IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 22 February 2024, accepted 10 April 2024, date of publication 16 April 2024, date of current version 23 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3390004

RESEARCH ARTICLE

Machine Interpretation of Ballet Dance: Alternating Wavelet Spatial and Channel Attention Based Learning Model

P. V. V. KISHORE[1], (Senior Member, IEEE), D. ANIL KUMAR[2], (Member, IEEE), P. PRAVEEN KUMAR[3], D. SRIHARI[4], N. SASIKALA[5], AND L. DIVYASREE[6]

[1]Department of Electronics and Communication Engineering, Biomechanics and Vision Computing Research Center, Koneru Lakshmaiah Education Foundation (Deemed to be University), Guntur 522502, India
[2]Department of Electronics and Communication Engineering, PACE Institute of Technology and Sciences, Ongole 523272, India
[3]Department of AI and DS, Koneru Lakshmaiah Education Foundation (Deemed to be University), Guntur 522502, India
[4]Department of Electronics and Communication Engineering, Sri Venkateswara College of Engineering and Technology, Chittoor 517127, India
[5]Department of Electronics and Communication Engineering, Kamala Institute of Technology and Science, Warangal 506009, India
[6]Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Guntur 522502, India

Corresponding author: P. V. V. Kishore (pvvkishore@kluniversity.in)

This work was supported by the Department of Science and Technology, Science and Engineering Research Board (SERB) Division, Start-up Research Grant (SRG) Scheme, India, under Grant SERB-SRG/2021/001637.

ABSTRACT 'Ballet' is a 15^{th}- century concert performing dance form that originated in Italy. Current AI models for ballet dance pose identification in live performance videos is challenging due to variational pixel distribution of human actions across backgrounds. Notably, their performance on online video datasets improved with both channel (CA) and spatial attention (SA) models but tend to generate over-smoothed Convolutional features due to feature averaging in the attention network. Alternatively, wavelet attention preserves both high and low frequency components in the features which improves the test accuracy. Applying CA and SA on wavelet features simultaneously resulted in hyper-refined features due to double averaging. To overcome this drawback, Alternating Wavelet Channel and Spatial Attention (AWCSA) across any learning network as backbone architecture is proposed. The global features across the residual connections in the backbone (ResNet50) are amplified exclusively with low and high-frequency local features across the channel and spatial dimensions alternatively one after the other. The Ballet online dance video dataset (BOVD23) evaluates the performance of the proposed AWCSA along with baseline action datasets. The end-to-end trained AWCSA has recorded a 6-8% higher performance metrics on BOVD23 dataset over the counterparts.

INDEX TERMS Ballet classical dance, deep feature fusion, multi-head attention, wavelet channel and spatial attention.

I. INTRODUCTION

Ballet is a physically intense type of performing art form developed in Italy in the early 15^{th} century. Later it was practiced and promoted extensively by France and Russia. Ballet demands the highest possible bodily strength, endurance and flexibility that compliments beauty. The dance moves in ballet are complex human actions and reproducing them by untrained learners is an impossible and dangerous task. This dance form is one of the most difficult performing arts which involves jumps, rotations, spins hunches, bending and aerobatic maneuvers. Professional ballet dancing requires precision and skill practiced over many years. To help learners practice professional ballet

The associate editor coordinating the review of this manuscript and approving it for publication was You Yang.

efficiently and effectively, this work proposes to initiate the development of a rapid feedback mechanism. This mechanism is a software solution that provides on-the-spot feedback on the performance of the ballet learner in real time.

However, the objective of this work is not to develop a complete tool, but rather initiate the first process by building a computer vision-based ballet dance classifier. The unavailability of online benchmark datasets has triggered to create one from ballet dance pose sequences available online. Consequently, this work establishes a 10-class 10 subject online ballet dance video dataset for recognition. The frames in most of the labels are blurry due to faster subject movements during a ballet performance. As a result, the automated feature representation layers in deep networks lose key information in the end layers causing poor training and testing of the dataset. To overcome this loss of key feature representation in the depth layers of CNN, attention is proposed as a solution that has indeed improved accuracy across multiple types of image and video datasets.

The goal of automated ballet dance pose recognition (BDPR) is to help performing art lovers get a deeper experience. As a result, the primary choice for implementing BDPR is Convolutional Neural Networks(CNN) [1]. Specifically, the visual attention models further divide into channel and spatial domains [2]. Channel attention (CA) computes the weighted average across all the filter channels and outputs a reduced dimensionality feature vector. This reduced channel attention feature fuses with backbone network generated image features in specific layers thereby producing dominating features from regular networks. Global average pooling (GAP) [3] is the most widely applied channel scalar given in SENet [4]. However, in the case of object recognition tasks, the averaging features across channels precipitate spatial information and this loss affects the overall outcome of the classifier. In order to prevent the spatial loss, global maximum pooling across channels is used in the convolutional block attention module (CBAM) [5] and global standard deviation pooling in the style-based recalibration module (SRM) [6]. Though simple, the results of CA are deprived of the necessary channel information that can only be extracted with proper weighing function between channels.

In contrast, the spatial domain attention (SA) computes weighted pooling on the image features in the form of maximum or average pooling [7]. This preserves the spatial information in the image thereby reducing the resolution of the image after the attention layer. Further, the SA compresses the image and retains the spatial relationships that help in reducing complexity and boosting recognition accuracy. However, the operating window size in maximum and average pooling affects the ability of SA in preserving important features. To further enhance the convolution feature capabilities in SA and CA, a multi scale feature fusion attention was proposed with coordinate attention (CA) mechanism [8] on a light weight bidirectional feature pyramid network.

The output of the attention layers is fused with that of convolutional layers spatially. Frequency domain compressive fusion is also practiced by transforming the channel scalar representation as in FcaNet [9]. This transformation was achieved using discrete cosine transform(DCT). Meanwhile, the DCT compresses the scalars as well as preserves the information encoded in the features. Though the methods using DCT as an attention model produce good accuracy, they have to accommodate information loss in the form of quantization during reconstruction. The other transformation that has the ability to nullify the information loss in DCT is the discrete wavelet transform (DWT). The conventional advantage of DWT lies in its ability to generate contextual image features in orthogonal space [10]. Specifically, 3D DWT was applied to extract relational features from video data [11].

In recent times wavelet-based attention has gained importance due to its ability to represent contextual features that have been fused with any backbone CNN features to maximize accuracy [12], [13], [14]. Specifically, the dual wavelet attention networks (DWAN) [15] has further increased the accuracy on complex datasets. The DWAN is a mixture of both channel and spatial attention. The DWT channel attention specifically compresses the features, and a unique channel scalar is provided as the weight for each channel. Consequently, DWT spatial attention provides structural components of the objects in the image. The DWT channel and spatial networks are joined in sequence [15]. They were primarily applied to the features obtained in the deep layers such as just before the dense or in the dense layers. Subsequently, this paper proposes Alternative Wavelet Channel and Spatial Attention (AWCSA) by following the work in [15]. The AWCSA applies wavelet channel and spatial attention modules across the backbone features over multiple resolutions. This will preserve contextual information across the video sequence for maximizing recognition. Conventionally, the subjects in the ballet dance videos move rapidly during a performance which induces scale changes in the required pose information for recognition. feature fusion at multiple resolutions retains information during scale changes. The proposed method experiments on our Ballet Online Dance Video dataset (BODV23) and benchmark person re-identification action datasets such as NTU RGB D [16], Kinetics-700 [17] and MPII Human Pose [18].

A more technical reason for selecting 'Ballet' dance recognition is to validate two challenges encountered in human action recognition using video data [19]. The 1^{st} challenge was to establish the fact that human motion is nonuniform across video frames. Now creating a full motion constrain human action dataset showing the same action at different speeds was found to be challenging. Alternatively, searching for actions with uneven distribution of human motion across a singular class, we discovered

Ballet. This is the reason why 'Ballet' was selected. The 2^{nd} challenge would be to construct a model that can effectively characterize these unsymmetrical motion features from online ballet dance videos.

On the whole, this work offers the following contributions:

1) Constructing and benchmarking a Ballet Online Dance Video dataset (BODV23) that has demonstrated effectiveness for training the proposed approach across multiple subjects, music, and background changes.
2) The proposed alternating wavelet spatial channel attention mechanism has enhanced backbone classification network's ability to capture spatial variations in complex human actions such as Ballet.
3) Established significance and substantiated the importance of learned features through alternating wavelet spatial channel attention on benchmark datasets through comparison with existing state - of - the - arts.

The rest of the manuscript is organized into 4 sections. The second section outlines the past research with strong and weak areas for further investigation. The methodology to examine the proposed hypotheses that led to the formulation of the above contributions is discussed in section III. The experiments conducted and results obtained were analyzed in section IV. Finally, the overall impact of the proposed work on the selected research problem is presented in section V.

II. LITERATURE REVIEW

The literature reviews show the past and current trends in research on the recognition of dance forms. The goal of this part of the work is to generate insights into the methods which in turn provide merits and demerits. Finally, we summarize these methods based on the tolerance to recognition accuracies. This section highlights four key components required for a BDP identification problem, dance data, feature engineering, model building and performance evaluation.

The dance video data selected in most of the previous works is quite skeptical. The video data used for experiments has been generated in controlled laboratory conditions where the dancer has no costume, and the background is constant across the entire video sequence. Traditional methods used feature representations using computer vision algorithms such as histogram of oriented optical flow [20] and histogram of gradients [21]. These features across each frame are collected temporally in the video sequence to generate a spatiotemporal representation of the dance poses. Subsequently, the Spatio-temporal features were classified with a multivariate support vector machine (SVM) or an Adaboost classifier [22]. Few research methods used multi-modal data such as depth [23] and skeletal datasets [24] recorded with a Kinect sensor. Even though the results reported on these multi-modal data such as depth and skeleton are encouraging, they critically underestimate the finger joint shapes, costumes, lighting, and viewing angles. The recognition accuracies of SVM with HoG features have excelled over others. The conclusion drawn is fairly inconsistent due to the datasets used for training. As discussed, most of the dance datasets were constructed by the researchers and are not publicly available to others. This is the first difference between the previous works and the work presented in this paper. This work proposes the online ballet dance dataset created largely from YouTube videos. These videos are in raw format with many anomalies such as costume variations, lighting inconsistency, camera source movements, occlusions, and video background changes. Under the given challenges in the BDP video dataset, the task is to discover the best machine-learning model for classification.

In order to come up with the best possible architecture for the BDP identification problem, we performed an independent assessment of the previous works. With the expansion of GPUs and deep learning architectures, the performance of these training algorithms also improved considerably. The first models to apply dance classification were developed using pre-trained neural networks on skeletal datasets [25]. Though there has been an improvement in test accuracy, they have some serious limitations. They fail to enumerate the actual physical characteristics of the dance like the costume, hand gestures, pose invariance, and missing joint movements due to view variations. These limitations were addressed by using RGB images and Kinect depth data on a convolutional neural network (CNN). Though the dataset is small, results obtained from the feature fusion of RGB and depth are ordinary [26]. This ordinary improvement is enhanced by 3D point clouds using the recurrence condition of neural networks [27]. Improvements were proposed by applying image pre-processing of dance video frames and then extracting features such as motion information [28], dancer parts [29], body shapes [1] and global automated features using conventional layers [29], [30]. All the above models used either CNN dense layers for classification or represented frame-level features as time series information using recurrent neural networks.

