

Audio-visual Speaker Recognition with a Cross-modal Discriminative Network

*Ruijie Tao*¹, *Rohan Kumar Das*¹ and *Haizhou Li*^{1,2}

¹Department of Electrical and Computer Engineering, National University of Singapore, Singapore ²Machine Listening Lab, University of Bremen, Germany

ruijie.tao@u.nus.edu.sg, {rohankd, haizhou.li}@nus.edu.sg

Abstract

Audio-visual speaker recognition is one of the tasks in the recent 2019 NIST speaker recognition evaluation (SRE). Studies in neuroscience and computer science all point to the fact that vision and auditory neural signals interact in the cognitive process. This motivated us to study a cross-modal network, namely voice-face discriminative network (VFNet) that establishes the general relation between human voice and face. Experiments show that VFNet provides additional speaker discriminative information. With VFNet, we achieve 16.54% equal error rate relative reduction over the score level fusion audio-visual baseline on evaluation set of 2019 NIST SRE.

Index Terms: Audio-visual speaker recognition evaluation, cross-modal verification, multimedia, SRE 2019

1. Introduction

Speaker recognition has enabled many real-world applications [1–4]. These systems are expected to perform effectively under adverse conditions. The NIST speaker recognition evaluation (SRE)s are organized to benchmark systems in different such scenarios [5]. Various robust systems are developed in the past that perform effectively and provides state-of-the-art [6, 7]. Unlike previous SREs, the 2019 NIST SRE investigated a new direction on audio-visual (AV) SRE [8]. The evaluation task deals with verifying the claimed identity of a person for a given pair of enrollment and test videos. In other words, it advocates the use of audio-visual cues for improved speaker recognition in real-world scenarios.

The significance of processing multimedia in other fields has increased in the recent years [9]. The latest audio-visual SRE can be viewed as one such outcome following this trend. Some of the other tasks considering multimedia instead of single modality using speech are automatic speech recognition (ASR) [10], speech separation [11] and speech diarization [12]. The studies in these works exploited the association of audio and visual cues adequately. For instance, in audio-visual ASR, lip language recognition is used to support ASR systems; in audio-visual speech separation, the movement of mouth can assist detecting who is speaking when.

While coming to audio-visual SRE, the simplest way to perform multimedia based speaker recognition is to have separate systems for audio and visual inputs, then combine the results of speaker and face recognition systems [8, 13]. We note separating audio-visual SRE into two sub-tasks is a straight forward approach to simplify the problem. However, the two subsystems are disjoint and one does not consider the knowledge from other. It is further worth emphasizing that the motivation to process multimedia for SRE is not only to add another visual system but also to explore the relationship between the audio and video. Therefore, disregarding the association between different modalities may result in the loss of some information.

The studies in neuroscience show that humans associate the voice and the face of a person in the memory [14]. While listening to a voice of an individual, one can select the right static face corresponding to the same person between two static faces at a higher than chance level and vice versa [15-18]. In computer science, there has been study on cross-modal biometric matching [19]. Further, various works use pair-wise loss-based methods to improve the cross-modal system performance [20, 21]. The learned associations between audio and visual cues are general identity features (such as gender, age and ethnicity) and appearance features (such as big nose, chubby and double chin) [22, 23]. The existing works utilize both joint and disjoint general information between two modalities to train the cross-modal verification network to determine if the given face and speech segment belongs to the same identity for verification tasks [24-26]. We believe that the general cross-modal discriminative features provide additional information in audio-visual speaker recognition.

In this work, we propose a cross-modal discriminative network, that is called voice-face network (VFNet), to learn the association between voice and face. VFNet is trained using speaker and face embeddings collectively that are extracted from separate systems. We consider x-vector and InsightFace based systems for extracting the speaker and face embeddings, respectively [27, 28]. Further, these two systems are used for SRE using audio as well as visual input based single systems, followed by their fusion for a baseline audio-visual system. The output of VFNet is used to represent the general association between voice and face. The speaker recognition studies are conducted on 2019 NIST SRE corpus. The contributions of this paper include the novel idea of cross-modal discriminative network, and its use in audio-visual speaker recognition study.

The rest of the paper is organized as follows. Section 2 describes the proposed VFNet based cross-modal verification system. In Section 3, we present the audio-visual speaker recognition with cross-modal verification. Section 4 and Section 5 reports the experiments and their results, respectively. Finally, Section 6 concludes the work.

