FUSION AND ORTHOGONAL PROJECTION FOR IMPROVED FACE-VOICE ASSOCIATION

Muhammad Saad Saeed[1], *Muhammad Haris Khan*[2], *Shah Nawaz*[3*], *Muhammad Haroon Yousaf*[1], *Alessio Del Bue*[3,4]

[1]Swarm Robotics Lab (SRL)-NCRA, University of Engineering and Technology Taxila
[2]Mohamed Bin Zayed University of Artificial Intelligence
[3]Pattern Analysis & Computer Vision (PAVIS) - Istituto Italiano di Tecnologia (IIT)
[4]Visual Geometry & Modelling (VGM) - Istituto Italiano di Tecnologia (IIT)

ABSTRACT

We study the problem of learning association between face and voice. Prior works adopt pairwise or triplet loss formulations to learn an embedding space amenable for associated matching and verification tasks. Albeit showing some progress, such loss formulations are restrictive due to dependency on distance-dependent margin parameter, poor runtime training complexity, and reliance on carefully crafted negative mining procedures. In this work, we hypothesize that enriched feature representation coupled with an effective yet efficient supervision is necessary in realizing a discriminative joint embedding space for improved face-voice association. To this end, we propose a light-weight, plug-and-play mechanism that exploits the complementary cues in both modalities to form enriched fused embeddings and clusters them based on their identity labels via orthogonality constraints. We coin our proposed mechanism as fusion and orthogonal projection (FOP) and instantiate in a two-stream pipeline. The overall resulting framework is evaluated on a large-scale VoxCeleb dataset with a multitude of tasks, including cross-modal verification and matching. Our method performs favourably against the current state-of-the-art methods and our proposed supervision formulation is more effective and efficient than the ones employed by the contemporary methods.

***Index Terms*—** Multimodal, Face-voice association, Cross-modal verification and matching

1. INTRODUCTION

Recently, Nagrani et al. [1, 2, 3] introduced the face-voice association task into vision community with the creation of a large-scale audio-visual dataset, comprising faces and voices of $1,251$ celebrities. Since then, the face-voice association task has gained significant research interest [1, 2, 4, 5, 6, 7, 8]. We are also witnessing the creation of new audio-visual datasets to study this novel task. For example, Nawaz et al. [9] introduced a Multilingual Audio-Visual (*MAV-Celeb*) dataset to analyze the impact of language on face-voice association task; it comprises of video and audio recordings of different celebrities speaking more than one language. Most existing works [1, 2, 5, 9] tackle face-voice association as a cross-modal biometric task. The two prominent challenges in developing an effective method for this task are learning of a common yet discriminative embedding space, where instances from two modalities are sufficiently aligned and instances of semantically similar identities are nearby. Often separate networks for face and voice modalities are leveraged to obtain the respective feature embeddings and Contrastive or Triplet loss formulations are employed to construct this embedding space. Although showing some effectiveness in this task, such loss formulations, however, are restrictive in following ways. First, they require tuning of a margin hyperparameter, which is hard as the distances between instances can alter significantly while training. Secondly, the run-time training complexity for contrastive and triplet losses are $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$, respectively, where n is the number of available instances for a modality. Finally, to mitigate the high run-time training complexity challenge, different variants of carefully crafted negative mining strategies are used, which are both time-consuming and performance sensitive.

A few methods e.g., [6] have attempted to replace the Contrastive/Triplet loss formulations by utilizing auxiliary identity centroids [10]. The training process alternates between the following two steps: 1) clustering embeddings around their identity centroids and pushing embeddings away from all other identity centroids, and 2) updating these centroids using the mini-batch instances. Such centroid based losses are used with traditional classification loss (i.e. softmax cross-entropy (CE)). However, their co-existence is ineffective because the former promotes margins in Euclidean space whereas latter implicitly achieves separability in the angular domain. In this work, we hypothesize that an enriched, unified feature representation, encompassing complementary cues from both modalities, alongside an effective yet efficient supervision formulation is crucial towards realizing a discriminative joint embedding space for improved face-voice (F-V)

[*] Current Affiliation: Deutsches Elektronen-Synchrotron DESY

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

association. To this end, we propose a light-weight, plug-and-play mechanism that exploits the best in both modalities through fusion and semantically aligns fused embeddings with their identity labels via orthogonality constraints. These constraints align well with the angular characteristic of the commonly used classification loss and are very efficient. We instantiate our proposed mechanism in the two-stream pipeline, which provides face and voice embeddings, and the resulting overall framework performs favourably (in both accuracy and efficiency) against the existing state-of-the-art methods on large-scale VoxCeleb dataset [3].