As the above networks are liable to the scarce input video data variations, recognition accuracies of dance lyrics have been improved through the use of multiple types of deep neural networks. The first model used hybrid particle swarm and grey wolf algorithms as the optimizers [31] during the training process instead of regular stochastic gradient descent (SGD) or Adam. Apart from influencing optimizers, the work in [32] applied reinforcement learning to impact dense layer outputs. This has improved the recognition accuracy due to the influence of pre-trained convolutional layers. Instead of focusing on automated feature classification with a dense layer of CNN, feature engineering has been initiated with automated CNN or CNN-RNN variations [33] which are then classified with machine learning methods such as k-nearest neighborhood, Bayes, fuzzy and SVM [34]. A slight enhancement in accuracy has been achieved by using a capsule network for training and testing on dance image data [35]. Capsule networks are part-based training algorithms that translate encoded features for recognition using hierarchical relationships between data samples. Though the results on the non-noisy dance video datasets have been satisfactory they could not be transferred to online or real-time BDP identification. The latest model [36] shows the semantics of dance pose as the underlying features that have enhanced the performance of backbone networks such as VGG and ResNet.

Finally, the proposed model is designed to overcome the most challenging problems from the previous works. Firstly, it overcomes the problem of highly structured dance datasets by creating a more robust BDP dataset from recorded live performances. Secondly, the drawbacks associated with 2D dance data will be diminished by applying an alternating wavelet channel and spatial attention (AWCSA) learning framework on a sequence of BDP frames. Lastly, the attention score is improved by applying channel and spatial attention modules alternatively across layers which is otherwise applied on the dense layer features.

III. METHODOLOGY

This article proposes an alternating wavelet channel and spatial attention (AWCSA) deep neural network with a ResNet50 backbone. The work on dual wavelet attention in [15] has been an inspiration for the proposed methodology. Wavelet channel attention (WCA) and wavelet spatial attention (WSA) are models incepted from the work in [15]. The WCA and WSA are attention models operating on wavelet coefficients. In WCA the approximate and detailed coefficients are averaged and pooled to construct features that are focused on a particular object of interest in the image. The channel-wise global averaging across the wavelet coefficients results in a numerical representation for each channel, which are learned to model attention. In contrast WSA operates averaging is performed on detailed wavelet coefficients and concentrated with average coefficients. As a result, WSA or WCA will ensure that future features generated from backbone layers are concentrated around the focused regions in the image. The previous works using wavelet attention use all the subbands directly or as a single map of averaged subbands [15], [37], [38], [39]. Moreover, data compression is the dominant advantage propagated by these works on top of attention, which comes as an integral part of the 1 × 1 convolutional network.

Three aspects create a difference between AWCSA and similar previous works.

1) AWCSA uses both types of attention maps one after the other, alternatively.
2) The alternative WCA and WSA ensure a good channel and spatial information selection to improve the overall learning of the classifier.
3) The multi-resolution attention ensures that the dominant structural and textural features in the primary layers of the backbone network sustain till the final layers.

The proposed method is illustrated in figure 1. The ResNet50 backbone model takes a batch of video frames with a resized resolution of 256 × 256 × 3 as input. The 2D discrete wavelet transform with biorthogonal filters maps into average and detailed components respectively. The attention map is constructed with low and high-frequency components to create dominant structural and textural features for recognition. Moreover, the attention maps are generated at multiple resolutions throughout the learning process to attain attention sustainability till the end of the feature generating network as can be seen in figure 1.

A. 2D DISCRETE WAVELET TRANSFORM

Given an image $I(x, y) \in \Im^{M\times N}$, where $\Im$ is the set of integers of size $M \times N$, the coefficients of 2D discrete wavelet transform (2D DWT) are extracted by applying a scale function

$$W_{\varphi}(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M} \sum_{y=0}^{N} I(x, y) \times \varphi_{j_0,m,n}(x, y) \quad (1)$$

where j_0 is the starting scale and $\varphi_{j_0,m,n}$ is the scaling function. Similarly, the detailed coefficients can be obtained by applying the following formulation as

$$W_{\psi}(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M} \sum_{y=0}^{N} f(x, y) \times \psi_{j,m,n}(x, y) \quad (2)$$

where $\psi_{j,m,n}$ is the wavelet function. The orthogonality principle is satisfied by the scale function resulting in $\varphi(x, y) = \varphi(x)\varphi(y)$, transforming eq'n(1) into

$$W_{\varphi}(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M} \sum_{y=0}^{N} I(x, y) \times \varphi_{j_0,m}(x) \times \varphi_{j_0,n}(y) \quad (3)$$

The above equation draws parallels with convolutional operation with scale kernels along x and y directions. In previous works, it has been proved that the HAAR wavelet transform can be formulated by the following relation $\varphi_{j_0,m,n}(x) \cdot \varphi_{j_0,m,n}(y) = 1$ in eq'n(3), which results in

$$\sum_{m,n\in 0,1,2,.....,2^{j-1}} W_{\varphi}(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M} \sum_{y=0}^{N} I(x, y) = \left(\sqrt{MN}\right) \mathrm{GAP}(I(x, y)) \quad (4)$$

The GAP is, Global Average Pooling. Consequently, the $W_{\varphi}(j_0, m, n)$ characterizes the approximate or low frequency components with HAAR wavelet basis. Therefore, it has been proved in [3], that the summation of low frequency components results in a functionality equivalent to GAP. This has been explored in many of the channel attention models using DWT [40], [41]. Since the online BDP video dataset is quite vibrant in pixel variations across frames, the experiments showed that bi-orthogonal wavelets have the ability to represent such transformations. Figure 2 shows the variation between Haar and Bior filters on a frame from BDP.

In general, bi-orthogonal wavelets eliminate the problem of phase distortions caused by the unsymmetrical nature of

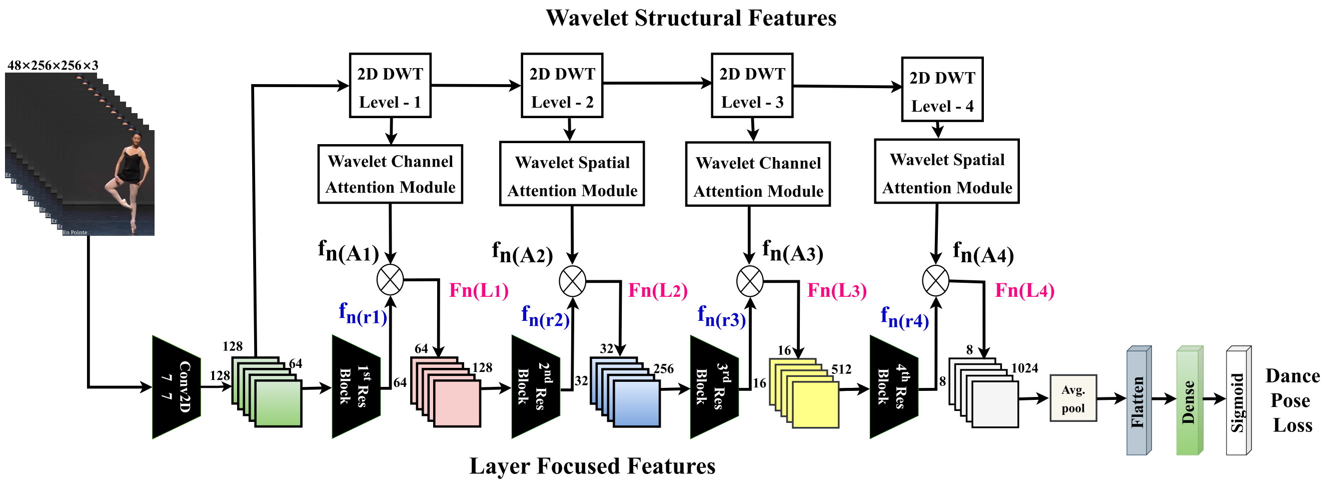


FIGURE 1. The complete end-to-end architecture of alternating wavelet channel and spatial attention (AWCSA) for Ballet dance recognition on online multi-source video data.

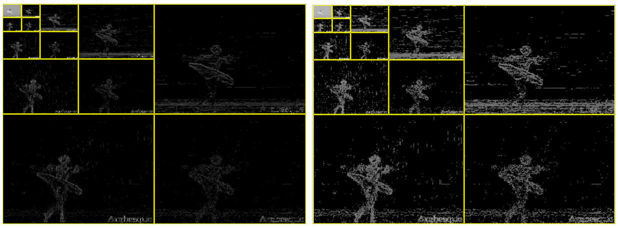

(a) HAAR Wavelet Filter (b) Bior1.1 Wavelet Filter

FIGURE 2. Visual comparison between HAAR and Bior1.1 wavelet filters on a ballet dance pose frame in a class label.

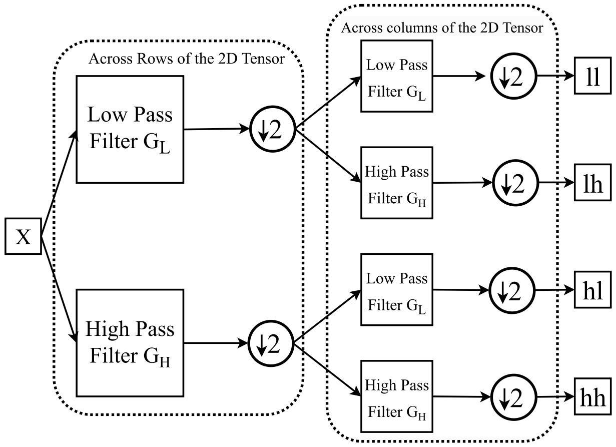


FIGURE 3. 2D Wavelet transform on a tensor.

orthogonal wavelets. The multiresolution analysis is considered the basis for bi-orthogonal wavelets. The approximate and detailed coefficients in bior 2D DWT are given as

$$W_{\psi}\left(j,m,n\right) = \sum_{x=0}^{M}\sum_{y=0}^{N} I\left(x,y\right)\psi_{j,m,n}\left(x,y\right) \tag{5}$$

and

$$W_{\tilde{\psi}}\left(j,m,n\right) = \sum_{x=0}^{M}\sum_{y=0}^{N} I\left(x,y\right)\widetilde{\psi}_{j,m,n}\left(x,y\right) \tag{6}$$

where the wavelets ψ and $\widetilde{\psi}$ are biorthogonal wavelets. This is the most suitable wavelet for representing the detailed components of BDP online video data as they contain vast amounts of non-linearity. In the following subsection, the attention mechanism and the involvement of wavelets is formulated.

