future internet

MDPI

Article

Relational Action Bank with Semantic–Visual Attention for Few-Shot Action Recognition

Haoming Liang [1], Jinze Du [1,*], Hongchen Zhang [1], Bing Han [1] and Yan Ma [2]

[1] School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China
[2] State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
* Correspondence: dujz05@lut.edu.cn; Tel.: +86-139-1925-0662

Abstract: Recently, few-shot learning has attracted significant attention in the field of video action recognition, owing to its data-efficient learning paradigm. Despite the encouraging progress, identifying ways to further improve the few-shot learning performance by exploring additional or auxiliary information for video action recognition remains an ongoing challenge. To address this problem, in this paper we make the first attempt to propose a relational action bank with semantic–visual attention for few-shot action recognition. Specifically, we introduce a relational action bank as the auxiliary library to assist the network in understanding the actions in novel classes. Meanwhile, the semantic–visual attention is devised to adaptively capture the connections to the foregone actions via both semantic correlation and visual similarity. We extensively evaluate our approach via two backbone models (ResNet-50 and C3D) on HMDB and Kinetics datasets, and demonstrate that the proposed model can obtain significantly better performance compared against state-of-the-art methods. Notably, our results demonstrate an average improvement of about 6.2% when compared to the second-best method on the Kinetics dataset.

Keywords: semantic attention; visual attention; relational action bank; few-shot learning; action recognition

Citation: Liang, H.; Du, J.; Zhang, H.; Han, B.; Ma, Y. Relational Action Bank with Semantic–Visual Attention for Few-Shot Action Recognition. *Future Internet* **2023**, *15*, 101. https://doi.org/10.3390/fi15030101

Academic Editor: Paolo Bellavista

Received: 5 February 2023
Revised: 23 February 2023
Accepted: 27 February 2023
Published: 3 March 2023

1. Introduction

In recent years, video action recognition has been well studied. However, conventional action recognition requires a large amount of labelled data for training, such as Sports-1M [1] and Kinetics-700 [2] datasets. Labeling these video data for training is a highly expensive process. Identifying methods that use a smaller amount of labelled data to train a better available model remains an open issue that needs to be solved urgently. For image classification, there are many classical methods to solve few example issues. The Siamese Network [3] enabled the model to generalize from few examples, through limiting assumptions on the input structure and acquiring features automatically. To eliminate the process of fine-tuning to adapt to new class types, ref. [4] designed Match Network that maps a small labelled support set and an unlabelled example to its label. The work in [5] introduced Prototype Network, a few example image-classification method that computes distances to prototype representations of each class in the learned metric space.

However, for video action recognition, there remains a lack of effective works. Ref. [6] attempted to reproduce the methods from image to video and proposed CMN method to solve the few example video action recognition issue, which improved the performance of action recognition compared against reproduced methods. Another method presented by [7] can simultaneously capture local and long-term spatial temporal information employing the proposed dilated dense network, whose blocks consist of densely connected dilated convolutions layers. Subsequently, Bishay et al. [8] introduced the temporal attentive relation network (TARN) to perform temporal alignment and learn a deep-distance measure on the aligned representations at the video segment level.

Although some researchers have studied few-example-based video action recognition, they all omit the existing prior knowledge. Consequently, they have suffered large losses in performance. As shown in Figure 1, we can see that the "shoot ball" action is related to the existing actions, especially the action "turn" and "shoot bow". As we all know, when humans memorize and recognize novel few example actions, we will use other related and familiar action categories to assist this process. How can we make CNNs perform few-shot action recognition like a human? How can CNNs be enabled as to exploit past knowledge? The naive method is that we train the CNNs on the past knowledge and fine-tune the network on novel few shot categories; however, the subsequent performance of the action recognition is poor.

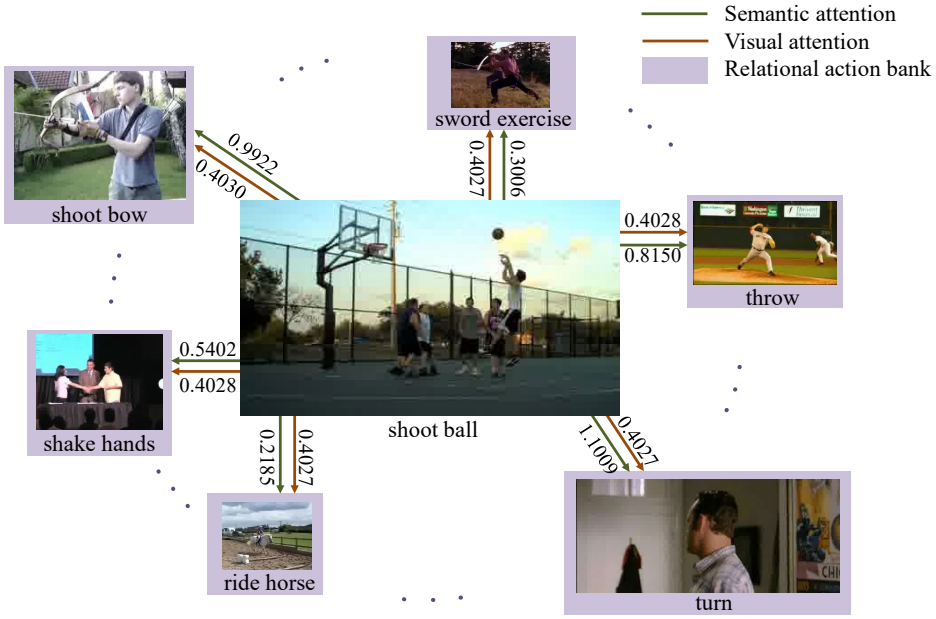


Figure 1. The relevancy between the few labelled action "shoot ball" and the actions in the relational action bank. The value on the green arrow is the output of the semantic attention submodule, which displays the semantic relevancy, and the value on the brown arrow is the output of the visual attention submodule, which displays the visual similarity.

To solve this problem, we propose a novel relational action bank with semantic–visual attention for few example action recognition. Our proposed RABSVA module consists of two submodules, relational action bank and semantic–visual attention. The action bank provides prior knowledge for the network and the semantic–visual attention mechanism generates attention weights to take advantage of the action bank. Thus, the RABSVA can utilize both the semantic attention and visual similarity attention simultaneously to acquire useful action information from the action bank. At the same time, our proposed action bank can learn the clip relations in each video and then use graph neural network (abbreviated as GNN) to cluster the clips of each class to obtain the representation of each category.