2. RELATED WORK

Face-voice Association. Nagrani et al. [1] leveraged audio and visual information to establish an association between faces and voices in a cross-modal biometric matching task. Similarly, some recent work [5, 2, 7, 9] introduced joint embeddings to establish correlation between face and voice of an identity. They extract audio and face embeddings and then minimize intra-identity distance and maximize inter-identity distance. Wen et al. [8] presented a disjoint mapping network to learn a shared representation for audio and visual information by mapping them individually to common covariates (gender, nationality, identity) without needing to construct pairs or triplets at the input. Similarly, Nawaz et al. [6] extracted audio and visual information with a single stream network to learn a shared deep latent representation, leveraging identity centroids to eliminate the need of pairs or triplets [1, 2]. Both Wen et al. [8] and Nawaz et al. [6] show that effective F-V features can be learned without pairs or triplets formation. In contrast, our method constructs enriched embeddings via exploiting complementary cues from the embeddings of both modalities via an attention-based fusion. Further, it clusters the embeddings of same identity and separates embeddings of different identities via orthogonality constraints. The instantiation of both proposals in a two-stream pipeline results in an effective and efficient F-V association framework.

3. OVERALL FRAMEWORK

To learn a discriminative joint face-voice embedding for F-V association tasks, we develop a new framework for cross-modal face-voice association (See Fig. 1) that is fundamentally a two-stream pipeline (sec. 3.1) and features a light-weight module that exploits complementary cues from both face and voice embeddings and facilitates discriminative identity mapping via orthogonality constraints (sec. 3.2).

3.1. Preliminaries

Problem Settings. Without losing generality, we consider cross-modal retrieval of bimodal data, i.e., for face and voice. Given that we have N instances of face-voice pairs, $\mathcal{D} = \{(x_i^f, x_i^v)\}_{i=1}^N$, where x_i^f and x_i^v are the face and voice examples of the i_{th} instance, respectively. Each pair of an instance (x_i^f, x_i^v) has an associated label $y_i \in \{0, 1\}$, where $y_i = 1$ if x_i^f and x_i^v belong to the same identity and $y_i = 0$ if x_i^f and x_i^v belong to a different identity. Both face and voice embeddings typically lie in different feature spaces owing to their different superficial statistics and are mostly unaligned semantically, rendering them incomparable for cross-modal tasks. Cross-modal learning aims at projecting both into a common yet discriminative feature space, where they are sufficiently aligned and instances from the same identity are nearby while from a different identity are far apart.

Two-stream pipeline. We employ a two-stream pipeline [2] to obtain the respective feature embeddings of both face and voice inputs. The first stream is a pre-trained convolutional neural network (CNN) on image modality. We take the penultimate layer's output, denoted as $\mathbf{b}_i$, of this CNN as the feature embeddings for an input face image. Likewise, the second stream is a pre-trained audio encoding network that outputs a feature embedding, denoted as $\mathbf{e}_i$, for an input audio signal (typically a short-term spectrogram). Existing approaches handling face-voice retrieval [2, 6], mostly resort to triplet and contrastive objectives with carefully crafted negative mining strategies, which significantly increases computational time and are performance-sensitive, to learn a discrminative embedding space. To this end, we introduce a light-weight mechanism that exploits complementary cues from both modality embeddings to form enriched fused embeddings and imposes orthogonal constraints on them for learning discriminative joint face-voice embeddings.

3.2. Learning Discriminative Joint Embedding

In this section, we first describe extracting complementary cues, via multimodal fusion, from both face and voice embeddings obtained through their respective pre-trained networks. We then discuss clustering fused embeddings belonging to the same identity and pushing away the ones with different identity via orthogonality constraints. Prior to multimodality fusion, we project the face embeddings $\mathbf{b}_i \in R^F$ to a new d-dimensional embedding space $\mathbf{u}_i \in R^d$ with a fully-connected layer. Similarly, we project the voice embedding $\mathbf{e}_i \in R^V$ to a similar d-dimensional embedding space $\mathbf{v}_i \in R^d$ with another fully-connected layer. We then L2 normalize both $\mathbf{u}_i$ and $\mathbf{v}_i$ which can now be fused to get $\mathbf{l}_i$, using the procedure described next.

