

# Audio-Visual Person Verification based on Recursive Fusion of Joint Cross-Attention

R. Gnana Praveen and Jahangir Alam  
Computer Research Institute of Montreal (CRIM), Canada

*Abstract*—Person or identity verification has been recently gaining a lot of attention using audio-visual fusion as faces and voices share close associations with each other. Conventional approaches based on audio-visual fusion rely on score-level or early feature-level fusion techniques. Though existing approaches showed improvement over unimodal systems, the potential of audio-visual fusion for person verification is not fully exploited. In this paper, we have investigated the prospect of effectively capturing both intra- and inter-modal relationships across audio and visual modalities, which can play a crucial role in significantly improving the fusion performance over unimodal systems. In particular, we introduce a recursive fusion of a joint cross-attentional model, where a joint audio-visual feature representation is employed in the cross-attention framework in a recursive fashion to progressively refine the feature representations that can efficiently capture the intra- and inter-modal relationships. To further enhance the audio-visual feature representations, we have also explored BLSTMs to improve the temporal modeling of audio-visual feature representations. Extensive experiments are conducted on the Voxceleb1 dataset to evaluate the proposed model. Results indicate that the proposed model shows promising improvement in fusion performance by adeptly capturing the intra- and inter-modal relationships across audio and visual modalities. Code is available at <https://github.com/praveena2j/RJCAforSpeakerVerification>

## I. INTRODUCTION

Person verification deals with the problem of verifying the identity of a person, which has a wide range of applications in various fields such as forensics, commercial, and law enforcement applications [1]. The task of person verification has been predominantly explored using faces [2] and speech [3] signals independently. With the advancement of deep learning models, both face- and speech-based methods have individually shown significant improvement in the performance of person verification [3]. However, relying on individual modalities may often deteriorate the performance of the system when face or speech-based signals are degraded by extreme background noise or intra-variations such as pose, low illumination, manner of speaking, etc. Therefore, leveraging the fusion of both faces and voices has been gaining momentum as multiple modalities are often expected to complement each other [4]. For instance, when speech modality is corrupted, we can rely on other modalities such as face to verify the identity of a person and vice-versa. Effectively leveraging both inter-modal complementary associations and intra-modal relationships among the audio and

visual modalities plays a crucial role in significantly outperforming unimodal approaches [5]. Inspired by the close association between faces and voices, several approaches have been explored for cross-modal verification tasks by learning a common representation for both modalities [6], [7]. Most of the existing approaches for person verification based on audio-visual (A-V) fusion focused on score-level fusion [8], [9] or early feature-level fusion [10], [11]. Score-level fusion refers to the fusion of scores obtained from vocal and facial embeddings of the individual modalities, whereas early feature-level fusion refers to the aggregation of the embeddings of audio and visual modalities using simple feature concatenation or averaging. Though these methods have improved the fusion performance over unimodal systems, they fail to leverage the rich complementary inter-modal relationships across the audio and visual modalities.

In recent years, attention-based models have been explored to efficiently capture the complementary inter-modal associations across faces and voices [12], [13]. Most of the existing attention-based models attempted to leverage the intra- and inter-modal relationships in a decoupled fashion [14]. Another line of approaches focused on dealing with noisy modalities using a weighted combination of audio and visual modalities [4], [15]. To effectively fuse audio and visual modalities, it is important to adeptly capture both intra-modal relationships (temporal dynamics of videos) and inter-modal relationships (complementarity of the modalities). Contrary to the prior approaches, we have explored joint cross-attentional fusion in a recursive fashion to simultaneously enhance the modeling of both intra- and inter-modal relationships to obtain robust A-V feature representations. Recursive attention has been previously explored successfully for emotion recognition [5] and event localization [16]. By recursively fusing the features of audio and visual modalities we can achieve more refined feature representations progressively in order to improve the performance of A-V fusion for person verification. The major contributions of the proposed approach can be summarized as follows: (1) A recursive fusion of joint cross-attentional model is introduced to efficiently capture both intra- and inter-modal relationships across faces and voices for person verification. (2) Bidirectional Long Short-Term Memory (BLSTM) networks are explored to further improve the modeling of temporal dynamics pertinent to A-V feature representations. (3) Extensive experiments are conducted on the Voxceleb1 dataset to evaluate the robustness of the proposed model.

