses several stacked convolutional layers with different kernels to interact modal-wise information by expanding the spatial information from the local view. Then the expanded feature \( f_{ml} \) can be described as follows,

\[
f_{ml} = \text{Conv}_1(\text{Conv}_2(\text{Conv}_3(f))),
\]

where \( f_{ml} \in \mathbb{R}^{m \times d \times 1} \), \( d_m \) is the spatial dimension after convolution transformation \((m \times d_m > n \times d) \). Here, \( f_{ml} \) can be also regarded as the single-modal feature set \( \{f_{ml}^i \in \mathbb{R}^{m/n \times d_m \times 1}\}_{i=1}^n \).

Modal-wise squeeze step. It utilizes \( f_{ml} \) to calculate the modal-wise representation \( f_{mlg} \in \mathbb{R}^{m \times 1} \) \((m < n \times d) \) from global view via the average pooling, as follows,

\[
f_{mlg} = \frac{1}{d_m \times 1} \sum_{i=1}^{d_m} \sum_{j=1}^{1} f_{ml}(i,j).
\]

Modal-wise excitation step. It utilizes the modal-wise representation \( f_{mlg} \) to learn the Modal-wise Expansion-Squeeze-Excitation Attention (M-ESEA) \( W_m \) for capturing the modal-wise dependence of features. Finally, the original input feature \( f \) can be updated to the modal-wise fused feature \( h_m \in \mathbb{R}^{m \times d \times 1} \) based on the following equations,

\[
W_m = \sigma(F_3(\text{Relu}(F_4(f_{mlg}))));
\]

\[
h_m = f \otimes W_m.
\]

Compared with SENet, it is noted that we additionally bring in the expanding of features as the partner of the squeezing. Expanding and squeezing provide the interactions of features in up-size and down-size ways, which can capture both of local and global dependencies of features.

D. Channel-fusion Net (C-Net)

In this section, we focus on our Channel-fusion Net (C-Net), as shown in figure 5 C-Net aims to learn C-ESEA for fusing multi-modal features in channel-wise way. Similar to M-Net, the implementation of C-Net can be also divided into the channel-wise expansion, channel-wise squeeze, and channel-wise excitation in turn, described in the following.

Channel-wise expansion step. It uses the convolutional layer to interact the channel-wise information by expanding spatial information from the local view. Then the expanded features \( h_{cg} \) can be calculated as follows,

\[
h_{cg} = \text{Conv}_4(h_m),
\]

where \( h_{cg} \in \mathbb{R}^{d \times n_1 \times 1} \), \( n_1 > n \).

Channel-wise squeeze step. It utilizes \( h_{cg} \) to calculate the channel-wise representation \( f_{cg} \) from the global view via the average pooling, as follows,

\[
f_{cg} = \frac{1}{n_1 \times 1} \sum_{i=1}^{n_1} \sum_{j=1}^{1} h_{cg}(i,j),
\]

where, \( f_{cg} \in \mathbb{R}^{d \times 1} \) can be also regarded as the single-channel feature set \( \{f_{cg}^i \in \mathbb{R}^{d}\}_{i=1}^d \).

Channel-wise excitation step. It utilizes the channel-wise representation \( f_{cg} \) to learn the Channel-wise Expansion-Squeeze-Excitation Attention (C-ESEA) \( W_c \) for capturing the channel-wise dependence of features, as follows,

\[
W_c = \sigma(F_5(\text{Relu}(F_6(f_{cg})))).
\]
Finally, we get the feature updated in channel-wise way, described as follows,

$$h_{mc} = h_m \otimes W_c.$$  

(17)

By comparing with the formulations of M-Net and C-Net, modal-wise fusion can be seen as a rough fusion of multi-modal features in modal-wise way, while channel-wise fusion can be seen as a fine-grained fusion of multi-modal features in channel-wise way. Both of them attentively aggregate the discriminative information of multi-modal features based on modal and channel-wise dependencies of features.

E. Multi-modal Loss (ML)

To keep the consistency between the single-modal features and the fused multi-modal features, we design a new Multi-modal Loss (ML) to additionally measure the difference between the prediction loss on single modalities and the prediction loss on the fused modality. The key idea of multi-modal loss is that we take the minimum prediction loss on single modalities to be consistent with the prediction loss on the fused modality. Formally, we define three types of recognition losses $\mathcal{L}_r$, $\mathcal{L}_s$, and $\mathcal{L}_{rs}$ corresponding to the RGB modality, skeleton modality, and fused modality, respectively. Then the multi-modal loss can be described as follows,

$$\mathcal{L} = \alpha \times \mathcal{L}_{rs} + \beta \times (\min(\mathcal{L}_r, \mathcal{L}_s) - \mathcal{L}_{rs}),$$

(18)

where $\alpha$ and $\beta$ are hyper-parameters (will be discussed in Section IV.C). In this work, the forms of $\mathcal{L}_r$, $\mathcal{L}_s$, and $\mathcal{L}_{rs}$ are cross entropy loss.

