MAiVAR: Multimodal Audio-Image and Video Action Recognizer

Muhammad Bilal Shaikh[1], Douglas Chai[1], Syed Mohammed Shamsul Islam[1] and Naveed Akhtar[2]

[1]Edith Cowan University, 270 Joondalup Drive, Joondalup, WA 6027, Perth, Australia

[2]The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Perth, Australia

[1]{m.shaikh,d.chai,syed.islam}@ecu.edu.au, [2]naveed.akhtar@uwa.edu.au

Abstract—**Currently, action recognition is predominately performed on video data as processed by CNNs. We investigate if the representation process of CNNs can also be leveraged for multimodal action recognition by incorporating image-based audio representations of actions in a task. To this end, we propose Multimodal Audio-Image and Video Action Recognizer (MAiVAR), a CNN-based audio-image to video fusion model that accounts for video and audio modalities to achieve superior action recognition performance. MAiVAR extracts meaningful image representations of audio and fuses it with video representation to achieve better performance as compared to both modalities individually on a large-scale action recognition dataset.**

I. INTRODUCTION

Action recognition techniques proposed in the last decade have predominantly used video representations, i.e., sequences of images to recognize an event or action in a video [1–3]. Visual modality contains spatial information, which is inherently helpful for CNN-based classification. To better capture the multimodality aspect of action data, a recent trend has been to combine information from different modalities such as optical flow, RGB-difference, warped-optical flow and others. These types of hybrid features can be applied to various action recognition techniques such as Temporal Segment Networks (TSN) [4], Temporal Relation Network (TRN) [5] and Temporal Shift Module (TSM) [6]. Motivated by the success of CNN-based models in object detection [7] and image classification [8], this paper investigates whether CNNs have the ability to understand complex image-based audio representations and to impact multimodal classification.

The audio modality has significant features that can contribute significantly in action recognition from videos: audio contains dynamics and rich contextual temporal information [9] and requires a lighter computational process as compared to video frames. For example, as shown in Fig. 1, within a short clip of action of Boxing a Punching Bag, a single audio-image frame includes most of the dynamic contextual information contained in the audio, i.e., the sound of a boxing glove hitting the bag, while the accompanying video clip contains useful cues of spatial dynamics. This prompts the question of: across an entire video, can audio be beneficial to select the critical features that are useful for recognizing actions? To answer this question, we introduce the Multimodal Audio-Image and Video Action Recognizer (MAiVAR), a CNN-based audio-image to video fusion-based model that can be applied to video and audio features for classifying actions at
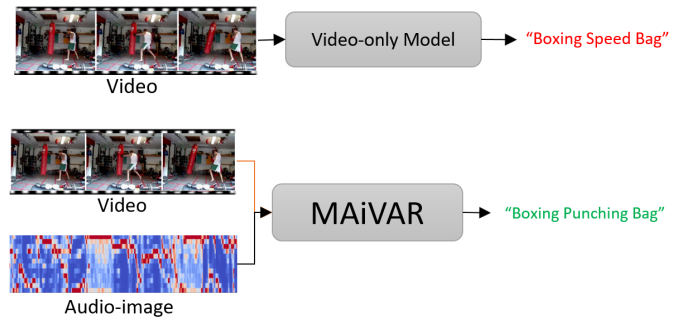

Fig. 1: Action recognition using unimodal video-only approach (top) versus **MAiVAR** with intermediate fusion of audio-visual features (bottom).

a better accuracy than the unimodal equivalent. Additionally, we propose an approach to extract meaningful image representations of audio, which can significantly improve classification scores when used with CNNs.

The benefits of MAiVAR are threefold. Firstly, MAiVAR has superior performance, outperforming the state-of-the-art model on the same data for video representation when evaluated based on the UCF-101 dataset. Secondly, MAiVAR naturally supports both audio and video inputs and can be applied to different tasks without any architectural change. Thirdly, as compared to state-of-the-art video-based CNN models, MAiVAR features a simpler and more intuitive architecture. To the best of our knowledge, MAiVAR is the first audio-image to video fusion-based action classification model. The key contributions of this work are as follows: (1) a new feature representation strategy is proposed to select the most informative candidate representations for audio-visual fusion; (2) collection of effective audio-image based representations that complements video modality for better action recognition are included; (3) a novel fusion architecture, MAiVAR is proposed for audio-visual fusion that supports different audio-image representations and can be applied to different tasks; and (4) state-of-the-art results for action recognition on audio-visual dataset have been reported.