Overall, our contributions can be summarized as follows:

1. To the best of our knowledge, we are the first to propose a relational action bank to assist the few-shot action recognition.
2. The proposed relational action bank can both enhance the representation ability of clips in each video and learn the feature vector representations for each class adaptively.
3. We propose a semantic–visual attention mechanism, which can utilize the relational action bank both in semantic categories and visual similarities.
4. Our proposed method obtains state-of-the-art performance in few-shot action recognition on Kinetics dataset, notably achieving an average improvement of 6.2%. Further-

more, we achieve the improved performance on the HMDB dataset compared to the original model.

The remainder of the paper is arranged as follows: A review of the related work is presented in Section 2. In Section 3, the detailed network structure design is elaborated upon. Related experiments are shown in Section 4. Lastly, concluding remarks are provided in Section 5.

2. Related Works

This work is related to many topics including video action recognition, attention mechanism, and few-shot learning. We will give an overview of these in this section.

2.1. Video Action Recognition

Video action recognition task has recently become a popular research topic in the computer vision community. With the great success of deep convolutional networks in computer vision, especially in the field of images, massive CNN-based approaches [9–16] have been introduced for video action recognition, gradually surpassing the accuracies of traditional approaches [17,18]. One approach attempts to learn spatial–temporal features directly from RGB frames [19–22] using 3D CNNs. C3D [20] is an early work that employs deep 3D CNNs to model spatio-temporal features and exposes the source code. However, the performance of C3D remains unsatisfactory due to the large number of parameters to optimize for 3D convolution and the lack of high-quality and large-scale assessment datasets. I3D [22] extends Imagenet's pre-trained 2D kernel to 3D to capture spatio-temporal features and uses another stream to model motion features. Furthermore, with the help of a high-quality and large-scale Kinetics dataset and a two-stream setup, I3D achieves a highly competitive performance in the benchmark dataset. To solve the problem whereby most conventional 3D networks can only use local spatio-temporal features, ref. [23] introduces the Spatio-Temporal Channel Correlation (abbreviated as STC) block to model the correlation between channels of 3D CNNs involving both temporal and spatial features. 3D-CNN is a natural extension of 2D-CNN for learning spatio-temporal features in videos and has long been used for action recognition. Due to the vast number of parameters, 3D-CNNs require a large number of videos to learn a promising representation. Since I3D [22], 3D-CNN has become the dominant method for video action recognition task. Since then, the action recognition community has proposed many advanced 3D-CNN models [24,25] that outperform I3D in both accuracy and efficiency. Ref. [25] proposes an available pseudo labeling algorithm, named Cross-Model Pseudo Labeling (abbreviated as CMPL), for video action recognition. Specifically, by introducing a lightweight auxiliary network on top of the backbone network and asking them to predict pseudo label to each other. In [24], the authors propose the PoseConv3D framework as a competitive alternative to the GCN-based approach for action recognition.

Recently, with the great success of transformer networks in the field of Natural Language Processing, many researchers have started to apply the technique to video understanding [26–31]. Since a complex action can consist of a sequence of atomic actions, ref. [26] designs a probabilistic model, referred to as an Uncertainty Guided Probabilistic Translator (abbreviated as UGPT), for the recognition of complex actions. To train a more powerful video vision transformer for the epic-kitchens-100 action recognition dataset, ref. [32] explores augmentations, resolutions and initialization techniques that combine the video vision transformer with some convolutional video networks to achieve good action recognition performance. In [33], the authors devise a spatial dimension sparse attention enhanced skeleton-based human video action recognition model, which has segmented linear attention in the temporal dimension of the data. To implement a convolution-free video classification, ref. [34] introduces "TimeSformer", which adapts the Transformer architecture to video by learning spatio-temporal features directly from frame-level patch sequences. Subsequently, researchers have proposed numerous refining works [16,35–41] on transformer by combining video spatio-temporal features.

2.2. Attention Mechanism

As deep learning continues to evolve, the attention mechanism [42–44] remains a widely used technique with the effect of enhancing some parts of the input data and weakening others, with the motivation that the network should devote more attention to the important parts of the data and take the edge off its less important parts. In the early stages, attention mechanisms were mainly implemented in the domain of Natural Language Processing (abbreviated as NLP) [45–51]. Later, with the successful application in NLP, an increasing number of researchers started to use the attention mechanism in the computer vision domain [52–57]. Ref. [52] combines a spatial attention mechanism and designs a Deep Recursive Attention Writing (abbreviated as DRAW) architecture for generating images with neural networks. Inspired by the classical non-local means approach in computer vision, ref. [54] proposes non-local operations as a family of generic building blocks for capturing long-range dependencies among different points in the feature maps. For image classification, ref. [58] presents an end-to-end trainable attention module that based on a convolutional neural network (abbreviated as CNN) architecture. Using a new contextual aggregation scheme to address the semantic segmentation task, ref. [59] proposes Ocnet, which focuses on strengthening the effect of object information. Focusing on channel relationships, ref. [60] proposed a classical "squeeze and excitation" (abbreviated as SE) block that adaptively recalibrates the feature responses of channels by explicitly modeling the interdependencies between them. For image super-resolution, ref. [61] proposes a very deep residual channel attention network (abbreviated as RCAN). Moreover, a second-order attention network (abbreviated as SAN) for powerful feature representation and feature correlation modeling was introduced by [62]. For semantic segmentation, to explore the impact of global context information, based on channel attention, ref. [63] introduces a context encoding sub-network, which captures the semantic context of a scene and selectively highlights class-related feature maps.

For computer vision, there are not only the aforementioned separate studies on spatial attention and channel attention, but also some studies on attention mechanisms that combine these two [64–66]. For image captioning, ref. [64] introduces a convolutional neural network, called SCA-CNN, which integrates spatial and channel attention in a CNN. In [65], the authors propose the convolutional block attention module (abbreviated as CBAM) that derives the attention map sequentially along two independent dimensions, named space and channel, for the given intermediate feature maps, and then multiplies the attention map by the input feature map to perform adaptive feature refinement. Focusing on the impact of attention in deep neural networks, ref. [66] devises the Bottleneck Attention Module (abbreviated as BAM), which derives attention maps along two independent paths, named space and channel, and can be embedded into any deep convolutional network.

2.3. Few-Shot Learning

Along with the improvement in requirements on sample utilization efficiency, few-shot learning [67–70] has become one of most popular topics in artificial intelligence. The main problem that needs to be solved is the classification of unseen categories with only a few labelled samples in each category from the training set. In recent years, the major approaches in few-shot learning are metric-based, meta-based and data-augmentation-based methods. The metric-based methods model the distance distribution of samples so that samples of the same category are close while those of different categories are far away. Siamese Network [3], Matching Network [4] and Prototypical Network [5] are three of the most renowned methods. Different from the metric-based method, the meta-based method can be summarized as learning to learn. In particular, it learns from the experience of how different deep learning models perform on a large number of learning tasks in order to learn new tasks faster [71]. The remarkable methods include [72,73]. In [72], the authors make use of a memory-augmented CNN network to rapidly absorb new data, leveraging which the model is able to make accurate predictions with only a few samples. By exploiting an LSTM-based meta-learner model to train another few-shot learner neural

network with the learned optimization algorithm, ref. [73] obtains a meta-learning model that is competitive with deep-metric-learning techniques. Recently, much attention has been focused on generating more samples from a small number of available samples, marking the emergence of data-augmentation-based methods, including [74,75] and so on.