IV. EXPERIMENTS

In this section, we conduct experiments to evaluate the performance of the proposed ESE-FN in terms of the elderly activity recognition task. Besides, we further conduct comparative experiments between the proposed ESE-FN and the other advanced methods in terms of the normal action recognition task.

A. Datasets

We evaluate the performance of ESE-FN in terms of the elderly activity recognition task on the ETRI-Activity3D dataset [51]. It is the currently largest elderly activity recognition dataset collected in real-world surveillance environments, which contains 112,620 samples performed by 100 persons including RGB videos, depth maps, and skeleton sequences. All videos are grouped into 55 classes of actions, including individual activities, human-object interactions, and multiperson interactions. The splitting of training and testing sets is based on person ID, namely the samples with person ID \{3, 6, 9, \ldots, 99\} for testing, and the samples with person ID \{1, 2, 4, 5, \ldots, 100\} for training.

| Modality | Backbone | Params | Accuracy (%) |
|----------|----------|--------|--------------|
| RGB      | ResNeXt18 | 15.60M | 87.1         |
|          | ResNeXt101| 48.16M | 93.5         |
| Skeleton | Shift-GCN | 0.32M  | 88.6         |

Fig. 6. The training loss and accuracy of ESE-FN on the ETRI-Activity3D dataset.

B. Implementation details

In the data pre-processing phase, we split each video into $T = 64$ clips, and randomly select one frame for each clip. Finally, we obtain a new video set, wherein each video contains 64 frames. Likewise, we also obtain a new skeleton set, wherein each skeleton sequence contains 64 frames.

In the feature extraction phase, we use ResNeXt18 or ResNeXt101 [50] as the RGB backbone, and Shift-GCN [12] as the skeleton backbone. For the pre-trained ResNeXt18 and ResNeXt101, we fine tune them with the standard SGD optimizer by setting the momentum, initial learning rate, weight decay, and total epochs as 0.9, 0.1, $10^{-3}$, and 120, respectively. Especially, the batch size is set to 128 and 32 for ResNeXt18 and ResNeXt101, respectively. For the pre-trained Shift-GCN, we fine tune it with the standard SGD optimizer by setting the momentum, initial learning rate, batch size, and total epochs as 0.9, 0.1, $10^{-4}$, 32, and 140, respectively. Subsequently, we use the fine-tuned RGB and skeleton backbone to extract RGB and skeleton features, respectively. Here, to test the representation performance of ResNeXt18, ResNeXt101, and Shift-GCN, we use these three models to extract single-modal features for training a softmax independently. For example, we use ResNeXt18 to extract RGB features and feed them into the softmax. The obtained performance for three models is listed in Table I. We can see that ResNeXt101 performs better than ResNeXt18. In this paper, we choose ResNeXt101 as the RGB backbone in default.

In the training phase of ESE-FN, we use the standard SGD optimizer to train it by setting the momentum, basic learning rate, weight decay, batch size, and total epochs as 0.9, 0.1, $10^{-4}$, 32 and 30, respectively. All above experiments are performed via the PyTorch deep learning framework on the Linux server equipped with Titan RTX GPU. The training loss and accuracy of ESE-FN as the number of epochs are shown in Figure 6. We can see that the training loss and accuracy of
ESE-FN reach a steady state after about 30 epochs.

C. Diagnostic Study

We conduct the diagnostic study to discuss the sensitiveness of the hyper-parameters α and β in Eq. (18). Here, to enhance the efficiency of the diagnostic study, we use ResNeXt18 instead of ResNeXt101 as the RGB backbone and set T = 16 without loss of generality. Specifically, we empirically tune them by α ∈ {0.3, 0.5, 0.7, 0.9} and β ∈ {0, 0.3, 0.6, 0.9}, respectively. The corresponding results are shown in Figure 7. It can be found that the best performance of ESE-FN is achieved when α = 0.7, and β = 0.3. Thus, we set α = 0.7, and β = 0.3 in experiments. Moreover, when β=0, the multi-modal loss L is degraded to a basic action recognition loss L_{rs}, where the performance is degraded significantly. This illustrates that the designed Multi-modal Loss is more effective than the basic loss.