The authors wish to acknowledge the funding from the Government of Canada's New Frontiers in Research Fund (NFRF) through grant NFRFR-2021-00338.

## II. RELATED WORK

Sari et al. [17] explored a multiview approach by transforming the individual feature representations into a common representation space and a shared classifier is used for both modalities for person verification. Chen et al. [18] leveraged the complementary information as a means of supervision to obtain robust A-V feature representations using a co-meta learning paradigm in a self-supervised learning framework. Tao et al. [4] also explored the complementary relationships between audio and visual modalities to clean noisy samples, where consistency among audio and visual modalities is used to discriminate between easy and hard samples. Another line of approaches is to deal with mitigating the impact of noisy modalities by leveraging complementary relationships. Shon et al. [15] proposed an attention mechanism to assign higher attention scores to the modality exhibiting higher discrimination. They leveraged the complementary nature across audio and visual modalities to mitigate the influence of noisy modalities. Hormann et al. [13] further extended the idea of [15] by introducing feature fusion of audio and visual modalities at intermediate layers to improve the quality of feature representations. Chen et al. [10] explored the prospect of obtaining robust feature representations by investigating various fusion strategies at the embedding level and achieved the best performance using gating-based fusion. They further exploited data augmentation strategy to deal with extremely corrupted or missing modalities.

All the above-mentioned approaches fail to leverage the cross-modal interactions to effectively capture the rich inter-modal relationships. Cross Attention (CA) has been successfully explored in several applications such as weakly supervised action localization [19], event localization [16], and emotion recognition [20], [21]. Recently, Mocanu et al. [22] explored CA based on cross-correlation across the audio and visual modalities to effectively capture the complementary inter-modal relationships for person verification. Liu et al. [14] explored CA by deploying cross-modal boosters in a pseudo-siamese structure to model one modality by exploiting the knowledge from another modality. However, they focus only on inter-modal relationships [22] or capture the intra- and inter-modal relationships in a decoupled fashion [14]. Unlike these approaches, Praveen et al. [23] introduced joint feature representation in the CA framework and improved fusion performance. In this work, we explored CA in a joint recursive fashion to simultaneously enhance the modeling of intra- and inter-modal relationships. Although [5] is closely related to our approach, the proposed approach differs primarily from [5] in two aspects: (1) In [5], RJCA is proposed for emotion recognition in the context of continuous regression, whereas we have explored it in the framework of binary classification using AAMSofmax [24] loss for person verification. (2) In [5], frame-level predictions are obtained, while in this work we have used Attentive Statistics Pooling [25] to effectively aggregate features across temporal dimensions to obtain robust A-V embeddings at the utterance level, achieving state-of-the-art results.

## III. METHODOLOGY

**A) Problem Statement:** Given an input video subsequence  $S$ , we uniformly sample  $L$  non-overlapping video segments and extract deep feature vectors from pre-trained models for audio and visual modalities. Let  $\mathbf{X}_a$  and  $\mathbf{X}_v$  denote the deep feature vectors of audio and visual modalities respectively for the given input video sub-sequence  $S$  of fixed size, which is expressed as  $\mathbf{X}_a = \{\mathbf{x}_a^1, \mathbf{x}_a^2, \dots, \mathbf{x}_a^L\} \in \mathbb{R}^{d_a \times L}$  and  $\mathbf{X}_v = \{\mathbf{x}_v^1, \mathbf{x}_v^2, \dots, \mathbf{x}_v^L\} \in \mathbb{R}^{d_v \times L}$ , where  $d_a$  and  $d_v$  represent the dimensions of the audio and visual feature vectors, respectively.  $\mathbf{x}_a^l$  and  $\mathbf{x}_v^l$  denote the audio and visual feature vectors of the video segments, respectively, for the  $l = 1, 2, \dots, L$  segments. The objective of the problem is to estimate the person verification model  $f: \mathbf{X} \rightarrow \mathbf{Y}$  from the training data  $\mathbf{X}$ , where  $\mathbf{X}$  denotes the set of audio and visual feature vectors of the input video segments and  $\mathbf{Y}$  represents the person identity of the corresponding subsequence  $S$ .