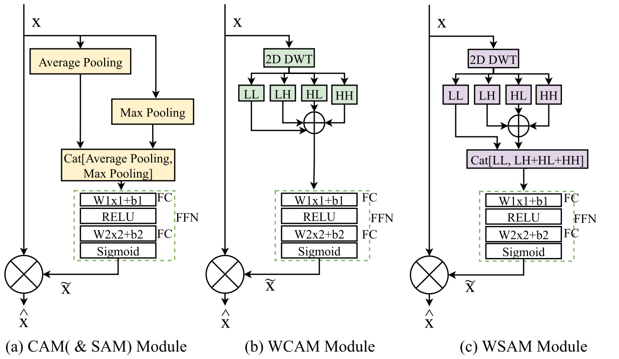
B. WAVELET ATTENTION MODELS (WAM)

The theoretical analysis leads to the formulation of wavelet-based attention mechanisms(WAM). The wavelet decompositions on a 2D tensor has been disclosed in figure 3. The outputs of figure 3 are applied in different combinations across the learning systems to induce attention into the convolutional features. However, to understand the difference between the regular channel or spatial attention models used previously to the wavelet-based channel and spatial attention, figure 4 is reproduced from works [15]. Figure 4(a) describes the process followed in channel-based attention across 2D tensor features $f^{(c)} = \left[f_1^{(c)}\ldots.f_z^{(c)}\ldots\ldots.f_Z^{(c)}\right]^T$ with $z \in 0-to-Z$ channel representations in a particular layer ℓ using global average pooling (GAP) [3] and the spatial attention based on the maximum pooling of features. The GAP is expressed as the average of $\chi_n^{(\ell)} \in R^{n\times n\times Z}$ features across all channels c at layer ℓ with Z learnable convolutional filters is

$$f_n^{(c)} = \frac{1}{\left|\chi_n^{(l)}\right|}\sum_{x\in\chi_n} x \in R^{\frac{n}{2}\times\frac{n}{2}\times Z} \tag{7}$$

where n gives the dimensionality of the feature matrix and is called as channel attention module (CAM) as shown in figure.4(a). Similarly, the spatial attention (SAM) in figure.4(a) is obtained as the maximum pooling across $\chi_n^{(\ell)} \in R^{n\times n\times Z}$ features in layer ℓ as

$$f_n^{(s)} = \arg\max_{x\in\chi_n^{(\ell)}} (x) \in R^{\frac{n}{2}\times\frac{n}{2}\times Z} \tag{8}$$



(a) CAM(& SAM) Module (b) WCAM Module (c) WSAM Module

FIGURE 4. Attention based networks. (a) Channel and Spatial attention mechanisms - CAM or SAM. (b) Wavelet based channel attention (WCAM) and (c) Wavelet Spatial attention mechanism(WSAM).

The features $f_n^{(c)}$ (channel) and $f_n^{(s)}$ (spatial) are being learned at the output of fully connected layers as shown in the figure.4(a). Finally, $f_n^{(c)}$, $f_n^{(s)}$ fuses multiplicatively with the original features to generate attentive features $\left(\hat{x}_c = x \bullet f_n^{(c)}\right)$ $\left(\hat{x}_s = x \bullet f_n^{(s)}\right)$. Interestingly, both channel and spatial $\left(\hat{x}_c = \left(x \bullet f_n^{(c)}\right) \bullet f_n^{(s)}\right)$ attentions have also been applied in series which has shown to improve the overall accuracy of the classifier. However, the average or maximum pooling decreases the intensity of the feature being learned by the fully connected attention layers. Consequently, this has been improved by applying the discrete wavelet transform (DWT) which has proved to compress and preserve the structural information. Simultaneously, these advantages empower the DWT to be used as an efficient attention generating network.

C. WAVELET CHANNEL ATTENTION MODEL (WCAM)

The WCAM takes inspiration from the channel attention module [42] shown in figure.4(a). The 2D input features of image $I(x, y) \in R^2$ or frame $I(x, y, t = T) \in R^2\ \forall\ T\ frames$ represented by x transforms into approximate low frequency coefficients denoted by (LL) and high frequency coefficients (LH, HL, HH).

The 2D DWT results in the coefficients

$$LL, LH, HL, HH = 2D\_DWT(x) \tag{9}$$

The above decomposition is a Level-2 decomposition and as the decomposition levels increases, the number of sub bands also get inflated with $2^{number\_of\_levels}$. For C channels in input feature $x \in R^{C\times M\times N}$, the 1×1 convolutions with C_1 filters will produce a statistical averaging at the output as $R^{C_1\times M\times N}$. The output of the wavelet channel attention module in figure.4(b) formulates into

$$\tilde{x} = \sigma\left(\Theta_{A_2}\left(relu\left(\Theta_{A_1}\left(\sum_{i=0}^{N/2}\sum_{j=0}^{N/2}(LL + LH + HL + HH)\right)\right)\right)\right) \tag{10}$$

where, $\left\{\Theta_{A_1}, \Theta_{A_2}\right\}$ are trainable parameters of the attention network on input features $\{LL, HL, HH, LH\}$ and σ is of the sigmoid function. The x input features are transformed into wavelet domain as low and high frequency components. Subsequently, in channel attention network in figure.4(b), they are averaged and channel scalars across each of the channels are learned by the combination of two fully connected and two activation layers. The output of the channel attention network for an input feature x is formulated as

$$\hat{x}_c = \arg\min_{\Theta_c} L_c\left(\Theta_c : w_x(x)\right) \bullet x \tag{11}$$

where the operator $(\bullet)$ indicates an element wise multiplication. The $w_x(x) = \sum_{i=0}^{N/2}\sum_{j=0}^{N/2}(LL + LH + HL + HH)$ wavelet features are trained with channel model parameters Θ_c using a loss function L_c. The output features $\hat{x}_c$ of the learned channel attention network are element wise multiplied with the original features x.

D. WAVELET SPATIAL ATTENTION MODEL (WSAM)

The spatial attention model using wavelets is shown in figure.4(c), which is compared with WCAM and CAM(& SAM) if figure's 4(a) and 4(b) respectively. The spatial wavelet features are

$$w_x(x) = \underset{0}{\overset{\infty}{\|}}\left(LL, \sum_{(M/2,N/2)}(LH, HL, HH)\right) \tag{12}$$

where $\underset{0}{\overset{\infty}{\|}}$ is a concatenation operator for all the elements in the wavelet low frequency (LL) and high frequency (LH,HL,HH) feature representations. The obtained spatial wavelet features are learned by the fully connected and activation layers. The output of the WSAM in figure.4(c) is formulated as

$$\tilde{x} = \sigma\left(\Theta_{A_2}\left(relu\left(\Theta_{A_1}\left(w_x(x)\right)\right)\right)\right) \tag{13}$$

The output of the spatial attention network for an input feature x is given as

$$\hat{x}_s = \arg\min_{\Theta_s} L_s\left(\Theta_s : w_x(x)\right) \bullet x \tag{14}$$

where the operator $(\bullet)$ indicates an element wise multiplication. The $w_x(x)$ spatial wavelet features are trained with channel model parameters Θ_s using a loss function L_s. The output features $\hat{x}_s$ of the learned attention network are element wise multiplied with the original features x. However, concatenating the low frequency wavelet coefficients has equivalence to GAP in figure.4(a). This in turn influences the attention values produced by the network.

E. ALTERNATING WAVELET CHANNEL AND SPATIAL ATTENTION MODEL (AWCSA)

Figure.5 shows previously used multi attention feature integration modules along with the proposed ones. The first row of figure.5 describes the models from the past works. The models in figure's 5(a) and (b) integrate either wavelet channel or spatial features in between the Resnet50 blocks. The WCA_R50 and WSA_R50 can either have attention layers across one ResNet50 block or at multiple blocks.

Both these models have shown good attention capabilities with respect to standard datasets such as CIFAR100, SVHN and WHURS-19. Further improvements were observed with the use of dual wavelet attention models [15] in figure's 5(c) and (d), where sequential channel and spatial attention layers were used for integration. Though WCSA_R50 and WSCA_R50 recorded highest accuracy on the above benchmark datasets, they failed to formulate structural and textural information for good recognition on ballet dance and human action online video datasets. Moreover, if all the blocks in the ResNet50 were inflated with attention layers, the training process of dual wavelet attention model became extremely complicated. To avoid the above shortcomings, we propose alternating wavelet channel and spatial attention model (AWCSA). The multiple block integration of attention layers is shown in figure's 5(e) and (f). This has been elevated by alternating the channel and spatial attention layers across the ResNet50 blocks. The proposed models AWCSA_R50 and AWSCA_R50 are shown in figure's 5(g) and (h) respectively.

The obtained attention maps for individual frames in a class label are fused with the mainstream Resnet50 features for classification during the training operation. Above, the classifier in figure.1 is shown to be ResNet50. However, any classifier can be used, and it would be interesting to find the usefulness of attention maps as a generalized attention provider. Moreover, the capabilities of ResNet50 as a feature extractor can also be challenged by using other standard networks. The attention maps generated from each of the layers in the global feature extractor ResNet are represented as

$$f_{na} = \left\{f_{n(A_1)}, f_{n(A_2)}, f_{n(A_3)}, f_{n(A_4)}\right\} \in R^{C_2\times M_l\times N_l}\forall n \subset [1, N] \tag{15}$$

where, the variable a gives the attention at the output of the Residual layer and n denotes the frame number. Here $\{A_1, A_3\}$ are channel attention modules described above and $\{A_2, A_4\}$ are spatial and vice versa as shown in figure's 5(g) and (h). The attention maps at the output of Residual layers have the same dimensions as that of the features in the classifier net denoted by $M_l \times N_l$ where l is the layer number. The dimensions will be reduced to half with each passing layer. The training of the AWCSA_R50 or AWSCA_R50 results in attention features $\left\{F_{n(L1)}, F_{n(L2)}, F_{n(L3)}, F_{n(L4)}\right\}$ across each of the residual layers in the ResNet backbone as

$$F_{n(Ll)} = \prod_{i=1}^{class} f_{n(Al)} .f_{n(rl)}^{class}\ \forall R^{C_2\times M_l\times N_l},$$
$$l \rightarrow number\ representing\ residual\ layers \tag{16}$$

The AWCSA_R50 or AWSCA_R50 model is trained with categorical cross entropy loss function and Adam optimizer. The classifier model in figure.1 trains on the local features from $F_{n(Ll)}(\tilde{x})$ using the trainable parameters Θ_{AWCSA} (Θ_{AWSCA}) by optimizing the loss function L_{AWCSA} (L_{AWSCA}) on the entire dataset as

$$\Theta_{AWCSA} = \arg \min_{\Theta_{AWCSA}} L_{AWCSA}\left(\Theta_{AWCAS} : F_{n(Ll)}(x), y\right) \tag{17}$$

Here class labels y denotes the lyrics of the song for which the ballet dance poses are recorded. The trainable parameters Θ_{AWCSA} are optimized using the cross-entropy loss L_{AWCSA} defined as

$$L_{AWCSA} = -\sum_{i=1}^{Class} (y_i \times \log(y_i) + (1-y_i) \times \log(1-y_i)) \tag{18}$$

where, *Class* is the total number of labels in the ballet dance song used during the training process. The trained model $M(\Theta_{AWCSA})$ outputs a set of spatial features x_l representing RGB BDP data at the end of each Residual block using the following function

$$x_l = \sum_{i=1}^{C}\sum_{j=1}^{C} I_n(i, j)K((k-i)(k-j)) \quad \forall\, l = 1\ to\ 4 \tag{19}$$

where k is the kernel size across each of the layers and x is the feature matrix in l^{th} Residual block. The final spatial feature at the input of the dense layer is of size $N \times k_{dense}$. The activation function used in convolutional layers is rectified linear unit defined as

$$R(z) = \max(0, z) \tag{20}$$

where z is the output of the neuron. Similarly, the dense layers have tanh and the SoftMax layer has sigmoid activations.

The proposed AWCSA or AWSCA is on end-to-end trainable model. The global BDP features x for all classes are extracted through the backbone ResNet50 architecture as shown in figure.1. There are 96 frames in each video sample. The learning rate for the entire network was fixed at 0.0001. Whenever the error rate of the classifier became constant for more than 10 epochs, the learning rate was decreased by 10%. The weights and biases are initialized randomly through zero mean unit gaussian distribution function. The momentum factor was kept at 0.84. All the models (Proposed and State-of-the-art) were trained with Adam optimizer on an 8GB NVDIA A4000 GPU with 16GB memory using TensorFlow 2.5 APIs.