II. RELATED WORK

Multimodal action recognition employs multi-stream approaches to incorporate different modalities. TSN [4], TRN [5] and TSM [6] are based on 2D CNNs. All three models make use of a two-stream approach that uses both RGB and optical flow. Besides RGB and optical flow streams, Temporal
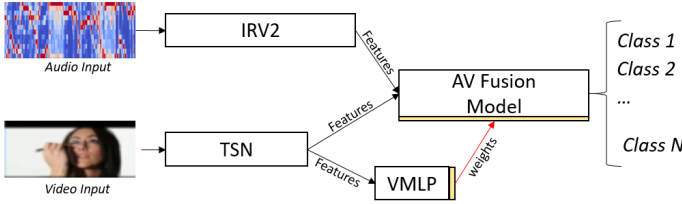
Fig. 2: The proposed MAiVAR architecture.

Binding Networks (TBN) [10] adds audio as another modality. SlowFast [11] uses two RGB streams with different resolutions and frame rates. The proposed MAiVAR is based on TSN [4] pre-trained on the Kinetics [12] dataset and Inception-ResNet2 (IRv2) pre-trained on ImageNet [8]. TSN was primarily proposed for video-level action recognition tasks, while IRV2 is a CNN architecture that builds on the inception family [13] of architectures but incorporates residual connections.

Recently, TSN has been adapted for the backbone module in video understanding scenarios [5, 14–17], where it is typically used in conjunction with a succeeding module. In [17], TSN was employed as 2D CNN backbones to learn motion dynamics in videos. In contrast, IRV2 has been used for feature extraction from images in [18], which helped in different image restoration tasks [19, 20]. Spatial-Temporal Network (STNeT) [21] adapts IRV2 to model local and global spatio-temporal features. The closest works to ours are [4, 22]. Takahashi *et al.* have classified audio events using 3D CNN with some representations of audio, while the TSN-based approach [4] uses optical flow variants with RGB.

In another work [23], a video skimming mechanism is applied for untrimmed video aided by audio to eliminate both short-term and long-term redundancies. IMGAUD2VID uses audio to extract dynamic scene information in a single frame, which captures most of the appearance information to form an audio-image pair. These pairs are then used to select key moments from the video for action recognition. Unlike IMGAUD2VID, our idea capture more spatial information along with the holistic dynamic information of a scene from the image representations of audio.

The proposed MAiVAR differs from these works in that it extracts audio features using IRV2 and video features using TSN, fusing both features using a multi-layer perceptron (MLP), which is initialized with weights from an MLP pre-trained with video features. We also conduct extensive experiments to show the design choice of MAiVAR on audio-image representations and fusion tasks.

III. METHODOLOGY

The proposed MAiVAR architecture is illustrated in Fig. 2. Below, we discuss the background of our method, feature representations and network architecture.

A. Background

Given the training set $T = \{x_1...x_n, y_1...y_n\}$, where x_i is the i-th training example and y_i is its true label, training on a single modality m (e.g. RGB frames or audio) means minimizing an empirical loss:

$$L(C(\phi_m(X)), y) \qquad (1)$$

where ϕ_m is normally a deep network with parameter $\ominus_m$, and C is a classifier, typically one or more fully-connected (FC) layers with parameter $\ominus_c$. For classification problems considered here, L is the cross-entropy loss. Minimizing (1) gives solution $\ominus*_m$ and $\ominus*_c$. We train a late-fusion model on M different modalities $(\{m_i\}_1^k)$. Each modality is processed by a different deep network ϕ_{mi} with parameter $\ominus_{mi}$, with their features fused and passed to a classifier C. Formally, training is done by minimizing the loss:

$$\mathcal{L}_{multi} = L(C(\phi_{audio} \oplus \phi_{video}), y) \qquad (2)$$

where $\oplus$ denotes a fusion operation (e.g. concatenation). Fig. 3 shows an illustration of a joint training between $audio$ and $video$ modalities.