2. Cross-modal Verification

This section describes the proposed VFNet for cross-modal verification. Figure 1 shows the architecture of VFNet that considers two inputs: a voice waveform and a human face. The output of the network is a confidence score to describe whether the voice and the face come from the same person. We now discuss the detailed pipeline of VFNet.

First, the speaker and the face embeddings are extracted by x-vector and InsightFace models, respectively [27,28]. As both these embeddings represent information from different modality, they are fed to a 256-D fully connected layer (FC1) with the rectified linear unit (ReLU) activation, then followed by an-



Figure 1: Architecture of the proposed cross-modal discrimination network, VFNet, that relates the voice and face of a person.

other 128-D fully connected layer (FC2) without the ReLU. These layers are introduced to lead the speaker and face embeddings for learning the cross-modal identity information from each other. Further, they help to project the embeddings from both modalities into a new domain, where their relation can be established.

For a given pair of speaker embedding e_v and a face embedding e_f , their transformed embeddings $T_v(e_v)$ and $T_f(e_f)$ are derived from VFNet, followed by the cosine similarity scoring $S(T_v(e_v), T_f(e_f))$ between them. We also have $1 - S(T_v(e_v), T_f(e_f))$ to represent the negative voice-face pair. By using softmax function based on these two scores, the output of VFNet is obtained as

$$p_1 = \frac{e^{S(T_v(e_v), T_f(e_f))}}{e^{S(T_v(e_v), T_f(e_f))} + e^{1 - S(T_v(e_v), T_f(e_f))}}$$
(1)

$$p_2 = \frac{e^{1-S(T_v(e_v),T_f(e_f))}}{e^{S(T_v(e_v),T_f(e_f))} + e^{1-S(T_v(e_v),T_f(e_f))}}$$
(2)

where final output p_1 is the score to describe the probability that the voice and the face belong to the same person, p_2 being the score depicting the probability that the voice and the face do not belong to the same person. Finally, we feed our predictions pand the ground truth verification labels \hat{p} to optimize the crossentropy loss $\mathcal{L}_{\hat{p}}(p)$ as follows

$$\mathcal{L}_{\hat{p}}(p) = -\sum_{i} \hat{p}_i \log(p_i) \tag{3}$$

3. Audio-visual Speaker Recognition with Cross-modal Verification

In audio-visual SRE, an enrollment video provides the target individual's biometric information (voice and face) and the assignment asks the model to automatically determine whether the target person is present in a given test video [8]. Figure 2 shows the proposed audio-visual speaker recognition framework with VFNet on the left panel, and the baseline, a voice-face score level fusion system, on the right panel. The given voice segments of the target speakers from the enrollment utterances and the entire test utterances are considered for extracting the speaker embeddings using x-vector system [27]. Similarly, the InsightFace system extracts the face embeddings for given faces of the target speakers from the enrollment videos and all detected faces from the test videos [28].

On the left panel, the VFNet system provides an association score between the target speaker voice in the enrollment and the detected faces from the test video. Matching pairs between voice and face will give rise to high association, while mismatches, such as age, gender, and ethnicity discrepancy, will do otherwise.

On the right panel, the audio and visual systems run in parallel to verify the claimed identity by computing match between the enrollment and the test embeddings. We note that the speaker recognition system considers probabilistic linear discriminant (PLDA) based likelihood scores, whereas the face recognition system computes cosine similarity scores. We consider the baseline system as a score level fusion between the two parallel systems.

We propose to fuse the VFNet score and the baseline audiovisual system as shown in Figure 2 for a final decision. The score level fusion is performed using logistic regression for various systems discussed in this work. We report performance of the VFNet, baseline systems, and the overall system separately in the experiments.

4. Experiments

In this section, the details of the audio and visual systems developed in this work are mentioned. The database, embedding extraction and audio-visual speaker recognition systems are described in the following subsections.

4.1. Database

We consider the original videos corresponding to VoxCeleb2 corpus to derive a set with voices and faces for cross-modal verification [29]. For each video, the entire audio is extracted to represent the voice of the speaker. On the contrary, we perform a face detection on each video and then consider the most prominent faces representing an individual. For cross-modal discriminative training, the positive trials are faces and voices come from the same identity, whereas the negative trials are obtained by shuffling the faces and voices belonging to different persons. A summary of VoxCeleb2 corpus used for cross-modal verification is shown in Table 1. VFNet learns the general association between voice and face from VoxCeleb2.

