Research Article

Attention-Based Temporal Encoding Network with Background-Independent Motion Mask for Action Recognition

Zhengkui Weng¹, Zhipeng Jin¹, Shuangxi Chen¹, Quanquan Shen¹, Xiangyang Ren^{2,3}, and Wuzhao Li⁴

¹Jiaxing Vocational and Technical College, Jiaxing, Zhejiang, China

²Medical 3D Printing Center, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China

³School of Electrical Engineering, Zhengzhou University, Zhengzhou, Henan, China

⁴Wenzhou Polytechnic, Wenzhou, Zhejiang, China

Correspondence should be addressed to Zhengkui Weng; zkweng19@163.com

Received 12 September 2020; Revised 24 February 2021; Accepted 12 March 2021; Published 30 March 2021

Academic Editor: Mario Versaci

Copyright © 2021 Zhengkui Weng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Convolutional neural network (CNN) has been leaping forward in recent years. However, the high dimensionality, rich human dynamic characteristics, and various kinds of background interference increase difficulty for traditional CNNs in capturing complicated motion data in videos. A novel framework named the attention-based temporal encoding network (ATEN) with background-independent motion mask (BIMM) is proposed to achieve video action recognition here. Initially, we introduce one motion segmenting approach on the basis of boundary prior by associating with the minimal geodesic distance inside a weighted graph that is not directed. Then, we propose one dynamic contrast segmenting strategic procedure for segmenting the object that moves within complicated environments. Subsequently, we build the BIMM for enhancing the object that moves based on the suppression of the not relevant background inside the respective frame. Furthermore, we design one long-range attention system inside ATEN, capable of effectively remedying the dependency of sophisticated actions that are not periodic in a long term based on the more automatic focus on the semantical vital frames other than the equal process for overall sampled frames. For this reason, the attention mechanism is capable of suppressing the temporal redundancy and highlighting the discriminative frames. Lastly, the framework is assessed by using HMDB51 and UCF101 datasets. As revealed from the experimentally achieved results, our ATEN with BIMM gains 94.5% and 70.6% accuracy, respectively, which outperforms a number of existing methods on both datasets.

1. Introduction

Understanding and the process to recognize actions of humans in realistic videos remains a challenge due to the diversified objects, ambiguous human motions, and complex relationship between scenes and human beings. From traditional handcraft features [1] to deep learning-based approaches [2], impressive progress has been made owing to the development of powerful pattern recognition approaches over the past few years. Meanwhile, the convolutional neural network (CNN) has become popular for human action recognition [3–5]. However, compared with

the progress in still image classification [6], CNN action recognition's performance improvements is relatively small.

Different with the still image classifying, a special property for action recognition refers to scattered background and dynamic perspective. According to Figure 1, some example frames with poor conditions are illustrated in the first row. Thus, some existing video action recognition methods seek the diminishment of the interfere of different visual appearances based on the detection of interest points of space and time [7], creating trajectories of motion [8], or splitting the object that moves [9] prior to their feeding to CNNs. Nevertheless, the mentioned approaches have two

FIGURE 1: Some sampled video frames and their corresponding categories. (a) BodyWeightSquats, (b) HighJump, (c) LongJump, (d) StillRings, (e) ApplyEyeMakeup, (f) ApplyLipStick, (g) CricketBowling, and (h) CliffDiving.

major drawbacks. On the one hand, several subtle motions carrying vital data are likely to be lost under the extremely sensitive threshold. On the other hand, discriminative characteristics are hard to extract from the taken areas with the relationship to action since the mentioned foreground areas are likely to show a spatial coherence [10].

In addition, the other distinct feature exhibited by the video refers to temporal dimension. It is difficult for 3D-CNN architectures [11] to obtain higher performance when compared with the sophisticated improved dense trajectory (IDT) [12] at the early stage because only a few frames rather than the whole video may be covered by comparatively small 3D convolution. More recently, researchers focus on combing the CNNs with RNNs such as LSTM [13] to model the temporal configuration within the video. The other one focuses on aggregating the frame-level descriptors through different feature encoding approaches (e.g., the average or maximal pooling along temporal dimension [14], vector of descriptors with local aggregation (VLAD) [15], and temporal pyramid pooling (TPP) [16]).

Nevertheless, in the absence of semantic data guiding at a high level, the representation of actions under the extraction with the use of the mentioned methods fail to have the expression of the frames with the higher relevance inside a video in a selective manner; accordingly, the residual noise is likely to increase unavoidably based on frames with any relevance [17]. The second row of Figure 1 gives some sampled video frames and their corresponding categories that are inappropriate for classification.

In view of the above observations and analysis, a novel framework is developed, termed as temporal encoding network (ATEN) on the basis of attention with background-independent motion mask (BIMM) for video action recognition (Figure 2). Inspired by the successful application of saliency technology in still images [18, 19], one motion object segmentation

approach on the basis of boundary prior receives the introduction through the combination of the minimal geodetic distance based on dynamic contrast segmenting strategic procedure in an undirected weighted graph. Subsequently, we build the BIMM for enhancing the object that moves based on the suppression of the background motion with relevance within the respective frame. The attention mechanism in ATEN does not process all sampled frames equally but automatically highlight the semantic vital frames.

The major contribution here is elucidated as follows:

- (i) We develop an overall automatic background-independent motion mask (BIMM) for enhancing the moving objects based on sophisticated background by complying with one emerging motion segmenting approach on the basis of boundary prior
- (ii) An attention-based temporal encoding network (ATEN) is developed for video action recognition. Instead of solving the features at the frame level in an equal manner, one uncomplicated and significant attention mechanism based on long ranges is proposed
- (iii) We conduct extensive experimental processes on two difficult datasets for demonstrating that the framework here statistically outclasses a number of existing methods by average 7.8% and 4.0%

2. Related Work

2.1. Approaches on the Basis of Handcraft Features. Prior to the extensive CNN application, methods under the handcrafting process are dominant in action-recognizing field. For instance, local spatiotemporal descriptors with the fisher vector (FV) [20] and classical bag of words (BoW) [21] model had shown good recognition performance.