**B) Recursive Joint Cross-Attention (RJCA):** It has been shown that the performance of unified multimodal training may decline over that of individual modalities due to the differences in learning dynamics, noise topologies, etc. [26]. Therefore, we have used fixed feature vectors of audio and visual modalities to train the proposed A-V fusion model. By deploying the joint feature representation in the CA framework in a recursive fashion, we are able to simultaneously enhance the modeling of both intra- and inter-modal relationships among audio and visual modalities. The block diagram of the proposed approach is shown in Fig. 1. The joint representation  $\mathbf{J}$  is obtained by concatenating the audio and visual feature vectors as  $\mathbf{J} = [\mathbf{X}_a; \mathbf{X}_v] \in \mathbb{R}^{d \times L}$ , where  $d = d_a + d_v$  denotes the feature dimension of concatenated features. The concatenated A-V feature representations ( $\mathbf{J}$ ) of the given video subsequence ( $S$ ) are now used in the CA framework to attend to the individual modalities. This helps to incorporate both intra- and inter-modal relationships in obtaining the attention weights of audio and visual modalities. Now the correlation across joint feature representation and the individual modalities are obtained as joint cross-correlation matrix, which is given by

$$\mathbf{C}_a = \tanh\left(\frac{\mathbf{X}_a^\top \mathbf{W}_{ja} \mathbf{J}}{\sqrt{d}}\right); \mathbf{C}_v = \tanh\left(\frac{\mathbf{X}_v^\top \mathbf{W}_{jv} \mathbf{J}}{\sqrt{d}}\right) \quad (1)$$

where  $\mathbf{W}_{ja} \in \mathbb{R}^{d_a \times d}$ ,  $\mathbf{W}_{jv} \in \mathbb{R}^{d_v \times d}$  represents learnable weight matrices of audio and visual modalities respectively.

The joint correlation matrices  $\mathbf{C}_a$  and  $\mathbf{C}_v$  for audio and visual modalities help to obtain the attention weights based on the semantic relevance of both across and within the modalities. The higher the correlation coefficient, the higher the correlation across the corresponding samples within the same modality as well as across the modalities. Now the joint cross-correlation matrices are used to obtain the attention maps of audio and visual modalities, which are given by

$$\mathbf{H}_a = \text{ReLU}(\mathbf{X}_a \mathbf{W}_{ca} \mathbf{C}_a) \quad (2)$$

$$\mathbf{H}_v = \text{ReLU}(\mathbf{X}_v \mathbf{W}_{cv} \mathbf{C}_v) \quad (3)$$