The focused information losses are minimized in our proposed AWCSA model during the training process. This is due to the alternating channel and spatial attention layers that control the flow of information passing through the network. On the other hand, the number of attention layers is halved when compared to dual attention networks WCSA_R50 or WSCA_R50 models [15]. Moreover, the feature combinations in AWCSA_R50 occur at multiple resolutions with alternating spatial and channel pooling which has produced accurate localizations on complex video datasets. Subsequent sections provide a detailed description of the results obtained through rigorous experimentation on various BDP video datasets to evaluate the performance of the proposed method against similar frameworks.

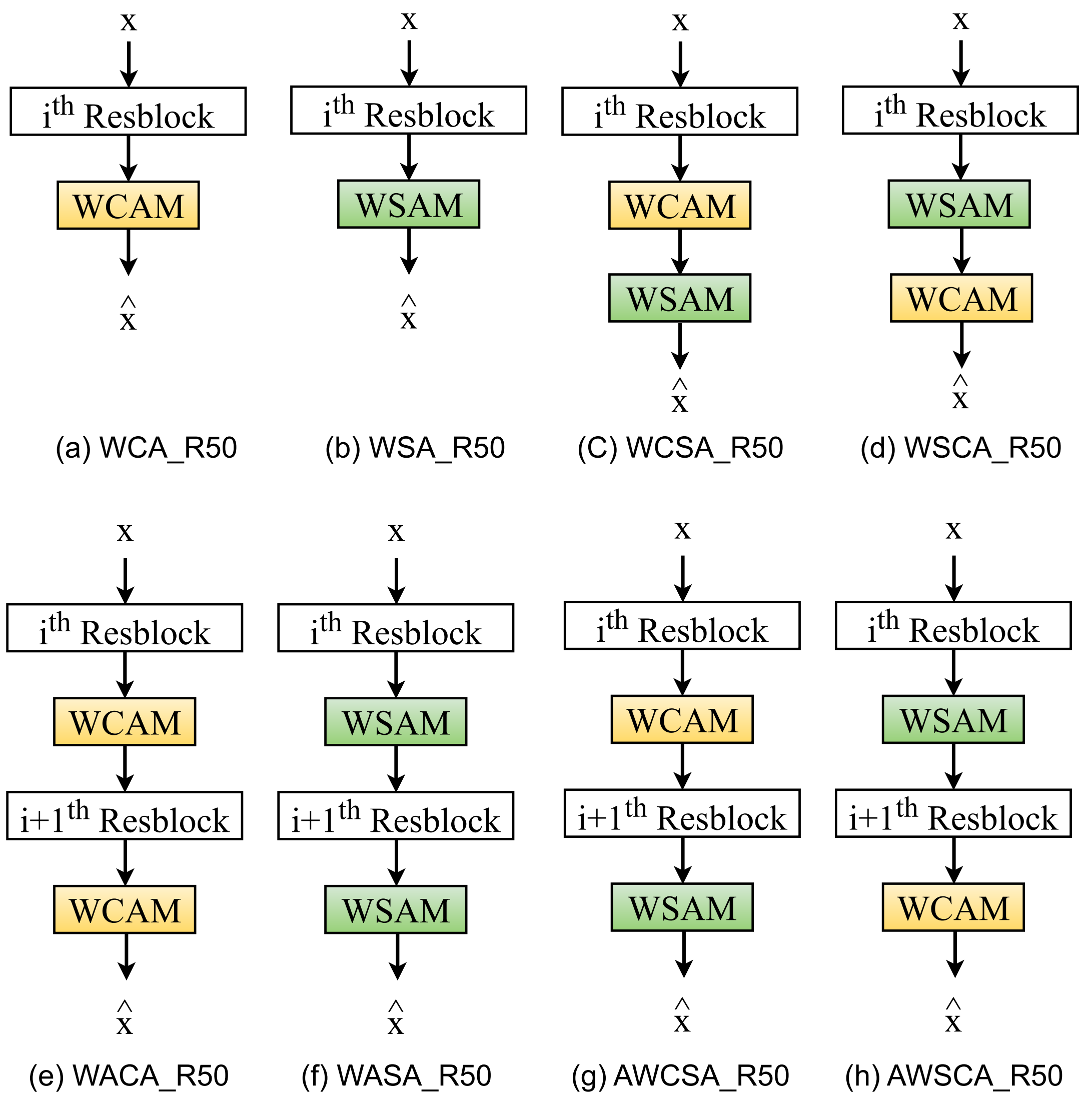


FIGURE 5. Attention feature integration modules into ResNet50 blocks. (a) Wavelet channel attention (WCA_R50) [4], (b) Wavelet Spatial Attention (WSA_R50) [43]. Dual wavelet attention models [15], (c) Wavelet channel spatial attention (WCSA_R50), (d) Wavelet spatial channel attention (WSCA_R50). Proposed Alternating wavelet channel spatial attention (AWSCA), (e) wavelet alternating channel attention (WACA_R50), (f) wavelet alternating spatial attention (WASA_R50),(g) alternating wavelet channel spatial attention (AWCSA_R50),(h) alternating wavelet spatial channel attention (AWSCA_R50).

IV. RESULTS AND DISCUSSION

Two architectures AWCSA_R50 and AWSCA_R50 are built and trained from scratch on online sourced ballet dance poses which are further validated with benchmark human action recognition (HAR) RGB video datasets. The BDP online videos are transformed into frames at specific intervals which are further split into train, validate and test data. The output labels for each of the video sequences are sourced from online ballet dance learning portals. Correspondingly, the attention mechanisms employed in AWCSA_R50(AWSCA_R50) are tested against the models from figure.5 on different backbone architectures to estimate their robustness. Additionally, the findings of AWCSA_R50 were validated against the other state - of - the - art on human action recognition methods.

The following research has broader implications in the fields of human computer interactions, automated training tools for dancers, and application domains of fine-grained human motion recognition. Intuitively, a real time interface can help dance learners with precision feedback on their movements for refining their technique and enhance spectator engagement statistics. Additionally, the automated dance interfaces can help prevent injuries during practice and at times help them fine tune their problem moves by comparing them with the best performances. On the other hand, machine automated the fine-grained human motions can

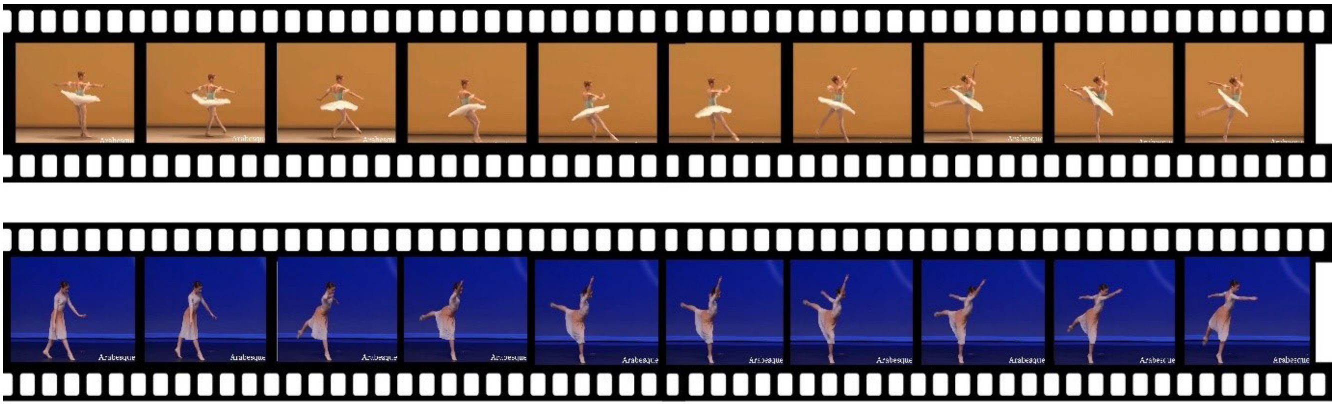


FIGURE 6. Ballet Online Dance Video dataset (BODV23), All the rows describe a lyrical video song named 'Arabesque'. Each column describes different samples from different dancers for the same lyrics or class labels. In some cases, as we can see, the samples are found to be missing in the online source. The dataset is unstructured.

benefit individuals with disabilities for gesture recognition using natural interfaces and aid in developing custom fit rehabilitation programs. Moreover, it can also be applied to improve sports performance as well as to enhance video surveillance for human detection and tracking.

A. BALLET ONLINE DANCE VIDEO DATASET (BODV23)

This work generates Ballet Online Dance Video dataset (BODV23), an online BDP video dataset with 10 classes. A set of 10 popular poses have been sourced from [44] and the corresponding videos of different dancers was downloaded from various online sources [45]. Specifically, YouTube has been the largest source of our BODV23 dataset produced at KL Biomechanics and Vision Computing Research Centre and is available for download https://github.com/pvvkishore/Ballet_Dance_Recognition_2023. Historically, no such BDP dataset is available for training and testing. Hence, to validate the AWCSA_R50 (AWSCA_R50) against the baselines, this work selected benchmark human action datasets such as NTU RGB D [16], Kinetics-700 [17] and MPII Human Pose [18].

Figure.6 shows multiple subjects from BODV23 dataset with class label 'Arabesque'. Each label consists of 10 samples distributed unevenly across number of video frames. Evenness in number of frames per video sample is guaranteed for training and testing on AWCSA_R50 by manually selecting frames of interest. The frames in BODV23 are restricted to 96 frames per / label which covers all the pose related information. The BODV23 embodies 10 classes per subject and each of these 10 classes has 10 samples from multiple sources and dancers. Consequently, there are 10 × 10 BDP videos with 96 frames per sample. Hence, the BODV23 online BDP dataset have $10(classes) \times 10(dancers) \times 96(frames) =$ 96000 frames. All the videos are downloaded from YouTube with output schema resolution of 780p. Based on the original uploaded data through multiple sources, there was difficulty in maintaining the set resolution. Therefore, all the videos are split into frames at a frame rate of 30fps and 96 frames are separated as labeled data. The 10 labelled frames are first standardized by manually cropping each frame to keep the dancing subject at the center of the cropped image. The cropped frames are standardized at 256 × 256 × 3. Annotating the dance poses based on labels to include multiple subjects was really challenging as their body movements varied along with the camera angles for capturing the same dance pose across multiple videos. At this point of time, the BODV23 was carefully created with dance poses that match a particular camera angle within a class label. More annotations such as segmentation masks and bounding boxes are the updates planned for the next version. The current version of BODV23 is a classification dataset with training, validation, and test labels. BODV23 is biased with respect to viewpoints, frame quality, temporals, and data sampling. Mostly only the best viewpoints of a subject within a particular label are selected. The frame quality is most of the videos was upgraded or downgraded to 256 for uniformity. The number of frames per video sequence was standardized to 30fps, even though they resulted in poor quality image frames. However, these were sampled and removed without compromising on the sequential nature of the dance performance. Further, a 9-fold data augmentation is performed during training to avoid overfitting on various backbone networks. Since there are no benchmark ballet dance pose datasets available publicly, this work used video human action datasets to test the performance of the proposed method against the state-of-the-arts. This is because of the closeness dance is a complicated version of human action that consists of large complex structural variation of human body as against small simple structures.