FIGURE 2: The framework of the temporal encoding network on the basis of attention here with background-independent motion mask.

Descriptors on the basis of histogram under the calculation following the trajectory [12, 22] achieved high-prospect outcomes in human action-recognizing task. In [22], histograms of optical flow (HOF) and oriented gradients (HOG) received the extraction in the respective spatial-temporal interest point to model local motion information. Wang et al. [22] conducted the extraction of dense trajectory based on the track of dense point frame by frame under a range of scales. Four descriptors, namely, HOG, HOF, motion boundary histogram (MBH) [22], and trajectory shape (TS) [22] received the extraction along every dense trajectory to represent the action. To eliminate the impact of camera motion, an improved dense trajectory (IDT) was proposed by using a human detector to extract trajectory more concentrated in the action-related areas [12]. In order to capture motion features more effectively, as suggested by Vig et al. [23], the recognition performance is able to be kept with just 30–40% descriptors chosen from dense trajectories. In this end, Yi and Lin [24] and Xu et al. [25] aimed to select dense trajectories from salient motion areas and encoded them into a compact video representation. Unlike [24, 25] utilizing video saliency to select action-related trajectory, the authors in [26] proposed a length-variable edge trajectory extracted from edge points to model different speeds of motion. These approaches commonly focused on capturing edge, corner, and motion features through trajectories and were successful in recognizing relatively simple actions. However, it remains challenging to recognize various long-term complex human actions.

2.2. Approaches on the Basis of Deep Learning. Progresses for deep learning over the past few years lead to impressive applications for video action recognition. To be specific, pretrained CNN transfer learning with limited training information shows great performance in [27–30]. For instance, Simonyan and Zisserman [27] pretrained a 2-stream CNN composed of spatial CNN and temporal CNN using the ImageNet dataset to capture features from single static frame and motions, respectively. Tran et al. [28] extended the 3D-CNN [11] by learning spatiotemporal features from adjacent frames with 3D convolution and pooling. Sun et al.

[4] introduced one emerging spatiotemporal CNN structure which factorized 3D convolution into 2D spatial convolution to reduce the computation complexity. In [29], a new three-stream CNN model was proposed, which is derived from two-stream CNN. Wang et al. [30] proposed a hybrid representation which combined the merits from both IDT [12] and two-stream CNN to form the trajectory-pooled deep convolution descriptor (TDD). Feichtenhofer et al. [31] discovered a deficiency in two-stream CNN and further improved it with convolutional fusion of two streams. Besides, the selection of informative areas to extract features refers to a reasonable method for promoting action recognition performance. Cheron et al. [14] proposed one CNN (P-CNN) on the basis of pose which was different from conventional two-stream architecture. In [14], the human pose was first detected and tracked over time, and then the CNN features were obtained separately from each part of body in each frame. However, the performance of P-CNN remains unsatisfied because pose estimation in natural images remains challenging. In [10], a fast and accurate video representation method based on salient areas of motion was proposed, which can extract the most useful features for classification. Rui et al. [32] developed one tube CNN (T-CNN) based on 3D convolution operation to recognize actions. However, most CNN-based methods are subject to false recognition and detail missing, since merely coarse global features at the frame level are adopted [3]. Moreover, the other limitation of the aforementioned CNN frameworks is that they usually focus more on capturing appearance and short-term features rather than multistaged temporal information. Considering that the temporal information is critical for video classification, Shi et al. [33] proposed an effective representation for long-term motions by extending two-stream framework with an informative representation named sequential trajectory texture (STT). Wang et al. [34] developed one video-level framework termed as the temporal segmenting network (TSN), aiming to utilize long-term temporal information by a sparse sampling strategy. Meanwhile, several works [13, 35, 36] introduced recurrent neural networks (RNNs) such as LSTM, which can preserve the previous states in memory cells for modelling the data in video in a long term. Donahue

et al. [13] designed a unified model called LRCN which connected LSTM with CNNs to describe actions. A hierarchical LSTM network was proposed in [36], which enables to incorporate the information from three aspects, namely, long-term temporal structure, short-term motion, and static spatial in terms of complex action classification.

Lately, applying models of attention arouses rising attention inside a range of aspects (e.g., image captioning [37] and machine translation [38]). Several methods on the basis of attention were designed to recognize actions. Li et al. [39] employed a dual attention ConvNet (DANet) to deal with the computational cost of two-stream framework. Li et al. [40] developed one end-to-end structure based on the adaptive spatiotemporal attention model. Liu et al. [41] proposed one unified spatiotemporal attention network (STAN) to recognize video actions, one general attention neural cell called AttCell was devised to estimate the attention probability in the respective space site and for the respective video segment inside one temporal sequence. Yang et al. [42] developed a spatial-temporal attentive convolutional neural network (STA-CNN), incorporating one spatial attention system and one temporal attention system inside one convolutional network under the unification for recognizing video action. However, majority of the presented attention models to recognize actions employ RNN structures for implementing attention systems. The main defect exhibited by RNN refers to its instable property within training. The ATEN here receives the implementation with no use of RNN. As opposed to the mentioned, neural network structure is adopted for building the temporal encoding module.

