applied sciences

MDPI

Article

Whole-Body Keypoint and Skeleton Augmented RGB Networks for Video Action Recognition

Zizhao Guo and Sancong Ying *

National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610017, China; 2018326040003@stu.scu.edu.cn
* Correspondence: yingsancong@scu.edu.cn

Abstract: Incorporating multi-modality data is an effective way to improve action recognition performance. Based on this idea, we investigate a new data modality in which Whole-Body Keypoint and Skeleton (WKS) labels are used to capture refined body information. Unlike directly aggregated multi-modality, we leverage distillation to adapt an RGB network to classify action with the feature-extraction ability of the WKS network, which is only fed with RGB clips. Inspired by the success of transformers for vision tasks, we design an architecture that takes advantage of both three-dimensional (3D) convolutional neural networks (CNNs) and the Swin transformer to extract spatiotemporal features, resulting in advanced performance. Furthermore, considering the unequal discrimination among clips of a video, we also present a new method for aggregating the clip-level classification results, further improving the performance. The experimental results demonstrate that our framework achieves advanced accuracy of 93.4% with only RGB input on the UCF-101 dataset.

Keywords: action recognition; aggregation function; multi-modality; Swin transformer

Citation: Guo, Z.; Ying, S. Whole-Body Keypoint and Skeleton Augmented RGB Networks for Video Action Recognition. *Appl. Sci.* **2022**, *12*, 6215. https://doi.org/10.3390/app12126215

Academic Editor: Byung-Gyu Kim

Received: 4 May 2022
Accepted: 16 June 2022
Published: 18 June 2022

1. Introduction

Action recognition has attracted a lot of attention because it [1–7] contains more complicated information than individual images. Since the development of 3D CNNs [1] and two-stream CNNs [2], increasingly advanced deep learning methods have been proposed. In addition to direct operations on RGB frames, multi-modality features have also been employed for classification, such as optical flow [2], skeleton-based [8], motion vector [9], motion history images [10], etc. The application of multi-modality data is considered to be an effective measure for performance improvement.

Optical flow is the most common method for obtaining motion information and improving recognition accuracy. However, computing optical flow is too expensive for real-time application, as it occupies most of the operation time [11]. Therefore, multi-modality data [9,12–15] are employed to replace the computationally expensive optical flow. For example, Shou, Lin [11] presented discriminative motion cue (DMC) representation to reduce noise in motion vector estimation, and Chaudhary and Murala [13] proposed a Weber motion history image method for obtaining temporal features. Skeleton-based action recognition has also been used to obtain robust features; based on this approach, Chen et al. [16] proposed channel-wise topology refinement graph convolution (CTR-GC) which is able to adaptively study various topologies and aggregate joint features, achieving remarkable results.

In this study, we introduce a new data modality: RGB images with Whole-Body Keypoint and Skeleton (WKS) labels, as shown in Figure 1. Skeleton-based methods are robust against changes in body scale, motion speed, and camera viewpoint [8]; however, the videos in available datasets contain a variety of contexts, with many videos showing only a half body, or only hands or head, so that skeleton labeling is insufficient for marking the whole dataset. Therefore, we attempt to employ MMPose [17] for whole-body marking, including the skeleton and keypoints of the face, hands, and body, as shown in Figure 1.

The marked samples increase the brightness of the keypoints and skeleton, and reduce the brightness of the background, thus increasing the contrast between the body and background. In addition, irrelevant information such as human clothing and appearance can be further filtered, making the model well-suited for classification of actions that need to focus on poses, such as the categories of FrontCrawl and BreastStroke.

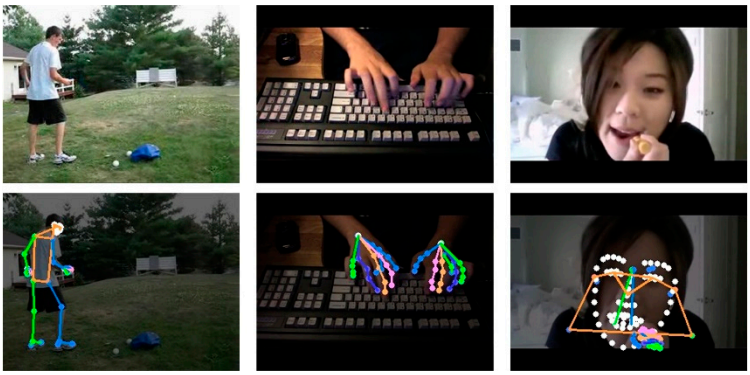

Figure 1. Comparison of original images (**top**) and WKS labeled images (**bottom**). From **left** to **right**: body, hands, face, respectively.

Most advanced methods generally utilize multi-modality data as a separate stream, and the final results are obtained through aggregating the multi-modality results with the RGB stream. In our work, we explore using the information from the WKS stream to train the RGB network, enabling it to possess the body information capturing ability of the WKS stream. For this purpose, we first trained a high-performing WKS network, which can make decisions based on human pose information. This means the high-level features in the WKS network are produced by that concerning body information. Therefore, if we take these high-level features as a teacher to instruct the learning process of the RGB network, the RGB network will be guided to classify actions focusing on body information. The knowledge extraction is based on distillation [18]. To the best of our knowledge, this study is the first to transfer WKS information to the RGB stream for action recognition.

The shifted windows (Swin) transformer [19] has demonstrated great potential for vision tasks, and has achieved a state-of-the-art performance in the fields of image classification, semantic segmentation, and object detection. However, it is difficult to extract motion information using the conventional 2D Swin transformer. To address this problem, we explore a novel method that takes advantage of both the 3D CNNs and Swin transformer.

Tran et al. [20] demonstrated that temporal modeling is a type of bottom-level operation; it can be extracted by 3D convolution at the bottom level, and performs plane calculations on a high level without accuracy reduction. Based on this theory, we use 3D convolution to extract spatiotemporal features from the bottom layers, then concatenate the temporal features to spatial form which a transformer can process. Finally, the results are obtained by the Swin transformer which operates on the produced features at the top layers. The network structure of CNNs lead to it only focusing on local information and being hard to capture and store a long-distance dependent relationship, while self-attention mechanisms in the transformer can effectively make up this weakness. Therefore, the high-level spatiotemporal features produced by 3D CNNs can be well analyzed in a Swin transformer. To the best of our knowledge, this is the first work to attempt to combine the advantages of 3D CNNs and a transformer.

Deep learning-based methods often require expensive hardware resources, and it is infeasible to feed the model with a whole video due to limited computational resources. In our framework, we take clip-based architecture to keep memory consumption manageable. This architecture aggregates clip-level results to video-level results, and effective aggregation function is crucial for accurate classification. Most common aggregating techniques, such as average pooling and max pooling, are simple and data-independent. They are not

well-suited for evaluating unequal discrimination of each clip. Some complicated aggregation methods [21–24] have also been proposed; for example, in study [23], a recurrent neural network (RNN) was designed to yield video-level scores. However, confidence of clip-level results is not well considered in these methods. Thus, we introduce a new aggregation function called the confidence-weighting aggregation function (CWAF). We aggregate clip-level prediction results by analyzing the confidence of each result and determine the weights, finally improving the accuracy by approximately 0.5% compared to average pooling.

The contributions of this study are summarized as follows:

1. We introduce a novel framework that utilizes WKS labeled images for training RGB network processes in body information capturing ability. The knowledge transferring is based on the concept of distillation. The evaluations show that this new data modality effectively improves the recognition ability of the RGB network. So far as we know, this is the first work to transfer body concerned features extracting ability to a RGB network.
2. We explore a novel architecture for concatenation that uses 3D convolution and the Swin transformer, which fully takes advantage of the two architectures.
3. By analyzing the confidence of clip-level results, we design an aggregation method to assign more rational weights to each clip output.

We organized our paper as follows. Discussion of related works are presented in Section 2. In Section 3, we describe the details of our method. In Section 4, experiments and analysis on popular action recognition datasets are represented. Finally, the conclusion and discussion are represented in Section 5.

2. Related Works

This section presents the prior work related to ours; the discussion about recent action classification methods is represented first, then distillation strategies and aggregation functions related to our work are described, respectively.

