symmetry

MDPI

Article

Hi-EADN: Hierarchical Excitation Aggregation and Disentanglement Frameworks for Action Recognition Based on Videos

Zeyuan Hu * and Eung-Joo Lee

Department of Information Communication Engineering, Tongmyong University, Busan 48520, Korea; ejlee@tu.ac.kr

* Correspondence: zyhu@tu.ac.kr

Abstract: Most existing video action recognition methods mainly rely on high-level semantic information from convolutional neural networks (CNNs) but ignore the discrepancies of different information streams. However, it does not normally consider both long-distance aggregations and short-range motions. Thus, to solve these problems, we propose hierarchical excitation aggregation and disentanglement networks (Hi-EADNs), which include multiple frame excitation aggregation (MFEA) and a feature squeeze-and-excitation hierarchical disentanglement (SEHD) module. MFEA specifically uses long-short range motion modelling and calculates the feature-level temporal difference. The SEHD module utilizes these differences to optimize the weights of each spatiotemporal feature and excite motion-sensitive channels. Moreover, without introducing additional parameters, this feature information is processed with a series of squeezes and excitations, and multiple temporal aggregations with neighbourhoods can enhance the interaction of different motion frames. Extensive experimental results confirm our proposed Hi-EADN method effectiveness on the UCF101 and HMDB51 benchmark datasets, where the top-5 accuracy is 93.5% and 76.96%.

Keywords: action recognition; excitation aggregation; squeeze-and-excitation hierarchical disentanglement; spatiotemporal features; long-short range motion modelling

Citation: Hu, Z.; Lee, E.-J. Hi-EADN: Hierarchical Excitation Aggregation and Disentanglement Frameworks for Action Recognition Based on Videoss. *Symmetry* **2021**, *13*, 662. https://doi.org/10.3390/sym13040662

Academic Editor: Jan Awrejcewicz

Received: 8 March 2021
Accepted: 9 April 2021
Published: 12 April 2021

1. Introduction

Action recognition [1] is a fundamental problem of video process tasks and has drawn significant attention in computer vision communities, such as autonomous driving and intelligent surveillance [2]. However, the early-stage approaches train classifiers based on the spatial-temporal nature of images. For instance, hand-made features are input into a support vector machine for classification [3]. In recent years, with the continuous development of deep learning technology, methods have been applied to video action recognition tasks, namely, learning deep features via convolutional neural networks (CNNs) [4,5] and obtaining state-of-the-art performance on HMDB51 and UCF101 action video datasets. The performance of these deep learning methods outperforms traditional methods However, they ignore important detailed information because these CNN-based methods [6–8] depend mainly on the discriminativeness of high-level semantic information to aggregate these features on the fully connected layer and finally achieve classification; namely, these features usually focus more on high-level semantic information but less on important detail information [9]. For example, ResNet [10] and Inception [11] are used for the high-level features of final classification, and their feature maps are small in size, which cannot preserve the local details of the action to the maximum.

To address the above problems, the middle layers of convolutional neural networks use action recognition in video. Compared with high-level CNN features, middle layer convolution features usually show more detailed information [12–14]. However, these methods still leave some limit, namely, how to model the spatial-temporal structure with

significant variations and complexities effectively and enhance the interaction power with neighbourhoods. For these problems, most existing methods adopt two-stream action recognition framework-based CNNs. These are processing images and optical flow by different branch CNNs, reducing the interaction between spatial and temporal features. Moreover, subtle motion and local discriminative information are ignored.

Thus, to address the above problems, we propose hierarchical excitation aggregation and disentanglement networks (Hi-EADNs), which integrate motion frame modelling into spatial-temporal information and calculate between adjacent frames to represent feature-level motion by a multiple frame excitation aggregation (MFEA) module. Then, these difference features are utilized to optimize weights, and the motion-sensitive information in the spatial-temporal features of frames can be processed with a series of squeezes and excitations. Moreover, the SEHD module can enhance the interaction of different motion frames and is forced to discover differentiated detail information and improve the capture ability of the informative spatial-temporal features. However, as the network deepens, repeating the use of a large number of local operations and redundant information leads to optimization difficulty of our proposed frameworks; thus, we squeeze-and-excitation block hierarchical embedding to DenseNet [15]. When the spatial-temporal information goes through the SEHD module, these features complete multiple information exchanges with adjacent frames and further increase the interaction of neighbours and reduce the use of redundant information, ultimately improving recognition accuracy for action in the video.

To summarize, we propose that Hi-EADN methods are complementary and cooperate; namely, they not only maximize the retention of important detailed information but also improve the modelling abilities of the short-long spatial-temporal range. In this paper, the main contributions of our Hi-EADN method are as follows.

- We propose a novel hierarchical excitation aggregation and disentanglement network (Hi-EADN) with a multiple frame excitation aggregation (MFEA) module to better model long-short range motion and calculate the feature-level spatial-temporal difference.
- The squeeze-and-excitation hierarchical disentanglement (SEHD) module excites motion-sensitive channels by a series of squeeze-and-excitation operations and enhances the interactivity of different motion frames. Moreover, the spatial-temporal information with neighbourhoods is aggregated and efficiently enlarges the modelling of short-long ranges.
- The two components of Hi-EADN are complementary and cooperate to realize multiple information exchanges with adjacent frames and increase the interaction of neighbours. We proposed methods that outperform other state-of-the-art methods on the UCF101 and HMDB51 benchmark datasets.

The rest of this paper is organized as follows. Section 1 introduces the related work. Section 2 details overviews of the proposed method. Section 3 reports and analyses the experimental results and finally gives conclusions in Section 4.

2. Related Work

In this section, we mainly overview video-based action recognition from two perspectives: traditional handcrafted features and deep learning methods.

Traditional methods for action recognition based on video: there are many traditional methods for making spatial-temporal features of video actions, such as histograms of gradient (HOG) [16] and scale-invariant feature transform (SIFT) [17] descriptors. Additionally, a series of improved methods, such as histograms of optical flow (HOF) [18] and improved dense trajectory (IDT) [19], are widely used in video action recognition tasks. Traditional human-computer interaction has been unable to meet people's needs. Many SVM-based action recognition algorithms have been developed, such as Liu and Zhi-Pan [20], who proposed an action recognition algorithm based on Kinect and SVM methods. Jagadeesh and Patil et al. [21] proposed that video-based human action recognition is addressed and performed on a public baseline dataset to conduct verification. Karpathy and Toderici et al. [22] combined local spatial-temporal feature information and

proposed multiresolution action recognition methods and improved network speeding of the training.

Deep learning methods for action recognition based on the video: with the tremendous success of deep learning methods on image recognition [23], object detection [24] and image segment [25] tasks, in recent years, a series of methods have been developed to learn deep-layer feature information by convolutional neural networks (CNNs) for video-based action recognition. The performance of deep learning methods outperforms traditional handcrafted methods. For instance, Simonyan and Zisserman [26] developed two-stream convolutional neural networks (CNNs) to capture spatial-temporal deep information on RGB frames and optical flows. Donahue et al. [27] used long-short-term memory (LSTM) networks to model the spatial-temporal relation by aggregating 2-dimensional coevolution features. Wang et al. [28] proposed a spatial-temporal segment network based on two-stream convolution neural networks (CNNs) to fuse video-clip representation into video representation. Feichtenhofer and Pinz et al. [29] presented a general ConvNet architecture for video action recognition based on multiplicative interactions of spacetime features. Yue-Hei Ng and Hausknecht et al. [30] developed recurrent neural networks (RNNs) with LSTM to combine frame-level features into video-level representations.