3. Background-Independent Motion Mask

3.1. Overview of the Boundary Prior-Based Motion Object Segmentation for BIMM. Figure 3 overviews the boundary prior-based motion segmentation for BIMM. Initially, the respective input frame receives the segmenting process to superpixels. One weighted graph that is not directed, with the respective superpixel recognized to be a node, is built inside the respective frame under the segmentation, and the nearby relationship of superpixels is presented by the edge between two nodes. Inspired by [19, 43], we define the geodesic distance as the cumulative weighted shortest path between two superpixels in the constructed graph. According to observation, the objects that move are constantly surrounded by the superpixels exhibiting great edge spatiotemporal value and the saliency probability receives the computation to be the minimal geodesic distance to the respective superpixel to the frame boundary. For extracting appearance and motion data in the area with the relationship to action, 2 kinds of the edge received the extraction from the frame: motion edges under the estimation based on optical flow via 2 consecutive frames and static apparent edges in the single static frame. Then, one adaptive threshold receives the design for generating the coarse saliency map based on the process to split the superpixels to foreground superpixel set as well as background superpixel set. Moreover, one

dynamic contrast segmenting strategic procedure receives the introduction for gaining the precise motion areas based on the computation of the geodesic distance of each 2 superpixels of foreground and background sets inside the graph. Foreground superpixels exhibiting greater mean geodesic distance to the background superpixels are considered to be pseudoforeground superpixels and introduced to the background set and vice versa. Lastly, the respective frame background receives the suppression based on halving of the brightness of pixels not within the precise motion area.

3.2. Spatiotemporal Edge Generation. Image preprocessing [44, 45] also plays an important role in BIMM. Given a frame sequence $F = \{F^1, F^2, \dots, F^k\}$, $P_i^k = (x_i^k, y_i^k)$ denotes the i -th pixel in frame F^k . We calculate the apparent edge probability map $E_{\text{static}}^k(P_i^k) = E_{\text{static}}^k(x_i^k, y_i^k)$ corresponding to k -th frame F^k and pixel P_i^k using Canny edge detector [46]. Let $\omega_i^k = (u_i^k, v_i^k)$ be the optical flow of i -th pixel in frame F^k , where u and v are the horizontal and vertical component, respectively, then we calculate the motion gradient $E_{\text{motion}}^k(\omega_i^k)$ of the optical flow [47] ω_i^k as

$$E_{\text{motion}}^k(\omega_i^k) = \|\nabla \omega_i^k\| = \left\| \frac{\partial \omega_i^k}{\partial u_i^k} + \frac{\partial \omega_i^k}{\partial v_i^k} \right\|. \quad (1)$$

Subsequently, based on SLIC, the present study segments the frame to superpixels [48]. To be specific, the color space of the image is transformed from RGB to LAB firstly, and a five-dimensional feature vector is composed of three components from LAB and the distance between two pixels x, y . Then, the K-means algorithm is used to cluster the pixels according to the certain distance measurement. Finally, the superpixels with relatively uniform size are generated. $Y^k = \{Y_1^k, Y_2^k, \dots, Y_n^k\}$ is denoted as the superpixel set in frame F^k . Based on the static edge map E_{static}^k , we calculate the superpixel Y_n^k edge probability to be the pixels on average with ten maximal edge probability values in Y_n^k . Likewise, the gradient magnitude of motion is obtained based on E_{motion}^k . Through the mentioned processes, two superpixel-based edge probability maps E_{static}^k and E_{motion}^k are developed. Next, one spatiotemporal edge probability map $E^k(Y^k)$ is able to be developed [19] to be the element-wise multiplying of E_{static}^k and E_{motion}^k :

$$E^k(Y^k) = E_{\text{static}}^k \circ E_{\text{motion}}^k. \quad (2)$$

The reason for computing $E^k(Y^k)$ is that both static and motion information capable of providing useful information for motion object segmentation is considered. If the foreground motion region is different from the background, optical flow gradient magnitude is required to achieve high value close to the foreground region compared with background. Moreover, the static edge probability provides one instructing element in terms of the apparent boundaries by complying with the appearance in the RGB frame. When both kinds of information fused using (2), the final spatiotemporal edge probability map $E^k(Y^k)$ is able to point out the location of the foreground moving object.

FIGURE 3: Overview of boundary prior-based motion segmentation for BIMM.

3.3. Boundary Prior Saliency Estimation. In terms of the respective frame F^k , one weighted graph that is not directed is built to be $g^k = \{Y^k, e^k\}$ with superpixels Y^k as nodes and the edges between pairs of adjacent superpixels as e^k . Under the built configuration, one $|Y^k| \times |Y^k|$ adjacency matrix of the graph $W^k(m, n)$ receives the development, with the definition of the element-wise multiplying of $e^k(m, n)$ and w_{mn}^k , where $e^k(m, n)$ represents the edge between adjacent superpixels Y_m^k and Y_n^k , if there is an edge between two superpixels, and $e^k(m, n)$ is 1; otherwise, $e^k(m, n)$ is 0. w_{mn}^k represents the weight of adjacent superpixels Y_m^k and Y_n^k :

$$W^k(m, n) = e^k(m, n) \circ w_{mn}^k, \quad (3)$$

$$w_{mn}^k = \|E^k(Y_m^k) - E^k(Y_n^k)\|,$$

where $E^k(Y_m^k)$ and $E^k(Y_n^k)$ express the spatiotemporal boundary probability of superpixels Y_m^k and Y_n^k , separately. For emphasizing the foreground motion objects with large spatiotemporal edge data or having the surrounding by areas with high spatiotemporal edge data, the present study adopts geodesic distance for computing one coarse object probability map. The geodesic distance $d_g(Y_m^k, Y_n^k, g^k)$ in graph g^k has the definition of the cumulative weighted shortest path over all possible paths between Y_m^k and Y_n^k :

$$d_g(Y_m^k, Y_n^k, g^k) = \min \left\{ \sum_{Y_n^k}^{Y_m^k} |W^k(m, n) \cdot e^k(m, n)| \right\}. \quad (4)$$