In contrast to few-shot learning in images, FSL research in the video realm has only just begun [7,76–78]. As one of the early studies, for few-shot video action recognition, ref. [6] presents a compound memory network (abbreviated as CMN) framework. Another method by [7], proposed a dense dilated network for few-shot action recognition, and the method can outperform the CMN when using a larger backbone. The drawbacks to these methods are that they collapse the order of frames at representation, so that the learned model is weakened when the sequential order of video is important.By developing a Temporal Alignment Module (abbreviated as TAM), the authors in [79] ameliorate the problem with a few-shot learning framework that specifically leverages the ordering information of the time in video data via temporal alignment. Later, the problem was investigated in [8] with a meta-based approach. In the literature, a novel temporal attention relation network (abbreviated as TARN) learns to contradistinguish representations of variable lengths of time. On the other side, the work in [80] addresses the task of few-shot video action recognition in metric-based method with a set of two-stream models. To avoid the bias of the classifier against the seen category in the above approaches, ref. [76] employs a novel ProtoGAN framework, a conditional generative adversarial network conditional on the category prototype vector, for synthesizing additional samples of novel categories.

3. Method

In this section, we will present the design of our relational action bank with semantic–visual attention (RABSVA) module. We first review the whole architecture and then elaborate the detailed design for each submodule in the following paragraphs.

3.1. The Framework

To recognize few labelled examples, we propose a novel relational action bank with semantic–visual attention module to use past or existing knowledge to assist current action recognition. Taking full advantage of action bank, we present the semantic and visual attention mechanism which considers the learned semantic attention and the visual similarity attention simultaneously. As Figure 2 shows, semantic attention uses the semantic information to mine the semantically related action representations so that get help from the existing actions, while visual attention uses the visual similarity to exploit the hidden useful information from the action bank. We use these two kinds of attention mechanism to mine semantically and visually related actions and form new useful action feature representation, which is used to obtain the new attention weights for weighting and enhancing the original feature maps $\mathbf{X}$.

In Figure 2, the input video clips pass through the backbone subnet (such as C3D and ResNet-50), which outputs the feature maps $\mathbf{X} \in \mathbb{R}^d$, where d is the dimension of the feature maps (here the reshape operation is omitted for simplicity, and if not specified the one dimensional vector is a row vector). Then the feature maps pass through our proposed RABSVA module and the output is the enhanced feature maps $\mathbf{Y} \in \mathbb{R}^d$, which are the input of the classifier. Our RABSVA module has two branches. The top of Figure 2 denotes the semantic attention branch and the bottom is the visual attention branch. We fuse the output of semantic and visual branches together and form our RABSVA module.

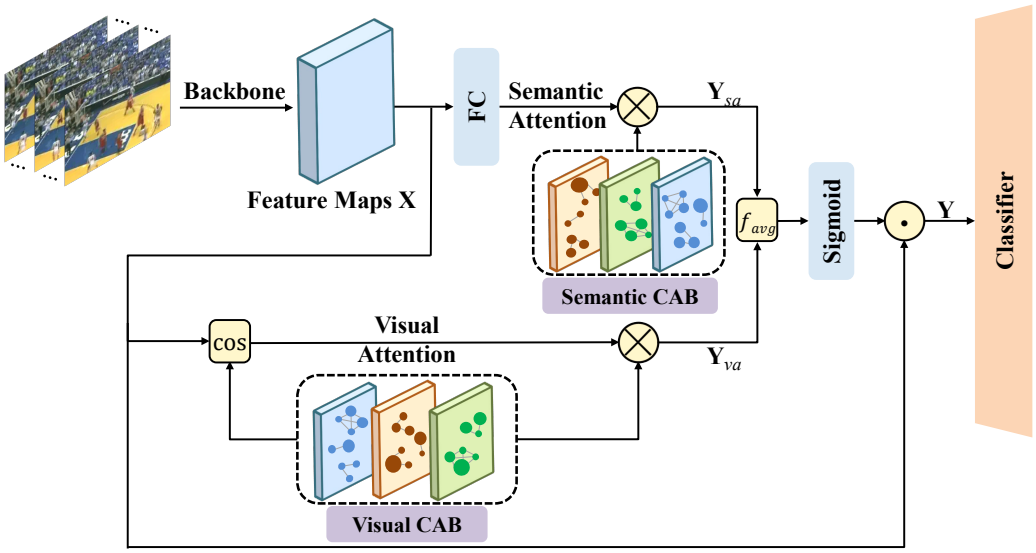


Figure 2. The architecture of our relational action bank with semantic–visual attention module. This module includes relational action bank and semantic–visual attention submodules. $\odot$ denotes the element-wise multiplication. $\otimes$ denotes matrix multiplication.

3.2. Relational Action Bank

The Action Bank is the bank that stores the feature vectors of existing actions. As we all know, we can employ the existing knowledge (the action representations in action bank) to assist the recognition of novel categories. Specifically, we can use the exclusive method as one of the straightforward methods. In fact, we use the features of base category video clips to enhance the modeling ability of the network for novel categories. We cluster video clips of each category into one feature vector as the general representation of corresponding categories. Then, we use these feature vectors to enhance the modeling ability of the network for novel categories.

Here, we first give the definition of relational action bank (RAB), which enables a more robust representation of learned class actions by modeling the relationships between different videos of the same class action. We extract the features of the base category video clips via backbone network and form the feature action bank $\mathbf{Z} = \{\mathbf{z}_{i,j,k} \in \mathbb{R}^d | i \in \{1, 2, \cdots, N\}, j \in \{1, 2, \cdots, v\}, k \in \{1, 2, \cdots, s\}\}$, where N, v and s denote the number of base categories, the number of videos per category, the number of clips in each videos, respectively. $\mathbf{z}_{i,j,k}$ is the output of one video clip passing through the backbone subnet. In fact, the number of videos v and the number of clips s are not the same for different categories and different videos, but we assume v and s are the same for different categories and different videos for simplicity (in practice, we can use the max value as the corresponding v and s, and use zeros to fill corresponding positions when the video or the clip does not exist). The output of RAB is $\widetilde{\mathbf{Z}} = \{\widetilde{\mathbf{z}}_i \in \mathbb{R}^d | i \in \{1, \ldots, N\}\}$. Thus, the RAB submodule only needs to cluster the feature vectors in each category into one feature vector. We propose to use graph neural network (GNN) to model relations of the action clips which are stored in the action bank and to aggregate video clips of the same category into one feature vector. Our relational action bank has N base categories, each category has v videos and each video has s clip vectors which are extracted via pretrained neural network from the base categories.

