INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Method of Audio-Visual Person Verification by Mining Connections between Time Series

Peiwen Sun[1], *Shanshan Zhang*[2], *Zishan Liu*[1], *Yougen Yuan*[2], *Taotao Zhang*[2], *Honggang Zhang*[1], *Pengfei Hu*[2]

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China [2] TEG AI, Tencent Inc, Beijing, China

[1]{sunpeiwen,shana_a,zhhg}@bupt.edu.cn,
[2]{susanszhang,yougenyuan,tttaozhang,alanpfhu}@tencent.com

Abstract

It has already been observed that audio-visual embedding is more robust than uni-modality embedding for person verification. But the relationship of keyframes in time series between modalities seems to be unexplored. Hence, we proposed a novel audio-visual strategy that considers connections between time series from a generative perspective. First, we introduced weight-enhanced attentive statistics pooling to extend the salience of the keyframe weights. Then, joint attentive pooling incorporating 3 popular generative supervision models is proposed. Finally, each modality is fused with a gated attention mechanism to gain robust embedding. All the proposed models are trained on the VoxCeleb2 dev dataset and the best system obtains 0.14%, 0.21%, and 0.37% EER on three official trial lists of VoxCeleb1 respectively, which is to our knowledge the best-published results for person verification.

Index Terms: person verification, audio-visual, generative model

1. Introduction

Biometrics-based person verification technologies are widely used in access control. In the wild, the speech segments are corrupted with real-world noise including laughter, music, and other sounds. Similarly, face images have variations in pose, image quality, and motion blur. It creates additional challenges for the verification system.

The development of verification in Voxceleb [1, 2] first appeared in speaker verification. The research [3] proposed attentive statistics pooling to focus on important frames of speaker verification and get higher discriminative ability than the traditional averaging method. In further research, ECAPA-TDNN [4] was based on blocks of TDNNs and Squeeze-Excitation(SE) [5] to reconstruct frame-level features. Meanwhile, The margin-based softmax loss originating from face recognition is widely used for training models [6, 7, 8, 9]. These losses were also customized to adapt to the speaker verification task [10, 11] and achieved SOTA at that time. Recent advances driven by large-scale pre-trained models have taken the task of speaker recognition to a new level. With pre-training, Superb [12], Unispeech [13], Wavlm [14], HuBERT [15] have achieved excellent performances on multiple sets.

The performance of speaker systems would degrade dramatically under wild circumstances [16]. To solve the problem, researchers have found that simply fusing the scores from speaker embeddings and facial embeddings can yield good results [17, 18]. For the first time, S. Shon [16] attempted to fuse audio-visual information with deep learning-based models to achieve better results. As the transformer has been proposed, it seemed more efficient [19] to fuse different modalities in large datasets. Previous studies have sought implicit feature expression and supervision patterns from the perspective of network structure. To further explore the explicit correlations, Y. Liang [20] gave person verification a point of view of the HGR maximal correlation. Most of the research on audio-visual works [21, 19] still focused on extracting frame-level embeddings and then simply averaging them as an aggregator to get segment-level embeddings.

Our network is reconsidered from the perspective that the weight of each frame in time series of the different modalities are different expression styles of the same facial action. As a prerequisite, we enhance the frame weight of the attention statistics pooling [3] to adjust the situation of the audio-visual training process. Then, 3 popular generative models inspired by Cycle Consistency [22], Patch NCE [23], Dual Diffusion Implicit Bridges(DDIB) [24] is introduced into the aggregator for transferring styles and monitoring inference. The method that combines weight-enhanced attentive statistics pooling and generative supervision is called joint attentive pooling. Each frame contributes according to its relevance to the final representation avoiding equal contributions and preventing accumulated errors. Finally, All three generative supervised methods demonstrate unique robustness in performance, which can be corroborated by mining correlations on different modal time series.