By complying with the definition of geodesic distance d_g , we can find that supposing a superpixel is located externally from the moving object, there possibly exists a path to the frame boundaries not passing through any superpixels which have high spatiotemporal edge value. For this reason, the geodesic distance of such superpixel is relatively small. In contrast, supposing a superpixel is inside the moving object, the superpixel should have the surrounding by the superpixels exhibiting large spatiotemporal edge value, increasing the geodesic distance to the frame boundaries. When the

superpixel is on the edge of the moving object, the path to one of the boundaries does not pass through the moving object (except itself) and the geodesic distance of the mentioned superpixel turns out to be comparatively low. However, our graph is very sparse, and although the superpixel is on the edge of the moving target, its own spatiotemporal edge value is much larger than those of the background superpixels. Likewise, the boundary prior saliency value S_n^k in terms of the respective superpixel Y_n^k is computed by

$$S_n^k = \min_{q \in Q^k} \{d_g(Y_n^k, q, g^k)\}, \quad (5)$$

where Q^k denotes the superpixels from four boundaries in each frame F^k . All saliency values in S_n^k are normalized to $[0, 1]$. Under the saliency value S_n^k , the coarse saliency map can be generated based on a self-adaptive threshold, dividing the superpixels into background set $\hat{B}^k$ and foreground set $\hat{O}^k$. The self-adaptive threshold θ^k for each frame F^k is computed by

$$\theta^k = \mu(S_n^k) = \frac{1}{n} \sum_{i \in n} S_i^k, \quad (6)$$

where $\mu(\cdot)$ denotes the mean value of all superpixels within frame F^k by saliency value S_n^k . Subsequently, the superpixels in the respective frame F^k can be cataloged to background and foreground set:

$$\begin{aligned} \hat{O}^k &= \{Y_n^k | S_n^k > \theta^k\}, \\ \hat{B}^k &= Y^k - \hat{O}^k. \end{aligned} \quad (7)$$

For obtaining the refined motion object region, we add one dynamic contrast segmentation strategy by comparing both inter- and intrageodesic distance between foreground and background sets. Three kinds of distances, namely, d_{OO} , d_{OB} , and d_{BB} are utilized as the measurement for the superpixels $Y^k = \hat{O}^k \cup \hat{B}^k$ in each frame F^k :

$$\begin{aligned}
d_{OO} &= \mu \left(\sum_{Y_n^k, Y_{n+1}^k \in \hat{O}^k} [d_g(Y_n^k, Y_{n+1}^k, g^k)] \right), \\
d_{OB} &= \mu \left(\sum_{Y_n^k \in \hat{O}^k, Y_{n+1}^k \in \hat{B}^k} [d_g(Y_n^k, Y_{n+1}^k, g^k)] \right), \\
d_{BB} &= \mu \left(\sum_{Y_n^k, Y_{n+1}^k \in \hat{B}^k} [d_g(Y_n^k, Y_{n+1}^k, g^k)] \right),
\end{aligned} \quad (8)$$

where d_{OO} denotes the mean geodesic distance between two foreground superpixels. The higher the geodesic distance value, the closer the relationship between the superpixels. Similarly, d_{OB} and d_{BB} are also adopted as the measurement to describe inter- and intrarelation among superpixels in one frame. Thus, we define the refined motion object areas O^k of k -th frame as

$$\begin{aligned}
O^k &= \{Y^k \in \hat{O}^k | d_{OO} > d_{OB}\} \cup \{Y^k \in \hat{B}^k | d_{BB} \leq d_{OB}\}, \\
B^k &= Y^k - O^k.
\end{aligned} \quad (9)$$

The main rationale behind (9) is both inter- and intradifferences between two superpixel sets are taken into consideration in the proposed dynamic contrast segmentation strategy. The intermediate results of BIMM are shown in Figure 4.

3.4. Background Suppression. By complying with the classified superpixels Y^k in each frame F^k , the corresponding BIMM is computed by binarizing the values S_n^k in the coarse saliency map. If $S_n^k(i, j) \in B^k$, then $S_n^k(i, j) = 0$; otherwise, $S_n^k(i, j) = 1$. Next, we suppress the background pixels in F^k to improve the importance of foreground motion objects. We first transform the F^k in RGB color space to HSI color space, halving the I component of background areas where $S_n^k(i, j) = 0$, and then do an opposite operation to reverse the frame back to RGB color space. As a result, a BIMM-based frame can be obtained and denoted as RGB-BIMM. Similarly, to construct the OF-BIMM, the optical flow w_i^k of each frame F^k is treated in the same way.

4. Video-Based Action Recognition Using ATEN

4.1. Temporal Structure Modelling. For each video V , we divide it into N segments along the temporal domain, i.e.,

$V = \{S_k\}_{k=1}^N$, where S_k is the k -th segment. Thus, a sequence of video snippets $T_1, T_2, T_3, \dots, T_N$ is obtained by randomly sampling each frame from the corresponding segment S_N . Then, the temporal structure of a video V can be modelled as follows:

$$TS(V) = C(\mathcal{F}(T_1; W), \mathcal{F}(T_2; W), \dots, \mathcal{F}(T_N; W)), \quad (10)$$

where $\mathcal{F}(T_N; W)$ is the function of a CNN with parameter W which represents the convolution results of snippet T_N . Here, $T_N = \{T_N^{\text{RGB-BIMM}}, T_N^{\text{OF-BIMM}}\}$, which represents the two modalities RGB-BIMM and OF-BIMM from the N -th snippet. C is the snippet-level aggregation function, which aggregates all the features learned from the base model via a temporal attention model to obtain a video-level representation.