Another video-based action recognition method adopts 3-dimensional convolutional neural networks (CNNs) or their variants [31] or attention networks [32]. For example, Kalfaoglu et al. [33] combined 3-dimensionality convolution with temporal modelling and replaced the global average pooling (TGAP) layer with the bidirectional encoder representations from transformers (BERT) layer. Anvarov Kim et al. [34] proposed the aggregation of squeeze-and-excitation and attention modules with 3-dimensional convolutional neural networks to analyse both long-short-term temporal actions efficiently. Jalal and Aftab et al. [35] proposed dual-stream spatial-temporal motion fusion with self-attention for action recognition. Purwanto and Pramono et al. [36] proposed three-stream human action recognition frameworks based on extreme low-resolution videos. Yu and Guo et al. [37] presented a high-level action representation method, namely, the framework mainly backbone as a joint spatial-temporal attention model.

However, the above methods can achieve battery recognition results in video-based action. However, they directly add the spatial-temporal feature information and motion frame flow encoding together, which loses much important detail information and locks the interaction between neighbourhoods and could be harmful to modelling between long- and short-range motion frames. In contrast, our proposed method utilizes motion features to recalibrate the spatial-temporal information to enhance the motion pattern and realize multiple information exchanges with neighbouring motion frames. Then, we extract optical flow information and learn the spatial-temporal feature-level motion representations by calculating the time difference of different motion frames. Aggregate motion frame encoding with spatial-temporal information to improve the interactivity of different motion frames and efficiently enlarge the modelling of short-long ranges.

3. Our Proposed Method

In this section, our Hi-EADN methods aim to improve interaction with neighbouring motion frames and reduce the use of redundant information while reducing the different modality information (image feature and optical flow), intramodality discrepancies and exciting motion patterns and a multiple spatial-temporal aggregation to build a long short-range spatial-temporal relationship. To complete this goal, we elaborate the basic knowledge for two key components, namely, MFEA and SEHD. The frameworks of our proposed method are illustrated in Figure 1.

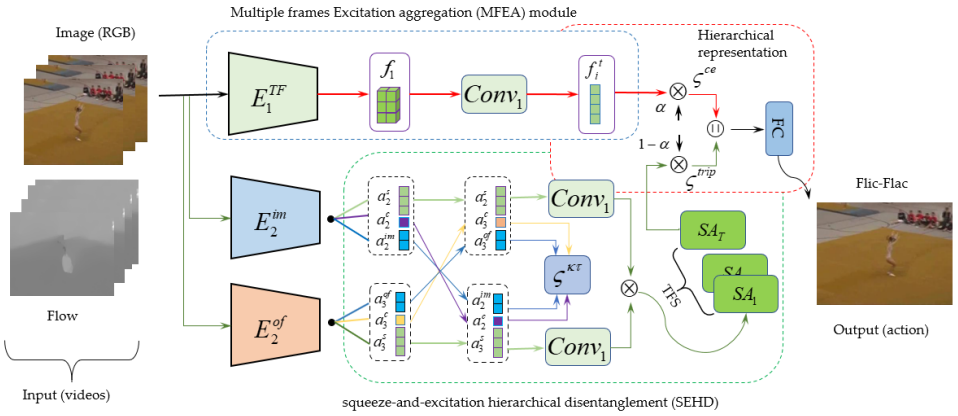

Figure 1. The frameworks of our proposed method for action recognition based on videos. MFEA indicates the multiple-frame excitation aggregation module, where E_1^{TF} indicates the decoded representation of a sequence frame and E_2^{im} and E_2^{of} indicate the encoding process of image and optical flows. $Conv_1$ indicates the convolution layer whose size is $1 * 1$. SA indicates dual self-guided attention module. SEHD indicates the module of squeeze-and-excitation hierarchical disentanglement. ς indicates the loss function. a^{of} and a^{im} indicate each optical and image flow, respectively. *ce* indicates the loss function of multi-class cross-entropy. $\varsigma^{\kappa\tau}$ indicates sharing information of cross-modality. ς^{trip} indicates the loss of reconstruction $\varsigma^{trip} = \varsigma_{rce}$.

3.1. Multiple Frames Excitation Aggregation (MFEA) Module

Most existing methods utilize motion representations for video-based action recognition [37]. Although these works have been achieving better accuracy of action recognition, the majority of works represent motion actions in the form of optical flow and mutually independent learning of motion representations and spatial-temporal information. Moreover, these methods typically adopt the local convolution kernel to process neighbouring motion frames, and long-distance spatial-temporal modelling can be performed in deep neural networks. However, the long-distance detail information is greatly weakened, and the utility of the redundant information is repeated as the network depths. In contrast, in our present component of multiple frame excitation aggregation (MFEA), the spatial-temporal information and corresponding local convolution layers are aggregated into a group of subspaces, namely, the subspace is integrated into dense connectivity blocks [15] (the component is inspired by DenseNet), which could accordingly reduce the utility of redundant information and improve the long-range modelling ability.

Formally, the input feature can be represented as X is $[N, T, C', H, W]$, where H and W are the spatial shapes, T and N indicate the temporal dimension and batch size, respectively, C' indicates the fragments by splitting the feature channel dimensions, namely, $C' = \frac{1}{4}C$. That is, the component consists of a series of local convolutions and spatial-temporal convolutions. Different from this, we divide the local convolution into multiple subconvolutions, one of which is a fragment, as the input feature. Then, the remaining three fragments are sequentially processed with one spatial subconvolution layer and another channelwise temporal subconvolution layer. Finally, the residual dense connection is added between neighbouring fragments and transforms the component into a hierarchical cascade form by a parallel architecture. The process can be presented as.

$$\begin{cases} X_i^0 = X_i, i = 1 \\ X_i^0 = Conv_{sp} * (Conv_{tem} * X_i), i = 2 \\ X_i^0 = Conv_{sp} * (Conv_{tem} * H_{de}([X_i, X_{i-1}^0])), i = 3, 4 \end{cases} \tag{1}$$

where $Conv_{sp}$ indicates the spatial convolution layers, $Conv_{tem}$ indicates the temporal convolution layers, $H_{de}([\bullet])$ indicates densely connection layers. In the component, the dif-

ferent fragments adopt different convolution kernel sizes. Namely, the X_1 initial fragment receptive field is $1 * 1$, namely, the size is $1 * 1$ of the convolution kernel, and the other fragment receptive fields are $3 * 3$. Then, aggregating information for each fragment and one of them can be expressed as f_1. Finally, the aggregated motion information of the t frame sequence is obtained by a simple concatenation strategy.

$$X_0 = [MP(X_1^0); MP(X_2^0); MP(X_3^0); MP(X_4^0)] \tag{2}$$

where $MP(\bullet)$ indicates max pooling layer. X_1^0 indicates the input videos (including image, optical and sequence frames), $X_i^0, i = 1, 2, 3, 4$ indicates different sequence frames.

In other words, the MFEA module we designed can not only effectively capture the interactivity between continuous actions in the video but also avoids the lack of information caused by the information flow in the transmission process, which leads to inaccurate recognition. Moreover, the obtained output feature information involves spatial-temporal representations capturing different temporal ranges. It is superior to the local temporal representations obtained by using a single local convolution neural network approach.