Multimodal fusion. We propose to extract complementary features from both modalities, some of which could be related to age, gender and nationality, to form an enriched unified feature representation which is crucial towards learning a discriminative joint embedding space. Inspired by [11, 12], we employ an attention mechanism to first compute the attention scores (affinity) between the embeddings of two modalities and then fuse these individual modality embeddings after

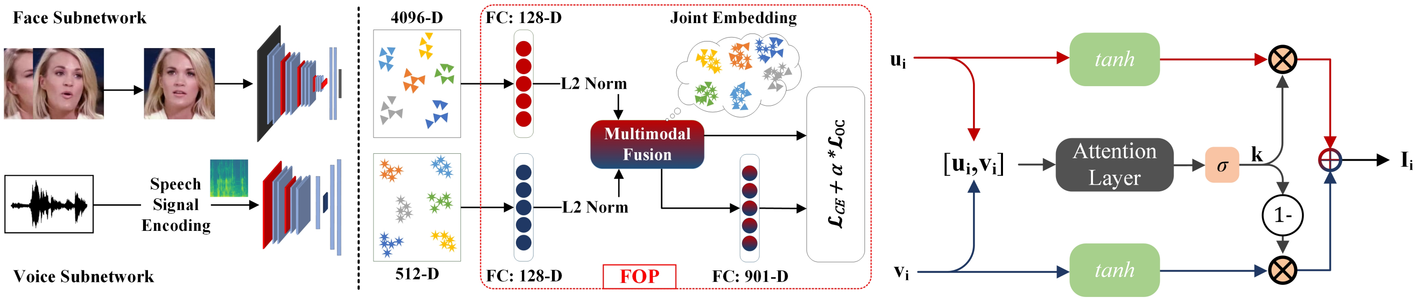

Fig. 1: (Left) Overall method. Fundamentally, it is a two-stream pipeline which generates face and voice embeddings. We propose fusion and orthogonal projection (FOP) mechanism (dotted red box). (Right) The architecture of multimodal fusion.

recalibrating them with the attention scores (see Fig. 1). We compute attention scores $\mathbf{k}$ between $\mathbf{u}_i$ and $\mathbf{v}_i$ as:

$$\mathbf{k} = \sigma(F_{att}([\mathbf{u}_i, \mathbf{v}_i]), \tag{1}$$

where σ is a sigmoid operator, and F_{att} are the attention layers. Finally, we fuse $\mathbf{u}_i$ and $\mathbf{v}_i$ after modulating them with the attention scores $\mathbf{k}$ to obtain the fused embeddings $\mathbf{l}_i$ as:

$$\mathbf{l}_i = \mathbf{k} \odot tanh(\mathbf{u}_i) + (1 - \mathbf{k}) \odot tanh(\mathbf{v}_i), \tag{2}$$

where $\odot$ is element-wise multiplication.

Supervision via orthogonality constraints. We want the fused embeddings to encapsulate the semantics of the identity. In other words, these embeddings should be able to predict the identity labels with good accuracy. This is possible if the instances belonging to the same identity are placed nearby whereas the ones with different identity labels are far away. A popular choice to achieve this is softmax cross entropy (CE) loss, which also allows stable and efficient training. Specifically, we use an identity linear classifier with weights denoted as $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_C] \in R^{d \times C}$ to compute the logits corresponding to $\mathbf{l}_i$. Where d is the dimensionality of embeddings and C is the number of identities. Now, identity classification loss with fused embeddings is computed as:

$$\mathcal{L}_{CE} = -log \frac{exp(\mathbf{l}_i^T \mathbf{w}_{y_i})}{\sum_{j=1}^{C} exp(\mathbf{l}_i^T \mathbf{w}_j)} \tag{3}$$