4.2. Attention-Based Temporal Encoding Network. We do not use the single RGB frame or optical flow stacks in our framework, and the proposed ATEN works through the snippets that receive the sampling process in a sparse manner from the overall video. The respective snippet in the sequence generates its own feature representation. The overall framework of the attention-based temporal encoding network is shown in Figure 5. We use $\mathcal{F}(T_N; W)$ to represent any given one from the above two modalities (RGB-BIMM and OF-BIMM). A soft attention e_i , $i = 1, \dots, N$ is computed for each segment through one multiple-layer perceptron in the output with the full connection, which is defined as

$$\begin{aligned}
w_{\text{att}} &= \text{ReLU}(W_{\text{att}} \mathcal{F}(T_i; W) + b_{\text{att}}), \\
e_i &= f_{\text{att}}(w_{\text{att}}, \mathcal{F}(T_i; W)) = w_{\text{att}}^T \mathcal{F}(T_i; W), \\
w_i &= \frac{\exp(e_i)}{\sum_{j=1}^N \exp(e_j)},
\end{aligned} \quad (11)$$

where w_{att} , W_{att} , and b_{att} are the parameters of the attention model and the soft attention coefficient e_i in each w_{att} is then normalized to guarantee the sum of e_i is 1. The obtained temporal attention weight w_i is used to characterize the importance of the i -th snippet. Specifically, a degenerate case appears if $w_i = 1/N$, i.e., all snippets have equal importance.

Therefore, the segment-level aggregation function C is formed as $C = \sum_{i=1}^N w_i \mathcal{F}(T_i; W)$ to obtain a video-level video representation. Finally, the output of the ATEN is modelled as follows:

$$ATEN(T_1, T_2, T_3, \dots, T_N) = \mathcal{H}(C(\mathcal{F}(T_1; W), \mathcal{F}(T_2; W), \dots, \mathcal{F}(T_N; W))). \quad (12)$$

C combines the output from each snippet to obtain a fix-length feature vector, and based on this vector, $\mathcal{H}$ aims to

predict the probability score of the respective class in terms of the videos. Accordingly, the common multiclass

FIGURE 4: Visualization of boundary prior-based motion object segmentation for BIMM. (a) Source frame F_k . (b) Optical flow ω^k of F_k . (c) Superpixel-based optical flow gradient magnitude E^k_{motion} . (d) Superpixel-based static edge map E^k_{static} . (e) Spatiotemporal edge probability map E_k . (f). Boundary prior saliency map S^k_n for each superpixel Y^n_k using geodesic distance to the four frame boundaries (g) BIMM generation from refined motion object areas by using dynamic contrast segmentation strategy. (h) Source frame F_k is decomposed into background set B^k and foreground set O^k after background suppression using BIMM, and the bright areas indicate the refined motion object areas. (i) The optical flow ω^k_i of each frame F_k is treated in the same way to form OF-BIMM.

FIGURE 5: The overall framework of the attention-based temporal encoding network.

activation function softmax is taken for $\mathcal{H}$. Based on the integration with the set softmax loss, we define the eventual loss function according to the aggregation function C as

$$\mathcal{L}(y, C) = - \sum_{i=1}^K y_i \log \left(\frac{\exp(C_i)}{\sum_{j=1}^K \exp(C_j)} \right), \quad (13)$$

where K is the number of action categories and y_i is the action label corresponding to category i . It is worth mentioning that the gradients of the loss function can freely backpropagate, and the whole framework can be trained in an end-to-end manner because the entire ATEN is differentiable with the attention model directly embedded.

5. Experiments

5.1. Datasets and Evaluation Protocol. The HMDB51 [49] dataset is collected from variety of sources. There are a total of 6766 videos distributed in 51 action categories, each containing a minimum of 100 clips. We follow the original protocol using the first train-test splits and report the accuracy.

The UCF101 [50] dataset consists of 13320 videos falling into 101 action categories. UCF101 provides the largest diversity in terms of actions and the videos from the same group may share similar background and viewpoint. We evaluate our framework on the first train-test splits and report the accuracy.

5.2. Implementation Details. At the data preprocessing phase, the TVL1 optical flow algorithm [47] is employed. Specifically, the optical flow receives the discretization to $[0, 255]$ for ensuring the information consistent property with RGB frames. Subsequently, the proposed BIMM is achieved by using both datasets for generating RGB-BIMM and OF-BIMM as the input of ATEN. Similar to [34], the temporal segment number is set to 6 empirically to guarantee the stability of the whole framework. In ATEN, we choose the public available Inception with Batch Normalization (BN-Inception) network as the base network here from [51] for its high efficient and accurate property. At the training stage, the mini-batch SGD optimizer is employed by setting batch size to 128 and setting momentum to 0.9. Specific to spatial ATEN, the network is initialized based on pretrained models from ImageNet [52] and the learning rates is first set to 0.001 and subsequently reduces to its 1/10 after per 1500 iterations. The iteration finally stopped at 3500. Specific to temporal ATEN, the cross-modality pretraining strategy proposed in [34] is employed based on the exploitation of the learned spatial models for initializing the temporal stream. The learning rate is set to 0.005, downregulated by a factor of 10 after 12K and 16K iterations. The max iteration is set to 18K. For data augmentation, we utilize the scale jittering and random horizontal flipping for reducing the risk of overfitting as advised in [34]. During the testing phase, we select 24 RGB-BIMM or OF-BIMM stacks from the action frames and then divide them into six snippets with each has four

sampled frames according to temporal segments $N=6$. Meanwhile, we crop 4 corners, 1 centre, and horizontal flipping of them from the sampled frames for assessing the ATEN. Finally, more credits are given to the temporal stream by setting its weight to 1.5 and that of the spatial stream to 1 for the fusion of two streams.

5.3. Evaluation on HMDB51. Here, the effect exerted by ATEN with BIMM on HMDB51 is shown in Table 1. We set TSN [34] as the baseline architecture in this section owing to its good performance. The performance of TSN here is implemented according to the reported settings (two modalities). We gain 59.7%, 62.2%, and 70.6% on spatial, temporal, and two-stream ATEN, respectively. The improvement of our ATEN (two modalities) over TSN is 2.1% on HMDB51 dataset which verifies the importance of modelling long-term temporal information. Furthermore, this study explores the improvement of performance of our ATEN over TSN on individual category. Table 2 lists the top-10 action classes with the biggest increment on HMDB51. It is not difficult to find that several action categories with the highest performance increment are easy-confused actions, such as “Run,” “Pullup,” and “Dive,” demonstrating the proposed framework to be effective as well. Moreover, the proposed ATEN gains better performance over TSN in majority of action categories (Figure 6). Consequently, the total average accuracy is higher than TSN.