Recent Action Recognition. In contrast to traditional hand-crafted action recognition methods [25–27], deep learning strategies, which have dominated the field of action recognition, have excellent modeling capacity and are capable of learning in an end-to-end manner [28]. Deep learning methods can be described in two categories: 3D CNN framework and multi-stream framework.

Simonyan and Zisserman [2] proposed a two-streams CNN framework that employs pre-computed optical flow to extract temporal information. Subsequently, Feichtenhofer et al. [29] improved the performance of the two-streams method by introducing a different fusion strategy. Shou, Lin [11] introduced the DMC representation to reduce noise in motion vectors, extracting motion information as an alternative technique to optical flow. Three-dimensional CNNs [1,30] have been employed as another efficient tool for modeling temporal information [21,31–33]. Some additional convolution-based methods have also been introduced [20,34], such as R(2 + 1)D [20], which replaces 3D convolutions with separate spatial and temporal convolutions. Furthermore, a number of state-of-the-art strategies [3,5,32] have developed frameworks by taking advantage of both mechanisms to achieve the best performance; however, although state-of-the-art accuracies have been achieved, this combination is computationally intensive.

In addition to the architectural researches, some action recognition methods [35–38] attempt to extract more refined motion features with object detection, object tracking, and pose detection methods. The emerging advanced tracking and detection methods [39–41] make these possible. For example, Cao, Simon [39] proposed an advanced approach to detect a 2D pose of multiple people by associating body parts; Dewi, Chen [40] presented a yolo-4-based high-performance detection framework, and Wachinger, Toews [41] introduced a novel whole-body segmentation method. With the help of these methods, more discriminative features are obtained and contribute to make accurate action recognition. For example, Brehar, Muresan [35] proposed a novel framework in which classification is per-

formed with information of pedestrian motion, distance of the pedestrian, and position of pedestrians, and the final results are obtained by a Long Short-Term Memory-based model aggregating temporal features. Yan, Hu [36] proposed a real-time human rehabilitation action recognition method based on a human pose. It fuses OpenPose with Kalman filter to track human targets. Verma, Meenpal [37] improved the performance of multiperson interaction recognition by extracting distance and angular relation features based on body keypoints. Based on the pose detection algorithm, Pandurevic, Draga [38] developed a motion sequences analysis method to help in training speed climbing athletes.

Distillation. The concept of distillation, first proposed by Hinton [18], is a training strategy that takes a complex model as a teacher to train a simple one. Category probabilities of the teacher model are the most common target to transfer knowledge, and these are called a "soft target". The training of our arithmetic is based on it.

After distillation was proposed, generalized distillation [42] was also introduced, it was designed based on distillation and privileged information [43]. In the field of action classification, several works have proposed to utilize distillation. Garcia et al. [44] developed a hallucination network that applied the feature of depth and RGB during the study process, as an inference course classifier only fed with RGB clips. Garcia et al. [45] proposed a DMCL framework to leverage the complementary information of multiple modalities. Crasto, Weinzaepfel [31] introduced a framework that distills the knowledge of flow stream into an RGB network, achieving state-of-the-art accuracy for the one-stream action recognition task. Our proposed framework is inspired by these advanced methods.

Aggregation function. The aggregation function is an important module for completing a video-level classification, and can directly influence the final results. Several approaches have been proposed to better utilize clip-level results. Kalfaoglu, Kalkan [21] concatenated a bidirectional encoder representation from the transformers (BERT) layer at the end of a 3D convolutional architecture for aggregation, achieving promising results. In study [23], an RNN was designed to yield video-level scores concerning all clip recognition results. Aiming at the defects of linear weighting schemes that lack concerning features, Wang, Xiong [22] proposed an adaptive weighting method to automatically assign weights to clip-level results. Wang and Cherian [24] introduced the concept of a positive bag and a negative bag to find useful features. In our approach, the judgement of confidence through analyzing the form of the category probabilities is performed, then weights for each clip-level result are determined by confidence scores.

Different from current detection-based or tracking-based action recognition methods, we attempt to guide the RGB network to pay attention to the body information without extra detection or a tracking process. We attempt to integrate this ability into the RGB network by using the distillation method. At inference, only RGB clips are taken for classification. Furthermore, as current clip-based action recognition methods lack consideration on the confidence differences among clip-level results, we directly took clip-level results to measure the discrimination of clips and, therefore, obtained a more rational aggregation function.

3. Proposed Method

We propose a framework to train an RGB network to learn the ability of a WKS network. An overview of this procedure is shown in Figure 2. In Section 3.1, we formally define the video-level and clip-level classification, then discuss the clip-level training and prediction course of our framework. The Swin transformer-based network architecture is described in Section 3.2. We describe the detailed training strategy in Section 3.3, and discuss the proposed CWAF in Section 3.4.

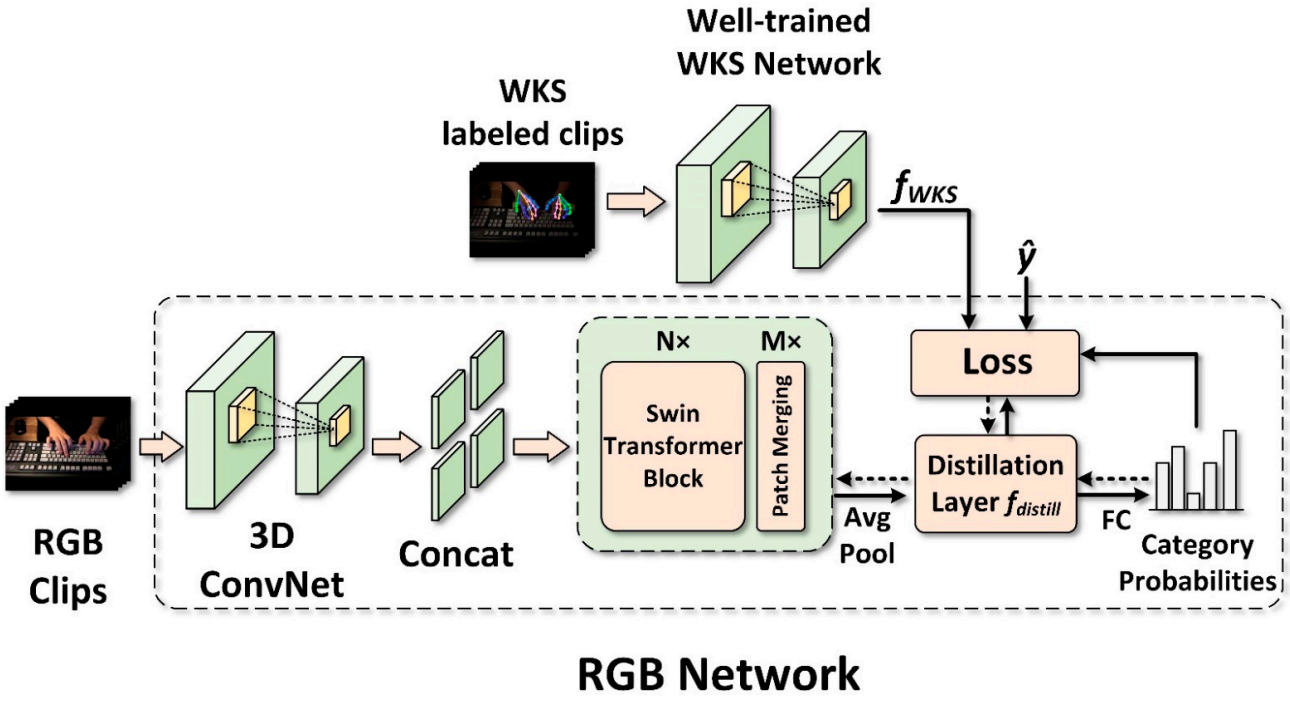

Figure 2. Framework of clip-level training and prediction. The architecture of our proposed RGB network is shown in the dotted box, which includes a 3D Conv module and Swin transformer module. The dotted arrows show the backpropagation path.