2. Method

The overall network (Fig.1) consists of the backbone of each modality (Sec.3), joint attentive pooling module, and fusion module (Sec.2.3). The joint attentive pooling module is the combination of weight-enhanced attentive statistical pooling in Sec.2.1 and generative supervision Sec.2.2

2.1. Weight-enhanced attentive statistical pooling

The data of each modality are pre-processed as stated in Sec.3 and sent to the respective encoders. But when visual information is added for joint training, there is a serious drop. To solve the problem, we introduce weight enhanced method of attentive statistical pooling. The detailed calculation process of weight enhanced method is as follows.

Through the original attentive statistical pooling mechanism [4], each frame will be given different weights.

$$e^{t,c} = (\boldsymbol{v}^c)^T tanh\left(\boldsymbol{W}\boldsymbol{h}^t + \boldsymbol{b}\right) + k^c, \tag{1}$$

where h^t denotes the weight of the last frame layer at time step t. The parameters $\boldsymbol{W} \in \mathbb{R}^{R\times C}$ and $\boldsymbol{b} \in \mathbb{R}^{R\check{\times}1}$ project the information for self-attention into a smaller $\mathbb{R}$-dimensional representation that is shared across all C channels to reduce the parameter count and risk of overfitting.

10.21437/Interspeech.2023-394

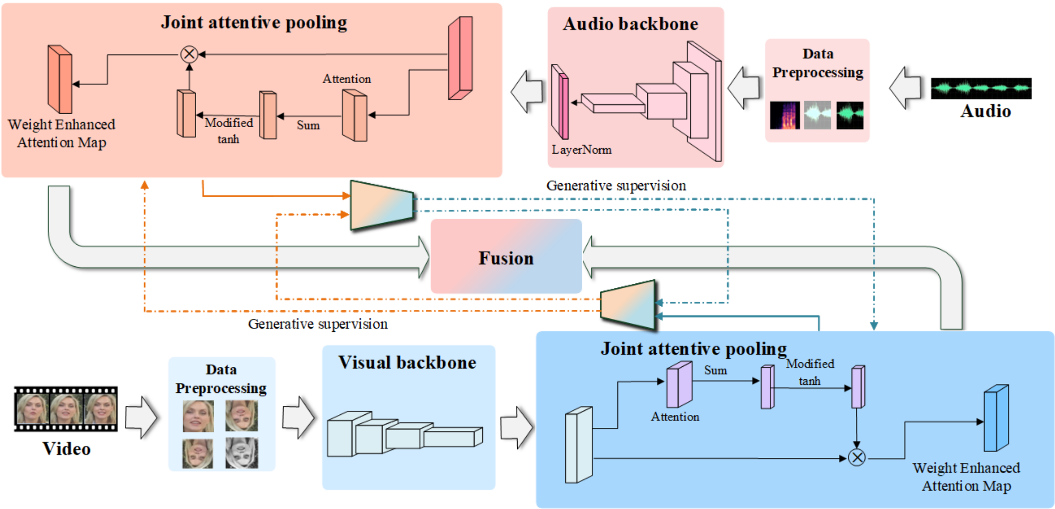

Figure 1: *Overall Network*

This information is transformed to a channel-dependent self-attention score through a linear layer with weights $\boldsymbol{v}^c \in \mathbb{R}^{R\times 1}$ and bias k^c.

However, here comes a problem. It is found that the value of temporal attention λ_t has a comparatively small standard deviation. That is to say, the keyframes in the time domain are not obvious. To make the attention more bifurcated and to help the learning of the subsequent layers, we enhance the weight of each frame by the following equation.

$$e^{t,c}_{tanh} = \frac{\lambda^t_{tanh} e^{t,c}}{\lambda^t}, \tag{2}$$

while the attention map $e^{t,c} \in \mathbb{R}^{T\times A}$ and the projected attention map $e^{t,c}_{tanh} \in \mathbb{R}^{T\times A}$ is given by

$$\lambda^t = sum_{temporal}(e^{t,c}), \tag{3}$$

$$\lambda^t_{tanh} = mu(\lambda^t) tanh(\frac{\lambda^t - mu(\lambda^t)}{std(\lambda^t)}) + mu(\lambda^t). \tag{4}$$

$mu(\cdot)$, $std(\cdot)$ is mean and standard deviation of the matrix, $\lambda^t \in \mathbb{R}^{T\times 1}$ denotes the temporal attention; The enhanced weights are projected into a more dispersed and polarized space $\mathbb{R}^{T\times 1}$ and preserve the mean. If not, fine-grained granularity brings the difficulties of convergence for further experiments. The enhanced method can avoid non-convergence, alleviate overfitting and be beneficial to the later learning of audio-visual weight interactions.