5.4. Evaluation on UCF101. As is given in Table 1, the recognition performance of ATEN on UCF101 with different modalities reaches 86.3%, 88.7%, and 94.5%, respectively. Interestingly, we observe that temporal stream in our framework achieves better performance on both two datasets compared with spatial stream and combining both of them leads to a higher accuracy, which is also consistent with [3, 27, 29]. To further investigate the performance of ATEN on UCF101, we report the individual category accuracy in Figure 7. Compared with TSN, the recognition accuracy on several typical actions such as “RopeClimbing,” “Swing,” and “TaiChi” is improved significantly. This improvement mainly comes from the use of the proposed framework. On one hand, the video features are extracted more concentrating from the action-related areas by BIMM, and on the other hand, the attention-based model in ATEN adaptively eliminates the temporal redundancy which makes the video representation more discriminative. Figure 8 shows the performance confusion matrix on two datasets. We can find that some actions are easily confused into wrong categories especially when the two actions are similar, such as “Throw” and “Wave,” “Eat” and “Drink.” In spite of that, our method still obtains promising performance on these easy-confused actions as well as on most of the actions.

5.5. Ablation Study. Here, for the verification of the effectiveness exhibited by the proposed background-independent motion mask (BIMM) and attention-based temporal encoding

TABLE 1: Accuracy (%) of proposed ATEN on HMDB51 and UCF101 datasets with different modalities.

Models	Accuracy (%)	
	HMDB51	UCF101
RGB-BIMM	59.7	86.3
OF-BIMM	62.2	88.7
Two modalities	70.6	94.5

TABLE 2: Top-10 performance improvement of ATEN over TSN [34] on HMDB51 and UCF101.

HMDB51			UCF101	
No.	Action category	Increment (%)	Action category	Increment (%)
1	Run	15.71	RopeClimbing	5.88
2	Pullup	14.29	Swing	5.62
3	Dive	12.86	TaiChi	5.56
4	Draw_sword	11.43	BreastStroke	5.48
5	Eat	10.00	HandstandWalking	5.19
6	Handstand	9.99	BabyCrawling	5.15
7	Wave	9.65	Basketball	5.05
8	Brush_hair	8.57	Rowing	4.95
9	Drink	8.24	Kayaking	4.76
10	Clap	7.14	ShavingBeard	4.24

FIGURE 6: Performance comparison between ATEN and TSN (two modalities) [34] on HMDB51.

network (ATEN), various ablation experiments were performed using a range of settings on 2 datasets: RGB + OF + average fusion (Setting 1, similar with TSN),

RGB + OF + ATEN (Setting 2), RGB-BIMM + OF-BIMM + average fusion (Setting 3), and RGB-BIMM + OF-BIMM + ATEN (Setting 4). Tables 3 and 4 separately present

(a)
FIGURE 8: Continued.

FIGURE 8: Performance confusion matrix on (a) HMDB51 and (b) UCF101 (two modalities).

TABLE 3: Ablation study on UCF101 (%).

Settings	BIMM	ATEN	Accuracy
Setting 1	✗	✗	93.0
Setting 2	✗	✓	94.1
Setting 3	✓	✗	93.3
Setting 4	✓	✓	94.5

TABLE 4: Ablation study on HMDB51 (%).

Settings	BIMM	ATEN	Accuracy
Setting 1	✗	✗	67.6
Setting 2	✗	✓	68.2
Setting 3	✓	✗	69.5
Setting 4	✓	✓	70.6

backbone architectures of STA-CNN. With a more novel architecture selected as the major ConvNet of ATEN, higher performance is likely to be achieved.

5.7. Visualization of BIMM and Learned Temporal Attention.

To verify the importance of BIMM for classification, Figure 9 presents the visualization of boundary prior-based motion segmentation for BIMM on two datasets, in which the respective sampled frame represents one action. We randomly select 15 actions from HMDB51 and UCF101 datasets in terms of visualizing process. To a certain extent, videos within the first two rows undergo the contamination with drastic background motion or other noises. The proposed

BIMM is implemented on all sampled frames, from which the motion areas are brighter than the background. This also indicates that the region with more visual information is preserved, because the convolution focuses more on these areas. It is worth noting that the segmented foreground always seems larger than the ground truth in each frame. The reason is possibly that the BIMM is computed from the frame column and thus cover the motion areas pertaining to overall frames inside one time window instead of a single static frame (Figure 9). The frames in the third row show actions with a static background, in which the moving objects are also well segmented. These visualized frames

TABLE 5: Comparison with existing on HMDB51 and UCF101.

HMDB51		UCF101	
Methods	Accuracy (%)	Methods	Accuracy (%)
IDT [12]	57.2	IDT [12]	85.9
ST [25]	58.9	C3D [11]	85.2
STMS [26]	58.2	Two-stream CNN [27]	88.0
Two-stream CNN [27]	59.4	TDD [30]	90.3
TDD [30]	63.2	MSR-CNN [10]	93.7
MSR-CNN [10]	66.9	T-CNN [32]	87.5
STT [33]	65.2	STT [33]	92.2
Three-stream CNN [29]	68.5	Three-stream CNN [29]	93.4
LTC [35]	64.8	LTC [35]	91.7
TSN (two modalities) [34]	68.5	TSN (two modalities) [34]	94.0
DANet-50 [39]	54.3	DANet-50 [39]	86.7
STA-CNN [42]	71.5	STA-CNN [42]	93.8
SAM [41]	69.1	SAM [41]	94.3
STAN [40]	—	STAN [40]	93.6
ATEN (two modalities)	70.6	ATEN (two modalities)	94.5

FIGURE 9: Visualization of motion segmenting approaches on the basis of boundary prior for BIMM on two datasets. (a) Biking, (b) Diving, (c) Lunges, (d) IceDancing, (e) JavelinThrow, (f) Hug, (g) Ride horse, (h) Run, (i) Jump, (j) ParallelBars, (k) Golf, (l) ApplyLipstick, (m) Archery, (n) PullUps, and (o) Haircut.

demonstrate that the motion segmenting approaches on the basis of boundary prior perform well in most cases.