3.2. Squeeze-and-Excitation Hierarchical Disentanglement (SEHD) Module

In the component, we coupled a hierarchical disentanglement module with motion excitation by sharing the squeeze-and-excitation block [17,22] and embedding layer of multi-head attention. This component enables both information encoders to extract the motion detail information between RGB images and dynamic optical flows. Moreover, this feature learning and disentanglement process are conducive to capturing intra- and external co-occurrence information from different modality image data (RGB and optical flows) and further enhances the interaction of neighbourhood motion frames.

Sharing the Squeeze-and-Excitation Layer

In sharing the squeeze-and-excitation (SE) layer, different channels capture different motion information. Some channels (channels) are mainly used to obtain static information related to the motion scene, that is, to model the motion scene, while most of the channels are used to explore the dynamic information between consecutive action frames. These two-channel construction methods are mainly to describe the temporal and spatial differences of the actions in the video. For action recognition based on videos, the SE is beneficial to enable the Hi-EADN frameworks to discover and then enhance neighbourhood motion-sensitive interactions and reduce the utility of redundant information. The squeeze-and-excitation block [27] can improve feature information representation by channel interconnection and increase the sensitivity to information for the recognition frameworks. However, we adopt global average pooling to squeeze each channel into a single numeric value. Then, to further utilize the detailed information in the squeeze operation, we conduct a second excitation operation to completely achieve channel-wise dependencies presented by the squeeze operation and use this squeezed-and-excitation feature information [27,30] as the input of a multi-head self-attention block, and reconstruction different scale information, which the process indicates as.

$$\begin{cases} F_{squeeze}(\iota_\theta) = \frac{1}{HWC} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{c=1} C\iota_\theta(i, j, c) \\ F_{excitation}(f_{squeeze}, W') = \delta(\varpi(W_1' \varphi(W_2' f_{squeeze}))) \\ F_{Att} = MHSGA_s(F_{squeeze}(\iota_\theta); F_{excitation}(f_{squeeze}, W')), s \in S = \{1, 2, 3, 4\} \end{cases} \tag{3}$$

where $MHSGA(;)$ indicates the dual self-guided attention module, where $\delta(\cdot)$, $\varphi(\cdot)$ and $\varpi(\cdot)$ indicate the ReLU activation function. s indicates the multi-scale features of the SE and the backbone networks of DenseNet. H, W, C indicate the height, width and channels. $F_{excitation}$ indicates the feature information of the channel excitation layer, $F_{squeeze}$ indicates the feature information of the channel squeeze layer.

The dual self-guided attention module (SA) is a module that calculates the response as a weighted sum of the features at all positions. The main idea of multi-head self-guided attention is to help convolutions capture long-range, full-level interconnections throughout the image domain. The SEHD module implemented with an SA and SE multi-scale hierarchical module can help determine images with small details connected with fine details in different areas of the image and frames at each position. Then, the SEHD module can further improve the interaction of neighbourhood frames or successive action sequences.

3.3. Hierarchical Integration Strategy

As illustrated in the hierarchical representation of Figure 1, our designed hierarchical representation learning (HR) module is coupled with multiple frame excitation aggregation (MFEA) by sharing squeeze-and-excitation hierarchical disentanglement (SEHD). This module enables both encoders to extract the common class attributes between motion frame sequences and successive actions. At the same time, this hierarchical representation process implicitly assists in separating intra- and cross-modality characteristics from successive motion frames and improves the quality of human recognition. The representation processing is defined as follows.

$$\begin{cases} F_{Total} = FC(F_{MFEA} \bigoplus F_{SEHD}) \\ F_{SEHD} = Conv(\sum_{\mu=1}^{S} F_{Att}^{\mu}), \mu \in S = \{1, 2, 3, 4\} \end{cases} \tag{4}$$

where $Conv(\cdot)$ indicates that the kernel size is $1 * 1$ of convolution layers. F_{Att}^{u} indicates the multi-scale feature information of dual self-guided attention layers, u and s indicate the size of the scale. F_{MFEA} indicates the fusion feature information of the MFEA, F_{SEHD} indicates the fusion feature information of the SEHD. $\oplus$ indicates a function concatenation.

The hierarchical representation module further strengthens the interaction between successive actions, describes actions from different levels, and strengthens the characterization ability of the features. At the same time, different levels of information can form a complementary relationship, highlighting the differences between different actions between classes or within classes.

3.4. Reconstruction of the Loss Function

In the future, we plan to capture the best multi-scale feature information and improve the performance of our proposed Hi-EAD action recognition based on videos. We employ multiple losses in all network layers during training, namely, the reconstruction (the process of multiscale features) loss and multi-classification cross-entropy loss, where the cross-entropy loss ς_{ce}, where ς_{ce} is useful to penalize the difference between the network output labels of the original sample labels. However, for the reconstruction loss function ς_{rec}, our primary strategy is to integrate a pair of cross-modality information by swapping the human action of two streams with the same id and same frames, where the stream includes image and optical flow. Formally, this cross-modality reconstruction loss function is indicated as follows

$$\varsigma_{rec_{im}} = E_{x_{im,of}^{i} \backsim \tau_{img,of}(x_{im,of}^{i})}[\|x_{im} - \kappa(\tau_2, a_2^s, a_2^{im})\|] \tag{5}$$

where $\varsigma_{rec_{im}}$ indicates the reconstruction loss of image flows. $E_{x_{im,of}^{i} \backsim \tau_{img,of}(x_{im,of}^{i})}$ indicates the encoder information of the image and optical flows. x_{im} indicates the input information of image flows. $\kappa(\tau_2, a_2^s, a_2^{im})$ indicates the process information of the cross-modality. a^s and a^{im} indicate the encoders multiscale features of optical and image flows.

The interactor learns how to conduct interaction and correlation of the same or similar human actions of the same successive frames from the cross-modality reconstruction loss function. To be clear, we only give the loss function for one image modality as $\varsigma_{rec_{im}}$. Thus, the reconstruction loss of optical flow $\varsigma_{rec_{of}}$ can be obtained by changing the parameters.

To improve the interaction quality, we proposed multiple additional reconstruction losses to capture the multi-scale information. In addition to the loss function of reconstructing images and optical flow [20,25] of different modalities, we apply multi-head self-guided attention to refine the feature map. This multi-modality reconstruction loss function plays a key role in regularization in the multi-scale refined feature map. Moreover, this loss includes both assumptions that relativity with short and opposite information should be preserved during the cross-modality reconstruction process and that interaction details and dependencies should be maintained during the same modality reconstruction process. Thus, the overall loss function for reconstruction is indicated as

$$\varsigma_{rec} = \alpha_1 \varsigma_{rec_{im}} + \alpha_2 \varsigma_{rec_{of}} + \alpha_3 \varsigma_{rec_{df}} + \alpha_4 \varsigma_{rec_{ms}} \tag{6}$$

where $\alpha_j, j \in \{1, 2, 3, 4\}$ indicates the importance factor of this reconstruction loss. $\varsigma_{rec_{ms}}$ indicates the refined reconstruction loss function of multiple scales. $\varsigma_{rec_{df}}$ indicates the reconstruction loss of the cross-modality information stream. In addition, the $\varsigma_{rec_{df}}$ are formulated as

$$\varsigma_{rec_{df}} = E_{x^i_{im,of} \backsim \tau_{img,of}(x^i_{im,of})}[\|a^s_{im} - a^s_{of}\|_1] + E_{x^i_{im,of} \backsim \tau_{img,of}(x^i_{im,of})}[\|\tilde{a}^s_{im} - \tilde{a}^s_{of}\|_1] \tag{7}$$

where $\widetilde{a}^s_{im}$ and $\widetilde{a}^s_{of}$ indicate the reconstruction feature information by cross-modality encoders of image and optical flows, respectively. $\| \bullet \|_1$ indicates the operation of a normal form.