3.1. Overview of Clip-Level Training and Prediction

To directly feed the classifier with a whole video is infeasible for our computer due to limited hardware resources. Therefore, normal strategies usually divided the video v into a set of clips $\{c^{(1)}, c^{(2)}, c^{(3)}, \ldots, c^{(i)}, \ldots\}$, and the classifier was fed with clips instead of a whole video. Given a clip classifier $\mathbf{g}$, the clip-level prediction can be denoted as $\mathbf{g}(c^{(i)};W)$, and video-level prediction can be formulated as follows:

$$Y_v = H(\mathbf{g}(c^{(1)};W);\mathbf{g}(c^{(2)};W);\ldots;\mathbf{g}(c^{(n)};W)), \quad (1)$$

where $\mathbf{g}(c^{(i)};W)$ is the function that represents the classifier $\mathbf{g}$ with weights W. We can infer from Equation (1) that the classification result of the given video is obtained by aggregating the clip-level results through function H.

The clip-level training and prediction course is shown in Figure 2. The classifier $\mathbf{f}$ operates on fixed-length clips of F frames with spatial resolution $H \times W$, and makes a prediction of classification probabilities.

Our goal is to train the RGB classifier $\mathbf{f} : \mathbb{R}^{F\times 3\times H\times W} \rightarrow [0,1]^C$ to possess more robust classification ability by focusing on body information when fed with clips that contain a noisy background. For this purpose, we first obtain a well-trained classifier $\mathbf{f}_{WKS}$ on the WKS labeled dataset, which is then used as the teacher network for training the RGB network based on the distillation strategy. We use f_{WKS} to indicate the feature map of the target distillation layer; it is localized at the high layer of classifier $\mathbf{f}_{WKS}$ and produced by fully concerning the body information. Normally, the extraction course from the input RGB frames to f_{WKS} should consist of two steps: locating the keypoints of input frames, and operating the classification function $\mathbf{f}_{WKS}$. In our method, this course is greatly simplified by using distillation, and high-level features f_{WKS} can be directly produced from RGB frames. Similarly, we use $f_{distill}$ to indicate the feature map of the distillation layer in the RGB network. During the training process, body information capturing ability is transferred to the RGB network as $f_{distill}$ gradually approximates to f_{WKS}. Finally, at inference, our RGB

network can distinguish the actions with the ability of concerning body information and do not require any pose detection processes.

Moreover, transformers lack some of the inductive biases inherent to CNNs, and therefore do not generalize well when trained on small datasets [46]. Distillation is an effective method for improving the performance of transformers [47]. Based on these considerations, our proposed transformer-based framework will further benefit from this distillation training method.

3.2. Swin Transformer-Based RGB Network

The Swin transformer is a powerful framework for the imaging task, but it has no ability to extract temporal features of videos. It first divides the input image with a size of $H \times W \times 3$ into non-overlapping patches with a size of $\frac{H}{4} \times \frac{W}{4} \times 48$; this part is referred to as "patch partition" in study [19]. Next, a liner projection function is applied on it, and the input matrix is shifted to a sequence that is calculated using a multi-head self-attention mechanism. To extract temporal information, we utilize a 3D convolutional module before the self-attention mechanism, yielding the module architecture shown in Figure 3. It is fed with fixed-length clips of F RGB frames with spatial resolution $H \times W$. We set F as 16; thus, after two down-sampling operations, four feature maps are produced in the temporal dimension: $\{f_1, f_2, f_3, f_4\}$, each $f_i \in \mathbb{R}^{96 \times \frac{H}{4} \times \frac{W}{4}}$. Then they are concatenated as one $f_W \in \mathbb{R}^{96 \times \frac{H}{2} \times \frac{W}{2}}$. Such operations map the temporal motion information into the spatial domain.

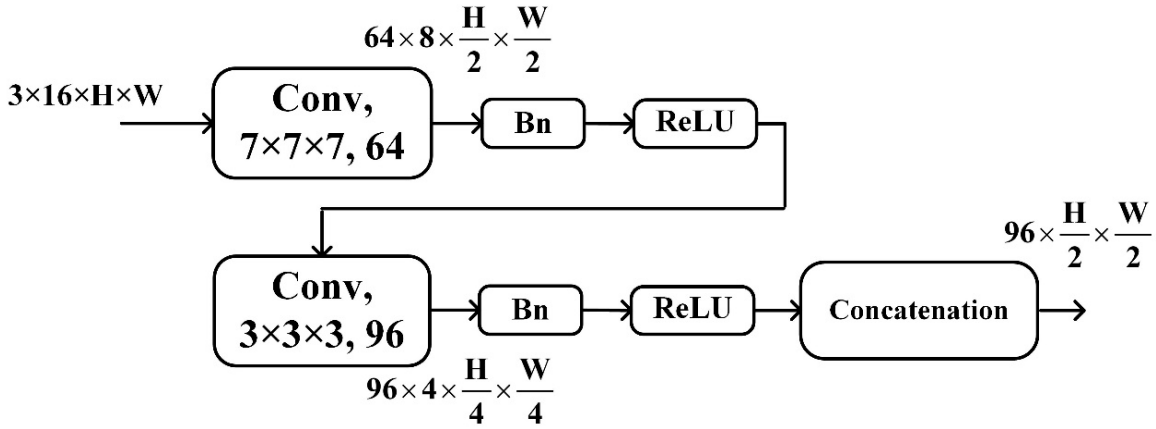

Figure 3. Architecture of a 3D convolutional module. Two 3D convolutional operations are included in the module. After layers of Bn and ReLU, feature map concatenation module concatenate four feature maps into one.

During the inference, two 3D convolution operations are performed on the input clip, and the temporal dimension is finally reduced to 1/4 of its original size. The convolution kernels' size in the temporal dimension are 7 and 3, respectively. Thus, the receptive field region in the temporal dimension of the final output feature map is of size 11. This means that the concatenated feature map f_w can access all of the temporal information.

It was demonstrated in study [19] that the computed objects of the self-attention mechanism in the Swin transformer are unlike those in ViT [46]; its calculations are performed in a window. The self-attention computations in the windows are defined as follows:

$$Attention(Q, K, V) = SoftMax(QK^T/\sqrt{d} + B)V, \quad (2)$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ are the *query*, *key*, and *value* matrices; d is the *query*/*key* dimension, and M^2 is the number of patches in a window. Q, K, V is obtained from a liner function which operated on the pixels in a window, and we can infer from Equation (2) that the spatial features in the window are fully compared and without constraints by the size of the

receptive field, which occurred in convolutional operation. The Swin transformer utilizes a window architecture to reduce the computational complexity, but this architecture lacks the computation between windows. Therefore, a method that shifts the window at the next layer is proposed, and by continuously changing the position of the window, the features in different area can be connected and compared. Benefitting from this architecture, in our situation, the temporal differences among $\{f_1, f_2, f_3, f_4\}$ can be effectively compared and captured. In addition, because the size of the feature map is reduced in the top layers, the receptive field of each window is relatively expanded, so that the temporal features are more comprehensively extracted.

The complete architecture of our proposed framework is illustrated in Figure 2, in which the produced features f_W are fed into the network that is formed by the stacked Swin transformer block and patch merging [19]. Considering that the features of the top layers represent high-level global information [31], we designed the distillation layer $f_{distill}$ to be located after the last Swin transformer block. The logit that is obtained by a fully connected layer over $f_{distill}$ is then converted by the SoftMax function, and the final category probabilities are obtained.

3.3. Training Strategy

The distillation strategy was first proposed by Hinton et al. [18], and aims to transfer knowledge from a cumbersome pretrained network to a simple lightweight network. We explore how to distill the prior knowledge of a teacher model to a student model on the target layer. As shown in Figure 2, f_{distll} is denoted as the distillation layer in the RGB classifier $\mathbf{f}$; it transfers action recognition ability from $\mathbf{f_{WKS}}$ to $\mathbf{f}$ by performing a mean squared error (MSE) loss function between f_{WKS} and f_{distll}:

$$\mathcal{L}_{WKS} = \frac{1}{M}\sum_{i=1}^{M} \| f^i_{WKS} - f^i_{distll} \|^2, \tag{3}$$

where $f^i_{distill}$ is obtained from $\mathbf{f}$ fed with the i-th RGB clip, and the same clip that is marked with WKS is fed to $\mathbf{f_{WKS}}$, producing f^i_{WKS}. As $\mathcal{L}_{WKS}$ decreases gradually, f_{distll} will approximate to f_{WKS}, and accordingly, the feature extraction method of f will be more similar to $\mathbf{f_{WKS}}$.