To analyse the effects of the model of attention and to visualize the key snippets from long-term temporal

segments proposed in ATEN, the temporal encoding results are visualized from video segments to reveal how the temporal encoding aggregation represents the target actions. Figure 10 illustrates which input frames from the snippets to

FIGURE 10: Illustration of the learned attentions on four actions over temporal dimension. (a) ParallelBars, (b) CleanAndJerk, (c) Basketball, and (d) LongJump.

cause the attention w_{att} in the temporal domain. The proposed ATEN prioritizes discriminative video frames (the most critical stages) by assigning higher attention values to the specific actions. Naturally, less informative frames are assigned with lower attention weights as expected. As shown in Figure 10(d), frames within the athlete jumping upon the sandpit are assigned higher attention values than that of the athlete running on the playground because the frames that contain both athlete and sandpit are more discriminative for classifying the action “LongJump.”

5.8. Evaluation of Computational Complexity. For evaluating the computational complexity of the proposed ATEN, we compute the average time consumption per video for our action recognition framework, including preprocessing, CNN-based feature extraction, and feature aggregation on the HMDB51 dataset. The experiment compared the proposed ATEN with TSN in Keras with 16G RAM and GTX1080ti, and the codes do not receive the optimization. The time consumption for the above framework is compared in Table 6. The preprocessing of ATEN is somewhat more computationally expensive than TSN. It is because the BIMM is used to enhance the moving objects from complicated background in ATEN. However, the time

TABLE 6: Comparison of the time consumption on average of each video.

Framework	Preprocessing	CNN-based feature extraction	Feature aggregation	Total
TSN [34]	78.6	13.3	1.2	93.1
ATEN	103.2	9.1	1.1	113.4

consumption for CNN-based feature extraction is much smaller than TSN, and we conjecture that the background in RGB-BIMM and OF-BIMM is relatively simple compared with the original frame, so the feature extraction computation can more significantly save time. Furthermore, the time consumption is not determined by model training and optical flow since the process of the mentioned models is consistent.

6. Conclusion

Here, the attention-based temporal encoding network (ATEN) based on attention is presented with background-independent motion mask (BIMM) for human action recognition which effectively reduces the side effect from both spatial and

temporal redundant information and noises. Various experimental processes on 2 thought-provoking datasets verified the superior performance exhibited by the action recognition framework here. It is largely ascribed to the temporal encoding for the sampled snippets, as well as the automatically foreground segmentation method designed here. The former can effectively and efficiently deal with the long-term dependency among frames, whereas the latter makes the convolution focus more on the action-related areas. In the future, we intend to explore the combination of ATEN with some recurrent models such as RNN or LSTM to further exploit the temporal structure in videos to recognize actions of humans.

Data Availability

The experimental data used to support the findings of this study were supplied by Zhengkui Weng under license and so cannot be made freely available. Requests for access to these data should be made to the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the support from the Zhejiang Provincial Natural Science Foundation of China under grant nos.LGF21F020007 and LGF20F020003 and the Foundation of Department of Education of Zhejiang Province under grant no.Y202043874. In addition, Zhengkui Weng especially wishes to thank Dr. Changqi Zhang for assistance with the experiments and to Mr. Zhetao Zhang and Mrs. Xunyin Hu for the contribution to improve the manuscript.

References

- [1] A. K. Wang, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of the CVPR*, pp. 3169–3176, Colorado Springs, CO, USA, June 2011.
- [2] J. M. Chaquet, E. J. Carmona, and A. Fern. available in, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.
- [3] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level CNN: saliency-aware 3-d CNN with LSTM for video action recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 510–514, 2017.
- [4] L. Sun, K. Jia, D. Y. Yeung, and E. S. Bertram, "Human action recognition using factorized spatio-temporal convolution networks," in *Proceedings of the ICCV*, pp. 4597–4605, Santiago, Chile, December 2015.
- [5] C. Lea, A. Reiter, R. Vidal, and G. D. Hager, "Gmental Spatiotemporal CNNs for fine-grained action segmentation," in *Proceedings of the ECCV*, pp. 36–52, Amsterdam, The Netherlands, October 2016.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolution neural networks," in *Proceedings of the NIPS*, pp. 1097–1105, Lake Tahoe, NV, USA, December 2012.
- [7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the CVPR*, pp. 1–8, Salt Lake City, UT, USA, June 2018.
- [8] Y. Yun and H. Wang, "Motion keypoint trajectory and covariance descriptor for human action recognition," *Vision Computing*, vol. 34, no. 3, pp. 391–403, 2018.
- [9] Z. Tu, W. Xie, Q. Qin et al., "Multi-stream CNN: learning representations based on human-related regions for action recognition," *Pattern Recognition*, vol. 79, no. 2, pp. 32–43, 2018.
- [10] Z. Tu, Y. Li, J. Cao, and B. Li, "MSR-CNN: applying motion salient region based descriptors for action recognition," in *Proceedings of the ICPR*, pp. 3524–3529, Cancun, Mexico, December 2016.
- [11] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.
- [12] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the ICCV*, pp. 3551–3558, Zurich, Switzerland, September 2014.
- [13] J. Donahue, L. Anne Hendricks, S. Guadarrama et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the CVPR*, pp. 2625–2634, Boston, MA, USA, June 2015.
- [14] G. Cheron, I. Laptev, and C. Schmid, "P-CNN: pose-based CNN features for action recognition," in *Proceedings of the ICCV*, pp. 3218–3226, Santiago, Chile, December 2015.
- [15] I. Mironică, I. C. Duță, B. Ionescu, and N. Sebe, "A modified vector of locally aggregated descriptors approach for fast video classification," *Multimedia Tools & Applications*, vol. 75, no. 15, pp. 9045–9072, 2016.
- [16] P. Wang, Y. Cao, C. Shen, L. Liu, and H. Shen, "Temporal pyramid pooling based convolution neural network for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 1, no. 1, p. 99, 2017.
- [17] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: deep networks for video classification," in *Proceedings of the CVPR*, pp. 4694–4702, Boston, MA, USA, June 2015.
- [18] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.
- [19] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 20–33, 2018.
- [20] K. Daniilidis, P. Maragos, and N. Paragios, "Improving the fisher kernel for large-scale image classification," in *Proceedings of the ECCV*, pp. 143–156, Crete, Greece, September 2010.
- [21] Y. Zhang, R. Jin, and Z. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1–4, pp. 43–52, 2010.
- [22] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.
- [23] E. Vig, M. Dorr, and D. Cox, "Space-variant descriptor sampling for action recognition based on saliency and eye movements," in *Proceedings of the ECCV*, pp. 84–97, Firenze, Italy, October 2012.