As the clips in each video have strong relations, inspired by Non-local [54], SENet [60] and bottleneck structure [81], we design our learnable relational action bank as Figure 3 shows, which consists of aggregate operations (f_1 and f_2) and expand operation (g). To learn the strong relations among clips in each video, we use graph neural network (GNN) to realize an aggregate-expand operation (f_1 and g). We aggregate the clips in each video and then resume the dimensions via expand operation. During this aggregate-expand process, our RAB module can learn the relations of clips in one video. At last, we use

GNN to aggregate the relationally enhanced clips of each category into one feature vector. As we have action clip feature vectors $\mathbf{Z}$, we consider every clip feature vector $\mathbf{z}_{i,j,m}$ as one node, and build the edges between nodes in the same video (we only build undirected fully connected graph for simplicity) to build the relations among the clips. We use GNN to accomplish aggregate operations, so here we first give the definition of GNN. Similar to [82,83], we define the GNN operations for our RAB module as follows:

$$\mathbf{z}_{i,j,m}^{(l)} = \sigma(w_m^{(l-1)}\mathbf{z}_{i,j,m}^{(l-1)} + \sum_{k\in\mathcal{N}(m)} w_k^{(l-1)}\mathbf{z}_{i,j,k}^{(l-1)}), \tag{1}$$

where l denotes the l-th layer, $w_m^{(l-1)}$ and $w_k^{(l-1)}$ are learnable parameters, σ is a non-linear function, $\mathcal{N}(m)$ denotes the neighbor nodes of node m (neighbor is defined as having connected relations with node m on the graph). f_1 and f_2 are aggregate operations and g is the expansion function which is the reverse process of aggregation in f_1. We can write the output of our RAB as follow (omit the non-linear function for simplicity):

$$\widetilde{\mathbf{Z}} = f_2 \circ g \circ f_1(\mathbf{Z}). \tag{2}$$

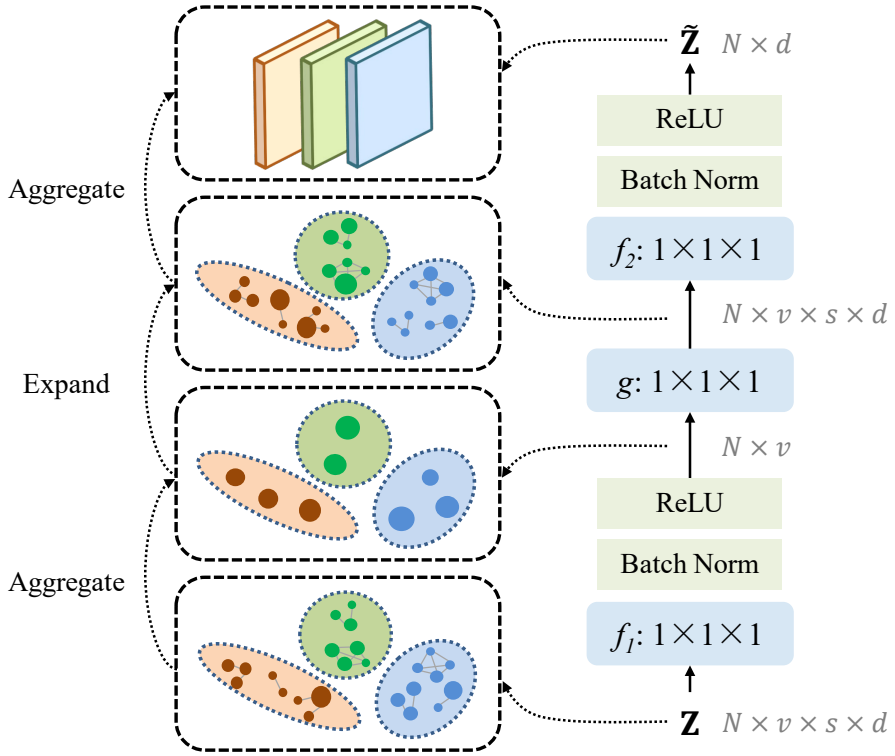


Figure 3. The detailed design of our proposed relational action bank module.

In practice, we use three $1 \times 1 \times 1$ 3D CNNs to realize the aggregation function (f_1 and f_2) and expansion function (g). As Equation (1) shows, we add BN and ReLU as non-linear function between f_1 and g to enhance the non-linear modeling abilities for clip relation modeling, and we add BN and ReLU following GNN operation f_2. As shown in Figure 3, we reshape the input feature maps $\mathbf{Z}$ and consider the $s \times d$ as the channel dimension of input into f_1 ($1 \times 1 \times 1$ 3D CNNs) to aggregate the $s \times d$ into 1. To expand to the original dimension, we still use g ($1 \times 1 \times 1$ 3D CNNs) to resume the channel dimension to $s \times d$ and obtain the clips relation enhanced features. Later, we use the aggregate operation f_2 ($1 \times 1 \times 1$ 3D CNNs) to aggregate all clips in each category into one feature vector. During the above process, we choose BN and ReLU as the non-linear function to enhance the non-linear representation ability.

Discussion The basic function of relational action bank is to cluster the clip feature vectors in each category into one feature vector to represent the corresponding category. The most direct idea is to use kmeans to cluster the clips of each category into one vector (we denote it as kmeans-RABSVA for convenient narrative). The kmeans-RABSVA can not dynamically learn the representation feature vector for the corresponding class adaptively and the feature vector is specified when the clips are specified. Another naive idea is that we can use one layer GNN to aggregate the clips in each class into one feature vector (We denote this method as naive-RABSVA). The naive-RABSVA can learn the representation vector adaptively. However, it omits the relations of clips in each video and suffer the much performance loss for that there are strong relations among these clips in each video.

3.3. Semantic–Visual Attention Mechanism

As Figure 2 shows, we propose two types of attention methods (semantic attention and visual similarity attention) to take advantage of relational action bank. We use these two attention mechanisms to acquire useful action knowledge from the bank, and then use this knowledge to form attention weights to weight the feature maps $\mathbf{X}$ and obtain the enhanced feature maps $\mathbf{Y}$. Next, we will describe these two types of attentions in detail.