Although the well-trained WKS model can provide knowledge that is learned from refined body information, they can also bring in noise interference, such as improper and low-discriminative features which can lead to confused recognition. Based on this consideration, we introduce a loss function to select the backpropagation path, which uses the following loss error to update the parameters:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{WKS} + \lambda_2 CrossEntropy(\mathbf{f}(c^{(i)}), \hat{y}) \tag{4}$$

In Equation (4), the RGB model's $\mathbf{f}$ is fed with clip $c^{(i)}$, $\hat{y}$ is ground truth label, and we use cross-entropy as our classification loss. Meanwhile, applying λ_1 and λ_2 to adjust the weights of two loss functions, λ_1 can be computed as follows:

$$\lambda_1 = \begin{cases} \theta_1 & if\ \mathbf{f}_{WKS}\left(c^{(i)}\right) = \hat{y}. \\ 0 & if\ \mathbf{f}_{WKS}\left(c^{(i)}\right) \neq \hat{y}, \end{cases} \tag{5}$$

We can infer from the Equation (5) that only the teacher model recognizes the actions successfully, namely $\mathbf{f}_{WKS}\left(c^{(i)}\right) = \hat{y}.$, then the features of a teacher model are participated in the backpropagation course through the parameter θ_1.

The updating of parameters connected to distillation f_{distll} can be described as follows:

$$W_d^+ = W_d - \eta(\lambda_1 \frac{\partial \mathcal{L}_{WKS}}{\partial W_d} + \lambda_2 \frac{\partial \mathcal{L}_{CE}}{\partial \phi} \cdot \frac{\partial \phi}{\partial W_d}), \tag{6}$$

where W_d are the parameters before updating, and the function map from distillation to final output is represented as $\phi : f_{distill} \rightarrow \mathbf{f}(c^{(i)})$. As observed from Equation (6), the parameters will be modified by $\mathcal{L}_{WKS}$ only when the $\lambda_1 > 0$, in this case, the $\mathbf{f}_{WKS}$ recognize the action correctly and target features are considered as sufficiently discriminative. From this method, we can effectively prevent the model from learning noisy information.

3.4. Confidence-Weighting Aggregation Function

We introduce the CWAF, a simple but effective method to adaptively assign weights for each output according to their confidence. The inspiration comes from the observation that more credible input is expected to produce more distinct probabilities over classes. A comparison of the output as the classifier which makes a right or wrong classification is shown in Figure 4. These examples were chosen from the videos v_Biking_g02_c02 and v_Haircut_g02_c04, which are classified correctly and incorrectly, respectively. We can easily determine that category probabilities corresponding to a correct recognition represent minor fluctuations, resulting in a more reliable output.

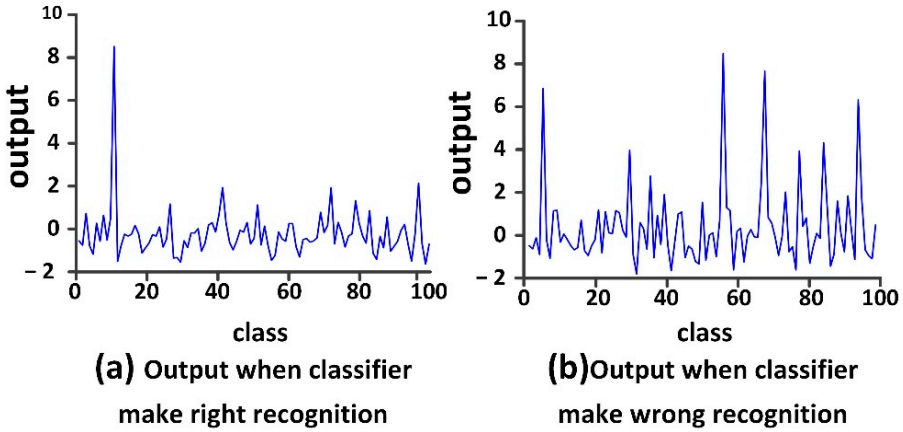

Figure 4. Comparison of the output when the classifier makes (**a**) right and (**b**) wrong classifications.

To assign weights for each clip-level result, we first need to design an estimator $\mathbf{h}(\mathbf{f}(c^{(i)})) \in [0, 1]$ to score the confidence; thus, we propose a confidence scoring network. The confidence of output $\mathbf{f}(c^{(i)})$ is mainly determined by the relative difference among the highest class scores and others. To reduce the distribution from the absolute positions of the values, we directly sort the output values in order from large to small, keeping only the relative difference information. We designed the confidence scoring network to be composed of multiple layers of a fully connected neural network. Several architectures have been examined regarding the balance of speed and accuracy, and a simple four-layer network has been found to be completely competent for this task.

For a video that consists of clips $\{c^{(1)}, c^{(2)}, c^{(3)}, \ldots, c^{(i)}\}$, we estimate each output $\mathbf{f}(c^{(i)})$ with confidence scoring network and produce scores $\{s_1, s_2, s_3, \ldots, s_i\}$ to represent confidence of clip-level results. Video-level results are obtained by:

$$Y_v = \sum_i w_i \mathbf{f}(c^{(i)}), \tag{7}$$

where w_i is the weight for each clip-level result. We attempt to assign higher weights for more confident clip-level results and enlarge the margin between the weight of lower-ranked and higher-ranked clip results. Therefore, a quadratic nonlinear function is applied to arrange the weight distribution:

$$w_i = \left(\frac{s_i - \min(\{s_i\}_{i=1}^N)}{\max(\{s_i\}_{i=1}^N) - \min(\{s_i\}_{i=1}^N)} \right)^2, \tag{8}$$

where the min and max functions are computed for the maximum and minimum values of the sequences, respectively. With this quadratic nonlinear function, the weight of lower-ranked and higher-ranked clip results will have a larger difference value.

For training the confidence scoring network, we introduce a new dataset $\mathcal{D}$, which shows the relationship between classifier output $\mathbf{f}(c^{(i)})$ and the predicted situation. It consists of M samples $\{(\mathbf{f}(c^{(i)}), y^{(i)})\}_{i=1}^{M}$, where $y^{(i)}$ is a scalar, and it denotes whether the classifier derives the correct result or not as it is fed with $c^{(i)}$. It can be formulated as follows:

$$y^{(i)} = \begin{cases} 1 & if\ \mathbf{f}\left(c^{(i)}\right) = \hat{y} \\ 0 & if\ \mathbf{f}\left(c^{(i)}\right) \neq \hat{y}, \end{cases} \tag{9}$$

Note that the form of the clip-level result $\mathbf{f}(c^{(i)})$ cannot strictly represent whether the clip is predicted correctly or not, that is, there are no forms that are definitively effective. However, it is still able to roughly help discriminate the reliable results.

4. Experimental Results

This section presents the experimental results of our proposed methods. The details of the datasets we used are firstly introduced, then we give the experimental details. Next, we report the distillation evaluation results of the proposed framework on the UCF-101 dataset of human actions [48] and the HMDB-51 human motion database [49]. The influence of the CWAF is discussed in Section 4.4. Lastly, we compare our results with some advanced methods.

4.1. Datasets

There are three datasets used for our experiments: UCF-101, HMDB-51, and Kinetics-400 [50]. The UCF-101 dataset collected 101 action categories. These categories are very diverse, which covers from single person to group, half-body to whole-body, and variations in body scale, motion speed, and camera viewpoint. There are 6766 videos in the HMDB-51 dataset, and it covers 51 categories which consist of videos that come from the real world and from various sources, such as social websites and movies. There are 400 action categories collected in Kinetics-400 for providing enough samples to train models, and each class has more than 400 videos; its volume is much larger than that of UCF-101 and HMDB-51.