Since softmax CE loss does not enforce margins between pair of identities, it is prone to constructing differently-sized class regions which affects identity separability [13, 14]. Some works attempt to include margin between classes in the Euclidean space [10, 15], which is not well synergized with the CE loss as it achieves separation in the angular domain. Therefore, we propose to impose orthogonality constraints on the fused embeddings to explicitly minimize intra-identity separation while maximizing inter-identity separability [16]. These constraints complement better with the innate angular characteristic of CE loss. Further, since they directly operate on mini-batches, they show greater training efficiency compared to the complex negative mining procedures required in contrastive and triplet loss formulations [2, 17] (sec. 4). Formally, the constraints enforce fused embeddings of different identities to be orthogonal and the fused embeddings with same identity to be similar:

$$\mathcal{L}_{OC} = 1 - \sum_{i,j \in B, y_i = y_j} \langle \mathbf{l}_i, \mathbf{l}_j \rangle + \left| \sum_{i,j \in B, y_i \neq y_k} \langle \mathbf{l}_i, \mathbf{l}_k \rangle \right|, \tag{4}$$

where $\langle ., . \rangle$ is the cosine similarity operator, and B represents the mini-batch size. The first term in Eq. 4 ensures intra-identity compactness, while the second term enforces inter-identity separation. Note that, the cosine similarity involves the normalization of fused embeddings, thereby projecting them to a unit hyper-sphere:

$$\langle \mathbf{l}_i, \mathbf{l}_j \rangle = \frac{\mathbf{l}_i . \mathbf{l}_j}{\|\mathbf{l}_i\|_2 . \|\mathbf{l}_j\|_2}. \tag{5}$$

Overall Training Objective. To train the proposed framework, we minimize the joint loss formulation, comprising of $\mathcal{L}_{CE}$ and $\mathcal{L}_{OC}$ as:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{OC}, \tag{6}$$

where α balances the contribution of two terms in $\mathcal{L}$. We empirically set α to 1.0 based on validation set performance. Note that, CE loss operates in logit space and orthogonal constraints are imposed in the embedding space, however, both of them synergizes well with each other owing to their common angular domain characteristic.

4. EXPERIMENTS

Training Details and Dataset. We train our method on Quadro P5000 GPU for 50 epochs using a batch-size of 128 using Adam optimizer with exponentially decaying learning rate (initialised to 10^{-5}). We extract face and voice embeddings from VGGFace [18] and Utterance Level Aggregation [19] respectively. Note that, we only backprop. through FOP module while the weights of face and voice subnetworks

Table 1: Cross-modal verification results for our (joint) loss and other losses under two configurations and two error metrics.

Method	EER	AUC	EER	AUC
	Seen-Heard		Unseen-Unheard	
CE Loss	21.8	86.6	26.8	81.7
Center Loss [10, 6]	19.8	88.6	29.7	77.5
Git Loss [15]	19.6	88.9	29.5	77.8
Contrastive Loss [2]	23.4	84.7	29.1	79.5
Triplet Loss [17]	20.7	88.0	27.1	81.4
Ours	**19.3**	**89.3**	**24.9**	**83.5**

Table 2: Theoretical/empirical training complexity of our (joint) loss and others. n: # training instances in a modality, and B is mini-batch size.

Method	Empirical Time (s)	Theoretical Worst Case
CE Loss	.02	$\mathcal{O}(n)$
Center Loss [10, 6]	6.8	$\mathcal{O}(n+\frac{n^2}{B})$
Git Loss [15]	6.2	$\mathcal{O}(n+\frac{n^2}{B})$
Contrastive Loss [2]	568.2	$\mathcal{O}(n^2)$
Triplet Loss [17]	619.7	$\mathcal{O}(n^3)$
Ours	0.7	$\mathcal{O}(n)$

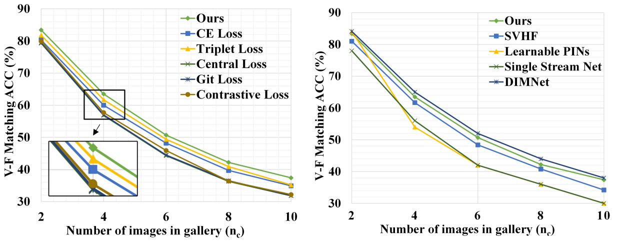

Fig. 2: Cross-modal matching results: (left) FOP vs other losses used in F-V methods. (right) Our method vs SOTA methods.

remains unaltered. We perform experiments on *cross-modal verification* and *cross-modal matching* tasks on the large-scale dataset of audio-visual human speech videos [3]. We follow the same train, validation and test split configurations as used in [2] to evaluate on *seen-heard* and *unseen-unheard* configurations.