D. Ablation Study

As mentioned before, Modal-fusion Net (M-Net), Channel-fusion Net (C-Net), and Multi-modal Loss (ML) are new components in the framework of the proposed ESE-FN. Thus we conduct ablation study to evaluate the superiority of M-Net, C-Net, and ML. We firstly set seven baselines, as follows,

- **B1**: Single modal for RGB employs RGB backbone to extract the RGB features, which are directly used to train the softmax classifier. This aims to test the basic recognition performance by using the single-modal features in RGB videos.
- **B2**: Single modal for skeleton employs skeleton backbone to extract the skeleton features, which are directly used to train the softmax classifier. This aims to test the basic recognition performance by using the single-modal features in skeleton sequences.
- **B3**: Multi modal for RGB and skeleton employs RGB backbone and skeleton backbone to extract the RGB features and skeleton features, which are concatenated and fed into the softmax classifier. This aims to test the basic recognition performance by simply combining multi-modal features.
- **B4**: ESE-FN w/o C-Net is a degraded version of ESE-FN by dropping out the channel-fusion module. This aims to show the superiority of M-Net.
- **B5**: ESE-FN w/o M-Net is a degraded version of ESE-FN by dropping out the modal-fusion module. This aims to show the superiority of C-Net.
- **B6**: ESE-FN w/o ML is a degraded version of ESE-FN by using the basic loss L_{rs} instead of ML L, namely setting α = 1, and β = 0. This aims to show the superiority of ML.
- **B7**: the proposed ESE-FN.

The comparison results among all baselines are shown in Table II. Compared with single-modal and multi-modal baselines, B3-B7 (using multi-modal features) outperform B1-B2 (using single-modal features), which validates that the multi-modal features contain more complementary information than the single-modal features. In particular, B4-B7 are equipped with the newly designed components, and perform better than B3 (simply combining multi-modal features). This validates that each component, i.e., M-Net, C-Net, or ML, is superior to show the superiority of C-Net.

Table II. Ablation study for ESE-FN with different components.

| Baseline | RGB | Skeleton | M-Net | C-Net | ML | Accuracy (%) |
|----------|-----|----------|-------|-------|----|--------------|
| B1       | ✓   | ✓        | ✓     | ✓     | ✓  | 95.9         |
| B2       | ✓   | ✓        | ✓     | ✓     | ✓  | 94.0         |
| B3       | ✓   | ✓        | ✓     | ✓     | ✓  | 95.3         |
| B4       | ✓   | ✓        | ✓     | ✓     | ✓  | 94.5         |
| B5       | ✓   | ✓        | ✓     | ✓     | ✓  | 94.0         |
| B6       | ✓   | ✓        | ✓     | ✓     | ✓  | 95.7         |
| B7       | ✓   | ✓        | ✓     | ✓     | ✓  | 95.9         |
### Comparison with Current Methods on ETRI-Activity3D Dataset

*Insert Table III here with the following content:*

| Methods                      | Modalities                  | Backbones                | Accuracy(%) |
|------------------------------|-----------------------------|--------------------------|-------------|
| Deep Bilinear Learning [53]  | RGB+Depth+Skeleton          | VGG16+RNN                | 88.4        |
| Evolution Pose Map [52]      | RGB+Depth+Skeleton          | VGG16+RNN                | 91.3        |
| c-ConvNet [54]               | RGB+Depth                   | VGG16                    | 91.7        |
| FSA-CNN [51]                 | RGB+Skeleton                | -                        | 93.7        |
| ESE-FN (Ours)                | RGB+Skeleton                | VGG16+Shift-GCN          | 93.4        |
|                              |                             | ResNeXt101+Shift-GCN     | 95.9        |

*Fig. 8. The confusion matrices obtained by Shift-GCN, ResNeXt101, and ESE-FN on the ETRI-Activity3D dataset. It should be noted that ETRI-Activity3D dataset does not provide the names of all classes.*

*Fig. 9. The confusion matrices obtained by ESE-FN on the NTU RGB+D dataset.*

In addition, the confusion matrices obtained by ESE-FN, Shift-GCN and ResNeXt101 are shown in Figure 8. First, the confusion matrix of ESE-FN shows that the color of the main diagonal is lighter than other spaces, which illustrates all classes of elderly activities can be well recognized by ESE-FN. Second, compared with the confusion matrices of ESE-FN, Shift-GCN, and ResNeXt101, we can find that the color of the main diagonal in Figure 8(c) is lighter than the color of the main diagonal in Figure 8(a) and (b). This shows that ESE-FN is more effective for recognizing the confusing elderly activities by capturing more discriminative multi-modal information.