3.3.1. Semantic Attention

Our semantic attention submodule aims to apply semantic attention to make use of the relational action bank. For this submodule, we need to use the network to learn semantic attention weights. Illuminated by [60] and the later layers of the CNNs is semantic related, so we can just use a FC layer to accomplish attention mechanism simply and efficiently.

For the input video clip, we use the FC layer to learn the semantic attention weights and apply these weights to the existing action bank. We use these attention weights and relational action bank to mine useful action knowledge $\mathbf{Y}_{sa} \in \mathbb{R}^d$ as follows:

$$\mathbf{Y}_{sa} = \mathbf{W}_{na}\widetilde{\mathbf{Z}}^{\mathrm{T}}, \tag{3}$$

where $\mathbf{W}_{na} \in \mathbb{R}^N$ is the output semantic attention weights of FC layer, and $\widetilde{\mathbf{Z}}$ is the output of relational action bank. Then, we use the useful knowledge $\mathbf{Y}_{sa}$ to form a new attention weights to weight the original feature maps $\mathbf{X}$.

3.3.2. Visual Attention

Except semantic attention, we can calculate the visual similarity between input feature maps $\mathbf{X}$ and the category representation vector $\widetilde{\mathbf{z}}_i$, $i \in \{1, 2, \cdots, N\}$, and then apply this physical visual similarity as another type of attention to utilize the action bank. We can write the process as follows:

$$\mathbf{Y}_{va} = \mathbf{W}_{va}\widetilde{\mathbf{Z}}^{\mathrm{T}}, \tag{4}$$

$$\mathbf{W}_{va} = \{w_i = \frac{\mathbf{X}\widetilde{\mathbf{z}}_i^{\mathrm{T}}}{\|\mathbf{X}\|\|\widetilde{\mathbf{z}}_i\|} | i \in \{1, 2, 3, \cdots, N\}\}, \tag{5}$$

where $\mathbf{Y}_{va} \in \mathbb{R}^d$ and $\mathbf{W}_{va} \in \mathbb{R}^N$ denote the acquisition from the action bank via visual attention and the visual similarity attention, respectively. We also omit the reshape operation in Equations (4) and (5).

3.3.3. Fusion

To obtain better performance for few-shot action recognition, we fuse semantic and visual attention branches together. As shown in Figure 2, we define $\mathbf{Y}$ as follows:

$$\mathbf{Y} = Sigmoid \circ f_{avg}(\mathbf{Y}_{sa}, \mathbf{Y}_{va}), \tag{6}$$

$$f_{avg}(\mathbf{Y}_{sa}, \mathbf{Y}_{va}) = \frac{1}{2}(\mathbf{Y}_{sa} + \mathbf{Y}_{va}). \tag{7}$$

As Equation (7) shows above, we simply average the output of $\mathbf{Y}_{sa}$ and $\mathbf{Y}_{va}$ to accomplish the fusion.

4. Experiments

In this section, we evaluate our proposed RABSVA module on two backbones (ResNet50 [84] and C3D [20]) and two widely used datasets (HMDB-51 and Kinetics [6]). In the following, we will first describe the datasets and the implementation details, and then give the ablation analysis for different components in our proposed RABSVA, followed by a comparison with state-of-the-art methods.

4.1. Datasets

4.1.1. HMDB-51

It contains 51 action categories. However, it has two categories in common with Kinetics-400 dataset (pushup and situp), so we omit these two categories from HMDB-51 and for that we use the pretrained model on kinetics to initialize our model training. We split these 49 categories into 40 base categories and 9 novel few example categories, and denote it as HMDB-49. Examples of the dataset can be seen in the top line of Figure 4.

Figure 4. Examples of the dataset visualization. The top line is the examples of HMDB-51 dataset, and the bottom line is the examples of Kinetics dataset.

4.1.2. Kinetics

The original Kinetics dataset includes three versions, Kinetics-400, Kinetics-600 and Kinetics-700, with 400, 600 and 700 action categories, respectively. For few shot learning, we follow the splits in [6] and select 100 classes from Kinetics-400. The training set, validation set and test set correspond 64, 12, 24 categories, respectively, and we denote it as Kinetics-100. The bottom line of Figure 4 shows the action examples.

4.2. Implementation Details

For HMDB-49, we randomly select 40 categories as base classes and the remainders are the novel classes for 3 times, respectively. We first train the model on base classes and obtain the well-trained model. Then, we extract clip features in the base classes via the well-trained model to form our relational action bank **Z**. For Kinetics-100 dataset, we use the splitted base classes and validation classes, which is the same splits as in [6], to train our base model, and then use it to extract clip features from the train set videos of base and validation splits. Furthermore, we organize these clip features into the action bank **Z**. For ResNet-50, we add our RABSVA module between layer4 and classifier, while for C3D network we add it into the classifier. To further reduce the dimensions of the feature maps and improve computing efficiency, we modify the dimension of the middle layer in classifier from 4096 to 1024 in practice. We fine-tune the model embedding our RABSVA based on the well-trained model on the base classes. We set the basic learning rate as 10^{-3} and initialize that of the new added layers with 10 times the basic learning rate. We divide it by 10 for every 200 epochs, and fine-tune the model for 300 epochs in total. Specially, we set the learning rate of the first three layers of ResNet50 with 0.

4.3. Ablation Analysis

To provide a detailed ablation analysis of various aspects of our RABSVA module, we employ ResNet-50 as a backbone and conduct an experiment on the HMDB-49 dataset.

4.3.1. Naive Aggregation, K-Means Aggregation and Relational Action Bank

As Table 1 shows, simply aggregate the clips in every category via a parameter matrix (Naive-RABSVA) can obtain better performance compared with baseline model in most cases. Meanwhile, we use the kmeans (Kmeans-RABSVA) to aggregate the clips of the same class into one feature vector, which can also acquire a better performance compared to baseline model. Overall, RABSVA can achieve the best action recognition performance, a phenomenon that illustrates the importance of modeling the relationship between different clips of each video, and further demonstrates the effectiveness of our proposed module.

Table 1. The top-1 performance of 5-shot using 3D ResNet-50 as backbone on HMDB-49.