Directly training on UCF-101 or HMDB-51 will lead to a serious overfitting due to the relatively small scale. Thus, the Kinetics-400 dataset is used for training the RGB network to complete transfer learning, but it is not applied for evaluation due to limited hardware resources. Both UCF-101 and HMDB-51 introduced three official dataset splits as a way for training and testing. Our experiments are primarily performed on the widely used first split of UCF-101 and HMDB-51.

4.2. Implementation Details

We designed our RGB network based on two Swin transformer architectures: Swin-S and Swin-B. Table 1 lists the detailed architectural parameters of our network. Both architectures are applied as RGB networks for training and evaluation. For the WKS teacher model, only the Swin-B-based network is considered for training and distillation. As the scales of UCF-101 and HMDB-51 are relatively small, insufficient data could lead to a serious overfitting problem. Therefore, we first trained two RGB networks on the Kinetics-400 dataset, and then fine-tuned them on UCF-101 and HMDB-51 by studying the large dataset to avoid the problem of overfitting.

Table 1. Detailed architecture of the proposed network.

Layer	Output Size	Swin-S	Swin-B
3D_ConvNet	56×56	conv, 7^3, 64-d conv, 3^3, 96-d concat	conv, 7^3, 64-d conv, 3^3, 128-d concat
Swin_layer1	28×28	concat 2×2, 192-d, LN $\begin{bmatrix} \text{win.sz.} 7 \times 7, \\ \text{dim } 192, \text{head } 6 \end{bmatrix} \times 2$	concat 2×2, 256-d, LN $\begin{bmatrix} \text{win.sz.} 7 \times 7, \\ \text{dim } 256, \text{head } 8 \end{bmatrix} \times 2$
Swin_layer2	14×14	concat 2×2, 384-d, LN $\begin{bmatrix} \text{win.sz.} 7 \times 7, \\ \text{dim } 384, \text{head } 12 \end{bmatrix} \times 18$	concat 2×2, 512-d, LN $\begin{bmatrix} \text{win.sz.} 7 \times 7, \\ \text{dim } 512, \text{head } 16 \end{bmatrix} \times 18$
Swin_layer3	7×7	concat 2×2, 768-d, LN $\begin{bmatrix} \text{win.sz.} 7 \times 7, \\ \text{dim } 768, \text{head } 24 \end{bmatrix} \times 2$	concat 2×2, 1024-d, LN $\begin{bmatrix} \text{win.sz.} 7 \times 7, \\ \text{dim } 1024, \text{head } 32 \end{bmatrix} \times 2$
Output	n_class	average pooling flatten fc $\times$ 2	average pooling flatten fc

We set the distillation layer at fc layer in both the Swin-B-based network and Swin-S-based network. It can be found that when the Swin-S based network is training, the dimensions between distillation layer $f_{WKS} \in \mathbb{R}^{1024}$ and $f_{distill} \in \mathbb{R}^{768}$ are not matching. Therefore, a liner layer is designed concatenated after the fc layer to connect the 768-dim layer to 1024-dim distillation layer, as shown in Table 1.

To obtain a well-trained WKS model, we first generated a WKS dataset by marking the UCF-101 and HMDB-51 datasets using MMPose [17]. The marking process followed the official instructions. It is noted that not all samples in the UCF-101 and HMDB-51 datasets can be well marked, and we filtered the noisy samples in which keypoints' mean confidence scores were below 0.4. At the training stage, only qualified samples were used.

For clip-level training and inference, to satisfy the requirement of memory consumption, we divided each video to a series of clips with 16 frames, and spatially resized to 112. At the training stage, to minimize overfitting, several data augmentation techniques were applied. From input frames, we first decided a base position and crop scale, then cropped the input frame in terms of the base position and the crop scale. The scale randomly sampled from set $\left\{1, \frac{1}{2^{1/4}}, \frac{1}{\sqrt{2}}, \frac{1}{2^{3/4}}, \frac{1}{2}\right\}$. Next, we randomly flipped half of the clips. These tricks are all operated in a spatial domain, and we also applied a temporal technique. In the temporal domain, we augmented data by random sample of 16 consecutive frames of videos. These tricks are applied for training both the RGB and WKS networks. We have trained the network with a batch size of 64 for 250 epochs at both pretraining stage and training stage. OpenCV-Python is applied for converting video to image sequences at 25 fps. We used the AdamW optimizer to train the network parameters, with the weight decay set to 1×10^{-3}, and initial learning rate set to 0.001. The proposed method was implemented on PyTorch 1.2 with Python 3.6, and all experiments were executed on a Tesla K40 GPU and an E5-2620 CPU.

For training the confidence scoring network, we introduced a new dataset $\mathcal{D}$, which consists of $\mathbf{f}(c^{(i)})$. Note that all clips $c^{(i)}$ come from the training dataset, and are not introduced by any testing dataset information. On the training dataset, the classifier is able to obtain remarkable classification accuracy, so that the negative samples of dataset $\mathcal{D}$ are quite scarce. To address this issue, we adopted data augmentation techniques to enlarge our negative samples. Tricks that were applied are similar to tricks of the RGB network, and we also deliberately reduced the number of positive samples to balance these two categories.

4.3. Evaluation of Distillation

We denote the Swin-B-based network and the Swin-S-based network as 3D-ConSwin-B and 3D-ConSwin-S, respectively. The evaluation on the UCF-101 and HMDB-51 datasets are listed in Table 2. All accuracies are obtained from split-1 of UCF-101 and HMDB-51. We compared Swin transformer-based models to advanced 3D CNNs, such as 3D-ResNext-101 [51], 3D-DenseNet-121 [51], and 3D-ResNet-152 [51], and our network achieved competitive results. Note that all our models are pretrained on Kinetics-400 and fine-tuned on the target dataset. We observed that a single stream of the WKS network or RGB network does not show remarkable recognition ability, but the combination of the two streams achieved an impressive accuracy of 94.1%. By utilizing the distillation method, the accuracy on UCF-101 is effectively boosted by approximately 2.1% and 1.4%, respectively, for 3D-ConSwin-B and 3D-ConSwin-S.

Table 2. Evaluation results on UCF-101 and HMDB-51.

	Architecture	**Top-1 Accuracy (%)**	
		UCF-101	**HMDB-51**
w/o distillation	3D-ConSwin-B (RGB)	90.9	63.7
	3D-ConSwin-S (RGB)	90.5	63.6
	3D-ResNext-101	90.7	63.8
	3D-DenseNet-121	87.6	59.6
	3D-ResNet-152	89.6	62.4
w/ distillation	3D-ConSwin-B (RGB)	93.0	66.7
	3D-ConSwin-S (RGB)	91.9	65.1
3D-ConSwin-B (HKS)		89.6	62.9
3D-ConSwin-B (RGB) + 3D-ConSwin-B (HKS)		94.1	69.3

We conducted an experiment to explore how the distillation method changes the RGB network, as illustrated in Figure 5. For deep learning, we found that what kinds of features are extracted for recognizing is very uncertain due to its "black box" characteristic. Therefore, extracted features may not be very discriminative to target action, such as in video v_ApplyLipstick_g04_c01, the RGB network may take the face of a female as a key feature and lead to confusion between categories of Apply Lipstick and Apply Eye Makeup, as both categories have a female face. However, if the hand pose and face keypoints are clearly marked, the network may capture the discriminative features of hand and lip overlapping, thus, making more precise classification based on these features. It can be observed in Figure 5, after distilled pose capturing ability to RGB network, some easily confused categories for the RGB network, such as Blow Dry Hair and Brushing Teeth, Nunchucks, and Archery which can be better distinguished by using pose features. With distillation, WKS augmented the RGB network which is able to classify these categories as more accurate.

A comparison of the confusion matrices is shown in Figure 6, which illustrates the confusion matrices of twenty categories having the lowest accuracy on the UCF-101 dataset; 3D-ConSwin-B(RGB) and 3D-ConSwin-B(RGB) with distillation are presented at the top and the bottom, respectively. We observed that after distillation, some confused categories have improved accuracy. As the distilled RGB network can extract features like the WKS network, it focuses more on the pose and pose motion of the input clips. Parts of the categories that share similar backgrounds are easily confused, but with the benefit of WKS labeling, they can be more effectively classified. For example, the category of FrontCrawl only obtains an accuracy of 45.9% with the RGB network, it is easily confused with the category of BreastStroke. They both have swimming pools as their background, but after distillation, the accuracy is improved to 59.5% because the pose is represented more clearly.