Multiclass cross-entropy loss: Given a set of training feature vectors with the class label f_i, Y_i, we use the multi-class, cross-entropy loss [10,19,31] for human action recognition learning

$$\varsigma_{ce} = E_{f \in F, Y \in \mathrm{Y}}[-log(p(Y|f))] \tag{8}$$

where $p(Y|f)$ indicates the predicted probability of the feature vector f belonging to the classes Y. $E_{f \in F, Y \in \mathrm{Y}}[\bullet]$ indicates the encoding process of the action recognition based on videos.

To further enhance the interaction of the action with neighbouring successive frames and improve the description ability of different levels of structural information, the spatial-temporal information with neighbourhoods is aggregated and efficiently enlarges the modelling of short-long ranges. The total loss function is indicated as

$$\varsigma_{Total} = \lambda_{rec} \varsigma_{rec} + \lambda_{mce} \varsigma_{ce} \tag{9}$$

where λ_{rec} and λ_{mce} indicate the importance factor of this loss; meanwhile, we train the proposed Hi-EAD framework by the total loss in an end-to-end manner and conduct iterative optimization.

In summary, our proposed Hi-EAD framework based on video action recognition first uses MFEA to perform static and dynamic modelling of continuous action frames to strengthen the interaction between action frames. At the same time, the SEHD module is used to achieve information complementation between different levels of features, strengthen the differences between different action features, refine the feature map by a multiscale dual self-guided attention layer to reduce the use of redundant information, and achieve effective channel and position information of each action in videos and build the best dependency relationship of different frames or successive actions, which improves the accuracy of action recognition. The MFEA and SEHD process of our proposed Hi-EAD framework for action recognition base on videos as follow Algorithm 1.

Algorithm 1: The recognition process of the proposed Hi-EAD framework for action based on videos

Input:the input feature sequence of MFEA module is X_i,where $X_i \in R^{N*T*H*W*C'}$; the input features information of SEHD module is X_i^t,where $t \in T$, $X_i^t \in R^{H*W*C'}$; the scale and sequence size is S and T, $C' = \frac{1}{4}C$, and C indicates channels.

for *i=1,l=1 to S and L* **do**

Calculate the initial layers output X^0 of the MFEA module by Equations (1) and (2),where $i = \{1, 2, 3, 4\}$ is the scale size of different features map, The final output features information is F_{MFEA} in the l^{th} ;

However, the output features information of SEHD module is F_{SEHD} according to Equation (3) ;

The total output information is definitions F_{Total} by Equation (4) ;

Calculate achieve the best features map by SA ;

Optimization the action recognition framework by our designed reconstruction loss function ς_{Total}, the loss are shown in Equation (9) ;

end

output:output optimization recognition results of videos action ;

4. Experimental Results and Related Discussion

In this section, we demonstrate the effectiveness of our proposed Hi-EADN methods for action recognition in videos by extensive experiments and then give a detailed analysis of the recognition results.

4.1. Dataset Preparation

UCF-101,ref-24: The dataset included 101 categories of human actions with 13,320 videos and mainly included camera operation, appearance changes, posture changes, object ratio changes, background changes, fiber changes.

HMDB-51,ref-25: The dataset contained 51 classes with 6776 videos of human action, where each category contained at least 101 clips. Similar to UCF101, it mainly included smiling, smoking, waving, hugging, dribbling.

We considered that there were instances of the number of classes and actions between the UCF-101 and HMDB-51 datasets and the difference in dynamic background and camera movement. Moreover, to ensure the consistency of subsequent experiments, we randomly divided these two data sets into three parts: training, testing and validation, and the proportion of each part was 40%, 40% and 20% respectively.

4.2. Training Parameters and Environments Configuration

Environments configuration: All the experiments in the paper were performed on torch version 1.7.1+cu110 and torch vision 0.8.2 with four NVIDIA RTX 3080 GPUs, where Python version 3.6.10 was employed. In addition, deep learning libraries such as NumPy, pandas, TensorFlow-GPU 1.4.0 and Keras 2.1.5 were used in our experiments.

Training parameters: For training, we uniformly divided a sample video into $N = 3$ video clips and randomly extracted T frames from each video clip by a sparse smoothing strategy. No emphasis was placed on $T = 4$ or $T = 8$ in our follow-up experiments. However, for the spatial-stream and temporal-stream networks, we took the input of a single RGB frame and optical flow stacks, respectively, where the optical flow was computed by the TVL1 optical flow algorithm [15].

We utilized 2D DenseNet-169 as the backbone, and for different branches, we used different modules as embeddings, namely, integrate the multiple modules we proposed into DenseNet-169. The Adam optimization algorithm was used to train and an adaptive learning rate in our proposed frameworks, where the initial learning rate was 0.001 and other hyperparameters such as $\beta_1 = 0.9$ and $\beta_2 = 0.995$. The batch size was set to 16, the number of units in the fully connected FC layers (hierarchical representations are shown in Figure 1) was 1024, and each self-attention guided layer contained a head size of 3. The epochs and dropout size were 1000 and 0.25, respectively.

4.3. Evaluation Metrics and Baseline Methods

Evaluation metrics: To verify the feasibility and correctness of the proposed $Hi - EAD$ framework, we adopted the Top-1 and Top-5 (accuracy) as the evaluation metrics to evaluate different action recognition models' performance.

Baseline methods: We compared the proposed $Hi - EAD$ to multiple state-of-the-art human action recognition approaches:

- GD-RN [10]. The frameworks integrated Gaussian mixture and dilated convolution residual network (GD-RN), and using Resnet-101 achieved the multi-scale feature, then conducted the recognition of action based on videos.
- T3D-ConvNet [11]. An end-to-end Two-stream convNet fusion networks. The frameworks can recognize actions based on videos and capture multiple features of action.
- SENet [9]. The methods utilize fully connected layers and SE blocks to produce modulation weights from original features and apply the obtained weights to rescale the features to improve the recognition accuracy;
- TS-ConvNet [12]. The method is a two-stream ConvNet architecture that incorporates spatial and temporal networks. This demonstrates that a ConvNet trained on multi-frame dense optical flow can achieve better performance with little training data.
- CNN+LSTM [13]. The approaches employ a recurrent neural network that uses long short-term memory (LSTM) cells connected to the underlying CNN's output and then achieve recognition of human action in videos.
- BCA [14]. Concatenates the output of multiple attention units applied in parallel via attention clusters (BACs) and introduces a shifting operation to capture more diverse signals.
- ISPA [15]. The methods are referred to as interaction-aware spatiotemporal pyramid attention networks. It can learn the correlation of local features at neighbouring spatial positions.
- ST Multiplier [16]. This method uses a general ConvNet architecture for video action recognition based on the multiplicative interaction of spacetime features. The appearance and motion pathways of a two-stream architecture are combined by motion gating and trained end-to-end.