- [24] Y. Yi and Y. Lin, "Human action recognition with salient trajectories," *Signal Processing*, vol. 93, no. 11, pp. 2932–2941, 2013.
- [25] Z. Xu, R. Hu, J. Chen et al., "Action recognition by saliency-based dense sampling," *Neurocomputing*, vol. 236, no. 2, pp. 82–92, 2017.
- [26] Z. Weng and Y. Guan, "Action recognition using length-variable edge trajectory and spatio-temporal motion skeleton descriptor," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, p. 8, 2018.
- [27] K. Simonyan and A. Zisserman, "Two-Stream convolution networks for action recognition in videos," in *Proceedings of the NIPS*, pp. 568–576, Montreal, Canada, December 2014.
- [28] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the ICCV*, pp. 4489–4497, Santiago, Chile, December 2015.
- [29] L. Wang, L. Ge, R. Li, and Y. Fang, "Three-stream CNNs for action recognition," *Pattern Recognition Letters*, vol. 92pp. 33–40, C, 2017.
- [30] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolution descriptors," in *Proceedings of the CVPR*, pp. 4305–4314, Boston, MA, USA, June 2015.
- [31] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional twostream network fusion for video action recognition," in *Proceedings of the CVPR*, pp. 1933–1941, Las Vegas, NV, USA, June 2016.
- [32] H. Rui, C. Chen, and M. Shah, "Tube convolutional neural network (T-CNN) for action detection in videos," in *Proceedings of the ICCV*, pp. 5823–5832, Venice, Italy, October 2017.
- [33] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.
- [34] L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment networks: towards good practices for deep action recognition," *Computer Vision—ECCV 2016*, pp. 20–36, 2016.
- [35] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.
- [36] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Hierarchical multi-scale attention networks for action recognition," *Signal Processing: Image Communication*, vol. 61, no. 2017, pp. 73–84, 2018.
- [37] K. Xu, J. Ba, R. Kiros et al., "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the International Conference on Machine Learning*, pp. 2048–2057, Lille, France, July 2015.
- [38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, <https://arxiv.org/abs/1409.0473>.
- [39] X. Li, M. Xie, Y. Zhang, G. Ding, and W. Tong, "Dual attention convolutional network for action recognition," *IET Image Processing*, vol. 14, no. 6, pp. 1059–1065, 2020.
- [40] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 416–428, 2019.
- [41] S. Liu, X. Ma, H. Wu, and Y. Li, "An end to end framework with adaptive spatio-temporal attention module for human action recognition," *IEEE Access*, vol. 8, pp. 47220–47231, 2020.
- [42] H. Yang, C. Yuan, L. Zhang et al., "STA-CNN: convolutional spatial-temporal attention learning for action recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 5783–5793, 2020.
- [43] M. Versaci, S. Calcagno, and F. C. Morabito, "Image contrast enhancement by distances among points in fuzzy hyper-cubes," in *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, pp. 494–505, Valletta, Malta, September 2015.
- [44] M. Versaci, F. C. Morabito, and G. Angiulli, "Adaptive image contrast enhancement by computing distances into a 4-dimensional fuzzy unit hypercube," *IEEE Access*, vol. 5, pp. 26922–26931, 2017.
- [45] M. Versaci, S. Calcagno, and F. C. Morabito, "Fuzzy geometrical approach based on unit hyper-cubes for image contrast enhancement," in *Proceedings of the IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 488–493, Malaysia, Kuala Lumpur, October 2015.
- [46] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-PAMI-8, no. 6, pp. 679–698, 1986.
- [47] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L1 optical flow," in *Proceedings of the Symposium on Pattern Recognition*, pp. 214–223, Heidelberg, Germany, September 2007.
- [48] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.
- [49] H. Kuehne, "HMDB: a large video database for human motion recognition," in *Proceedings of the ICCV*, pp. 2556–2563, Barcelona, Spain, November 2011.
- [50] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: a data set of 101 human actions classes from videos in the wild," Technical report CRCV-TR-12-01, UCF Center for Research in Computer Vision, Orlando, FL, USA, 2012.
- [51] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the ICML*, pp. 448–456, Lille, France, July 2015.
- [52] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the CVPR*, pp. 248–255, Miami, FL, USA, June 2009.