So compared to the original attention map, it is using the new map $e^{t,c}_{tanh}$ instead of the original map $e^{t,c}$. This scalar score $e^{t,c}_{tanh}$ then normalized over all frames by applying the softmax function channel-wise across time:

$$\alpha^{t,c} = \frac{\exp\left(e^{t,c}_{tanh}\right)}{\sum_{\tau}^{T} \exp\left(e^{t,c}_{tanh}\right)}. \tag{5}$$

The self-attention score $\alpha_{t,c}$ represents the importance of each frame given the channel and is used to calculate the weighted statistics of channel c. Then the weighted mean vector and weighted standard deviation vector can be calculated as

$$\tilde{\mu}^c = \sum_{t}^{T} \alpha^{t,c} h^{t,c}, \tag{6}$$

$$\tilde{\sigma}^c = \sqrt{\sum_{t}^{T} \alpha^{t,c} (h^{t,c})^2 - (\tilde{\mu}^c)^2}. \tag{7}$$

The remainder of the calculations is identical to [4]. The weight-enhanced attentive statistical pooling is proved to play a positive role in many aspects in the later comparison experiment of the whole network.

2.2. Generative supervision

From the perspective of face verification, researchers such as [25] hope to eliminate the impact of facial expressions on face verification. Meanwhile, from the perspective of speaker verification, some speech information of audio interval is not helpful for speaker verification. Then naturally, if the weights of keyframes of two modalities are different styles of expression of the same facial action, generative models can be used to achieve this style transition.

Essentially, we want to build a block where keyframes (weight of time series denoted as $X = \{\mathbb{A}, \mathbb{V}\}$, where $\lambda^{t,a}_{tanh} \in \mathbb{A}$ and $\lambda^{t,v}_{tanh} \in \mathbb{V}$) between different modalities can be derived from each other. So obviously, there is a style transition between the visual domain as $\mathbb{V}$ and the audio domain as $\mathbb{A}$. Here we define the corresponding modality domain that belongs to the same utterance as X as $\bar{X} = \{\mathbb{V}, \mathbb{A}\}$). The individual generators with different modalities of X as input are denoted as $G_X(\cdot)$. Three different generative supervision methods are applied here.

- Cycle consistency: The cycle loss follows the previous paper [22] to serve a similar purpose. It ensures each encoder can evolve smoothly, which means the encoder can produce more realistic temporal weights. The cycle loss ensures the consistency of the distribution of keyframes from different domains. Mapping loss and cycle consistency loss only involve the mapping of temporal attention λ^t_{tanh}, and we express the objective as:

$$\mathcal{L}_{\text{mapping}}(G_X, X) = \mathbb{E}_{\boldsymbol{x}\sim X}\{2 - \sum_{X}^{\mathbb{A},\mathbb{V}} < \bar{x}, G_X(x) >\}, \tag{8}$$

$$\mathcal{L}_{\text{cycle}}(G_X, X) = \mathbb{E}_{\boldsymbol{x}\sim X}\{2 - \sum_{X}^{\mathbb{A},\mathbb{V}} < x, G_X(G_{\bar{X}}(x)) >\}, \tag{9}$$

$$\mathcal{L}_1 = \mathcal{L}_{\text{mapping}}(G_X, X) + \mathcal{L}_{\text{cycle}}(G_X, X). \tag{10}$$

while $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{cycle}}$ denote mapping loss and cycle consistency loss, $< \cdot\ ,\ \cdot >$ is the cosine similarity operator. From the perspective of space, mapping loss is essentially a mutual mapping of weights between two modalities. Cycle-consistency loss is actually a cyclic mapping of a single modality through an intermediary space. Only through mapping loss can the intermediary space be given actual meaning.