### Extended Experiment on Normal Action Recognition

To test the generalization of the proposed ESE-FN, we also conduct comparative experiments between the proposed ESE-FN and the other advanced methods in terms of the normal action recognition task. We use the NTU RGB-D dataset [55] as the benchmark. The NTU RGB-D dataset is a large-scale dataset collected by three Kinect cameras from different views concurrently, and has been widely used for evaluating RGB-based or skeleton-based action recognition methods. It includes 56,880 skeleton sequences and 60 different classes from 40 distinct subjects. NTU RGB-D dataset provides two standard evaluation protocols, i.e., Cross-Subject (CS) setting, and Cross-View (CV) setting. In the CS setting, the training set includes 40,320 samples performed by 20 subjects, and the testing set includes 16,560 samples performed by the other 20
In this work, we propose a novel Expansion-Squeeze-Excitation Fusion Network (ESE-FN) to address the problem of elderly activity recognition by learning modal and channel-wise Expansion-Squeeze-Excitation (ESE) attentions for attentively fusing the multi-modal features in the modal and channel-wise ways, respectively. Overall, the framework of ESE-FN consisting of the feature extractor module, modal-fusion module, channel-wise module, and multi-modal loss can be summarised as two main insights. First, Modal-fusion Net (M-Net) and Channel-fusion Net (C-Net) can capture the modal and channel-wise dependencies among features for enhancing the discriminative ability of features via Modal and channel-wise ESEs. Second, Multi-modal Loss (ML) can enforce the consistency between the single-modal features and the fused multi-modal features by additionally bringing in the penalty of difference between the minimum prediction losses on the single modalities and the prediction loss on the fused modality. Experimental results on both elderly activity recognition and normal action recognition tasks demonstrate that the proposed ESE-FN achieves the SOTA performance compared with the other competitive methods. To the best of our knowledge, ESE-FN is the first work that adopts the combination of expanding and squeezing to fully interacting features from the local and global views for learning nonlinear attention.

## References

[1] X. Shu, L. Zhang, Y. Sun, and J. Tang, “Host-parasite: Graph Istm-in-Istm for group activity recognition,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 663–674, 2020.

[2] C. Li, Q. Zhong, D. Xie, and S. Pu, “Collaborative spatiotemporal feature learning for video action recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7872–7881.

[3] C. Feichtenhofer, H. Fan, I. Malik, and K. He, “Slowfast networks for video recognition,” in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6202–6211.

[4] X. Wang and A. Gupta, “Videos as space-time region graphs,” in *European conference on computer vision (ECCV)*, 2018, pp. 399–417.

[5] J. Tang, X. Shu, R. Yan, and L. Zhang, “Coherence constrained graph Learn for group activity recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019 (publish online DOI: 10.1109/TPAMI.2019.2928540).

[6] J. Lin, C. Gan, and S. Han, “Temporal shift module for efficient video understanding, 2019 iee,” in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7082–7092.

[7] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.

[8] X. Shu, J. Tang, G. Qi, W. Liu, and J. Yang, “Hierarchical long short-term concurrent memory for human interaction recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1110–1118, 2021.

[9] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, “View adaptive neural networks for high performance skeleton-based human action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1963–1978, 2019.

[10] A. Banerjee, P. K. Singh, and R. Sarkar, “Fuzzy integral based cnn classifier fusion for 3d skeleton action recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2206 – 2216, 2020.

[11] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 7444–7452.

[12] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, “Skeleton-based action recognition with shift graph convolutional network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 183–192.

[13] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, “Spatiotemporal attention recurrent neural networks for human-skeleton motion prediction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021 (publish online DOI: 10.1109/TPAMI.2021.305918).

[14] S. Albawi, T. A. Mohammad, and S. AlZawi, “Understanding of a convolutional neural network,” in *International Conference on Engineering and Technology (ICET)*, 2017, pp. 1–6.

[15] W. Zaremba, I. Sutskever, and O. Vinyals, “Recurrent neural network regularization,” arXiv preprint arXiv:1409.2329, 2014.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[17] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” arXiv preprint arXiv:1609.02907, 2016.

[19] X. Liu, J. Qian, F. Wen, X. Zhu, W. Li, and P. Liu, “Action recognition based on 3d skeleton and rgb frame fusion,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 258–264.
[64] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Two-stream adaptive graph convolutional networks for skeleton-based action recognition,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12026–12035.

[65] ——, “Skeleton-based action recognition with directed graph neural networks,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7912–7921.

[66] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, “Disentangling and unifying graph convolutions for skeleton-based action recognition,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 143–152.