4.4. Experimental Results and Analysis

4.4.1. Comparison with Other State-of-the-Art Methods

In this section, to prove that our proposed *Hi-EAD* model can more effectively capture the differences of different actions in this video and can also better extract the correlation of neighbouring frames with the same video sequence or segment, we provide detailed experimental results and a comparison of our results with other state-of-the-art human action recognition methods. Table 1 gives the experimental results of the different human action recognition methods on the UCF-101 and HMDB-51 baseline datasets.

However, to show that our proposed model has good convergence and robustness, we give the loss curves of all recognition frameworks in Figure 2.

Table 1. The experimental results of different recognition methods. "$Hi-EAD(T=8)$" indicates our proposed model. The black value indicates the best recognition result.

Model-Data	UCF-101		HMDB-51	
	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)
GDRN	68.49	91.45	49.8	72.95
T3D-ConvNet	67.92	88.51	47.16	71.33
SENet	66.63	87.12	42.73	68.83
TS-ConvNet	59.91	80.88	41.98	67.11
CNN+LSTM	47.21	71.89	33.78	58.41
BCA	59.01	80.46	39.01	63.34
ISPA	54.85	75.83	35.83	59.42
ST Multiplier	65.37	84.5	39.64	65.72
Hi-EAD (T = 8)	**71.97**	**93.5**	**58.78**	**76.96**

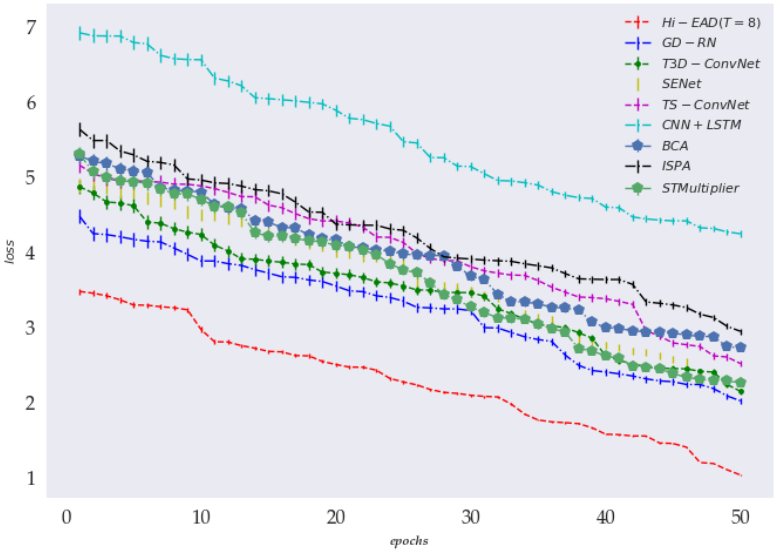

Figure 2. The loss of different recognition frameworks, where *epochs* indicates the number of iterations and *loss* indicates the loss function.

The following conclusions can be attained according to the experimental results presented in Table 1 and Figure 2.

(1) For the UCF-101 and HMDB-51 baseline datasets, our proposed Hi-EAD human action recognition framework achieved the optimal performance. For instance, the top-5 of the $Hi-EAD$ model was higher than that of GDRN by 2.05% and 4.01%, respectively, on the UCF-101 and HMDB-51 datasets; simultaneously, the loss value was in the lowest state. The experimental results indicated that our proposed Hi-EAD could better capture the correlation of the action between the frames. In addition, the high-order semantic information of the action and frame sequence was distinctly formulated and combined with the low-order important detail information by multiple frame excitation aggregates (MFEA) and the squeeze-and-excitation hierarchical disentanglement (SEHD) module. Moreover, the multiscale and hierarchical structure information were aggregated through this module, which could facilitate human action recognition.

(2) For all the experimental results on the UCF-101 and HMDB-51 baseline datasets, the end-to-end models containing 'ConvNet' and 'SE' components outperformed the other methods in terms of the Top-5 and Top-1 on human action recognition. For instance, the action recognition Top-1 of the BCA and CNN+LSTM models was smaller than that of the T3D-ConvNet by 8.91% and 20.71% on the UCF-101 datasets. The main reason is that the 'ConvNet' and 'SE' components could effectively excavate the depth feature formation of the action image. Moreover, the interaction between the human action of neighbour frame

sequences was strengthened. This finding demonstrated the superiority of the 'ConvNet' and 'SE' components on human action recognition.

4.4.2. Discussion of the Component

To further demonstrate the effectiveness of our proposed Hi-EAD framework of multiple frame excitation aggregates (MFEA), squeeze-and-excitation hierarchical disentanglement (SEHD), multi-head self-attention guided and other components, we provided and analyzed the experimental results for each component on the UCF-101 and HMDB-51 benchmark datasets in the human action recognition task. The experiment results are shown in Table 2.

Table 2. The experimental results of different components. '$No-$' indicates that this structure is not included in the models, '$Hi-EAD(T=8)$' indicates our proposed human action frameworks, 'SA' indicates multi-head self-attention guided layers, and 'SE' indicates the block of squeeze-and-excitation.

Model-Data	UCF-101		HMDB-51	
	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)
MFEA	57.37	76.9	40.16	58.08
SEHD	58.91	77.27	40.38	59.53
DenseNet(201)	60.06	78.57	42.91	61.57
DenseNet(121)	62.47	81.26	43.44	63.27
DenseNet(161)	64.75	83.39	43.65	66.19
Hi-EAD (No-MFEA)	68.75	84.64	44.55	66.49
Hi-EAD (No-SEHD)	68.97	84.93	48.98	66.5
Hi-EAD (No-SE)	69.97	85.54	50.75	72.01
Hi-EAD (No-SA)	70.28	86.21	55.81	73.53
Hi-EAD (T = 16)	70.65	87.44	56.16	74.01
Hi-EAD (T = 4)	71.11	90.70	56.42	75.33
Hi-EAD (T = 8)	**71.97**	**93.5**	**58.78**	**76.96**

The following conclusions can be obtained considering the experimental results in Table 2.

(1) In terms of the DenseNet module, DenseNet(161) achieved better performance than those of the other DenseNet models in terms of human action recognition, with 83.39% and 66.19% (Top-5) values on the two baseline datasets.

We can also clearly observe that the SEHD component outperformed MFEA by 0.37% and 1.45% (top-5), respectively. We found that both squeeze-and-excitation and multi-head self-attention guided layers were successfully implemented and worked well enough in our backbone networks (DenseNet-169). These observations were compatible with our assumption that multi-head self-guided attention layers were useful in capturing structural information of action 0 and the long-distance dependence of dynamic frame sequences.

(2) For the integration component of Hi-EAD, the $Hi-EAD$ (No-SA) and $Hi-EAD(No-SE)$ models outperformed the $Hi-EAD(No-MFEA)$ and $Hi-EAD(No-SEHD)$ on the UCF-101 and HMDB-51 datasets. The recognition performance indicated that the integration representation from multiple components was more important than the single structural embedding. The main reason is that the integration module (our proposed framework) could build interaction and complementary information via different components and could reduce the use of irrelevant, redundant information.