Method	1st Run	2nd Run	3rd Run	AVG
Baseline	77.4	77.0	71.9	75.4
RABSA	78.9	78.9	72.6	76.8
RABVA	79.3	79.3	72.2	76.9
Naive-RABSVA	**80.7**	77.0	71.5	76.4
Kmeans-RABSVA	**80.7**	77.0	**75.2**	77.6
RABSVA	**80.7**	**80.7**	**75.2**	**78.9**

4.3.2. Semantic Attention, Visual Attention and Semantic–Visual Attention

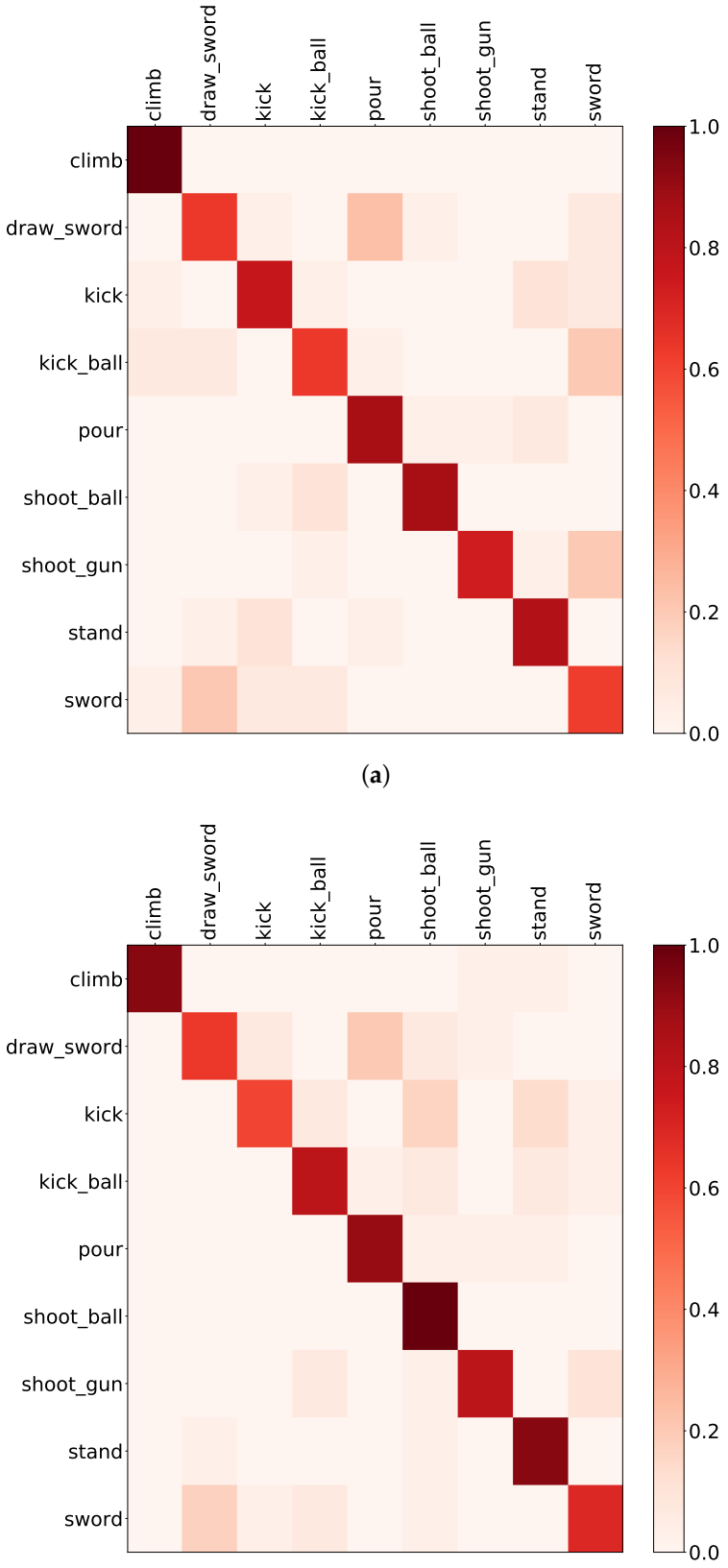
From Table 1, it is easy to see that Semantic attention branch (RABSA) or Visual attention branch (RABVA) can obtain an improved performance for 9-way 5-shot learning compared to original model (Baseline), with average performance increasing from 75.4 to 76.8 and 76.9, respectively, an increase of 1.4 and 1.5. At the same time, fusing semantic and visual attention branches together (RABSVA) can achieve consistent best top-1 accuracy (From 75.4 to 78.9, an increase of 3.5) for few-shot action recognition, which indicates that although semantic attention and visual attention can strengthen network features from two different perspectives, these two enhancements bear complementary effects on each other.

4.3.3. With and without the RABSVA Module

To understand the RABSVA module better, we use ResNet-50 model as backbone and conduct more experiments on HMDB-49 dataset. Embedding our RABSVA module, the ResNet-50 model obtains improved top-1 action recognition performance from 1 to shot to 5-shot as Table 2 shows. This is because our RABSVA module can exploit the action bank to assist the network to understand the action in novel categories. Furthermore, we use confusion matrix to visualize the top-1 accuracies of 9-way 5-shot action recognition. With the confusion matrix in Figure 5, we can see for 9-way 5-shot setting, by embedding our RABSVA module, the ResNet-50 model can achieve better top-1 performance on most of these 9 novel categories. Meanwhile, as shown in Figure 5, we can observe that RABSVA-ResNet-50 works relatively well on the "shoot_ball" category, and relatively poorly on "draw_sword" and "kick".

Table 2. The top-1 performance from 1 to shot to 5-shot using 3D ResNet-50 as backbone on HMDB-49.

Method	1-Shot	2-Shot	3-Shot	4-Shot	5-Shot
Baseline	54.3	64.0	67.9	73.8	75.4
RABSVA	**54.8**	**68.4**	**68.4**	**74.7**	**78.9**



(**a**)

(**b**)

Figure 5. Confusion matrix of RABSVA-ResNet-50 on HMDB-49 datasets for 5-shot recognition. (**a**) Baseline 9-way 5-shot; (**b**) Ours 9-way 5-shot.

4.4. Comparison with State-of-the-Art Methods

Because there are not so many works for few-shot learning in video action recognition, there is a scarcity of compared datasets except for Kinetics-100 dataset. To compare our proposed RABSVA module in detail, we report 1-shot, 2-shot, 3-shot, 4-shot and 5-shot results on the 5-way video action recognition to evaluate our model. We choose C3D module as our backbone network, which is pretrained on Sports-1M dataset. Because our proposed method is time-consuming and needs to be fine-tuned 300 epochs, the results in Table 3 show the top-1 mean accuracy by randomly sampling 100 episodes for all shot experiment.

Table 3. The top-1 accuracies of C3D embedding our RABSVA module compared with state-of-the-art methods for 1-shot to 5-shot, the average of which is listed in AVG.