B. Feature Representations

1) Visual Feature Representation: The duration of video input is t seconds, then converted into a frame-based representation that is compatible with TSN. This is fed as input to the TSN feature extractor. Each video clip represents an action example. We extract the feature from TSN using the *AvgPool2d* layer while pruning the original average consensus layer. TSN produces embeddings of size 1024×25, to allow the model to capture the spatio-temporal structure of the RGB video. The resulting sequence is then fed as input into the fusion model. The advantages of this setup are: 1) the standard TSN architecture is easy to implement and reproduce as it is off-the-shelf in TensorFlow and PyTorch, and 2) we intend to apply transfer learning for Fusion MLP, where a standard architecture makes transfer learning easier. We have adapted the MMAction2 [24] interface for using TSN as a feature extractor. TSN is used as a feature extractor based on its temporal pooling of frame-level features, which has been rigorously applied as an efficient video-level feature extractor in different problems. For example, when TSN is plugged into GSM [25], a growth of 32% accuracy is gained. Yang et al. [26] has used TSN with soft attention mechanism to capture important frames from each segment, where these frames are key candidates to volume-based features. Zhang [27] has used the TSN model as a feature extractor with ResNet101 for efficient automated pigs behavior recognition.
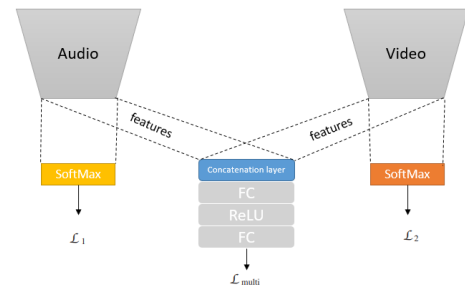

Fig. 3: Multimodal architecture schematics. The audio and video models output an independent action prediction $\mathcal{L}_1$ and $\mathcal{L}_2$ based on their respective input datasets. The features from the *AvgPool2d* layers of the two models are combined to feed the fusion module, which outputs the predicted action $\mathcal{L}_{multi}$.

2) Audio Feature Representation: The audio input has been transformed into different image-based representations. This results in $(224 \times 224 \times 3) \times n$ (n is the total number of audio action instances) audio-image as input to the IRV2 backbone. Accordingly, each audio-image represents a whole action instance (equivalent to video clip of t secs). We extract the feature from IRV2 using the *AvgPool2d* layer while pruning the original classification layer. IRV2 produces embeddings of size 1536 to allow the model to capture spectral information. The resulting sequence is then fed as input to the fusion model. The advantage of this simple setup is that the standard IRV2 architecture performs better on audio-image representations as compared to other CNN-based models (see Section IV-D).

C. Network Architecture

1) Visual Network: For simplicity, we adopt an established CNN-based visual feature extractor TSN [4] which projects visual embeddings to the fusion model. In TSN, the original BNInception backbone is used. The consensus layer is removed to expose features from *AvgPool2D* layer of the TSN model. The features from TSN are fed as input to the visual MLP. The weights from the trained visual network are then used to initialize the Fusion MLP.

2) Audio Network: The audio network uses an IRV2-based feature extractor initialized with weights from ImageNet for creating audio embeddings for the audio MLP. In IRV2, we froze all gradients and removed the last layer to expose features from the average pooling layers of the model. The features from IRV2 are fed as input to the fusion MLP.

3) Fusion Network: The fusion network concatenates the audio and visual features and classifies the combined feature into action labels. As discussed in Section III-C1, the visual network produces higher classification accuracy. Therefore, weights from the trained visual network were used to initialize the fusion model to fine-tune the fusion model.

All hidden layers, except the last fully-connected layer, are equipped with Rectified Linear Unit (ReLU) non-linearity. The fusion network was trained by minimizing the cross-entropy loss L with l_1 regularization using back-propagation:

$$\arg \min_W \sum_{i,j} L(x_j^i + z_j^i, y_j^i, W) \tag{3}$$

where x_j^i and z_j^i is the j-th input vector from the respective audio and video features, y_j is the corresponding class label and W is the set of network parameters. For the fusion model, we used the initialization weights from the video classification model (see Algorithm 1). Empirically, the weights from the video model boost the performance of the fusion model.