Figures.7, 8, & 9 shows the HAR benchmark datasets from NTU RGB-D [16], Kinetics-700 [17] and MPII Human Pose [18]. In NTU RGB-D, we have 120 action classes with 114,480 video samples. In this work, we restricted the datasets to 864000 frames with 120 action classes. Kinetics-700 is a large scale human action dataset with 10K videos in 700 classes with 10 to 40 samples per class. In this work we restricted to 120 classes with 10 samples per class at a maximum of 942500 frames. MPII human pose is an online YouTube video dataset with 25000 frames covering 410 actions. In this work only 120 classes with 9 data

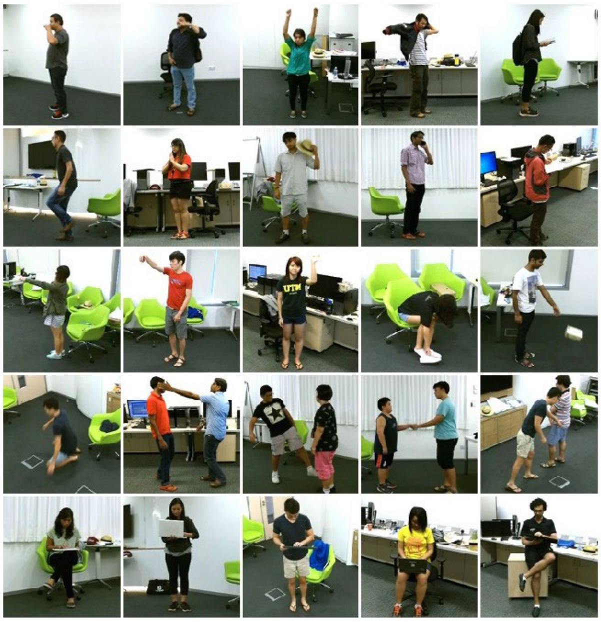


FIGURE 7. The benchmark NTU RGB D dataset used in this work to test the performance of the proposed method.

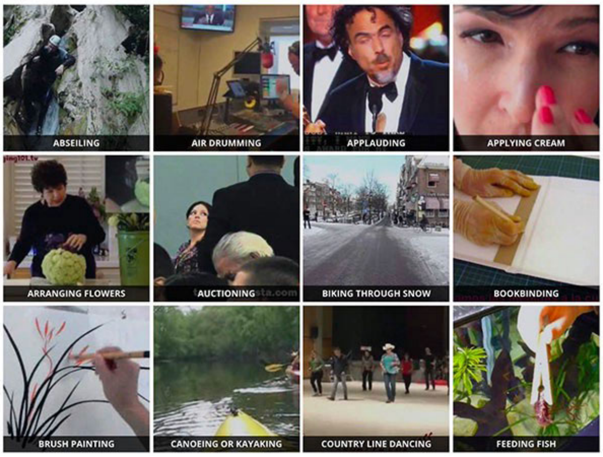


FIGURE 8. Kinetics-700 dataset sample frames.

augmentation were used to generate around 725k frames for training, validation and testing. However, the proposed work has normalized the use of image resolution to 256 × 256 in all the datasets and subsequently across the models used in this work.

B. MODELS AND EVALUATION CRITERIA

Since the architecture and configuration of the ResNet50 backbone network has been extensively discussed in section III, this subsection presents the training parameters across the considered datasets. Firstly, the frame resolution was made constant across datasets to 256 × 256 × 3. Secondly, apart from ResNet50, other standard architectures such as VGG-16, VGG-19, ResNet101 and regular CNN models were tested. In all these models, the weights and bias initializations were fixed using a zero mean and unit variance Gaussian function. The initial learning rate for the has been fixed at 0.00001 throughout training process. The pose loss function for all the baseline networks has been categorical cross entropy with an Adam optimizer. The momentum factor for the network in figure.1 has been selected as 0.85. The combination of filters and other layers were used in accordance with the works cited in the comparison phase. The major difference between the proposed AWCSA and other wavelet attention models is in the distribution of features across all layers with a progressive resolution across layers in the backbone classifier network. The previously proposed wavelet attention models induce CA & SA attention features one after the other at a fixed resolution as against all layers progressively with adaptive resolution.

The proposed AWCSA_R50 (AWSCA_R50) were evaluated with the help of four experiments. The first one evaluates the performance of the proposed method on the BODV23 with mean average precision(mAP) over the entire dataset and reports it in two folds. The 1-fold mAP is the average precision after the 1^{st} positive testing of the model and the 5-fold mAP is the maximum average precision across 5 successful runs of the model on the test data. This experiment also reports the performance of individual classes in the BODV23. Second, the impact of the wavelet attention mechanism proposed in this work is evaluated against the previous models. Thirdly, the proposed work is compared against the state-of-the-art networks on the benchmark datasets and BODV22 respectively. Finally, ablation studies are performed to identify the inflicting parameters while testing the network. Another parameter called Cumulative Matching Characteristics (CMC) is computed with the expectation of finding the correct match for a test sample in top n - matches. This parameter is a measure of efficiency in recognition tasks on unseen test data.

C. EVALUATING THE PROPOSED AWCSA_R50 AND AWSCA_R50 MODELS

The 1^{st} experiment trains the model in figure.1 with attention model in figure.5(g) as AWCSA_R50. The 96K video frames in BODV23 are annotated manually using a predefined bounding box model. Out of the 96K frames distributed across 10 classes and 10 samples, 67.2K will be used for training and 14.4K each for validation and testing. The performance metrics are averaged over the test data. Table 1 records mAP performance metrics computed on the test data using the trained AWCSA_R50 model. Similarly, AWSCA_R50 results were also presented in table 1.

The metrics calculated are 1-fold and 5-fold mAP's. The highest 1-fold is obtained for the class 'Attitude'. Table.1 also shows mAP's obtained without attention model. Undoubtedly the wavelet attention has a strong input on the classifier performance in identifying shrewd pose formation in BDP database. Further, the attention mechanism has indeed highlighted the spatial content in the frame sequences that provided highly effective shape and textural features for classification.

In order to validate the mAP's obtained in table 1, sample specific testing is initiated through inferencing the

FIGURE 9. MPII Human Pose dataset sample frames.

TABLE 1. mAP's of the proposed method trained on BODV23 dataset with 1 and 5-fold testing.

S.No	Class Labels in BPD	without Attention	AWCSA_R50		AWSCA_R50	
			1-Fold	5-Fold	1-Fold	5-Fold
1	Arabesque	0.744	0.894	0.934	0.874	0.902
2	Assemblé	0.751	0.901	0.929	0.882	0.914
3	Attitude	0.802	0.958	0.979	0.942	0.961
4	En Pointe	0.726	0.886	0.907	0.853	0.886
5	Fouetté	0.724	0.842	0.873	0.851	0.869
6	Grand Jeté	0.697	0.841	0.864	0.819	0.842
7	Penché	0.670	0.808	0.836	0.787	0.807
8	Pirouette	0.713	0.867	0.895	0.838	0.891
9	Tour de reins	0.736	0.852	0.902	0.864	0.911
10	Tour en l'air	0.618	0.774	0.792	0.726	0.762

trained AWCSA_R50 model with data from each class singularly. Table 2 presents the mAPs in 1-Fold and 5-Fold recorded during the testing of unseen dance poses from labels 'Arabesque', 'Attitude' and 'Tour en l'air'. It can be seen that the label 'Tour en L'air' is least accurate of all the classes. This is due to the paced movements within the classes which are challenging for the feature extractor. However, in the next section it is proposed to compare the non-attention(ResNet50), attention (CA_R50 and SA_R50) and wavelet attentions in figure.5.

Interestingly, table 1 also shows comparison between AWCSA_R50 and AWSCA_R50. The majority of the mAPs from table 1 points to the finding that channel attention followed by spatial attention preserves more contextual information related to a video frame when compared to spatial followed by channel. This is because of the GAP in the channel attention module. It averages across all the channel features which results in loss of dominant features. In AWCSA_R50, this channel information loss is replenished in the consecutive residual layer having wavelet spatial attention. However, if the channel attention is after the spatial attention, this loss is quantified significantly. To justify this fact that the AWCSA_R50 or AWSCA_R50 generates highly structural and largely discriminating features during the training process, we validated the proposed attentions with similar models in the following section.

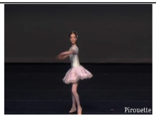
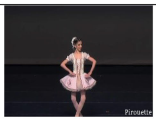
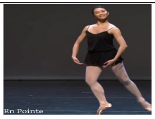
D. AWCSA_R50 VS STATE-OF-THE-ART ATTENTION MECHANISMS

This part of the section accentuates the significance of the proposed alternating wavelet channel and spatial attention (AWCSA) block in the overall BDP recognition. Consequently, experiments were conducted by replacing the proposed attention block in our network with eight state-of-the-art attention models as shown in figure.5. Out of which 2 attention models are based on channel (CA_R50) [4] and spatial attention (SA_R50) [5] and the remaining are based on wavelet attention. There is one model without attention (Res50) [46]. Remaining all are sourced from figure.5. They are wavelet channel attention (WCA_R50) [4], WSA_R50 [43], WCSA_R50 [15], WSCA_R50 [15], along with the proposed WACA_R50, WASA_R50, WACSA_R50 and WASCA_R50.

The results of the above experiment are quantitatively presented through the attention visualizations as shown in figure.10. The attention maps are shown for top 4 poses in the class 'Attitude'. Training and testing of all the attention models was carried out with Resnet 50 backbone feature network with no deviation in hyperparameters. Figure.10 has 10 columns, with the first column representing the original video frame from 4 top mAP class labels. The first column figure.10(a) gives four original frames from the class 'Attitude' pose in ballet dance dataset. The column figure.10(b) is the ResNet50 architecture with no attention mechanism induced into its layers. This map is constructed by plotting the features from the last residual block in ResNet50. Since there is no attention module, the feature distribution is distorted in all the frames. The column 10(c) uses a channel attention(CA_R50) in the final convolution layers [4], which missed key spatial features that were part of the initial layers. Subsequently, the results in column 10(d) has shown improved attentions due to spatial attention (SA_R50) [5]. The biggest drawback of these models is their inability to produce focused features during the training process that preserves the structural integrity across features for pose estimation. This drawback was successfully extenuated by integrating wavelet features into the attention model. Further, figure.10(e) uses a wavelet channel attention model (WCA_R50) in the earlier layers [4] of the global feature extractor. This has enabled the attention module to focus on improved distribution of convolution features across the frames.

This is further increased by applying WSA in WSA_R50 as can be seen in figure.10(f). The wavelet channel attention (WCA) [4] shown in column 10(e) did well in this regard but was consumed by global averaging of approximate and

TABLE 2. mAP metrics obtained during the testing of the proposed model on Individual classes of the song 'Arabesque', 'Attitude' and 'Tour en l'air'.