Method	1-Shot	2-Shot	3-Shot	4-Shot	5-Shot	AVG
Matching Net [4]	53.3	64.3	69.2	71.8	74.6	66.6
MAML [85]	54.2	65.5	70.0	72.1	75.3	67.4
CMN [6]	60.5	70.0	75.6	77.3	78.9	72.5
TARN [8]	66.6	74.6	77.3	78.9	80.7	75.6
OTAM [86]	**73.0**	-	-	-	85.8	-
TRX [87]	-	-	-	-	85.9	-
RABSVA-C3D (Ours)	71.0	**79.0**	**85.8**	**85.4**	**87.8**	**81.8**

As Table 3 shows, we compare our module to all the state-of-the-art methods, which include the classic image methods and the video-based methods. Matching Net [4] and MAML [85] are two classic methods in few-shot image classification. Zhu and Yang [6] reproduce these two classic methods for few example video action recognition. We can see the image-based method can not get better performance compared to other video-based methods. This is because the common image-based methods do not consider the dynamic information in videos. Meanwhile, we can see that as the training examples increase, the accuracy is improved in most cases. The 2-shot results of our module is almost the same as the 5-shot results of others, and is obviously better than other methods in 4-shot, 3-shot, 2-shot and 1-shot. We can also see our proposed module achieves the best performance on every shot compared with state-of-the-art methods. Moreover, our model can get the larger improvement at the higher shot. Compared with the second-best model, our model achieves an average improvement of 6.2% improvement. Moreover, we can observe that compared to the recent OTAM [86] and TRX [87], our proposed RABSVA-C3D is still able to achieve a performance improvement of about 2%.

5. Conclusions

In this paper, we proposed a novel relational action bank with a semantic–visual attention module for few example video action recognition. The proposed relational action bank can use existing or past actions for current few-shot action recognition. Furthermore, we introduce a semantic and visual attention mechanism to exploit the relational action bank. The extended experiments demonstrate that embedding our proposed RABVSA module can obtain state-of-the-art performance on a common kinetics dataset. In particular, we achieved a 6.2% improvement in top-1 accuracy compared to the current second-best method.

Author Contributions: Methodology, J.D. and H.L.; Investigation, H.Z.; Data curation, B.H.; Writing-original draft, H.L.; software, H.L.; Visualization, Y.M.; Funding acquisition, J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Planning Project of Gansu Province, China (Grant No. 20JR10RA185).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale Video Classification with Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
2. Carreira, J.; Noland, E.; Hillier, C.; Zisserman, A. A short note on the kinetics-700 human action dataset. *arXiv* **2019**, arXiv:1907.06987.
3. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML Deep Learning Workshop, Lille, France, 6–11 July 2015; Volume 2.
4. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3630–3638.
5. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4077–4087.
6. Zhu, L.; Yang, Y. Compound memory networks for few-shot video classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 751–766.
7. Xu, B.; Ye, H.; Zheng, Y.; Wang, H.; Luwang, T.; Jiang, Y. Dense dilated network for few shot action recognition. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, Yokohama, Japan, 11–14 June 2018; pp. 379–387.
8. Bishay, M.; Zoumpourlis, G.; Patras, I. TARN: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition. *arXiv* **2019**, arXiv:1907.09021.
9. Wang, L.; Qiao, Y.; Tang, X. Latent hierarchical model of temporal structure for complex activity classification. *IEEE Trans. Image Process.* **2013**, *23*, 810–822. [CrossRef] [PubMed]
10. Niebles, J.C.; Chen, C.W.; Fei-Fei, L. Modeling temporal structure of decomposable motion segments for activity classification. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 392–405.
11. Liu, F.; Xu, X.; Zhang, T.; Guo, K.; Wang, L. Exploring privileged information from simple actions for complex action recognition. *Neurocomputing* **2020**, *380*, 236–245. [CrossRef]
12. Thatipelli, A.; Narayan, S.; Khan, S.; Anwer, R.M.; Khan, F.S.; Ghanem, B. Spatio-temporal relation modeling for few-shot action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 19958–19967.
13. Wu, J.; Zhang, T.; Zhang, Z.; Wu, F.; Zhang, Y. Motion-Modulated Temporal Fragment Alignment Network for Few-Shot Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 9151–9160.
14. Yang, L.; Huang, Y.; Sugano, Y.; Sato, Y. Interact Before Align: Leveraging Cross-Modal Knowledge for Domain Adaptive Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 14722–14732.
15. Yang, J.; Dong, X.; Liu, L.; Zhang, C.; Shen, J.; Yu, D. Recurring the Transformer for Video Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 14063–14073.
16. Truong, T.D.; Bui, Q.H.; Duong, C.N.; Seo, H.S.; Phung, S.L.; Li, X.; Luu, K. Direcformer: A directed attention in transformer approach to robust action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 20030–20040.
17. Wang, H.; Klser, A.; Schmid, C.; Liu, C.L. Action recognition by dense trajectories. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2019.
18. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2019.
19. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. 2012. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=3c86dfdbdf37060d5adcff6c4d7d453ea5a8b08f (accessed on 4 February 2023).
20. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
21. Stroud, J.C.; Ross, D.A.; Sun, C.; Deng, J.; Sukthankar, R. D3D: Distilled 3D Networks for Video Action Recognition. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020.
22. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
23. Diba, A.; Fayyaz, M.; Sharma, V.; Arzani, M.M.; Yousefzadeh, R.; Gall, J.; Van Gool, L. Spatio-temporal channel correlation networks for action classification. In Proceedings of the the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 284–299.

24. Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; Dai, B. Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 2969–2978.
25. Xu, Y.; Wei, F.; Sun, X.; Yang, C.; Shen, Y.; Dai, B.; Zhou, B.; Lin, S. Cross-model pseudo-labeling for semi-supervised action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 2959–2968.
26. Guo, H.; Wang, H.; Ji, Q. Uncertainty-Guided Probabilistic Transformer for Complex Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 20052–20061.
27. Plizzari, C.; Cannici, M.; Matteucci, M. Skeleton-based action recognition via spatial and temporal transformer networks. *Comput. Vis. Image Underst.* **2021**, *208*, 103219. [CrossRef]
28. Patrick, M.; Campbell, D.; Asano, Y.; Misra, I.; Metze, F.; Feichtenhofer, C.; Vedaldi, A.; Henriques, J.F. Keeping your eye on the ball: Trajectory attention in video transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12493–12506.
29. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6836–6846.
30. Herzig, R.; Ben-Avraham, E.; Mangalam, K.; Bar, A.; Chechik, G.; Rohrbach, A.; Darrell, T.; Globerson, A. Object-region video transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 3148–3159.
31. Qiu, H.; Hou, B.; Ren, B.; Zhang, X. Spatio-Temporal Tuples Transformer for Skeleton-Based Action Recognition. *arXiv* **2022**, arXiv:2201.02849.
32. Huang, Z.; Qing, Z.; Wang, X.; Feng, Y.; Zhang, S.; Jiang, J.; Xia, Z.; Tang, M.; Sang, N.; Ang Jr, M.H. Towards training stronger video vision transformers for epic-kitchens-100 action recognition. *arXiv* **2021**, arXiv:2106.05058.
33. Shi, F.; Lee, C.; Qiu, L.; Zhao, Y.; Shen, T.; Muralidhar, S.; Han, T.; Zhu, S.C.; Narayanan, V. Star: Sparse transformer-based action recognition. *arXiv* **2021**, arXiv:2107.07089.
34. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? In Proceedings of the International Conference on Machine Learning, Online, 19–20 July 2021; Volume 2, p. 4.
35. Mazzia, V.; Angarano, S.; Salvetti, F.; Angelini, F.; Chiaberge, M. Action Transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognit.* **2022**, *124*, 108487. [CrossRef]
36. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 3202–3211.
37. Tong, Z.; Song, Y.; Wang, J.; Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv* **2022**, arXiv:2203.12602.
38. Qing, Z.; Zhang, S.; Huang, Z.; Wang, X.; Wang, Y.; Lv, Y.; Gao, C.; Sang, N. MAR: Masked Autoencoders for Efficient Action Recognition. *arXiv* **2022**, arXiv:2207.11660.
39. Yan, S.; Xiong, X.; Arnab, A.; Lu, Z.; Zhang, M.; Sun, C.; Schmid, C. Multiview transformers for video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 3333–3343.
40. Wei, C.; Fan, H.; Xie, S.; Wu, C.Y.; Yuille, A.; Feichtenhofer, C. Masked feature prediction for self-supervised visual pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 14668–14678.
41. Wu, C.Y.; Li, Y.; Mangalam, K.; Fan, H.; Xiong, B.; Malik, J.; Feichtenhofer, C. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 13587–13597.
42. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
43. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 2048–2057.
44. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
45. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in Neural Information Processing Systems Montreal, QC, Canada, 8–13 December 2014; Volume 27.
46. Luong, M.T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arXiv* **2015**, arXiv:1508.04025.
47. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
48. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
49. Daniluk, M.; Rocktäschel, T.; Welbl, J.; Riedel, S. Frustratingly short attention spans in neural language modeling. *arXiv* **2017**, arXiv:1702.04521.

50. Zhao, S.; Zhang, Z. Attention-via-attention neural machine translation. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
51. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6688–6697.
52. Gregor, K.; Danihelka, I.; Graves, A.; Rezende, D.; Wierstra, D. Draw: A recurrent neural network for image generation. In Proceedings of the International Conference on Machine Learning, Miami, FL, USA, 9–11 December 2015; pp. 1462–1471.
53. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
54. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
55. Li, K.; Wu, Z.; Peng, K.C.; Ernst, J.; Fu, Y. Tell me where to look: Guided attention inference network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9215–9223.
56. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2285–2294.
57. Wang, C.; Zhang, Q.; Huang, C.; Liu, W.; Wang, X. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 365–381.
58. Jetley, S.; Lord, N.A.; Lee, N.; Torr, P.H. Learn to pay attention. *arXiv* **2018**, arXiv:1804.02391.
59. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. Ocnet: Object context network for scene parsing. *arXiv* **2018**, arXiv:1809.00916.
60. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
61. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 286–301.
62. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11065–11074.
63. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7151–7160.
64. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
65. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
66. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
67. Huang, Q.; Ren, H.; Leskovec, J. Few-shot Relational Reasoning via Connection Subgraph Pretraining. *arXiv* **2022**, arXiv:2210.06722.
68. Wang, S.; Chen, C.; Li, J. Graph few-shot learning with task-specific structures. *arXiv* **2022**, arXiv:2210.12130.
69. Jiang, Z.; Dai, Y.; Xin, J.; Li, M.; Lin, J. Few-shot non-parametric learning with deep latent variable model. *arXiv* **2022**, arXiv:2206.11573.
70. Hermann, M.; Saha, S.; Zhu, X.X. Filtering Specialized Change in a Few-Shot Setting. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2023**, *16*, 1185–1196. [CrossRef]
71. Vanschoren, J. Meta-learning: A survey. *arXiv* **2018**, arXiv:1810.03548.
72. Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; Lillicrap, T. One-shot learning with memory-augmented neural networks. *arXiv* **2016**, arXiv:1605.06065.
73. Ravi, S.; Larochelle, H. Optimization as a Model for Few-Shot Learning. 2016. Available online: https://openreview.net/pdf?id=rJY0-Kcll (accessed on 4 February 2023).
74. Zhang, H.; Zhang, J.; Koniusz, P. Few-shot Learning via Saliency-guided Hallucination of Samples. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2770–2779.
75. Alfassy, A.; Karlinsky, L.; Aides, A.; Shtok, J.; Harary, S.; Feris, R.; Giryes, R.; Bronstein, A. LaSO: Label-Set Operations Networks for Multi-Label Few-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
76. Dwivedi, S.; Gupta, V.; Mitra, R.; Ahmed, S.; Jain, A. ProtoGAN: Towards Few Shot Learning for Action Recognition. *arXiv* **2019**, arXiv:1909.07945.
77. Zhu, X.; Toisoul, A.; Perez-Rua, J.M.; Zhang, L.; Martinez, B.; Xiang, T. Few-shot action recognition with prototype-centered attentive learning. *arXiv* **2021**, arXiv:2101.08085.
78. Peng, K.; Roitberg, A.; Yang, K.; Zhang, J.; Stiefelhagen, R. Delving Deep into One-Shot Skeleton-based Action Recognition with Diverse Occlusions. *IEEE Trans. Multimed.* **2023**. [CrossRef]

79. Cao, K.; Ji, J.; Cao, Z.; Chang, C.; Niebles, J. Few-shot video classification via temporal alignment. *arXiv* **2019**, arXiv:1906.11415.
80. Careaga, C.; Hutchinson, B.; Hodas, N.; Phillips, L. Metric-Based Few-Shot Learning for Video Action Recognition. *arXiv* **2019**, arXiv:1909.09602.
81. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
82. Hamilton, W.L.; Ying, R.; Leskovec, J. Representation learning on graphs: Methods and applications. *arXiv* **2017**, arXiv:1709.05584.
83. Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W.L.; Lenssen, J.E.; Rattan, G.; Grohe, M. Weisfeiler and leman go neural: Higher-order graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 4602–4609.
84. Hara, K.; Kataoka, H.; Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6546–6555.
85. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 1126–1135.
86. Cao, K.; Ji, J.; Cao, Z.; Chang, C.Y.; Niebles, J.C. Few-shot video classification via temporal alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10618–10627.
87. Perrett, T.; Masullo, A.; Burghardt, T.; Mirmehdi, M.; Damen, D. Temporal-relational crosstransformers for few-shot action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 475–484.