IV. EXPERIMENTAL SETUP AND RESULTS

Here we discuss the dataset, implementation details, our comparison with state-of-the-art models and the ablation study.

A. Dataset

We conducted experiments on UCF101 dataset [28] which is a standard benchmark for action recognition. UCF-101 contains $13,320$ videos of 101 human action categories, such

---

Algorithm 1 Weights assignment algorithm

Input: Initial weights ϕ^0, ρ^0 for audio and video respectively, N Number of epochs
Output: ψ^0
$\qquad\qquad\qquad\qquad\qquad\qquad\quad$ ▷ Weights after training

1: **procedure** WEIGHTS(ϕ^0, ρ^0, N)
2: $\quad$ Train ϕ^0 with $\{Wa_i\}_{i=1}^k$
3: $\quad$ **for** $i = 1, 2, ..., N$ **do**
4: $\quad\quad$ process batch
5: $\quad\quad$ update weights ϕ_i
6: $\quad$ **end for**
7: $\quad$ Train ρ^0 with $\{Wv_i\}_{i=1}^k$
8: $\quad$ **for** $i = 1, 2, ..., N$ **do**
9: $\quad\quad$ process batch
10: $\quad\quad$ update weights ρ_i
11: $\quad$ **end for**
12: $\quad$ $\psi^0 \leftarrow \rho^N$;
13: $\quad$ Fusion model weights ψ^0 initialized with weights from trained video model ρ^N.
14: $\quad$ **return** ψ^0 $\qquad\qquad$ ▷ Weights for fusion model
15: **end procedure**

---

as Apply Eye Makeup, Blow Dry Hair and Table Tennis with an average length of 180 frames per video. We observed that half of the dataset had no audio channel. Thus, to focus on the effect of audio features, we used the subset with audio channel [22]. This resulted in 6837 videos across 51 categories. Whilst this led the dataset to be significantly reduced, the distribution of the audio dataset was similar to the video dataset. We used the first train-test split setting provided with this dataset, which resulted in 4893 training and 1944 testing samples.

B. Implementation Details

We have considered two input modalities: video and audio. For video, we used RGB frames as input with 25 as the number of segments for TSN. We followed [4] for visual pre-processing and data augmentation. For audio, we used Librosa [29] to generate audio-image representations. Audio-image representations were normalized as per ImageNet configuration, with random horizontal and vertical flips. For video, [4] was followed. For feature extraction, the TSN-based [4] feature extractor used the BNInception backbone for visual data and IRV2 [13], using residual inception blocks for audio feature extraction.

For training and testing, all networks were implemented on a machine with an Intel XEON® CPU and $8\times$ GeForce GTX 1080 Ti GPUs. The models were implemented using the PyTorch [30]. Similar to [22], the data was randomly split as per split 1 on UCF-101 [28]. We use Adam [31] optimizer, where the optimal choices for learning rates α were $3e^{-4}$ for audio and video models, while for the fusion model $1e^{-4}$ a batch size of 16 and 128 for audio-image representation and the fusion model was used, respectively.

C. Comparing MAiVAR Against Other Methods on UCF-101

We show a comparison of the performance of our model against state-of-the-art methods in Table I. The benchmarks were reproduced using accuracy over the standard train and

TABLE I: Comparison of the performance of MAiVAR with the state-of-the-art models on UCF-101 dataset.

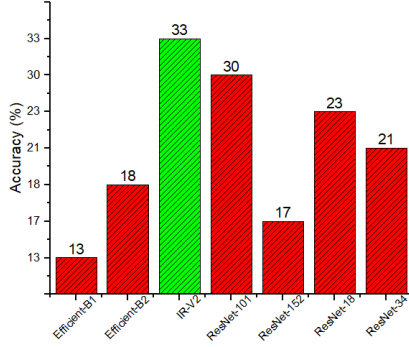| Method	Accuracy (%)
TSN (RGB) [4]	60.77
IMGAUD2VID [23]	81.10
C3D [32]	82.23
STA-Net (RGB)[26]	83.42
C3D+AENet [22]	85.33
MAiVAR-CH	**87.91**

Fig. 4: Comparison of audio-image only feature-extractors. Green bar highlights the best performing model.

test split. Following the evaluation protocol of [22], we used the accuracy to evaluate the performance of the models.