Assemble dance Poses				Pirouette dance poses				En Pointe dance poses			
Dance Poses		1-Fold	5-Fold	Dance Poses		1-Fold	5-Fold	Dance Poses		1-Fold	5-Fold
	Pose-1	0.892	0.912		Pose-1	0.875	0.899		Pose-1	0.917	0.931
	Pose-2	0.954	0.977		Pose-2	0.891	0.917		Pose-2	0.874	0.895
	Pose-3	0.876	0.907		Pose-3	0.857	0.892		Pose-3	0.937	0.952
	Pose-4	0.841	0.884		Pose-4	0.912	0.942		Pose-4	0.928	0.947
	Pose-5	0.889	0.914		Pose-5	0.928	0.957		Pose-5	0.924	0.951

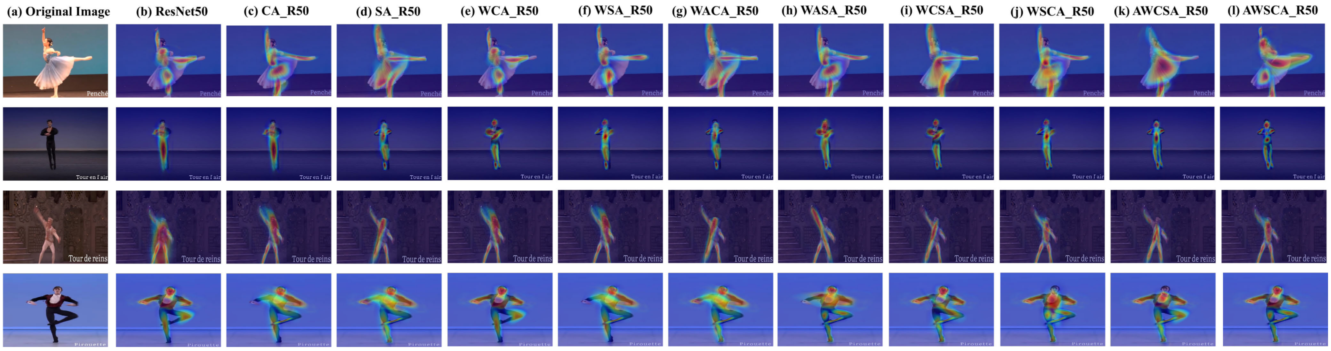


FIGURE 10. Comparison of the proposed attention model with the state-of-the-arts attention layers.(a) The Original poses one from the 'Tour en l'air' class of BDP, (b) ResNet50 with no attention [46], (c) CA_R50 [4], (d) SA_R50 [5], (e) WCA_R50 [4],(f) WSA_R50 [43],(g) WACA_R50 [15],(h) WASA_R50, (i) WCSA_R50, (j) WSCA_R50, (k) AWCSA_R50 and (l) AWSCA_R50.

detailed components. However, the spatial wavelet attention (WSA) [43] in column 10(f) preserved the spatial information to an extent by concatenating the average and sum of detailed components. This however gets effected when the spatial information is distributed nonuniformly across the video frames as in case of online Ballet dance poses. However, if multiple attention blocks are connected as in figure.5(e) and (f), we get a little improvement from previous models as seen in figure's 10(g) and (h). Interestingly, the dual wavelet attention model (WDAM) [15] works on the above disadvantages and improves the attention features as shown in column 10(i) and 10(j). The training process becomes computationally inefficient with two attention models in series between each ResNet block.

The above disadvantage is surpassed in the proposed alternating wavelet channel and spatial attention (AWCSA) or wavelet spatial and channel attention (AWSCA) where the attention network compensates for the information loss and reduced dimensionality. For example, the wavelet channel attention efficiently handles reduced dimensionality but fails to control the loss of information. This information loss is retrieved through intermediate global features selected by channel attention and enhanced further by the wavelet spatial attention layers. The AWCSA and AWSCA in columns 10(k) and 10(l) has dual advantage of reducing dimensionality and preserving spatial relationships with improved feature clarity for recognition. The model has shown ability to learn intricate relationships related to textural information in the detailed components across the global features. The proposed attention is computed across multiple resolutions alternatively as the features travels through the network making a human like attention mechanism for improved focus. Thus, the features from AWCSA_R50 characterize super attention which are capable of representing structural as well as textural information in multi sourced video data. This is proved qualitatively with the help of recognition accuracies computed across the test dataset of BODV23. Confusion matrices were plotted in figure.11 for each of the representation in figure.10.

To evaluate the universality of AWCSA_R50, this work selects human action datasets that are similar to our BODV23. Consequently, three benchmark human action datasets were applied to validate the robustness of AWCSA_R50 against the state-of-the-art backbone architectures in the following section.

E. PROPOSED VS STANDARD BACKBONE ARCHITECTURES

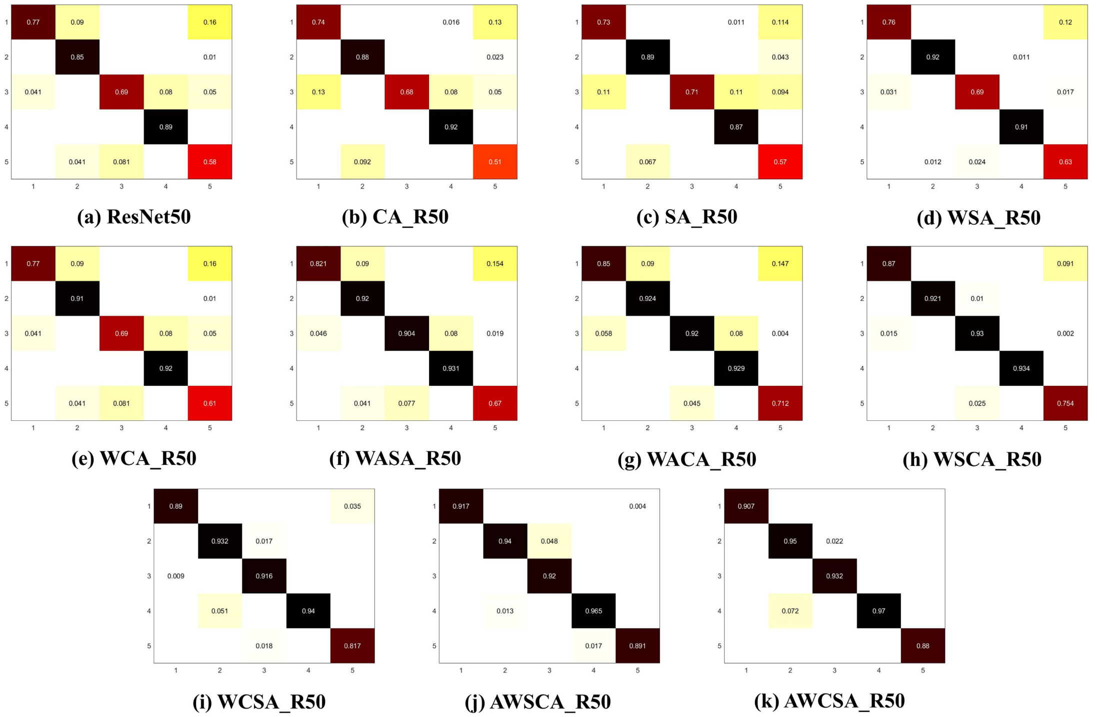
The plot in figure.12 shows the mAP computations across benchmark online video based human action and our BODV23 dataset processed on standard backbone architectures. The baseline backbone networks are incepted from AAM [47], DFL-CNN [48],ResNet34 [49], TASN [50] and VGGNet [51]. Importantly, the hyper parameter selection across these networks is flexed in accordance to the generate maximum precision and minimum loss during training. The 1- and 5-fold mAP values are presented as a range plot, where the variations in 1-fold mAP during multiple testing are plotted against the top 5-fold mAP. All the above models got an additional set of layers with AWCSA across 4 layers at different resolutions. For example, the VGG 16 is divided into four blocks with 2 convolution and 1 maximum pooling layer in each block. The outputs from each of these 4 blocks are applied to AWCSA module for generating attention maps. The plots in figure.12 show that the proposed attention maps are capable of generalization across baseline deep architectures for HAR online video data.

It shows that NTU RGB-D have resulted in better mAP's than BODV23 dataset. This is because of the large contrast and brightness variations in the online ballet dance videos as they are captured in closed environments and are collected from multiple online sources as compared to NTU RGB-D human action dataset. However, the performance of the AWCSA_R50 on BODV23 is better than the benchmark online human action datasets MPII and Kinematics-700 as shown in figure.12. Overall, there is an improvement in the performance of the standard models with AWCSA attention module across multiple layers. The wavelet attention module in this work have proved valuable in learning the local variations across frames for identification of structural and textural features for classification. The models were also tested by fusing the attention features at only one location in the backbone network. The computed mAPs are found to be always less than the attention across multiple layers. It is customary to judge that the increasing attention layers adds to the complexity of the overall classifiers. However, the results in the tables were averaged across the training dataset. Out of the 10 class labels, AWCSA reached the maximum precision across all classes on the test set. The improvements were large for individual classes in BODV23 (Ballet Online Dance Video 2023 dataset) and other Human Action Datasets. However, averaging on the number of samples across test dataset, this has further reduced. Table 3 shows the most used two backbone architectures ResNet and VGG along with the CAM and SAM.

Experimentation on the ballet dataset has been initiated to identify the computational complexity, model performance and robustness of the proposed and considered attention mechanisms. Table 3 reports the computational complexity and table 4 gives the performance and computational complexity of the models in table 3 along with the other state-of-the-art video-based classification techniques. Any attention model in spatial domain on ballet dance dataset must deal with noises such as blurring and background variations. These are considerably reduced due to DWT which are then applied to compute the attention in our proposed AWSCA and AWCSA. These additional wavelet layers have added a good amount of computational complexity to the backbone network as shown in table 3 below. However, this is far less than the transformer based self-attention models which use a 8 multi head attention layers to more than 6 self-attention layers to compute the attention scores. Though these transformer-based attentions are computationally complex, they generated good accuracies on the considered ballet dance dataset. The CViT has mAP almost equal to our proposed AWCSA model. Table 4 records the accuracies of all the considered datasets on the state - of - the - art models. The proposed AWCSA is more robust to changes in datasets and the model maintains a rather uniform reactions to test data which is not observed across other models. This is explainable as the wavelet-based attention are pixel regions are well refined in both edges and regions from source induced noises. Further the layers after the attention learn these focused regions which assist in making maximally correct classifications in the dense layers. We can visually confirm this fact by observing the attention maps across multiple layers as shown in fig.10.

F. VALIDATING AWCSA_R50(AWSCA_R50) THROUGH COMPARISON WITH STATE-OF-THE-ARTS

A chronicle validation has been conducted to judge AWCSA_R50(AWSCA_R50) against state-of-the-art human action recognition methods. These comparison networks are trained from scratch on the selected human action datasets and the hyperparameters are adjusted to extract maximum average precision (mAP). The results of this experiment are tabulated in Table 4. Contemplating on table 4, AWCSA_R50(AWSCA_R50) has precipitated maximum mAP when compared to previous similar methods on human action datasets. Since most of the models in table 4 are build using standard CNN architectures. The reconstruction and trains from scratch was effortlessly executed on a 8GB A4000 NVIDIA GPU. Accordingly, BOVD23 data was also trained and tested to validate AWCSA_R50 against the top rated HAR models. Table 4 concludes that ResNet50 with AWCSA(AWSCA) has produced good representations on 65536 input pixels because of its depth and residual connections. As the network depth increases (Resnet101) or



(a) ResNet50 (b) CA_R50 (c) SA_R50 (d) WSA_R50

(e) WCA_R50 (f) WASA_R50 (g) WACA_R50 (h) WSCA_R50

(i) WCSA_R50 (j) AWSCA_R50 (k) AWCSA_R50

FIGURE 11. Confusion matrices to show the attention maps obtained in the figure.10 is valid for all the classes in the dataset of (a) ResNet50 with no attention [46], (b) CA_R50 [4], (c) SA_R50 [5], (d) WCA_R50 [4],(e) WSA_R50 [43],(f) WACA_R50 [15],(g) WASA_R50, (h) WCSA_R50, (i) WSCA_R50, (j) AWCSA_R50 and (k) AWSCA_R50.