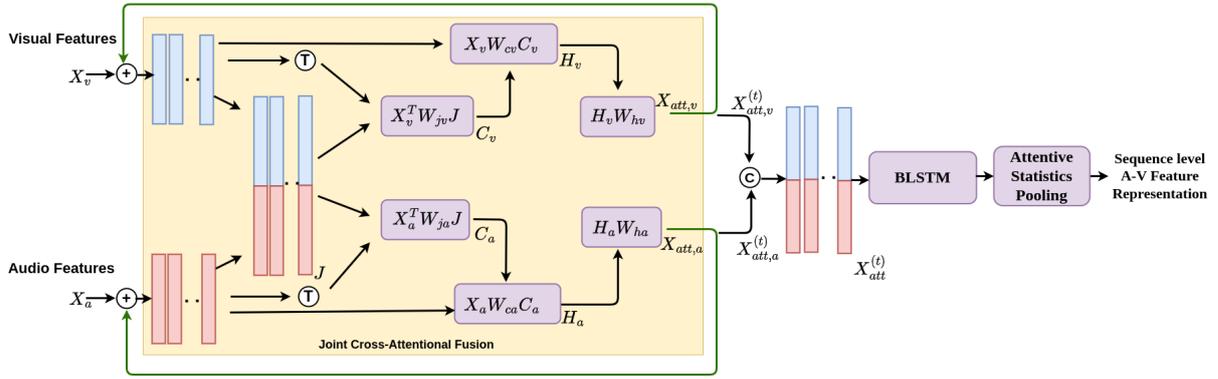


Fig. 1. Block Diagram of the Recursive Joint Cross-Attention model for A-V fusion

where  $\mathbf{W}_{ca} \in \mathbb{R}^{L \times L}$ ,  $\mathbf{W}_{cv} \in \mathbb{R}^{L \times L}$  denote learnable matrices of audio and visual modalities respectively. These attention maps are used to obtain the attended features of audio and visual modalities as follows:

$$\mathbf{X}_{att,a} = \mathbf{H}_a \mathbf{W}_{ha} + \mathbf{X}_a \quad (4)$$

$$\mathbf{X}_{att,v} = \mathbf{H}_v \mathbf{W}_{hv} + \mathbf{X}_v \quad (5)$$

where  $\mathbf{W}_{ha} \in \mathbb{R}^{L \times L}$  and  $\mathbf{W}_{hv} \in \mathbb{R}^{L \times L}$  denote the learnable weight matrices for audio and visual modalities respectively. To obtain more refined feature representations, the attended features are again fed as input to the joint cross-attentional model, which is given by

$$\mathbf{X}_{att,a}^{(t)} = \mathbf{H}_a^{(t)} \mathbf{W}_{ha}^{(t)} + \mathbf{X}_a^{(t-1)} \quad (6)$$

$$\mathbf{X}_{att,v}^{(t)} = \mathbf{H}_v^{(t)} \mathbf{W}_{hv}^{(t)} + \mathbf{X}_v^{(t-1)} \quad (7)$$

where  $\mathbf{W}_{ha}^{(t)} \in \mathbb{R}^{L \times L}$  and  $\mathbf{W}_{hv}^{(t)} \in \mathbb{R}^{L \times L}$  denote the learnable weight matrices of audio and visual modalities respectively and  $t$  refers to the recursive step.

Finally the attended audio and visual features,  $\mathbf{X}_{att,a}^{(t)}$  and  $\mathbf{X}_{att,v}^{(t)}$  obtained from the recursive fusion model are concatenated and fed to the BLSTMs followed by attentive statistics pooling (ASP) [25] to obtain semantic utterance-level representation of the A-V feature vectors. In this work, BLSTMs are used to enhance the temporal modeling of A-V feature representations. The utterance-level A-V feature representations are used to obtain the scores, where additive angular margin softmax (AAMSofmax) [24] loss function is used to optimize the parameters of the proposed model.

#### IV. RESULTS AND DISCUSSION

**A) Dataset and Evaluation Metrics:** We have evaluated the proposed model on the Voxceleb1 dataset [27], obtained from YouTube videos under challenging environments. The dataset consists of 1,48,642 video clips, captured from 1251 speakers, of which 55% of the speakers are male. Each video clip has a duration of 4 to 145 seconds, and the speakers are chosen to cover a diverse range of ethnicities, accents, professions, and ages. For our experiments, we have divided the voxceleb1 development set, which has 1211 speakers into training and validation sets. The training and

validation splits were randomly selected as 1150 and 61 speakers, respectively, and our results are reported on both the validation split and Vox1-O (Voxceleb1 original) test set for performance evaluation. It is worth mentioning that the models are trained only on the Voxceleb1 dataset.

The performance of the proposed approach is evaluated using the Equal Error Rate (EER) and minimum Detection Cost Function (minDCF), which has been widely used for speaker verification in the literature [3], [22]. Compared to EER, DCF offers control over the contribution of false positives and false negatives based on their prior probabilities [28]. In our experiments, we considered the parameters of the DCF as  $P_{target} = 0.05$ ,  $C_{miss} = 1$  and  $C_{falsealarm} = 1$  similar to that of VoxSRC-20 [29].