- Patch NCE loss: A similar operation is applied as CUT [23].The goal of the model is to match the short time series of the corresponding position of the input and output, and other short time series blocks of the same modality are used as negative samples. Similar to the CUT model, but using $G_X()$ instead of the StyleGAN2-based generator and 1 layer of MLP (G_X) instead of 2 layers to fit simple data in one dimension. The temporal weights are passed through the encoder to obtain a series of features $\{z_l\}_L$=$\{H_X(G_X^l(x))\}_L$ and $\{\bar{z}_l\}_L$=$\{H_{\bar{X}}(G_{\bar{X}}^l(\bar{x}))\}_L$, where $G_X^l(x)$ denotes the l-th layer of the feature. We index into layers $l \in \{1, 2, 3, \ldots, L\}$ and denote $s \in \{1, 2, \ldots, S_l\}$, where S_l is the number of spatial locations in each layer. We refer to the corresponding feature as $\boldsymbol{z}_l^s$ and the other features as $\boldsymbol{z}_l^{S\setminus s}$. Overall Patch NCE loss is calculated as

$$\mathcal{L}_{\text{PatchNCE}}(G_X, H_X, X) = \mathbb{E}_{\boldsymbol{x}\sim X} \sum_{l=1}^{L}\sum_{s=1}^{S_l} \ell\left(\bar{\boldsymbol{z}}_l^s, \boldsymbol{z}_l^s, \boldsymbol{z}_l^{S\setminus s}\right), \tag{11}$$

$$\mathcal{L}_2 = \mathcal{L}_{\text{PatchNCE}}(G_X, H_X, X) + \mathcal{L}_{\text{GAN}}(G_X, X). \tag{12}$$

while $\ell\left(\boldsymbol{v}, \boldsymbol{v}^+, \boldsymbol{v}^-\right)$ and $\mathcal{L}_{\text{GAN}}()$has the same definition of CUT [23]. Compared to the cycle consistency approach, this method introduces the concept of contrastive learning and relaxes the assumption that a bijection relationship is required for different modal domains. Also based on maximizing mutual information, the multi-layer encoder is used to obtain features at different layers and use these features themselves for comparative learning.

- Diffusion loss: Inspired by [24], DDIM [26] models are trained independently on each domain ($\mathbb{A}, \mathbb{V}$). A 1-D diffusion model is implemented to adapt to this task. As in [24], we use the forward ODE of the DDIM of the source domain to transform the feature of the source domain X to its representation in the latent space x_{medium}, and then directly use this as the input of the reverse ODE of the DDIM of the target domain$\bar{X}$ to obtain the output of the corresponding target domain.

$$x_{medium} = \text{ODESolve}\left(x; G_X, 0, T\right) \tag{13}$$

$$x_{target} = \text{ODESolve}\left(x_{medium}; G_{\bar{X}}, T, 0\right). \tag{14}$$

As with DDIB, exact cycle consistency is enforced when training diffusion, eliminating the need to introduce additional cycle loss as in CycleGAN [22]. However, since supervision through the generative block is required, the main network needs to be updated using cycle consistency. Overall Diffusion loss is calculated as

$$\mathcal{L}_3 = \mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\boldsymbol{x}\sim X}\{2 - \sum_{X}^{\mathbb{A},\mathbb{V}} < \bar{x}, x_{target} >\}, \tag{15}$$

This approach nicely bypasses the pairs of positive and negative samples inside contrast learning and optimizes the ability to obtain potential encodings within a single domain. In addition, we have tried to modify the conditional diffusion [27, 28], but the result was not that good.

In summary, all three models have similar objectives that all models use generative models to achieve this style transition. If each model will obtain better robustness than the baseline, this gives supporting evidence for the presence of the assumption of style transitions between weights of time series.

2.3. Fusion strategy

Simple Soft Attention Fusion [29] is implemented. But we use the fully connected layer instead of the transformer layer in the original model to achieve the best fusion effect of the model.