(3) We can find that for the shallower model, using backbone of DenseNet(169) showed better performance compared to the more complex ones. The Hi-EAD action recognition models exhibited different performances when we changed the frame sequence of T, such as $T=4$ or $T=16$. Compared to the model without the multi-head self-guided attention ($Hi-EAD(No-SA)$) and squeeze-and-excitation (Hi-EAD (No-SE)) technique, the framework with the two modules (Hi-EAD ($T=16$)) achieved a higher performance; how-

ever, this human action recognition framework still exhibited a competitive performance compared to that of the other state-of-the-art methods. The main reason is that hierarchical multiscale and multi-frame aggregation could effectively obtain the deep semantic and detailed information of human action by aggregating and transferring neighbour frame information. Moreover, the interaction between the adjacent action was used to strengthen the dependence between the action and frames. In addition, this information flow was gradually refined through multi-head self-guided attention, a series of components, thereby reducing the use of redundant information.

4.4.3. Analysis Interaction of Hierarchical and Cross-Modal Information

To further prove the effectiveness of hierarchical feature extraction and cross-modal interaction strategies, we present a series of detailed experiments, and the experimental results are presented in Table 3.

Table 3. The interaction results of hierarchical and cross modal information.

Model-Data	UCF-101		HMDB-51	
	Top-1(%)	Top-5(%)	Top-1(%)	Top-5(%)
Hi-EAD (No-CMI)	65.2	88.36	48.23	68.89
Hi-EAD (No-Hi)	66.6	89.66	49.38	70.63
Hi-EAD(No-HR)	67.18	90.96	51.15	72.39
Hi-EAD(MLP+FC)	69.64	91.14	53.13	73.66
Hi-EAD (Conv1+GAP+FC)	70.7	92.73	55.75	74.26
Hi-EAD(T=8)	**71.97**	**93.5**	**58.78**	**76.96**

In Table 3, 'Hi-EAD (No-CMI)' indicates the model without cross-modal interaction information. 'Hi-EAD (No-Hi)' indicates the model without the operation of hierarchical feature extraction. That is, we only used ordinary multiple branch stream processing of these action recognition frameworks. 'Hi-EAD(No-HR)' indicates the networks without the component of hierarchical representation. '$Hi-EAD(MLP+FC)$' indicates that we use the Multi-Layer Perceptron (MLP), and fully connected layers replace the hierarchical representation layer and multi-head self-guided attention layers. '$Hi-EAD(Conv1+GAP+FC)$' means that we used a convolutional layer with a convolution kernel of $1*1$, global average pooling and a fully connected layer to replace the layered presentation layer and multi-head self-guided attention mechanism.

We can draw the following relevant conclusions via Table 3.

(1) In terms of the HMDB-51 datasets (Top-5(%)), the Hi-EAD(No-HR) methods outperformed the $Hi-EAD(No-CMI)$ and $Hi-EAD(No-Hi)$ by 3.5% and 1.73%, respectively. The main reason is that cross-modal interaction could establish an effective dependency relationship between different information flows, strengthen the information interaction between different frames, reduce the difference in information flow between different actions, and improve the complementarity between them. The hierarchical feature extraction mechanism could capture different feature information levels and describe human actions from different perspectives. This also further proved the effectiveness of the hierarchical feature extraction and cross-modal interaction mechanism used in this article.

(2) From the results of $Hi-EAD(Conv1+GAP+FC)$ and $Hi-EAD(MLP+FC)$, we can clearly observe that Hi-EAD (Conv1+FC) was better than Hi-EAD (MLP+FC) on the UCF-101 and HMDB-51 baseline datasets. The main reason is that the $1*1$ convolutional layer and global average pooling layer effectively refined the feature information and further integrated different feature information streams.

4.5. Ablation Study and Discussion

This section provides an ablation study on the influence of different numbers of training samples, SE blocks, and SA blocks. At the same time, frame sequences of different lengths and related visualization results are included.

4.5.1. The number of Training Samples

The performance of the two human action recognition models under different settings for training with different numbers of training samples is shown in Figure 3. In the experiment, we considered four training settings, that is, 1%, 10%, 20%, and 30% of the number of training samples.

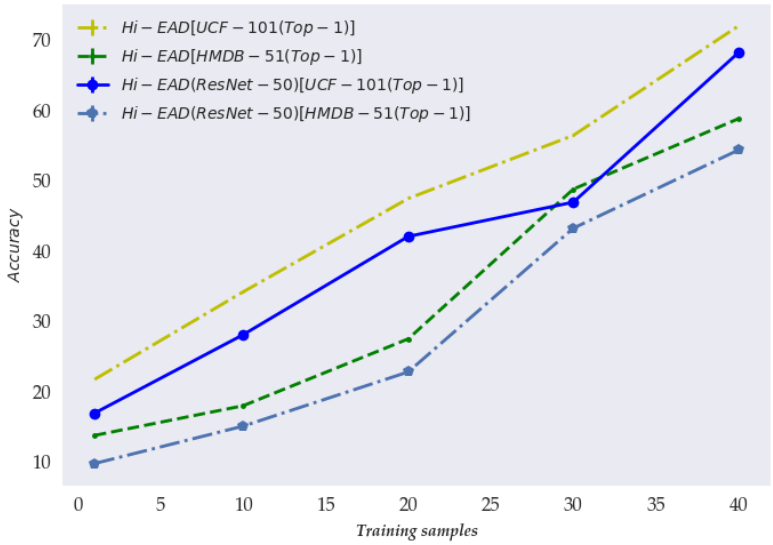

Figure 3. The experimental results of different human action recognition frameworks on different training samples, where 'ResNet-50' indicates the backbone and the networks are ResNet-50.

According to the experimental results shown in Figure 3, the backbone networks of DenseNet-169 (our proposed model) achieved a better performance of human action recognition on the UCF-101 and HMDB-51 baseline datasets. This finding suggesteed that the multi-scale feature extractor and reuse played a more important role. Specifically, this technique could assist the multiscale DenseNet-169 to aggregate the neighbourhood information and produce the optimal representation for the important detail information.

In contrast, as the number of training samples increased, the performance gradually increased on all baseline datasets and recognition models. The highest performance corresponded to the Hi-EAD when using 1% of the training samples, and the Top-1 was 21.71%, corresponding to an improvement of 4.76% for the Hi-EAD(ResNet-50) methods on one UCF-101 dataset. These experimental results showed that our proposed action recognition framework of Hi-EAD was more effective when using small-scale datasets.

4.5.2. The Number Size of SE and SA Blocks

This section discusses the performance of the different numbers for SE and SA blocks. Experimental verification was performed on the HMDB-51 dataset, and the results are shown in following Figures.

We can see that according to the experimental results in Figure 4, as the number of SE and SA blocks increased, the performance gradually increased on the UCF-101 baseline datasets. These experimental results further suggested the superiority of the embedding representation of SE blocks and multi-head self-guided attention technique in capturing and aggregating the important detail semantic information of the action from neighbour frame sequences.

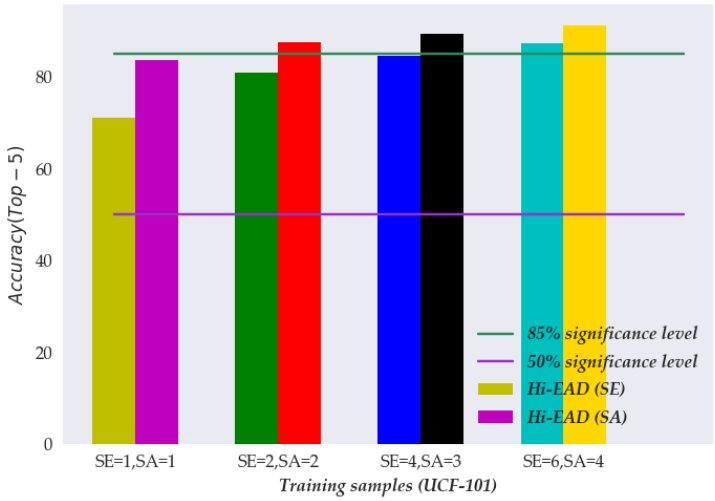

Figure 4. The experimental results of the different numbers for SE and SA blocks, where $SE = 1, 2, 4, 6, 8$ and $SA = \{1, 2, 3, 4\}$ indicate the embedding number of SE and SA blocks. However, $SE = 8$ and $SA = 4$ were used in our proposed Hi-EAD frameworks.