**B) Ablation Study:** The results are reported based on the average of three runs for statistical stability. To obtain the audio and visual feature vectors, we have used Resnet-18 [30] for visual modality and ECAPA-TDNN [31] for audio modality similar to that of [4] in order to have a fair comparison. We have conducted a series of experiments to analyze the performance of the proposed approach by comparing it with some of the widely used fusion strategies shown in Table 1. First, we have implemented a simple score-level fusion, where scores are obtained from individual modalities and then fused to verify the identity of a person. We also explored another widely used fusion strategy of feature concatenation, where the features of the audio and visual modalities are concatenated and used to obtain the final score. We can observe that the fusion performance has been improved over simple score-level fusion. To analyze the impact of the proposed approach, we further compared it with some of the widely explored attention mechanisms. By employing a self-attention mechanism on the concatenated features of individual modalities, the fusion performance has been further improved by leveraging the intra-modal relationships. We also explored inter-modal relationships across the modalities using CA and found further improvement in fusion performance over prior fusion strategies. Now, we employed joint cross-attention, where joint feature representation is integrated into the CA framework to simultaneously capture both intra- and inter-modal relationships. Finally,

TABLE I

PERFORMANCE OF VARIOUS FUSION STRATEGIES ON THE VALIDATION SET

Fusion Method	Validation Set	
	EER ↓	minDCF ↓
Score-level Fusion	2.521	0.217
Feature Concatenation	2.489	0.193
Self-Attention	2.412	0.176
Cross-Attention	2.387	0.149
Joint Cross-Attention	2.315	0.135
RJCA (w/o BLSTMs)	1.946	0.128
RJCA (w/ BLSTMs)	<b>1.851</b>	<b>0.112</b>

TABLE II

PERFORMANCE OF PROPOSED MODEL WITH VARYING NUMBER OF ITERATIONS AND BLSTM ON THE VALIDATION SET

Fusion Method	Validation Set	
	EER ↓	minDCF ↓
<b>RJCA Fusion w/o Recursion</b>		
RJCA Fusion w/o BLSTMs	2.315	0.135
RJCA Fusion w/ BLSTMs	2.195	0.132
<b>RJCA Fusion w/ Recursion</b>		
RJCA Fusion w/o BLSTMs, $t = 3$	1.946	0.128
RJCA Fusion w/ BLSTMs, $t = 2$	2.029	0.119
RJCA Fusion w/ BLSTMs, $t = 3$	<b>1.851</b>	<b>0.112</b>
RJCA Fusion w/ BLSTMs, $t = 4$	1.982	0.124
RJCA Fusion w/ BLSTMs, $t = 5$	2.159	0.137

we have introduced the proposed recursive fusion with joint CA to obtain the attended features of individual modalities, and we can observe that recursive fusion helps in obtaining more refined feature representations and achieves the best performance among all the fusion strategies.

To further analyze the contribution of the individual components of the proposed approach, we conducted another series of experiments with BLSTMs and varying the number of iterations of the recursive mechanism. First, we performed two experiments without recursion, one with BLSTM and another without BLSTM to understand the contribution of BLSTM to the proposed approach. Table II shows the results of the experiments on the validation set of the Voxceleb1 dataset. We can observe that BLSTMs improve the fusion performance by taking advantage of the temporal modeling of the A-V feature representations. To understand the impact of the recursive fusion, we performed experiments by varying the number of iterations and obtained the best performance at 3 iterations. Beyond that, we observe a decline in the fusion performance, which can be attributed to the fact that though recursion helps in improving fusion performance initially, it results in overfitting with more iterations, resulting in a performance decline on the validation set.