For the joint embedding generated by the network, the convergence can be accelerated by some tricks during the training process. We follow [30] to impose an orthogonality constraint on the fused embeddings. We hold the point of view that it is mainly used to accelerate convergence.

3. Experiment Setup

VoxCeleb1&2 [1, 2] are used in our experiment. VoxCeleb is an audio-visual dataset consisting of 2,000+ hours of short human speech clips extracted from interview videos on YouTube. For model training, we use the development set of VoxCeleb2, which contains 5,994 speakers. To better evaluate the performance of our network, we adopt 3 trials in VoxCeleb1. For audio data, 80-dimensional Fbank features are extracted with a 25 ms window and 10 ms frameshift, and augmentation with the random mask is added along both the time and frequency domain. Then we do the cepstral mean on the Fbank features. The MUSAN [31] and RIR Noise datasets [32] are used as noise sources and room impulse response functions, respectively. For each video segment, we extracted 25 fps in VoxCeleb [1, 2] datasets. Then use the similarity transformation to map the face region to the same shape (1 × 128 × 128), which means that we use grey images instead of RGB. Finally, we normalize each image's pixel value to reside in the range of [-0.5, 0.5]. The advanced face detection methods [33, 34] and datasets [35, 36] are wilder and more fine-grained. So face detection and face alignment are not employed during pre-processing since rough face detection has been performed in VoxCeleb datasets [1, 2].

For the audio encoder, we use ECAPA-TDNN [4] into which Fbank feature is fed to extract speaker embedding. We made only a few changes to ECAPA-TDNN, which adds additional scale information to the second layer. For the visual encoder, the face feature is extracted by the IResNet18 backbone same as [6].

The training process is divided into two stages. Two different modalities are trained separately first and then finetune in the overall network. The main network is trained using AAM softmax loss with a margin of 0.5 and a scaling factor of 30 with 32 NVIDIA Tesla P40s. We use the SGD optimizer with an initial learning rate of 0.01 and decrease the learning rate by 50% every 5 epochs. Weight decay is set to 5e-4 to avoid overfitting and the global batch size is 320.