4.5.3. Visual Demonstration of Hi-EAD

In this section, we provide detailed visual results of our proposed Hi-EAD frameworks, where the results are illustrated in Figure 5.

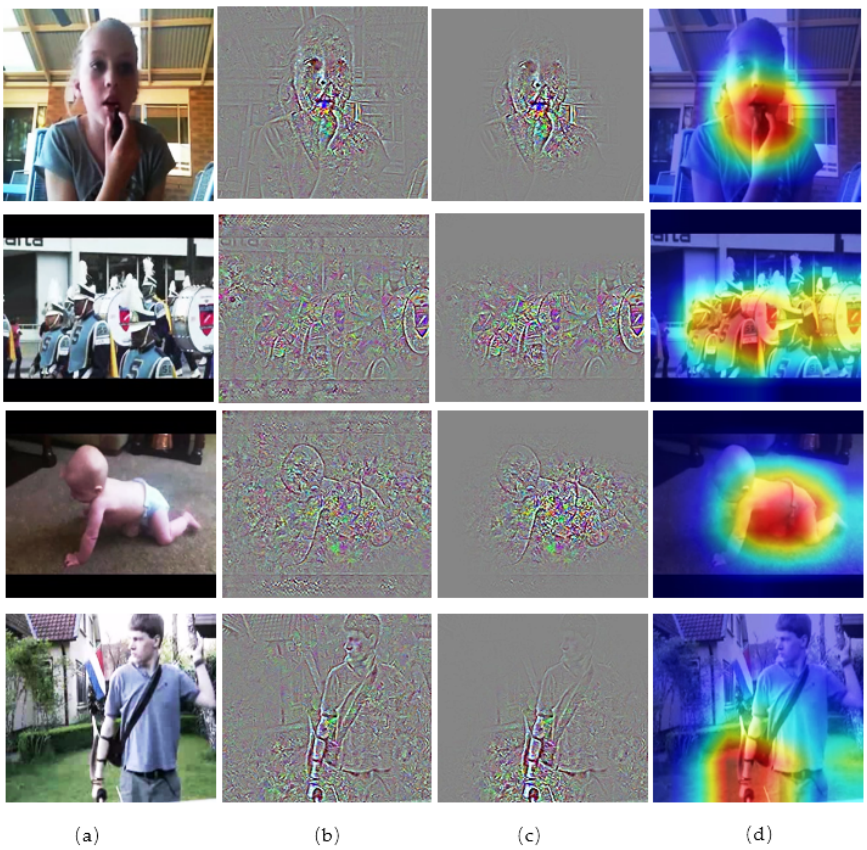

Figure 5. The visual results of the proposed Hi-EAD model. Where (**a**) indicates the initial RGB image. (**b**,**c**) indicate a middle feature map of the action. (**d**) indicates the feature map via multi-head self-guided attention.

The class activation map indicates the discriminative image regions used by the proposed Hi-EAD frameworks to identify an action. From the results, we can see that the

main focus was on the action, and there was comparably less attention to other aspects of the image. Since the idea with SA was to learn the entire picture, the network started learning every single detail around the action, and the final decision on action recognition was made based on the aggregation of action and the environment. From qualitative results, we can conclude that the network's ability to identify the action region increased. All of these results proved that our trained model could focus on important and meaningful actions.

4.6. Effectiveness of the Loss Function

To verify the effectiveness of the designed loss function, we use the UCF-101 and HMDB-51 baseline datasets. Figure 6 shows the experimental results.

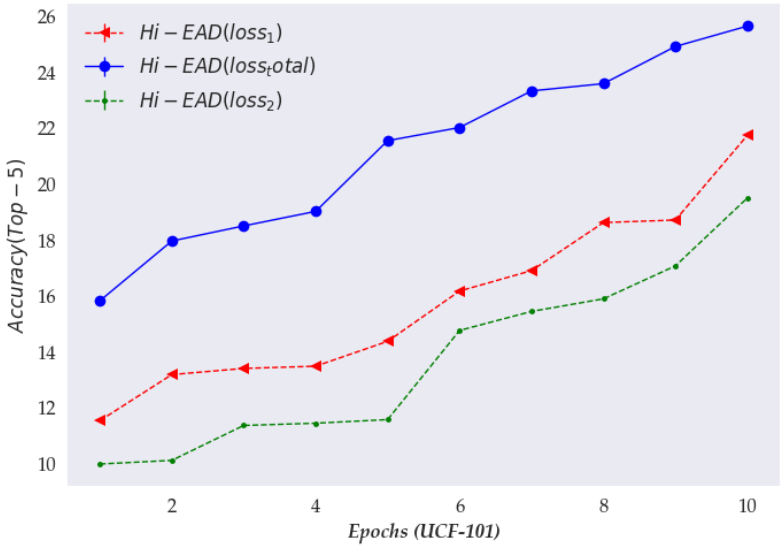

Figure 6. The experimental results of different loss functions. $loss_1$ indicates the multiclass cross-entropy loss ς_{ce}; $loss_2$ indicates the reconstruction loss ς_{rec}; $loss_{total}$ indicates the overall loss ς_{Total}.

According to Figure 6, the performance of the model using the total loss outperformed the model using the binary cross-entropy and multiclass cross-entropy loss, specifically on the UCF-101 datasets. In addition, in the figure, we can also clearly observe that when using two-class and multi-class cross-entropy to optimize the training process, the error was large; that is, the performance of this method in video action recognition was limited, and the difference between consecutive frames or actions weakened the ability to interact. This finding proved the effectiveness of the designed loss function.

5. Conclusions

We propose a hierarchical excitation aggregation and disentanglement network (Hi-EAD) to realize action cognition in video. The framework can effectively capture the possible relation between action and frames through multiple frame excitation aggregation (MFEA) and a feature squeeze-and-excitation hierarchical disentanglement (SEHD) module. The experimental results show that our proposed Hi-EAD frameworks of action recognition based on videos outperform the other traditional approaches based on CNNs on the UCF-101 and HMDB-51 published baseline datasets. Moreover, the effectiveness and reliability of our proposed Hi-EAD framework is verified for action recognition tasks based on videos.

Future research will aim to improve the network structure and enhance feature information extraction. Specifically, we aim to design a simple and efficient semantic framework to accurately extract the feature information. Network structure: Design a more robust graph structure or use graph attention networks to reduce the loss of information flow passing between the nodes and layers. External feature information: Design an effective feature extraction network, such as by improving the DensNet structure, such as embedding

hierarchical squeeze-and-excitation and self-attention, and retain the most discriminative detail information of different actions or frames.

Author Contributions: Z.H. designed the experiment to evaluate the performance and wrote the paper. E.-J.L. supervised the study and reviewed the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank the anonymous reviewers for their comments and helpful suggestions.