We consider 2019 NIST SRE audio-visual corpus as the speaker recognition application [8]. The evaluation set has provided manually marked diarization labels for voice and keyframe indices along with target speaker's face bounding boxes in the enrollment videos. However, no such information is provided for the test segments. A summary of this corpus is also shown in Table 1. We note that to maximize the usage of cross-modal information in this corpus, we extract the speaker and face embeddings of the target person from the enrollment segments of development set, then combine them with that of VoxCeleb2 to retrain the VFNet.



Figure 2: Block diagram of proposed audio-visual (AV) speaker recognition framework with VFNet, where VFNet provides voice-face cross-modal verification information that strengthens the baseline audio-visual speaker recognition decision.

Table 1: Summary of VoxCeleb2 and 2019 SRE audio-visual (AV) corpora.

VoxCeleb2	Train	Test	
# identity	5,994	108	
# faces	1,088,047	36,166	
# voices	1,092,009	36,237	
# cross-modal trials	2,176,094	72,332	
2019 SRE (AV)	Development	Evaluation	
# enroll segments	52	149	
# test segments	108	452	
# target trials	108	452	
# non-target trials	5,508	66,896	

4.2. Embedding Extraction

We use an x-vector based system to extract the speaker embeddings [27]. The speech utterances are processed with an energy based voice activity detection to remove the non-silence regions and 30-dimensional mel frequency cepstral coefficient (MFCC) features are extracted. In addition, a short-time cepstral mean normalization is applied over a 3-second sliding window. The x-vector extractor is trained using VoxCeleb1-2 corpora and the detailed settings of x-vector network architecture can be found in [30].

For extracting the face embeddings, we first use the ResNet-50 RetinaFace model trained using WIDER FACE database [31] for detecting the faces [32] followed by multi-task cascaded convolutional network (MTCNN) [33] to align them. We then use InsightFace to obtain highly discriminative features for face recognition by using the additive angular margin loss [28]. In addition, it consists of ResNet-100 extractor model trained on cleaned MS1MV2 database [34] to extract the face embeddings.

The dimension of both speaker and face embeddings are

kept as 512 in our studies. The speaker and face embeddings for VFNet to perform cross-modal verification follow the same pipeline discussed above.

4.3. Audio-visual Speaker Recognition

Although the dimensions of speaker and face embeddings are the same, the back-end scoring for respective individual system is different. We use linear discriminant analysis (LDA) on speaker embeddings for channel/session compensation and reduce the dimension of x-vectors to 150. Finally, PLDA is used as a classifier to get the final speaker recognition score. On the other hand, cosine similarity between face embeddings from enrollment video and that from the detected faces in the test video are computed. Finally, the average of top 20% scores of the number of face embeddings in the test video are taken to derive the final face recognition score.

We now focus on the back-end of audio-visual SRE with VFNet. The VFNet back-end computes the likelihood score between the speaker embedding of the target speaker and all the face embeddings of detected faces in the test video as given by Equation (1). Finally, the average of top 20% scores is taken, which is then combined with the scores generated from audio and visual systems by logistic regression. It is to be noted that cross-modal verification can also be done by considering the all the given faces in the test audio. However, it requires an additional speaker diarization module to detect the voice belonging to different speakers in the test audio. Therefore, we follow the former approach for audio-visual SRE with VFNet.

We use Bosaris toolkit [35] to calibrate and fuse the scores of different systems. The performance of systems are reported in terms of equal error rate (EER), minimum detection cost function (minDCF) and actual detection cost function (actDCF) following the protocol of 2019 NIST SRE [8].

Table 2: Comparative study between the proposed VFNet and other systems in cross-modal verification.

Model	EER (%)	AUC (%)	Database
DIMNet [25]	24.56	NA	VoxCeleb, VGGFace
SSNet [26]	29.5	78.8	VoxCeleb
Pins [24]	29.6	78.5	VoxCeleb
VFNet	22.52	85.4	VoxCeleb2

Table 3: Comparative study between the proposed VFNet and other systems in cross-modal matching. V-F and F-V refer to cases that consider reference modality as voice and face, respectively.