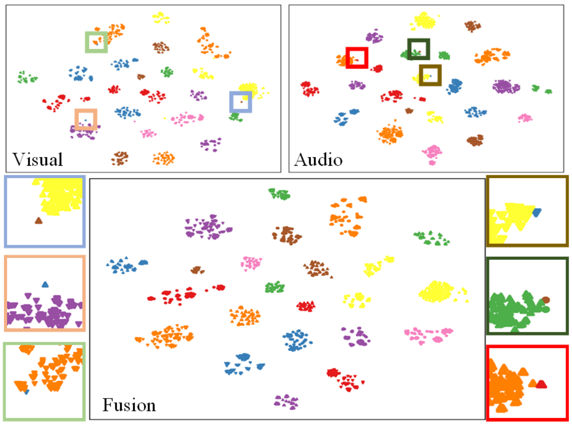

Figure 2: *Sample visualization of outliers*

Table 1: *Performance of proposed network*

Type	Architecture	#Modality	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER(%)	**MinDCF**	**EER(%)**	**MinDCF**	**EER(%)**	**MinDCF**
Unimodal	ECAPA-TDNN[4]	A	0.87	0.107	1.12	0.132	2.12	0.210
	SimAM-Resnet[37]	A	0.64	0.067	0.84	0.089	1.49	0.146
Unimodel Pretrain (only finetune classifier)	Hubert Base[14, 15]	A	0.99	-	1.07	-	2.22	-
	Hubert Large[14, 15]	A	0.81	-	0.82	-	1.68	-
	WavLM Base+[14]	A	0.84	-	0.93	-	1.76	-
	WavLM large[14]	A	0.62	-	0.67	-	1.32	-
Unimodel Pretrain (finetune pretrain modal and classifier)	Hubert Large[14, 15]	A	0.59	-	0.65	-	1.34	-
	WavLM large[14]	A	0.38	-	0.48	-	0.99	-
Audio-visual	Z. Chen[21]	①A	2.31	-	2.23	-	3.78	-
		②V	2.26	-	1.54	-	2.37	-
		③fusion	0.59	-	0.43	-	0.74	-
		ensemble(①,②)	0.51	-	0.43	-	0.78	-
		ensemble(①,②,③)	0.50	-	0.38	-	0.68	-
	Y. Qian[19]	A	1.62	-	1.75	-	3.16	-
		V	3.04	-	2.18	-	4.23	-
		fusion	0.71	-	0.48	-	0.85	-
	Ours	A w/o enhanced†	0.98	0.140	1.24	0.163	2.30	0.264
		④A	0.99	0.140	1.24	0.163	2.27	0.264
		V w/o enhanced†	2.59	0.214	1.88	0.198	3.39	0.297
		⑤V	1.44	0.147	1.28	0.157	2.14	0.230
		fusion	0.22	0.022	0.27	0.035	0.52	0.058
		fusion(cycle)	0.18	0.017	0.26	0.035	0.49	0.058
		fusion(diffusion)	0.16	0.014	0.24	0.033	0.45	0.053
		⑥fusion(NCE)	0.16	0.014	0.23	0.031	0.42	0.050
Audio-visual	Ours	ensemble(④,⑤,⑥)	**0.14**	**0.012**	**0.21**	**0.028**	**0.37**	**0.046**

† "w/o" means "without"

The weights of the 3 losses of generative supervision are set to 0.5, 1, and 0.5. The generator G_X (including $G_{\bar{X}}$) mentioned in all three generation methods has 3 or 4 layers of MLP composition, which is determined by the dimensionality and complexity of the time series weights. And the generators of different modalities of different methods are independent of each other without sharing weights.

We use cosine distance with adaptive s-norm [38] for scoring. Then we report the Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) with $P_{target} = 0.01$ and $C_{FA} = C_{Miss} = 1$ for performance evaluation.

4. Result Analysis

Weight-enhanced attentive statistical pooling is effective from multiple angles. From table 1, we observe that the weight-enhanced method for unimodality will not bring obvious changes to speech while using the weight-enhanced in the face alone can be a relief for the generalization pressure of attentive statistic pooling. And the later ablation experiments show that the weight-enhanced operation on a single modality brings better performance on the joint audio-visual model. In terms of vision alone, compared with the results of works [19, 21], our proposed method uses less image information (grey image instead of RGB), fewer data preparation (without face detection and face alignment), a shallower extractor (IResNet18 instead of ResNet34 [19] or SE-ResNet50 [21]). This method enables us to gain a competitive reduction of EER to 63.1% of the previous baseline of vision by adopting a higher sampling rate.

Fusion with generative supervision reaches the SOTA of the dataset. The application of cycle consistency is a basic application of generative models for supervision and has limited performance improvement. NCE loss as a strong substitute [39] for cycle-consistency loss does have better training supervision. However, it seems that the generative power of diffusion can only be demonstrated in scenarios with much larger amounts of data. From the training perspective, the generative supervision losses will finally converge to the level of $1e$-3. As the losses converge and each encoder is analyzed separately, the transformation from vision to audio is easier to converge than the transformation from audio to visual. This shows that the supervision by transitions of style is mainly reflected in "vision supervises audio". From the results in Table 1, the use of each type of generative supervision loss indeed enhances the weight of important time frames. Compared to other networks in audio-visual person verification, the fusion network alone can greatly reduce the EER down 35.7% of the previous baseline. The simple score ensemble of multiple systems is still valid for this experiment. The best results are achieved after a simple score ensemble. In addition, our proposed model outperforms other networks across the board compared to all the various methods that have achieved good results on this dataset.

The assumption that the weights of keyframes for both modalities are different expression styles for the same facial action **works in audio-visual task**. These approaches reduce EER by 36.4% relatively compare with the normal fusion, but it is enough to prove that this supervision method is effective. As a result, this approach can be further adapted for recognition tasks covering both movement and sound.

About 20 hard samples were selected from the VoxCeleb1 for t-SNE visualization. It can be seen from Fig.2 that each diagram of a single modality can easily find 5-6 outliers. But our network is more robust to these outliers. This phenomenon well demonstrates the robustness of the proposed network.