**C) Comparison to state-of-the-art:** We have used the Voxceleb1 dataset to validate the proposed approach and compared it with the relevant state-of-the-art methods using A-V fusion for person verification. To have a fair comparison, we have reimplemented the state-of-the-art methods using our experimental setup due to the differences in their experimental protocols. Table III shows the comparison of the proposed approach with state-of-the-art methods and

TABLE III

PERFORMANCE OF THE PROPOSED APPROACH IN COMPARISON TO STATE-OF-THE-ART ON THE VALIDATION SET AND VOX1-O SET

Fusion Method	Validation Set		Vox1-O Set	
	EER ↓	minDCF ↓	EER ↓	minDCF ↓
Visual	3.720	0.298	3.779	0.274
Audio	2.553	0.253	2.529	0.228
Tao et al [4]	2.476	0.203	2.409	0.198
Sari et al [17]	2.438	0.178	2.397	0.169
Chen et al [10]	2.403	0.163	2.376	0.161
Mocanu et al [22]	2.387	0.149	2.355	0.156
Praveen et al [23]	2.173	0.126	2.214	0.129
RJCA (w/ BLSTMs)	<b>1.851</b>	<b>0.112</b>	<b>1.975</b>	<b>0.116</b>

individual modalities on both the validation split of the Voxceleb1 and Vox1-O partitions. First, we have conducted experiments with the individual modalities and found that audio performs relatively better than visual modality.

Tao et al. [4] explored complementary relationships between the audio and visual modalities as supervisory information to mitigate the impact of noisy labels. We have used the clean samples obtained using [4] to train our models as well as other methods [10], [17], [22], [23]. Most of the existing approaches [4], [10], [17] rely on a single image (randomly chosen), thereby focusing on the inter-modal relationships, failing to retain the intra-modal relationships. Both of these approaches [17] and [10] are able to improve the fusion performance better than that of [4], where [17] uses a shared representation space using a shared classifier and [10] explored the gating mechanism. Mocanu et al. [22] also focused on capturing the inter-modal relationships based on cross-correlation across the modalities. Unlike these approaches, Praveen et al. [23] showed improvement by capturing both intra- and inter-modal relationships simultaneously using joint feature representation in the CA framework. We have further improved the fusion performance by introducing recursive fusion to obtain more refined feature representations. Since the proposed RJCA model leverages both intra- and inter-modal relationships effectively, we can observe that the fusion performance has been improved over other state-of-the-art methods.

## V. CONCLUSION

In this paper, we have presented a novel approach of recursive joint CA fusion for person verification by effectively exploiting both inter- and intra-modal relationships across audio and visual modalities. Specifically, we have explored the joint feature representation in the CA framework recursively to obtain more refined A-V feature representations. To further enhance the intra-modal relationships of A-V feature representations, we have used BLSTMs on the refined A-V feature representations obtained from the proposed RJCA model. We showed that the fusion performance can be improved by effectively capturing the intra- and inter-modal relationships. The performance of the proposed approach can be further enhanced by training with the large-scale Voxceleb2 dataset as it can improve the generalization ability of the proposed approach.