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

Abbreviations

The following abbreviations are used in this manuscript:

Hi-EAD	Hierarchical Excitation Aggregation and Disentanglement Networks
MFEA	Multiple Frames Excitation Aggregation
SEHD	Squeeze-and-Excitation Hierarchical Disentanglement
GAP	Global Average Pooling
SA	multi-head self-guided Attention

References

1. Yang, C.; Xu, Y.; Shi, J.; Dai, B.; Zhou, B. Temporal pyramid network for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 591–600.
2. Saponara, S.; Greco, M.S.; Gini, F. Radar-on-chip/in-package in autonomous driving vehicles and intelligent transport systems: Opportunities and challenges. *IEEE Signal Process. Mag.* **2019**, *36*, 71–84. [CrossRef]
3. An, F.P. Human action recognition algorithm based on adaptive initialization of deep learning model parameters and support vector machine. *IEEE Access* **2018**, *6*, 59405–59421. [CrossRef]
4. Yang, H.; Yuan, C.; Li, B.; Du, Y.; Xing, J.; Hu, W.; Maybank, S.J. Asymmetric 3d convolutional neural networks for action recognition. *Pattern Recognit.* **2019**, *85*, 1–12. [CrossRef]
5. Chen, X.; Weng, J.; Lu, W.; Xu, J.; Weng, J. Deep manifold learning combined with convolutional neural networks for action recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *29*, 3938–3952. [CrossRef]
6. Jing, C.; Wei, P.; Sun, H.; Zheng, N. Spatiotemporal neural networks for action recognition based on joint loss. *Neural Comput. Appl.* **2020**, *32*, 4293–4302. [CrossRef]
7. Li, J.; Liu, X.; Zhang, W.; Zhang, M.; Song, J.; Sebe, N. Spatio-temporal attention networks for action recognition and detection. *IEEE Trans. Multimed.* **2020**, *22*, 2990–3001. [CrossRef]
8. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010.
9. Peng, Y.; Shu, T.; Lu, H. Weak integration of form and motion in two-stream CNNs for action recognition. *J. Vis.* **2020**, *20*, 615. [CrossRef]
10. Lin, Y.; Chi, W.; Sun, W.; Liu, S.; Fan, D. Human Action Recognition Algorithm Based on Improved ResNet and Skeletal Keypoints in Single Image. *Math. Probl. Eng.* **2020**, *2020*, 6954174. [CrossRef]
11. Bose , S.R.; Kumar, V.S. An Efficient Inception V2 based Deep Convolutional Neural Network for Real-Time Hand Action Recognition. *IET Image Process.* **2019**, *14*, 688–696. [CrossRef]
12. Li, W.; Feng, C.; Xiao, B.; Chen, Y. Binary Hashing CNN Features for Action Recognition. *TIIS* **2018**, *12*, 4412–4428.
13. Rahman, S.; See, J.; Ho, C.C. Deep CNN object features for improved action recognition in low quality videos. *Adv. Sci. Lett.* **2017**, *23*, 11360–11364. [CrossRef]
14. Cherian, A.; Gould, S. Second-order Temporal Pooling for Action Recognition. *Int. J. Comput. Vis.* **2019**, *127*, 340–362. [CrossRef]
15. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
16. Seemanthini, K.; Manjunath, S.S. Human Detection and Tracking using HOG for Action Recognition. *Procedia Comput. Sci.* **2018**, *132*, 1317–1326.
17. Chen, W.Q.; Xiao, G.Q.; Tang, X.Q. An Action Recognition Model Based on the Bayesian Networks. *Appl. Mech. Mater.* **2014**, *513*, 886–1889. [CrossRef]

18. Tran, D.T.; Yamazoe, H.; Lee, J.H. Multi-scale affined-HOF and dimension selection for view-unconstrained action recognition. *Appl. Intell.* **2020**, *50*, 1–19. [CrossRef]
19. Wang, L.; Koniusz, P.; Huynh, D.Q. Hallucinating Bag-of-Words and Fisher Vector IDT terms for CNN-based Action Recognition. *arXiv* **2019**, arXiv:1906.05910.
20. Wang, L.; Zhi-Pan, W.U. A Comparative Review of Recent Kinect-based Action Recognition Algorithms. *arXiv* **2019**, arXiv:1906.09955.
21. Jagadeesh, B.; Patil, C.M. Video based action detection and recognition human using optical flow and SVM classifier. In Proceedings of the 2016 IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT), Bangalore, India, 20–21 May 2016; IEEE: New York, NY, USA, 2016.
22. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; pp. 1725–1732. Available online: https://dl.acm.org/doi/10.1109/CVPR.2014.223 (accessed on 10 March 2021)
23. Patil, G.G.; Banyal, R.K. Techniques of Deep Learning for Image Recognition. In Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Pune, India, 29–31 March 2019; IEEE: New York, NY, USA, 2019.
24. Kang, B.R.; Lee, H.; Park, K.; Ryu, H.; Kim, H.Y. BshapeNet: Object Detection and Instance Segmentation with Bounding Shape Masks. *Pattern Recognit. Lett.* **2020**, *131*, 449–455. [CrossRef]
25. Sungheetha, A.; Rajesh, S.R. Comparative Study: Statistical Approach and Deep Learning Method for Automatic Segmentation Methods for Lung CT Image Segmentation. *J. Innov. Image Process.* **2020**, *2*, 187–193. [CrossRef]
26. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. *arXiv* **2014**, arXiv:1406.2199.
27. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term Recurrent Convolutional Networks for Visual Recognition and, Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [CrossRef]
28. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
29. Feichtenhofer, C.; Pinz, A.; Wildes, R.P. Spatiotemporal Residual Networks for Video Action Recognition. 2017. Available online: https://papers.nips.cc/paper/2016/file/3e7e0224018ab3cf51abb96464d518cd-Paper.pdf (accessed on 10 March 2021).
30. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
31. Li, C.; Zhong, Q.; Xie, D.; Pu, S. Skeleton-based Action Recognition with Convolutional Neural Networks. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017.
32. Liao, X.; He, L.; Yang, Z.; Zhang, C. Video-based Person Re-identification via 3D Convolutional Networks and Non-local Attention. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2018.
33. Kalfaoglu, M.E.; Kalkan, S.; Alatan, A. Late Temporal Modeling in 3D CNN Architectures with BERT for Action Recognition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020.
34. Anvarov, F.; Kim, D.H.; Song, B.C. Action Recognition Using Deep 3D CNNs with Sequential Feature Aggregation and Attention. *Electronics* **2020**, *9*, 147. [CrossRef]
35. Jalal, M.A.; Aftab, W.; Moore, R.K.; Mihaylova, L. Dual stream spatio-temporal motion fusion with self-attention for action recognition. In Proceedings of the 22nd International Conference on Information Fusion, Ottawa, ON, Canada, 2–5 July 2019.
36. Purwanto, D.; Pramono, R.R.; Chen, Y.T.; Fang, W.H. Three-Stream Network with Bidirectional Self-Attention for Action Recognition in Extreme Low-Resolution Videos. *IEEE Signal Process. Lett.* **2019**, *26*, 1187–1191. [CrossRef]
37. Yu, T.; Guo, C.; Wang, L.; Gu, H.; Xiang, S.; Pan, C. Joint Spatial-Temporal Attention for Action Recognition. *Pattern Recognit. Lett.* **2018**, *112*, 226–233. [CrossRef]