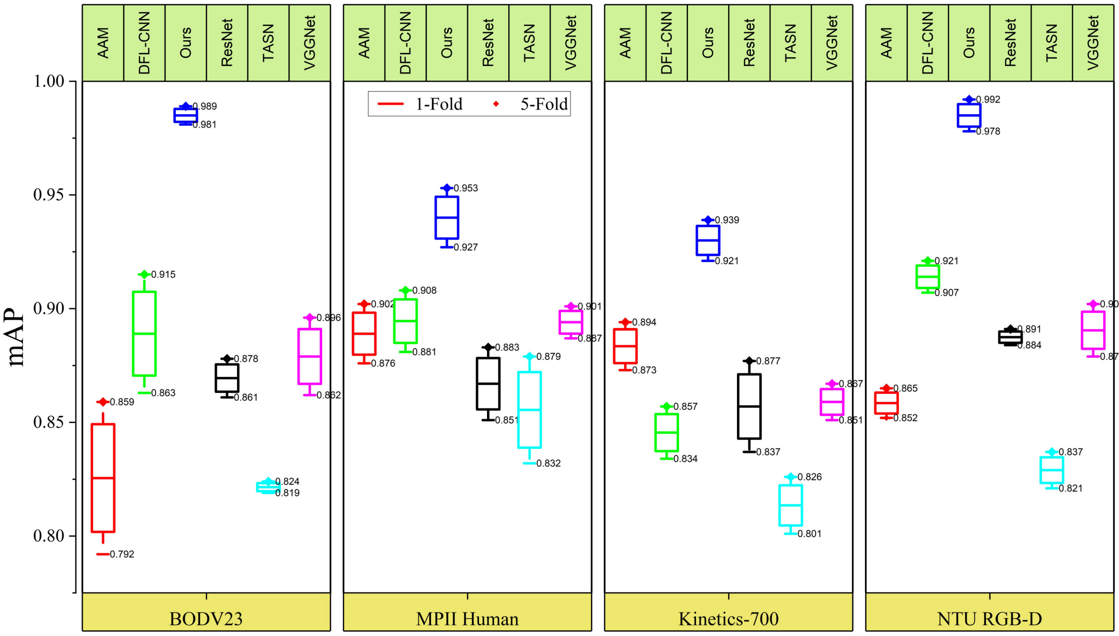


FIGURE 12. Plots show variation of mAP between 1- and 5-fold ranks on our BODV23 and three frequently used online video based human action datasets. It also shows the variations with respect to various backbone networks such as AAM [47], DFL-CNN [48],ResNet34 [49], TASN [50] and VGGNet [51].

decreases (Resnet34), the deep or shallow layers are deprived of good discriminating features for the set input image resolution. In case of ResNet101, increasing the input image resolution and decreasing it in case of ResNet 34 improves mAP.

The last two columns in table 4 indicates that the dance dataset representation needs improvement in the areas of contrast, background reduction and dancer body resolutions. In future endeavours, we are working on creating a 3D motion capture-based ballet dance pose dataset for real time dance

TABLE 3. Implementation details of the network in figure.1 along with two standard backbone architectures.

Model	Backbone Network	Input Size	Learning Rate	Trainable Parameters	Hyperparameter Initialization	Allowable Inferencing Loss	mAP
Alternating wavelet channel and spatial attention (AWCSA) (Fig.1)	ResNet 50	256	0.0001	26.1M	Gaussian Distribution with mean 0 and variance 1.	0.01	91.7
alternating wavelet spatial and channel attention (AWSCA)		256	0.0001	26.1M		0.01	91.3
CAM (SAM)		256	0.001	25.34M		0.05	87.4
Spatial Transformer Networks (STN)		256	0.001	20.02M		0.002	88.1
Squeeze – and – Excitation Network (SENet)		256	0.0001	24.31M		0.003	89.7
Self-Attention Vision Transformer (ViT)		256	0.000001	54.22M		0.05	87.6
Convolutional ViT (CViT)		256	0.0001	42.87M		0.04	87.9
Alternating wavelet channel and spatial attention (AWCSA) (Fig.1)	VGG16	256	0.00001	141M		0.05	89.9
alternating wavelet spatial and channel attention (AWSCA)		256	0.00001	141M		0.05	89.1
CAM (SAM)		256	0.0001	140.1M		0.1	86.2
Spatial Transformer Networks (STN)		256	0.001	161M		0.006	85.9
Squeeze – and – Excitation Network (SENet)		256	0.0001	162.8M		0.008	87.3
Self-Attention Vision Transformer (ViT)		256	0.000001	211M		0.06	84.7
Convolutional ViT (CViT)		256	0.0001	195.6M		0.05	84.2

TABLE 4. Comparison of the proposed AWCSA against the state-of-the-art Human Action Recognition methods trained from scratch on the considered datasets.

Method	Backbone	RAiD			Partial-iLIDS			Market-1501			RPIfield			BODV23		
		5-Fold	mAP	CMC	5-Fold	mAP	CMC	5-Fold	mAP	CMC	5-Fold	mAP	CMC	5-Fold	mAP	CMC
VGGNet	VGG19	0.814	0.764	0.843	0.758	0.683	0.807	0.751	0.647	0.801	0.729	0.681	0.797	0.823	0.784	0.85
MA-CNN	VGG19	0.799	0.759	0.815	0.769	0.713	0.799	0.769	0.711	0.794	0.715	0.593	0.749	0.845	0.769	0.858
DFL-CNN	VGG16	0.791	0.724	0.786	0.744	0.729	0.727	0.759	0.689	0.783	0.721	0.642	0.728	0.814	0.728	0.794
ResNet	ResNet50	0.883	0.816	0.893	0.861	0.798	0.859	0.807	0.813	0.813	0.829	0.764	0.838	0.897	0.834	0.912
TASN	ResNet50	0.867	0.854	0.861	0.845	0.789	0.849	0.822	0.735	0.842	0.792	0.673	0.791	0.874	0.849	0.855
AAM	VGG16	0.783	0.733	0.798	0.758	0.675	0.763	0.771	0.712	0.797	0.767	0.678	0.807	0.88	0.764	0.898
MGN	ResNet-50	0.881	0.821	0.907	0.871	0.859	0.874	0.879	0.795	0.883	0.829	0.755	0.861	0.911	0.815	0.927
BOT	ResNet-50	0.876	0.856	0.897	0.864	0.797	0.875	0.867	0.817	0.879	0.826	0.803	0.837	0.917	0.872	0.923
S-MIL-T	CNN	0.833	0.714	0.819	0.829	0.738	0.847	0.832	0.767	0.837	0.792	0.734	0.817	0.934	0.886	0.945
HAN	CNN	0.856	0.804	0.867	0.822	0.781	0.838	0.851	0.797	0.868	0.819	0.761	0.839	0.928	0.881	0.921
IANet	CNN	0.849	0.759	0.858	0.827	0.746	0.821	0.844	0.753	0.842	0.831	0.793	0.837	0.937	0.898	0.951
AGW	ResNet-50	0.875	0.812	0.882	0.857	0.806	0.863	0.884	0.813	0.894	0.827	0.775	0.839	0.941	0.904	0.939
FPR	FCN	0.853	0.798	0.862	0.862	0.793	0.876	0.846	0.772	0.873	0.855	0.786	0.867	0.929	0.899	0.948
PGFA	ResNet-50	0.915	0.852	0.911	0.869	0.816	0.872	0.907	0.816	0.911	0.878	0.811	0.893	0.914	0.907	0.936
PAM	ResNet101	0.912	0.842	0.909	0.881	0.793	0.903	0.892	0.827	0.907	0.879	0.803	0.881	0.952	0.912	0.957
PAM	ResNet50	0.919	0.873	0.904	0.876	0.813	0.884	0.914	0.797	0.893	0.892	0.816	0.911	0.948	0.909	0.951
STN	ResNet50	0.869	0.816	0.880	0.834	0.793	0.851	0.864	0.809	0.881	0.831	0.772	0.852	0.931	0.881	0.935
SENet	ResNet50	0.871	0.808	0.878	0.853	0.802	0.859	0.880	0.809	0.890	0.823	0.771	0.835	0.934	0.897	0.934
ViT	ResNet50	0.865	0.851	0.872	0.851	0.801	0.842	0.867	0.792	0.851	0.819	0.774	0.829	0.921	0.876	0.929
CViT	ResNet50	0.869	0.855	0.876	0.855	0.805	0.846	0.871	0.796	0.855	0.823	0.778	0.833	0.926	0.879	0.934
Proposed	ResNet50	0.942	0.893	0.957	0.904	0.846	0.927	0.931	0.838	0.938	0.923	0.902	0.941	0.969	0.917	0.974

performance estimation and learning applications. Despite these defects, the proposed AWCSA_R50(AWSCA_R50) on our BODV23 has shown to learn the structure and texture feature representations which improved the intra class discriminations.

G. ABLATION STUDY OF THE AWCSA_R50(AWSCA_R50)

The AWCSA_R50(AWSCA_R50) mainly alternates between the channel and spatial attention across all layers. Therefore, it becomes necessary to evaluate the impact on the order and length of channel and spatial wavelet attention blocks on the overall performance of the backbone architecture. Consequently, the ablation study focuses on estimating the performance of the classifiers AWCSA_R50(AWSCA_R50) based on the number of attention blocks and the order in which they alternate between channel and spatial layers.

Illustrations on the results in section D conclude the superiority of WCA_R50, WSA_R50 and dual Wavelet

TABLE 5. Results of ablation experiments to determine the number of wavelet channel and spatial attention pairs along with their locations in the backbone network.

Name of the Network	Location of SA@Resolution	Number of Attention Modules	NTU RGB D		Kinects-700		MPII Human		BODV23	
			1-Fold	5-Fold	1-Fold	5-Fold	1-Fold	5-Fold	1-Fold	5-Fold
AWCSA_R50E	64-32	2	0.889	0.923	0.838	0.868	0.853	0.899	0.867	0.921
AWCSA_R50M	32-16	2	0.878	0.912	0.827	0.857	0.842	0.887	0.856	0.910
AWCSA_R50L	16-8	2	0.819	0.850	0.771	0.799	0.785	0.827	0.798	0.848
AWCSA_R50	64-32-16-8	2	0.935	0.971	0.881	0.913	0.897	0.945	0.912	0.969
AWSCA_R50E	64-32	2	0.847	0.879	0.798	0.827	0.812	0.856	0.826	0.878
AWSCA_R50M	32-16	2	0.831	0.863	0.783	0.812	0.798	0.840	0.811	0.862
AWSCA_R50L	16-8	2	0.689	0.717	0.650	0.674	0.662	0.697	0.673	0.715
AWSCA_R50	64-32-16-8	4	0.854	0.887	0.804	0.834	0.819	0.863	0.833	0.885
DWCA_R50E	64-32	2	0.699	0.732	0.655	0.686	0.668	0.700	0.680	0.723
DWCA_R50M	32-16	2	0.738	0.766	0.695	0.720	0.708	0.746	0.720	0.765
DWCA_R50L	16-8	2	0.712	0.740	0.671	0.696	0.683	0.720	0.695	0.738
DWCA_R50	64-32-16-8	4	0.825	0.857	0.778	0.806	0.792	0.834	0.805	0.855
DWSA_R50E	64-32	2	0.789	0.819	0.733	0.767	0.746	0.783	0.761	0.809
DWSA_R50M	32-16	2	0.775	0.805	0.730	0.757	0.743	0.783	0.756	0.803
DWSA_R50L	16-8	2	0.751	0.780	0.708	0.734	0.721	0.760	0.733	0.779
DWSA_R50	64-32-16-8	4	0.841	0.882	0.789	0.826	0.804	0.843	0.819	0.871

attention(WCSA_R50)(WSCA_R50) over the traditional attention models. Based on the finding shown above, AWCSA_R50 has recorded improved performance metrics over the above discussed wavelet attention models. Consequently, this improved performance of AWCSA_R50 is due to multiple attention blocks which insists a deeps study on the requirements of these attention blocks and their fusion locations on the overall metrics.