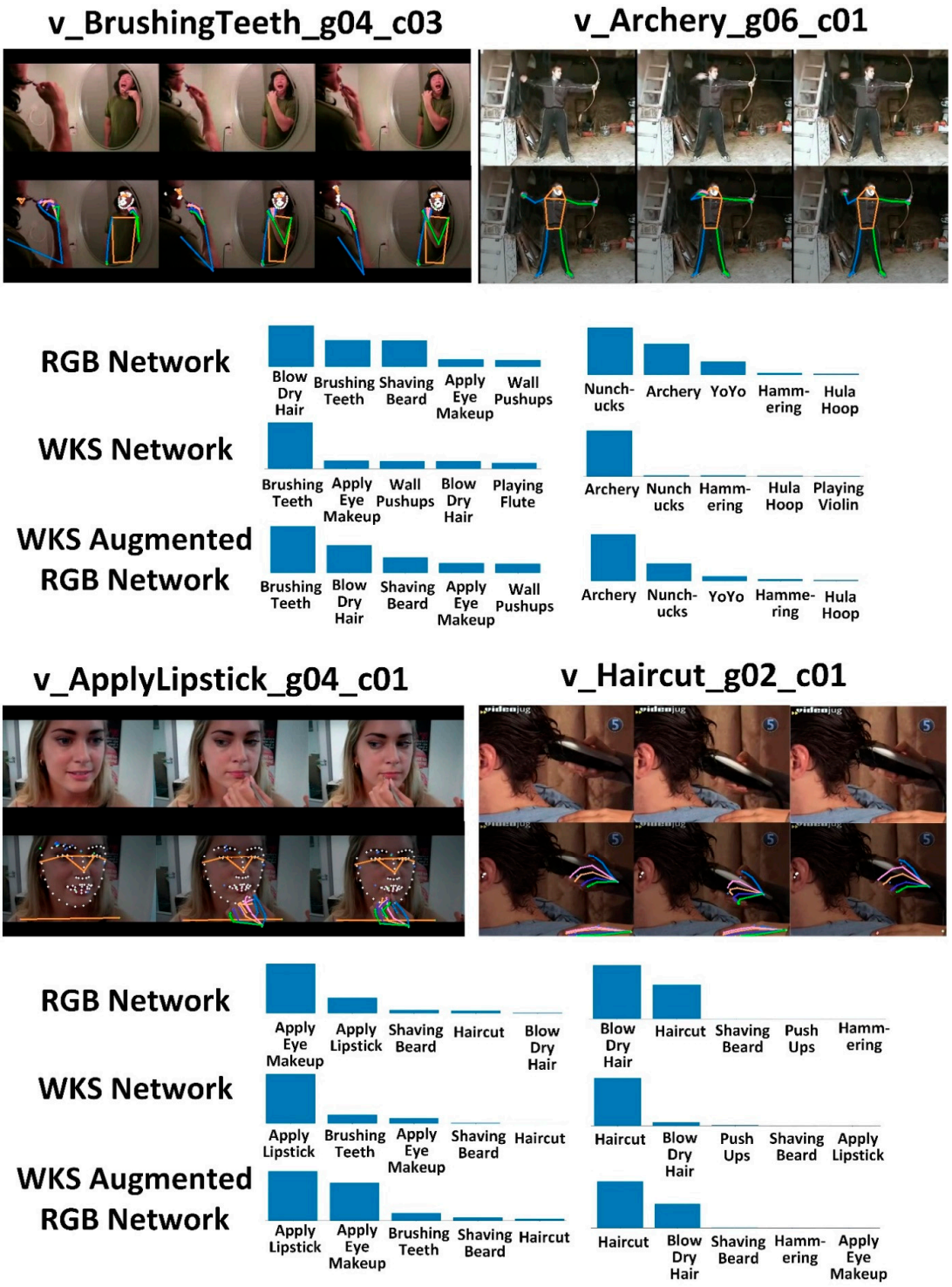

Figure 5. Comparison of the possibility of the Top 5 categories when the RGB network, WKS network, and WKS augmented RGB network are fed with clips of v_BrushingTeeth_g04_c03, v_Archery_g06_c01, v_ApplyLipstick_g04_c01, and v_Haircut_g02_c01, respectively. The original images and WKS labeled images are shown in the first line and the second line, respectively. Histograms represent the possibility of the Top 5 categories. Histograms of WKS augmented RGB network show how the output changed after transferring knowledge from the WKS network to RGB network.

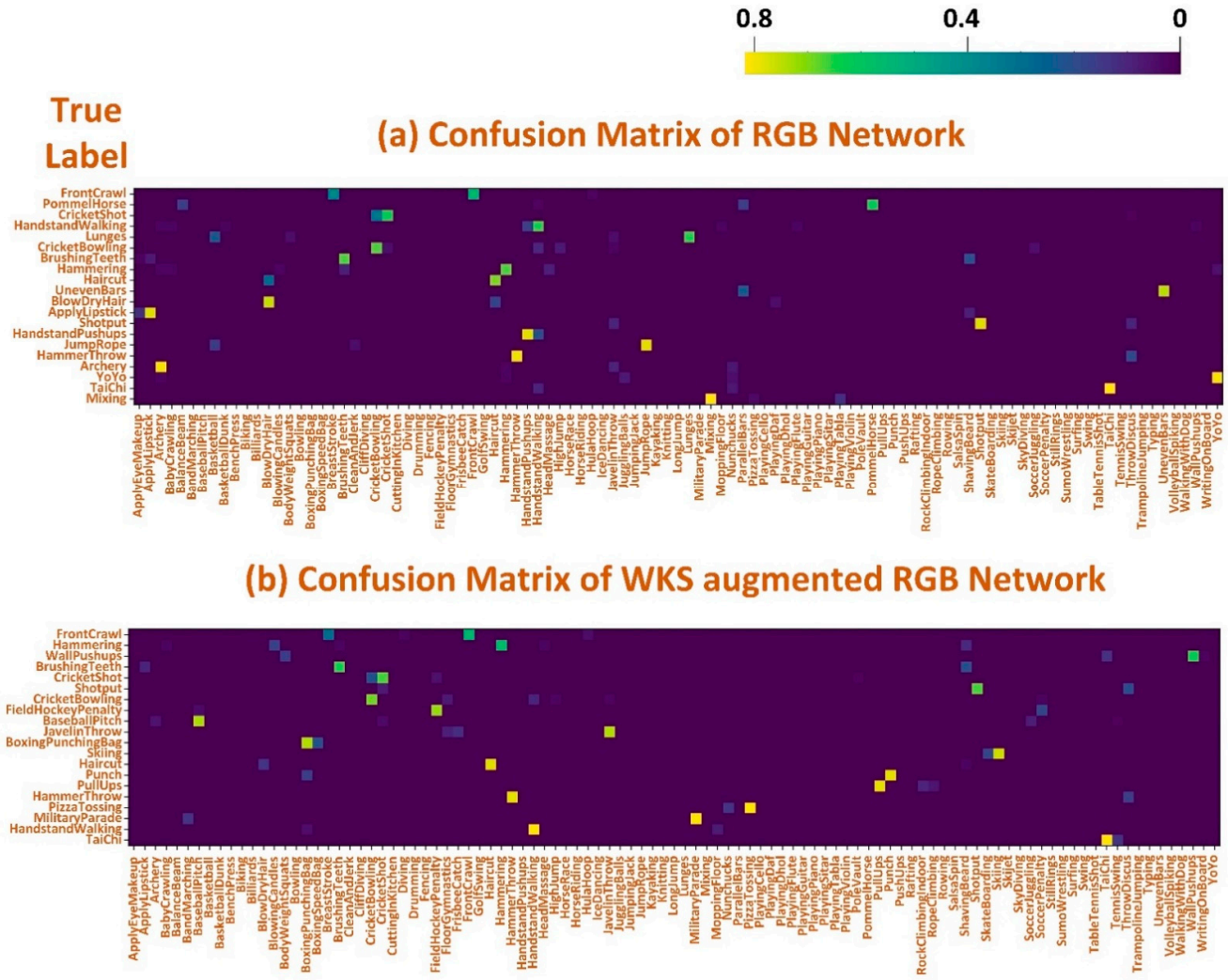

Figure 6. Confusion matrix comparison of RGB network (**a**) without and (**b**) with distillation. The *y*-axis shows the categories having the lowest accuracy. For a given true category, in which the category is often misclassified with, can be found in the figure in terms of the distribution of square icons.