D. Ablation Experiments

1) Audio-image Feature Extraction: Several experiments were conducted to determine the best settings for the proposed model, which can be broken down into selecting feature extractors for audio-image representations and choosing optimal audio-image representation and hyper-parameter settings. We tried several CNN-based feature extractors for audio-image representations including smaller as well as deeper variants of ResNet [33], EfficientNet [34] and InceptionResNet [13]. Compared to IRV2, ResNet-101 and ResNet-18 showed relatively better performance (see Fig. 4).

2) Convergence of Audio Representations: We evaluated performance of six different audio-image representations (shown in Fig. 5) for fusion with video features. Before fusion, as the structure of the chromagram is well-suited for CNN-based models, chromagram-based representations performed better than other competitor representations. Convergence of each audio-image representation after fusion with video modality is illustrated in Fig. 6. However, it has been observed that representation that produces better accuracy within unimodal learning is not the optimal candidate for achieving better accuracy after fusion. As presented in Table II, MFCCs Feature Scaling-based representation produced the best audio-only accuracy, but after fusion with video, the chromagram-based representation produced the best accuracy.

V. CONCLUSION

We propose MAiVAR, a fusion-based, end-to-end model for audio-video classification that features both simple architec-
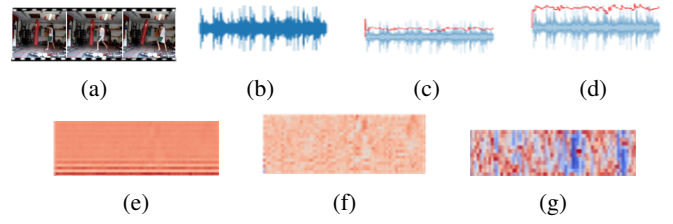

Fig. 5: (a) Segmented video input and six different audio-image representations of the same action: (b) Waveplot, (c) MFCC, (d) MFCC feature scaling, (e) Spectral Centroids, (f) Spectral Rolloff and (g) Chromagram.

TABLE II: Comparing audio-image representations before (audio MLP) and after fusion on the basis of accuracy in percentage. *Note: video-only accuracy is 75.67% .*

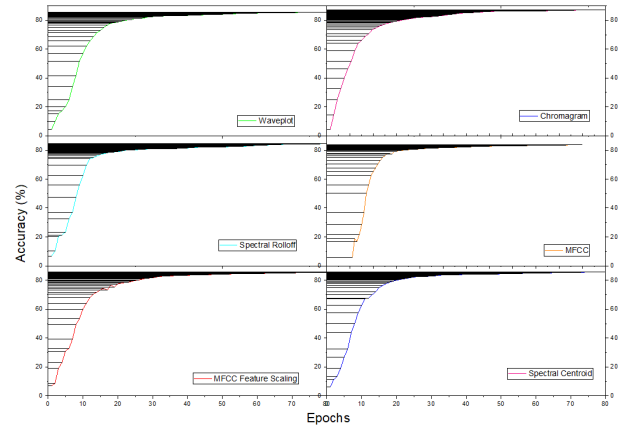| Representation	Audio	Fusion (Audio+Video)
Waveplot	12.08	86.21
Spectral Centroids	13.22	86.26
Spectral Rolloff	16.46	86.00
MFCCs	12.96	83.95
MFCCs Feature Scaling	**17.43**	86.11
Chromagram	15.48	**87.91**

Fig. 6: Convergence of the audio representations.

ture and superior performance. The experimental results show that our audio-image to video fusion-based model performs well in comparison with state-of-the-art models. The fusion module is significant in extracting joint information among different modalities. This is because it is evident that pre-training on a larger video dataset improves results substantially. Our framework can be applied to large-scale multimodal action recognition datasets, such as Kinetics 400/600/700 [35].

REFERENCES

[1] M. B. Shaikh and D. Chai, "RGB-D data-based action recognition: a review," *Sensors*, vol. 21, no. 12, p. 4246, 2021.

[2] C. Li *et al.*, "Collaborative spatiotemporal feature learning for video action recognition," in *Proc. of the CVPR*, 2019, pp. 7872–7881.