Model	Accura V-F	acy (%) F-V	Database
SSNet [26]	78	NA	VoxCeleb
Horiguchi's [20]	78.1	77.8	VoxCeleb, MS-Celeb-1M
Kim's [22]	78.2	78.6	VoxCeleb
SVHF [19]	81.0	79.5	VoxCeleb, VGGFace
Pins [24]	84	NA	VoxCeleb
DIMNet [25]	84.12	84.03	VoxCeleb, VGGFace
VFMR [21]	84.48	NA	VoxCeleb2, VGGFace2
VFNet	85.39	86.12	VoxCeleb2

5. Results and Analysis

5.1. Cross-modal Verification Studies

We evaluate the performance of proposed VFNet on VoxCeleb2 corpus for cross-modal verification studies and compare with some of the existing systems. The performance comparison is shown in Table 2, where EER and area under the ROC curve (AUC) are considered as the performance metrics.

We observe that VFNet performs effectively for crossmodal verification. Further, it is to be noted that we do not claim from this study that VFNet outperforms other systems as the results are evaluated on different corpora. We rather try to show that VFNet alone is comparable to the existing systems for cross-modal verification. For brevity, we do not go through the details of various systems [24–26] considered for cross-modal verification.

Further, to measure VFNet performance more comprehensively, we extend the studies for cross-modal matching task. For a given human voice and two static faces, this task aims to find the more inclined face to the voice, and vice versa. We note that this task also relates to 2019 NIST audio-visual SRE as there are multiple speakers present in the test videos that have to be matched with the target speaker in the enrollment video. For this cross-modal matching study, we add one more shared weights sub-branch to the original VFNet model for the selection requirements. The performance of VFNet thus obtained and its comparison to some of the other systems for cross-modal matching task in terms of accuracy is shown in Table 3. We observe that the effectiveness of VFNet holds good for crossmodal matching task as well and the performance is comparable to other systems.

5.2. Audio-visual SRE with VFNet Studies

We now study audio-visual speaker recognition with VFNet. To show the effect of VFNet, we first fuse the single modality Table 4: Performance comparison of various systems on 2019 NIST SRE audio-visual corpus.

Development Set						
System	EER (%)	minDCF	actDCF			
Speaker Recognition with VFNet	$08.62 \\ 09.82$	$0.367 \\ 0.365$	$0.399 \\ 0.393$			
Face Recognition with VFNet	$04.52 \\ 03.85$	$0.349 \\ 0.324$	$0.371 \\ 0.355$			
Audio-visual SRE with VFNet	03.70 03.20	0.141 0.141	0.166 0.141			
Evaluation Set						
Speaker Recognition with VFNet	$06.36 \\ 05.79$	$0.326 \\ 0.317$	$0.339 \\ 0.320$			
Face Recognition with VFNet	$01.77 \\ 01.66$	$0.074 \\ 0.073$	$0.098 \\ 0.094$			
Audio-visual SRE with VFNet	01.33 01.11	0.050 0.049	0.068 0.062			

speaker, face recognition systems with VFNet. Table 4 reports the performance comparison of various systems and with and without VFNet.

Examining the effect on single modality systems, we find that the contribution of VFNet is more evident for speaker recognition system on the evaluation set. Further, the VFNet is also able to enhance the audio-visual baseline system performance that suggests usefulness of associating audio and visual cues by cross-modal verification for audio-visual SRE. We obtain relative improvements of 16.54%, 2.00% and 8.83% in terms of EER, minDCF and actDCF, respectively.

6. Conclusions

In this work, we propose a novel framework for audio-visual speaker recognition with cross-modal discrimination network. The VFNet based cross-modal discrimination network finds the relation between a given pair of human voice and face to generate a confidence score if they correspond to the same person. While VFNet can perform comparable to the existing state-ofthe-art cross-modal verification systems, the proposed framework of audio-visual speaker recognition with VFNet outperforms the baseline audio-visual system. This highlights the importance of cross-modal verification, in other words, the relation between audio and visual cues for audio-visual speaker recognition.

7. Acknowledgements

This research work is partially supported by Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain), and in part by Human-Robot Interaction Phase 1 (Grant No. 192 25 00054) from the National Research Foundation, Prime Minister's Office, Singapore under the National Robotics Programme.