4.4. Evaluation of CWAF

We explore the CWAF method to more rationally aggregate clip-level results. The results of this strategy are listed in Table 3. The metric of video mean score indicates average confident scores for a given video. A comparison between the accuracy and number proportion when the video mean scores are located in sections (0.8, 1], (0.6, 0.8], (0.4, 0.6], [0, 0.4], respectively, is presented in Table 3.

Table 3. Evaluation of the accuracy and number proportion of videos when their Video Mean Scores are located in different sections.

Video Mean Score	Proportion of Videos (%)	Average Pooling Acc (%)	CWAF Acc (%)
0.8~1	50.4	99.9	99.9
0.6~0.8	30.9	92.4	92.3
0.4~0.6	9.1	82.5	84.5
<0.4	9.6	68.0	70.8
0~1	100	93.0	93.4

From this comparison, we know that a substantial part of the videos obtained high confident scores, which means that they are very likely to be classified successfully. The

experiment also verifies that videos with mean video scores above 0.8 are recognized almost completely correctly, whereas the videos containing several noisy clips, often with low scores, are misclassified. Some videos that cannot be validly classified by the densely averaging function can be successfully recognized by lowering the weights of the noisy clip-level results. This experiment also demonstrates that our confidence scoring network effectively assigns proper confidence scores for clip-level results, that is, videos with higher video mean scores obtain higher accuracy.

The experimental results for the recognition accuracies of each category are shown in Figure 7. The red bars represent the improved accuracy of our CWAF, and green bars denote reduced accuracy. We sort the category-level accuracies in order from low to high, and the results are divided into two parts; we demonstrate the fifty categories with the lowest accuracies.

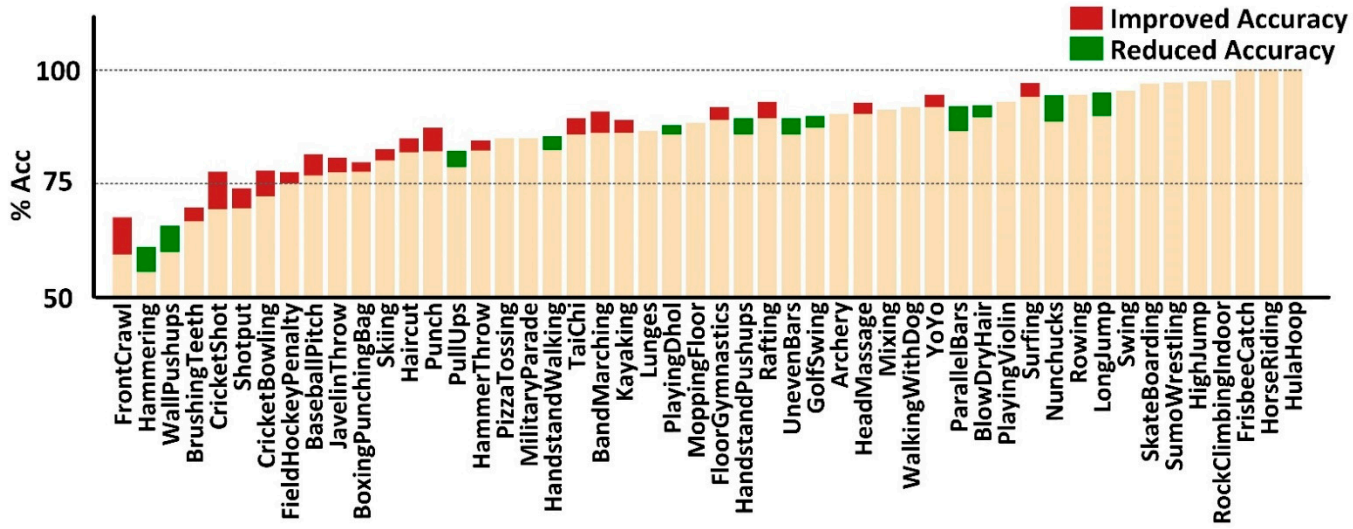

Figure 7. Category recognition accuracies of 3D-ConSwin-B and the changed accuracies contributed by CWAF. The labels of the *x*-axis are sorted by the original accuracy, which increased from left to right.

For the categories in which 3D-ConSwin-B can make correct classifications, no significant improvement is obtained from our aggregation function. However, we achieved good performance in categories where the accuracy is lower, which can be observed in Figure 6. In these categories, weighted clips exhibit a more remarkable effect. During the training process of confidence scoring in the network, we introduced a new training set $\mathcal{D}$, which is composed of the large number of negative samples that 3D-ConSwin-B had falsely classified on the UCF-101 training dataset. This ensures our confidence scoring network can achieve better results in the error-prone category.

4.5. Comparison with Existing Methods

We compare our proposed approach against the advanced methods on two challenging datasets, HMDB-51 and UCF-101. In Table 4, the experimental results are illustrated, where we make a comparison with both one-stream-based methods and multi-stream-based methods. One-stream-based method only adopts RGB frames as input, and multi-stream-based methods are operated on RGB frames and stacked optical flows or other data.

Table 4. Comparison with existing methods on UCF-101 and HMDB-51.

	Method	Pretrain Dataset	Top-1 Accuracy (%)	
			UCF-101	HMDB-51
One-Stream	C3D [30]	Sports-1M	82.3	51.6
	Temporal Segment Network (RGB) [22]	ImageNet	85.2	51.0
	3D-ResNet-152 [51]	Kinetics	89.6	62.4
	R(2+1)D [20]	Sports-1M	93.6	66.6
	3D CNN Architecture [52]	None	87.4	-
	CoViAR [12]	ILSVRC 2012-CLS	90.4	59.1
	Ours (3D-ConSwin-B-base)	Kinetics	92.2	66.0
	Ours (3D-ConSwin-B-full)	Kinetics	93.4	67.2
Multi-Stream	Two-stream CNN [29]	ImageNet	88.0	59.4
	I3D [3]	ImageNet + Kinetics	97.8	80.9
	ResNext101 BERT [21]	Kinetics	97.9	83.6
	DMC-Net [11]	ILSVRC 2012-CLS	92.3	71.8

We denote the classifier without distillation and CWAF as 3D-ConSwin-B-base, and 3D-ConSwin-B-full represented results are obtained under distillation and CWAF methods. Our framework achieves accuracies of 93.4% and 67.2% on UCF-101 and HMDB-51, respectively. Compared to the existing one-stream methods, our method achieved competitive accuracy. Our models are pretrained on the Kinetics-400 dataset. Among these methods, 3D-ResNet-152 is also pretrained on the Kinetics dataset. Finally, our framework exceeded the 3D-ResNet-152by about 3.8%.

The multi-stream-based approaches achieve a higher classification accuracy, but also have a higher time cost. According to the results obtained in [31], the average time cost per video on the optical flow is approximately 130 times that of the RGB input alone. We present the results of classifying only with the input RGB clips. In contrast to the multi-stream-based approaches, our method saved a great deal of computational cost.

5. Conclusions and Discussion

We introduced a new data modality, which is marked with WKS labels, strengthening the difference between the body and background. We also introduced a novel framework that attempts to use whole-body labeled images for model training, and transfers this recognition ability to the RGB network. The experimental results show that the data provided more refined body information for classification, and by distilling the high-level semantic features based on refined body information to the RGB model, the original RGB model can capture more comprehensive action information. Irrelevant information such as human clothing and appearance can be further filtered, which makes the model well-suited for the classification of actions that focus on pose. Some categories which shared a similar background can be classified more accurate by focusing on pose features, such as FrontCrawl and BreastStroke.