[3] C.-Y. Wu *et al.*, "Compressed video action recognition," in *Proc. of the CVPR*, 2018, pp. 6026–6035.

[4] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. of the ECCV*, 2016, pp. 20–36.

[5] B. Zhou *et al.*, "Temporal relational reasoning in videos," in *Proc. of the ECCV*, 2018, pp. 803–818.

[6] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. of the ICCV*, 2019, pp. 7083–7093.

[7] S. Ren *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. of the NeurIPS*, vol. 28, 2015, pp. 91–99.

[8] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[9] W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[10] E. Kazakos *et al.*, "EPIC-Fusion: Audio-visual temporal binding for egocentric action recognition," in *Proc. of the ICCV*, 2019, pp. 5492–5501.

[11] C. Feichtenhofer *et al.*, "Slowfast networks for video recognition," in *Proc. of the ICCV*, 2019, pp. 6202–6211.

[12] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? a new model and the kinetics dataset," in *Proc. of the CVPR*, 2017, pp. 6299–6308.

[13] C. Szegedy *et al.*, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. of the AAAI*, 2017, pp. 4278–4284.

[14] J. Lei *et al.*, "Less is more: Clipbert for video-and-language learning via sparse sampling," in *Proc. of the CVPR*, June 2021, pp. 7331–7341.

[15] R. Girdhar *et al.*, "ActionVLAD: Learning spatio-temporal aggregation for action classification," in *Proc. of the CVPR*, 2017, pp. 971–980.

[16] Y. Li *et al.*, "VLAD3: Encoding dynamics of deep features for action recognition," in *Proc. of the CVPR*, 2016, pp. 1951–1960.

[17] H. Kwon *et al.*, "Learning self-similarity in space and time as generalized motion for video action recognition," in *Proc. of the ICCV*, Oct. 2021, pp. 13 065–13 075.

[18] X. Mei *et al.*, "Artificial intelligence–enabled rapid diagnosis of patients with COVID-19," *Nature Medicine*, vol. 26, no. 8, pp. 1224–1228, 2020.

[19] J. Gu *et al.*, "NTIRE 2021 challenge on perceptual image quality assessment," in *Proc. of the CVPR*, 2021, pp. 677–690.

[20] C. Yan *et al.*, "Precise no-reference image quality evaluation based on distortion identification," *ACM TOMM*, vol. 17, no. 3s, pp. 1–21, 2021.

[21] D. He *et al.*, "Stnet: Local and global spatial-temporal modeling for action recognition," in *Proc. of the AAAI*, vol. 33, no. 01, 2019, pp. 8401–8408.

[22] N. Takahashi, M. Gygli, and L. Van Gool, "AENet: Learning deep audio features for video analysis," *IEEE TMM*, vol. 20, no. 3, pp. 513–524, 2017.

[23] R. Gao *et al.*, "Listen to look: Action recognition by previewing audio," in *Proc. of the CVPR*, 2020, pp. 10 457–10 467.

[24] "Welcome to mmaction2's documentation!" [Online]. Available: https://mmaction2.readthedocs.io/en/latest/

[25] S. Sudhakaran, S. Escalera, and O. Lanz, "Gate-shift networks for video action recognition," in *Proc. of the CVPR*, 2020, pp. 1102–1111.

[26] G. Yang *et al.*, "STA-TSN: Spatial-temporal attention temporal segment network for action recognition in video," *PloS one*, vol. 17, no. 3, pp. 1–19, 2022.

[27] K. Zhang *et al.*, "Automated video behavior recognition of pigs using two-stream convolutional networks," *Sensors*, vol. 20, no. 4, p. 1085, 2020.

[28] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[29] B. McFee *et al.*, "Librosa: Audio and music signal analysis in python," in *Proc. of the Python in Science Conf.*, vol. 8, 2015, pp. 18–25.

[30] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. of the NeurIPS*, 2019, pp. 8024–8035.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[32] D. Tran *et al.*, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. of the ICCV*, 2015, pp. 4489–4497.

[33] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. of the CVPR*, 2016, pp. 770–778.

[34] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019, pp. 6105–6114.

[35] L. Smaira *et al.*, "A short note on the kinetics-700-2020 human action dataset," *arXiv preprint arXiv:2010.10864*, 2020.