8. References

- K. A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, "Joint application of speech and speaker recognition for automation and security in smart home," in *Proc. of Interspeech*, 2011, pp. 3317–3318.
- [2] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," in *SLTC Newsletter*, February 2013.
- [3] R. K. Das, S. Jelil, and S. R. M. Prasanna, "Development of multilevel speech based person authentication system," *Journal of Signal Processing Systems*, vol. 88, no. 3, pp. 259–271, 2017.
- [4] S. Jelil, A. Shrivastava, R. K. Das, S. R. M. Prasanna, and R. Sinha, "SpeechMarker: A voice based multi-level attendance application," in *Proceedings Interspeech*, 2019, pp. 3665–3666.
- [5] C. S. Greenberg, L. P. Mason, S. O. Sadjadi, and D. A. Reynolds, "Two decades of speaker recognition evaluation at the National Institute of Standards and Technology," *Computer Speech & Language*, vol. 60, 2020.
- [6] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, R. Dehak, L. P. García-Perera, D. Povey, P. A. Torres-Carrasquillo, S. Khudanpur, and N. Dehak, "State-of-the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST SRE18," in *Proc. of Interspeech*, 2019, pp. 1488–1492.
- [7] K. A. Lee, V. Hautamäki, T. H. Kinnunen, H. Yamamoto, K. Okabe, V. Vestman, J. Huang, G. Ding, H. Sun, A. Larcher, R. K. Das, H. Li, M. Rouvier, P.-M. Bousquet, W. Rao, Q. Wang, C. Zhang, F. Bahmaninezhad, H. Delgado, and M. Todisco, "14U submission to NIST SRE 2018: Leveraging from a decade of shared experiences," in *Proc. of Interspeech*, 2019, pp. 1497– 1501.
- [8] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. Mason, and J. Hernandez-Cordero, "The 2019 NIST audio-visual speaker recognition evaluation," in *Proc. of Odyssey*, 2020, pp. 266–272.
- [9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
- [10] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.
- [11] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," ACM Transactions on Graphics, vol. 37, no. 4, pp. 112:1–112:11, 2018.
- [12] K. Hoover, S. Chaudhuri, C. Pantofaru, M. Slaney, and I. Sturdy, "Putting a face to the voice: Fusing audio and visual signals across a video to determine speakers," *CoRR*, vol. abs/1706.00079, 2017.
- [13] R. K. Das, R. Tao, J. Yang, W. Rao, C. Yu, and H. Li, "HLT-NUS submission for 2019 NIST multimedia speaker recognition evaluation," in *Submitted to APSIPA ASC*, 2020.
- [14] K. v. Kriegstein, A. Kleinschmidt, P. Sterzer, and A.-L. Giraud, "Interaction of face and voice areas during speaker recognition," *Journal of cognitive neuroscience*, vol. 17, no. 3, pp. 367–376, 2005.
- [15] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "'Putting the face to the voice': Matching identity across modality," *Current Biology*, vol. 13, pp. 1709–1714, 2003.
- [16] L. W. Mavica and E. Barenholtz, "Matching voice and face identity from static images." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 2, pp. 307–312, 2013.
- [17] H. M. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey, "Matching novel face and voice identity using static and dynamic facial images," *Attention, Perception, & Psychophysics*, vol. 78, no. 3, pp. 868–879, 2016.

- [18] —, "Concordant cues in faces and voices: Testing the backup signal hypothesis," *Evolutionary Psychology*, vol. 14, no. 1, pp. 1–10, 2016.
- [19] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2018, pp. 8427–8436.
- [20] S. Horiguchi, N. Kanda, and K. Nagamatsu, "Face-voice matching using cross-modal embeddings," in *Proc. of the ACM International Conference on Multimedia*, 2018, pp. 1011–1019.
- [21] C. Xiong, D. Zhang, T. Liu, and X. Du, "Voice-face crossmodal matching and retrieval: A benchmark," arXiv preprint arXiv:1911.09338, 2019.
- [22] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Proc. of Asian Conference on Computer Vision (ACCV)*, 2018, pp. 276– 292.
- [23] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2Face: Learning the face behind a voice," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7539–7548.
- [24] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable PINs: Cross-modal embeddings for person identity," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2018, pp. 71–88.
- [25] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2019.
- [26] S. Nawaz, M. K. Janjua, I. Gallo, A. Mahmood, and A. Calefati, "Deep latent space learning for cross-modal mapping of audio and visual signals," in *Digital Image Computing: Techniques and Applications (DICTA)*, 2019, pp. 1–7.
- [27] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [28] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019, pp. 4690–4699.
- [29] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. of Interspeech*, 2018, pp. 1086– 1090.
- [30] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5796–5800.
- [31] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 5525– 5533.
- [32] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *arXiv* preprint arXiv:1905.00641, 2019.
- [33] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
- [34] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc.* of the European Conference on Computer Vision (ECCV), 2016, pp. 87–102.
- [35] N. Brümmer and E. de Villiers, "The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF," *CoRR*, vol. abs/1304.2865, 2013.