Inspired by the success of transformers for vision tasks, we designed an architecture that takes advantage of both the 3D CNNs and Swin transformer to extract spatiotemporal features, which has demonstrated advanced performance. By analyzing the confidence of the clip-level results, we designed an aggregation method to assign more rational weights to each clip output. The experimental results demonstrate that our methods achieve favorable accuracy on the UCF-101 and HMDB-51 datasets. Of note, with limited hardware resources, we temporarily have no ability to perform our experiment on large datasets, such as Kinetics, therefore, we will investigate the performance of our strategy on them in the future.

Author Contributions: Conceptualization, Z.G.; Funding acquisition, S.Y.; Project administration, S.Y.; Writing—original draft, Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: The authors disclose the following financial support for the research, authorship, and/or publication of this article. This work was supported by the Major Special Science and Technology Project of Sichuan Province, grant number 2018GZDZX0024 and the Sichuan Science and Technology Program, grant number 2020YFG0288.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are openly available at https://www.crcv.ucf.edu/data/UCF101.php, accessed on 3 May 2022 and https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/#Downloads, accessed on 3 May 2022, reference number [48,49], respectively.

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

References

1. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]
2. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv* **2014**, arXiv:1406.2199.
3. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
4. Li, D.; Qiu, Z.; Pan, Y.; Yao, T.; Li, H.; Mei, T. Representing videos as discriminative sub-graphs for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
5. Wang, L.; Tong, Z.; Ji, B.; Wu, G. TDN: Temporal difference networks for efficient action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1895–1904.
6. Wu, L.; Yang, Z.; Jian, M.; Shen, J.; Yang, Y.; Lang, X. Global motion estimation with iterative optimization-based independent univariate model for action recognition. *Pattern Recognit.* **2021**, *116*, 107925. [CrossRef]
7. Gharahbagh, A.A.; Hajihashemi, V.; Ferreira, M.C.; Machado, J.J.; Tavares, J.M.R.J.A.S. Best frame selection to enhance training step efficiency in video-based human action recognition. *Appl. Sci.* **2022**, *12*, 1830. [CrossRef]
8. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with directed graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7912–7921.
9. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-time action recognition with deeply transferred motion vector cnns. *IEEE Trans. Image Process.* **2018**, *27*, 2326–2339. [CrossRef]
10. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267. [CrossRef]
11. Shou, Z.; Lin, X.; Kalantidis, Y.; Sevilla-Lara, L.; Rohrbach, M.; Chang, S.-F.; Yan, Z. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1268–1277.
12. Wu, C.Y.; Zaheer, M.; Hu, H.; Manmatha, R.; Smola, A.J.; Krähenbühl, P. Compressed video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6026–6035.
13. Chaudhary, S.; Murala, S. Deep network for human action recognition using Weber motion. *Neurocomputing* **2019**, *367*, 207–216. [CrossRef]
14. Mishra, S.R.; Mishra, T.K.; Sanyal, G.; Sarkar, A.; Satapathy, S.C.J.P.R.L. Real time human action recognition using triggered frame extraction and a typical CNN heuristic. *Pattern Recognit. Lett.* **2020**, *135*, 329–336. [CrossRef]
15. Liu, Z.; Li, J.; Gao, G.; Qin, A.K. Temporal memory network towards real-time video understanding. *IEEE Access* **2020**, *8*, 223837–223847. [CrossRef]
16. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 13359–13368.
17. OpenMMLab Pose Estimation Toolbox and Benchmark. Available online: https://github.com/open-mmlab/mmpose (accessed on 4 May 2022).
18. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
19. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 10012–10022.
20. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
21. Kalfaoglu, M.E.; Kalkan, S.; Alatan, A.A. Late temporal modeling in 3d cnn architectures with bert for action recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 731–747.

22. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2740–2755. [CrossRef]
23. 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, Boston, MA, USA, 23–28 August 2020; pp. 4694–4702.
24. Wang, J.; Cherian, A. Discriminative video representation learning using support vector classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 420–433. [CrossRef]
25. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558.
26. Sadanand, S.; Corso, J.J. Action bank: A high-level representation of activity in video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1234–1241.
27. Scovanner, P.; Ali, S.; Shah, M. A 3-dimensional sift descriptor and its application to action recognition. In Proceedings of the 15th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 357–360.
28. Chen, C.-F.R.; Panda, R.; Ramakrishnan, K.; Feris, R.; Cohn, J.; Oliva, A.; Fan, Q. Deep analysis of cnn-based spatio-temporal representations for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6165–6175.
29. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
30. 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.
31. Crasto, N.; Weinzaepfel, P.; Alahari, K.; Schmid, C. Mars: Motion-augmented rgb stream for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7882–7891.
32. Kataoka, H.; Wakamiya, T.; Hara, K.; Satoh, Y. Would mega-scale datasets further enhance spatiotemporal 3D CNNs? *arXiv* **2020**, arXiv:2004.04968.
33. Wang, J.; Peng, X.; Qiao, Y. Cascade multi-head attention networks for action recognition. *Comput. Vis. Image Underst.* **2020**, *192*, 102898. [CrossRef]
34. Li, X.; Wang, J.; Ma, L.; Zhang, K.; Lian, F.; Kang, Z.; Wang, J. Sth: Spatio-temporal hybrid convolution for efficient action recognition. *arXiv* **2020**, arXiv:2003.08042.
35. Brehar, R.D.; Muresan, M.P.; Mariţa, T.; Vancea, C.C.; Negru, M.; Nedevschi, S. Pedestrian street-cross action recognition in monocular far infrared sequences. *IEEE Access* **2021**, *9*, 74302–74324. [CrossRef]
36. Yan, H.; Hu, B.; Chen, G.; Zhengyuan, E. Real-time continuous human rehabilitation action recognition using OpenPose and FCN. In Proceedings of the 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Shenzhen, China, 24–26 April 2020; pp. 239–242.
37. Verma, A.; Meenpal, T.; Acharya, B. Multiperson interaction recognition in images: A body keypoint based feature image analysis. *Comput. Intell.* **2021**, *37*, 461–483. [CrossRef]
38. Pandurevic, D.; Draga, P.; Sutor, A.; Hochradel, K. Analysis of competition and training videos of speed climbing athletes Using feature and human body keypoint detection algorithms. *Sensors* **2022**, *22*, 2251. [CrossRef]
39. Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
40. Dewi, C.; Chen, R.-C.; Jiang, X.; Yu, H. Deep convolutional neural network for enhancing traffic sign recognition developed on Yolo V4. *Multimed. Tools Appl.* **2022**, 1–25. [CrossRef]
41. Wachinger, C.; Toews, M.; Langs, G.; Wells, W.; Golland, P. Keypoint transfer for fast whole-body segmentation. *IEEE Trans. Med. Imaging* **2018**, *39*, 273–282. [CrossRef]
42. Lopez-Paz, D.; Bottou, L.; Schölkopf, B.; Vapnik, V. Unifying distillation and privileged information. *arXiv* **2015**, arXiv:1511.03643.
43. Vapnik, V.; Izmailov, R. Learning using privileged information: Similarity control and knowledge transfer. *J. Mach. Learn. Res.* **2015**, *16*, 2023–2049.
44. Garcia, N.C.; Morerio, P.; Murino, V. Modality distillation with multiple stream networks for action recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 103–118.
45. Garcia, N.C.; Bargal, S.A.; Ablavsky, V.; Morerio, P.; Murino, V.; Sclaroff, S. Dmcl: Distillation multiple choice learning for multimodal action recognition. *arXiv* **2019**, arXiv:1912.10982.
46. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
47. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10347–10357.
48. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
49. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.

50. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
51. 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, 18–23 June 2018; pp. 6546–6555.
52. Vrskova, R.; Hudec, R.; Kamencay, P.; Sykora, P. Human Activity Classification Using the 3DCNN Architecture. *Appl. Sci.* **2022**, *12*, 931. [CrossRef]