The listed values in the table 5 conclude that the early fusions across the first residual blocks have greater impact on the overall performance of the classifier when compared to late fusion. This is due to greater structural integrity found across higher frame resolutions in starting layers than the deeper layers of the backbone network. In this subsection the number of wavelet channel and spatial attention blocks required to maximize test mAP is evaluated. Specifically, we will have 4 combinations for evaluation. In AWCSA_R50, the WSA follows WCA. Accordingly there are 4 places ResNet50 or VGG16 backbone networks where the WSA and WCA can fit one after the other. As a result, we have four combination namely early(AWCSA_R50E), mid(AWCSA_R50M), late(AWCSA_R50L) and our proposed AWCSA_R50 with alternating WCSA. Consequently, we have 3 new networks that will evaluate the proposed AWCSA_R50 where the attention mechanism is progressive. All considered datasets were used for training and testing with the previously considered hyper parameters. The results are tabulated in table 5 for 1-fold and 5-fold mAP metric. Similar analysis is conducted for AWSCA_R50 and the work in DWA [15].

V. CONCLUSION

The AWCSA_R50 proposes to enhance the feature space of ballet online dance video data thereby reducing inferencing error during testing. The framework is built on the Wavelet based attention layers in both channel and spatial dimensions that have been integrated alternatively across the global features. This process have improved the feature representation of online Ballet dance video dataset sourced from YouTube in both structure and texture. The BODV23 challenges the existing attention based deep learning models in generating decent accuracy due to imbalances in spatial attention across the global features. Accordingly, attention at different global feature resolutions using alternative wavelet based attention modes provides a human-like focus which resulted in excellent structural and texture information of the dancer in the video sequence. The AWCSA enables high and low frequency features to add focus to the weakly distributed global pose features. The experiments on BODV23 and human action recognition benchmarks has proved that the proposed AWCSA_R50(AWSCA_R50) improved the overall accuracy by around 7%. However, these models are burdened computationally with more trainable parameters than the traditional models. Moreover, the attention mechanism designed in this work is task specific and might need hyperparameter tuning for different datasets.

VI. ACKNOWLEDGMENT

The dataset was made available from the Biomechanics and Vision Computing research Laboratory, Koneru Lakshmaiah Education Foundation.

REFERENCES

[1] P. V. V. Kishore, K. V. V. Kumar, E. K. Kumar, A. S. C. S. Sastry, M. T. Kiran, D. A. Kumar, and M. V. D. Prasad, "Indian classical dance action identification and classification with convolutional neural networks," *Adv. Multimedia*, vol. 2018, pp. 1–10, Jul. 2018.

[2] Y. Hao, S. Wang, P. Cao, X. Gao, T. Xu, J. Wu, and X. He, "Attention in attention: Modeling context correlation for efficient video classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7120–7132, Oct. 2022.

[3] S. Gao, L. Duan, and I. W. Tsang, "DEFEATnet—A deep conventional image representation for image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 494–505, Mar. 2016.

[4] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[5] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[6] H. Lee, H.-E. Kim, and H. Nam, "SRM: A style-based recalibration module for convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1854–1862.

[7] H. Zhang, T. Lu, and S. Jia, "Vehicle re-identification based on multi-view and convolutional block attention," in *Proc. 4th Int. Conf. Artif. Intell. Pattern Recognit.*, Sep. 2021, pp. 225–231.

[8] X. Li, Y. Guo, W. Pan, H. Liu, and B. Xu, "Human pose estimation based on lightweight multi-scale coordinate attention," *Appl. Sci.*, vol. 13, no. 6, p. 3614, Mar. 2023, doi: 10.3390/app13063614.

[9] Z. Qin, P. Zhang, F. Wu, and X. Li, "FcaNet: Frequency channel attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 763–772.

[10] E. Amiri, M. Rahmanian, S. Amiri, and H. Y. Praee, "Medical images fusion using two-stage combined model DWT and DCT," *Int. Adv. Researches Eng. J.*, vol. 5, no. 3, pp. 344–351, Dec. 2021.

[11] H. Bi, L. Xu, X. Cao, Y. Xue, and Z. Xu, "Polarimetric SAR image semantic segmentation with 3D discrete wavelet transform and Markov random field," *IEEE Trans. Image Process.*, vol. 29, pp. 6601–6614, 2020.

[12] H. Bi, R. Santos-Rodriguez, and P. Flach, "Polsar image classification via robust low-rank feature extraction and Markov random field," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 708–711.

[13] C. He, S. Li, Z. Liao, and M. Liao, "Texture classification of PolSAR data based on sparse coding of wavelet polarization textons," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 8, pp. 4576–4590, Aug. 2013.

[14] S. Yousefi, M. T. M. Shalmani, J. Lin, and M. Staring, "A novel motion detection method using 3D discrete wavelet transform," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3487–3500, Dec. 2019.

[15] Y. Yang, L. Jiao, X. Liu, F. Liu, S. Yang, L. Li, P. Chen, X. Li, and Z. Huang, "Dual wavelet attention networks for image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1899–1910, Apr. 2023.

[16] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[17] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A short note on the kinetics-700–2020 human action dataset," 2020, *arXiv:2010.10864*.

[18] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 588–595.

[19] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2023.

[20] S. Samanta, P. Purkait, and B. Chanda, "Indian classical dance classification by learning dance pose bases," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2012, pp. 265–270.

[21] K. V. V. Kumar and P. V. V. Kishore, "Indian classical dance mudra classification using HOG features and SVM classifier," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 5, p. 2537, Oct. 2017.

[22] K. V. V. Kumar, P. V. V. Kishore, and D. A. Kumar, "Indian classical dance classification with AdaBoost multiclass classifier on multifeature fusion," *Math. Problems Eng.*, vol. 2017, pp. 1–18, Aug. 2017.

[23] M. Devi and S. Saharia, "A two-level classification scheme for single-hand gestures of sattriya dance," in *Proc. Int. Conf. Accessibility Digit. World (ICADW)*, Dec. 2016, pp. 193–196.

[24] S. Saha, S. Ghosh, A. Konar, and A. K. Nagar, "Gesture recognition from Indian classical dance using Kinect sensor," in *Proc. 5th Int. Conf. Comput. Intell., Commun. Syst. Netw.*, Jun. 2013, pp. 3–8.

[25] A. Mohanty, P. Vaishnavi, P. Jana, A. Majumdar, A. Ahmed, T. Goswami, and R. R. Sahay, "Nrityabodha: Towards understanding Indian classical dance using a deep learning approach," *Signal Process., Image Commun.*, vol. 47, pp. 529–548, Sep. 2016.

[26] A. D. Naik and M. Supriya, "Classification of Indian classical dance 3D point cloud data using geometric deep learning," in *Computational Vision and Bio-Inspired Computing*. Cham, Cham: Springer, 2021, pp. 81–93.

[27] S. Dewan, S. Agarwal, and N. Singh, "A deep learning pipeline for Indian dance style classification," *Proc. SPIE*, vol. 10696, pp. 265–273, Apr. 2018.

[28] N. Jain, V. Bansal, D. Virmani, V. Gupta, L. Salas-Morera, and L. Garcia-Hernandez, "An enhanced deep convolutional neural network for classifying Indian classical dance forms," *Appl. Sci.*, vol. 11, no. 14, p. 6253, Jul. 2021.

[29] A. D. Naik and M. Supriya, "Classification of Indian classical dance images using convolution neural network," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Jul. 2020, pp. 1245–1249.

[30] S. Biswas, A. Ghildiyal, and S. Sharma, "Classification of Indian dance forms using pre-trained model-VGG," in *Proc. 6th Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2021, pp. 278–282.

[31] J. R. Challapalli and N. Devarakonda, "A novel approach for optimization of convolution neural network with hybrid particle swarm and grey wolf algorithm for classification of Indian classical dances," *Knowl. Inf. Syst.*, vol. 64, no. 9, pp. 2411–2434, Sep. 2022.

[32] C. J. Rani and N. Devarakonda, "Indian classical dance forms classification using transfer learning," in *Computational Intelligence and Data Analytics*. Cham, Switzerland: Springer, 2023, pp. 241–255.

[33] R. J. Raj, S. Dharan, and T. T. Sunil, "Optimal feature selection and classification of Indian classical dance hand gesture dataset," *Vis. Comput.*, vol. 39, no. 9, pp. 4049–4064, Sep. 2023.

[34] S. Liaqat, K. Dashtipour, K. Arshad, K. Assaleh, and N. Ramzan, "A hybrid posture detection framework: Integrating machine learning and deep neural networks," *IEEE Sensors J.*, vol. 21, no. 7, pp. 9515–9522, Apr. 2021.

[35] S. Shailesh and M. V. Judy, "Capsule networks for classifying conflicting double-handed classical dance gestures," in *Data Engineering and Communication Technology*. Cham, Switzerland: Springer, 2021, pp. 29–37.

[36] S. Shailesh and M. V. Judy, "Understanding dance semantics using spatio-temporal features coupled GRU networks," *Entertainment Comput.*, vol. 42, May 2022, Art. no. 100484.

[37] S. Xue, W. Qiu, F. Liu, and X. Jin, "Wavelet-based residual attention network for image super-resolution," *Neurocomputing*, vol. 382, pp. 116–126, Mar. 2020.

[38] Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Wavelet attention embedding networks for video super-resolution," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7314–7320.

[39] X. Zhao, P. Huang, and X. Shu, "Wavelet-attention CNN for image classification," *Multimedia Syst.*, vol. 28, no. 3, pp. 915–924, Jan. 2022.

[40] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks," 2018, *arXiv:1805.08620*.

[41] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.

[42] H.-H. Yang, C. H. Yang, and Y. F. Wang, "Wavelet channel attention module with a fusion network for single image deraining," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 883–887.

[43] X. Song, D. Zhou, W. Li, H. Ding, Y. Dai, and L. Zhang, "WSAMF-Net: Wavelet spatial attention-based MultiStream feedback network for single image dehazing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 575–588, Feb. 2023.

[44] *Popular Songs Source*. Accessed: Sep. 16, 2019. [Online]. Available: https://medium.com/@info_70178/arabesques-and-art-histories-of-a-position-82428af8c204

[45] *Downloaded Online Sources*. Accessed: Apr. 13, 2011. [Online]. Available: https://www.youtube.com/watch?v=SmRrfm1ihGg

[46] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *Int. J. Cognit. Comput. Eng.*, vol. 2, pp. 57–64, Jun. 2021.

[47] X. Wang, J. Shi, H. Fujita, and Y. Zhao, "Aggregate attention module for fine-grained image classification," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 7, pp. 8335–8345, Jul. 2023.

[48] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4148–4157.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[50] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5007–5016.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.