## REFERENCES

- [1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.
- [2] M. Wang and W. Deng, "Deep face recognition: A survey," *Neuro-computing*, vol. 429, pp. 215–244, 2021.
- [3] Y. Tu, W. Lin, and M.-W. Mak, "A survey on text-dependent and text-independent speaker verification," *IEEE Access*, vol. 10, pp. 99038–99049, 2022.
- [4] R. Tao, K. A. Lee, Z. Shi, and H. Li, "Speaker recognition with two-step multi-modal deep cleansing," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.
- [5] R. G. Praveen, E. Granger, and P. Cardinal, "Recursive joint attention for audio-visual fusion in regression based emotion recognition," in *IEEE ICASSP*, pp. 1–5, 2023.
- [6] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *IEEE CVPR*, 2018.
- [7] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proc. of ECCV*, 2018.
- [8] Seyed, C. Greenberg, E. Singer, D. Olson, L. Mason, and J. Hernandez-Cordero, "The 2019 nist audio-visual speaker recognition evaluation," *The Speaker and Language Recognition Workshop: Odyssey 2020*, Tokyo, -1, 2020-05-18 2020.
- [9] J. Alam, G. Boulianne, L. Burget, M. Dahmane, M. S. Diez, O. Glembeek, M. Lalonde, A. D. Lozano, P. Matějka, P. Mizera, L. Mošner, C. Noisieux, J. Monteiro, O. Novotný, O. Plchot, A. J. Rohdin, A. Silnova, J. Slavíček, T. Stafylakis, P.-L. St-Charles, S. Wang, and H. Zeinali, "Analysis of abc submission to nist sre 2019 cmn and vast challenge," in *Proceedings of Odyssey 2020 The Speaker and Language Recognition Workshop*, vol. 2020, pp. 289–295, International Speech Communication Association, 2020.
- [10] Z. Chen, S. Wang, and Y. Qian, "Multi-modality matters: A performance leap on voxceleb," in *Proc. Interspeech*, pp. 2252–2256, 2020.
- [11] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1079–1092, 2021.
- [12] P. Sun, S. Zhang, Z. Liu, Y. Yuan, T. Zhang, H. Zhang, and P. Hu, "A Method of Audio-Visual Person Verification by Mining Connections between Time Series," in *Proc. INTERSPEECH 2023*, pp. 3227–3231, 2023.
- [13] S. Hörmann, A. Moiz, M. Knoche, and G. Rigoll, "Attention fusion for audio-visual person verification using multi-scale features," in *IEEE FG*, pp. 281–285, 2020.
- [14] M. Liu, K. A. Lee, L. Wang, H. Zhang, C. Zeng, and J. Dang, "Cross-modal audio-visual co-learning for text-independent speaker verification," in *IEEE ICASSP*, pp. 1–5, 2023.
- [15] S. Shon, T.-H. Oh, and J. Glass, "Noise-tolerant audio-visual online person verification using an attention-based neural network fusion," in *IEEE ICASSP*, pp. 3995–3999, 2019.
- [16] B. Duan, H. Tang, W. Wang, Z. Zong, G. Yang, and Y. Yan, "Audio-visual event localization via recursive fusion by joint co-attention," in *IEEE WACV*, pp. 4012–4021, 2021.
- [17] L. Sari, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A multi-view approach to audio-visual speaker verification," in *IEEE ICASSP*, pp. 6194–6198, 2021.
- [18] H. Chen, H. Zhang, L. Wang, K. A. Lee, M. Liu, and J. Dang, "Self-supervised audio-visual speaker representation with co-meta learning," in *IEEE ICASSP*, pp. 1–5, 2023.
- [19] J.-T. Lee, M. Jain, H. Park, and S. Yun, "Cross-attentional audio-visual fusion for weakly-supervised action localization," in *Proc. of the ICLR*, 2021.
- [20] R. G. Praveen, E. Granger, and P. Cardinal, "Cross attentional audio-visual fusion for dimensional emotion recognition," in *IEEE FG*, pp. 1–8, 2021.
- [21] R. G. Praveen, P. Cardinal, and E. Granger, "Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pp. 1–1, 2023.
- [22] B. Mocanu and T. Ruxandra, "Active speaker recognition using cross attention audio-video fusion," in *Proc. of the EUVIP*, pp. 1–6, 2022.
- [23] G. P. Rajasekhar and J. Alam, "Audio-visual speaker verification via joint cross-attention," in *Speech and Computer: 25th International Conference, SPECOM 2023, Dharwad, India, November 29 – December 2, 2023, Proceedings, Part II*, p. 18–31, 2023.
- [24] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE/CVF Conference on CVPR*, pp. 4685–4694, 2019.
- [25] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech*, pp. 2252–2256, 2018.
- [26] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?," in *CVPR*, pp. 12692–12702, 2020.
- [27] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.
- [28] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006.
- [29] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "Voxsrc 2020: The second voxceleb speaker recognition challenge," *arXiv preprint arXiv:2012.06867*, 2020.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.
- [31] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, pp. 3830–3834, 2020.