5. Conclusion

Generally, we proposed a novel audio-visual strategy that considers temporal relations from a generative perspective. Compared with previous aggregators, we proposed joint attentive pooling based on generative supervision as a generic aggregator for the first time. It significantly reduces the EER down to 28.0% of the popular systems. And analysis shows the robustness of our network and the existence of possible relations. In future work, more audio-visual tasks will be explored based on such thoughts.

6. References

[1] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[2] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[3] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *INTERSPEECH*, 2020, pp. 3830–3834.

[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, June 2019.

[7] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *Proc. CVPR*, 2020, pp. 5901–5910.

[8] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. CVPR*, 2017, pp. 212–220.

[9] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proc. CVPR*, 2018, pp. 5265–5274.

[10] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. ICASSP*. IEEE, 2018, pp. 4879–4883.

[11] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *APSIPA ASC*. IEEE, 2019, pp. 1652–1656.

[12] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[13] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *Proc. ICASSP*. IEEE, 2022, pp. 6152–6156.

[14] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE. TASLP*, vol. 29, pp. 3451–3460, 2021.

[16] S. Shon, T.-H. Oh, and J. Glass, "Noise-tolerant audio-visual online person verification using an attention-based neural network fusion," in *Proc. ICASSP*. IEEE, 2019, pp. 3995–3999.

[17] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, J. Hernandez-Cordero *et al.*, "The 2019 nist audio-visual speaker recognition evaluation." in *Odyssey*, 2020, pp. 259–265.

[18] J. Alam, G. Boulianne, L. Burget, M. Dahmane, M. D. Sánchez, A. Lozano-Diez, O. Glembek, P.-L. St-Charles, M. Lalonde, P. Matejka *et al.*, "Analysis of abc submission to nist sre 2019 cmn and vast challenge." in *Odyssey*, 2020, pp. 289–295.

[19] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE. TASLP*, vol. 29, pp. 1079–1092, 2021.

[20] Y. Liang, F. Ma, Y. Li, and S.-L. Huang, "Person recognition with hgr maximal correlation on multimodal data," in *Proc. ICPR*. IEEE, 2021, pp. 1–8.

[21] Z. Chen, S. Wang, and Y. Qian, "Multi-modality matters: A performance leap on voxceleb." in *INTERSPEECH*, 2020, pp. 2252–2256.

[22] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. CVPR*, 2017, pp. 2223–2232.

[23] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. ECCV*, 2020.

[24] X. Su, J. Song, C. Meng, and S. Ermon, "Dual diffusion implicit bridges for image-to-image translation," in *Proc. ICLR*, 2023.

[25] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proc. CVPR*, 2020, pp. 5710–5719.

[26] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. ICLR*.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proc. NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[28] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.

[29] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. ACL*, vol. 2019. NIH Public Access, 2019, p. 6558.

[30] M. S. Saeed, M. H. Khan, S. Nawaz, M. H. Yousaf, and A. Del Bue, "Fusion and orthogonal projection for improved face-voice association," in *Proc. ICASSP*, 2022, pp. 7057–7061.

[31] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[32] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*. IEEE, 2017, pp. 5220–5224.

[33] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proc. CVPR*, 2020.

[34] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," in *Proc. ICLR*, 2021.

[35] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proc. CVPR*, 2016.

[36] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proc. CVPR*, June 2016.

[37] X. Qin, N. Li, C. Weng, D. Su, and M. Li, "Simple attention module based speaker verification with iterative noisy label detection," in *Proc. ICASSP*. IEEE, 2022, pp. 6722–6726.

[38] P. Matejka, O. Novotnỳ, O. Plchot, L. Burget, M. D. Sánchez, and J. Cernockỳ, "Analysis of score normalization in multilingual speaker recognition." in *INTERSPEECH*, 2017, pp. 1567–1571.

[39] T. Li, Y. Liu, A. Owens, and H. Zhao, "Learning visual styles from audio-visual associations," in *Proc. ECCV*. Springer, 2022, pp. 235–252.