Comparison with other F-V losses. Table 1 reveals that our loss performs better than others, including *Center Loss* [10, 6], *Git loss* [15], *Contrastive Loss* [2], and *Triplet Loss* [1], across configurations and error metrics. Table 2 shows that our (joint) loss formulation is superior than others in terms of both theoretical and empirical training efficiency. We then validate the effectiveness of our (joint) loss formulation by examining the effect of Gender (G), Nationality (N), Age (A) and its combination (GNA) separately, which influence both face and voice verification (Table 3). It achieves consistently better performance on G, N, A and the combination (GNA) in both *seen-heard* and *unseen-unheard* configurations than other loss formulations. Further, we compare our (joint) loss formulation against other losses on a cross-modal matching task, $1 : n_c$ with $n_c = 2, 4, 6, 8, 10$ in Fig. 2 (left). It outperforms them for all values of n_c.

Table 3: Cross-modal biometrics results under varying demographics for *seen-heard* and *unseen-unheard* configurations.

Demographic	Rand.	G	N	A	GNA	Rand.	G	N	A	GNA
	Seen-Heard					Unseen-Unheard				
CE	86.6	78.0	85.0	86.3	77.3	81.7	65.9	53.6	76.0	52.8
Center	88.6	79.2	87.0	88.2	78.1	77.5	62.4	51.7	72.5	**54.2**
Git	88.9	**79.7**	87.4	**88.6**	**78.5**	77.9	62.6	51.8	72.8	**54.2**
Contrastive	84.7	69.7	83.7	84.5	69.2	79.5	61.0	53.5	74.7	51.8
Triplet	88.0	76.3	86.7	87.6	75.6	81.7	65.5	53.4	76.3	52.2
Ours	**89.3**	76.7	**87.9**	**88.6**	76.6	**83.5**	**68.8**	**54.9**	**78.1**	**54.2**

Comparison with state-of-the-art. Under *unseen-unheard* protocol, our method outperforms all competing approaches, including DIMNet [8], Learnable Pins [2], MAV-Celeb [9], Single Stream Network [6], and under *seen-heard* configuration, it achieves the second best performance (see Table 4). For cross-modal matching task, involving $1 : n_c$ matching tasks, our method outperforms [2] while achieves competitive performance against DIMNet [8] (Fig. 2 (right)).

Table 4: Cross-modal verification. Ours vs SOTA methods.

Methods	EER	AUC	EER	AUC
	Seen-Heard		Unseen-Unheard	
DIMNet [8]	-	-	**24.9**	-
Learnable Pins [2]	21.4	87.0	29.6	78.5
MAV-Celeb [9]	-	-	29.0	78.9
Single Stream Network [6]	**17.2**	**91.1**	29.5	78.8
Ours	19.3	89.3	**24.9**	**83.5**

Ablation study and analysis. Table 5 reveals that our method's performance is mostly robust to the choice of α, which is a hyperparameter to balance the contribution of two losses in (Eq. 6). We also show that on replacing gated multimodal fusion with a much simpler linear fusion, the performance of our method significantly drops (Table 6). Finally, in Fig. 3, we find that the proposed OC with CE loss, in comparison to CE loss alone, enhances (overall) feature discrminability with orthoganlity constraints, and enforces stronger intra-identity compactness and inter-identity separation in the joint F-V embedding space.

Table 5: Cross-modal verification results when varying α.

	$\mathcal{L}_{CE}+\alpha\mathcal{L}_{OC}$					
α	0.0	0.1	0.5	1.0	2.0	5.0
EER	26.8	26.1	25.8	**24.9**	25.9	26.0
AUC	81.7	82.4	82.8	**83.5**	82.7	82.6

Table 6: Cross-modal verification results with linear and gated fusion strategies.

Fusion Strategy	EER	AUC
Linear Fusion	25.6	82.7
Gated Fusion	**24.9**	**83.5**

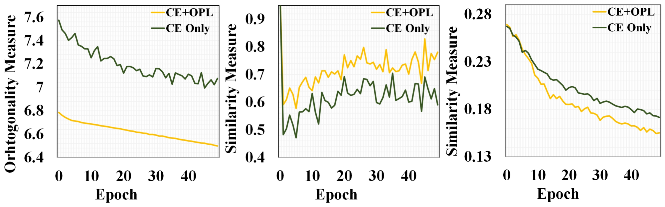

Fig. 3: (a) Feature Orthogonality (↓) (b) Similarity of same class features (↑) (c) Similarity of different class features (↓).

5. CONCLUSION

We presented a new module (FOP) for F-V association that performs attention-based fusion and clusters the fused embeddings on their identity-labels via orthogonality constraints. We instantiated this module in a two-stream F-V pipeline and the resulting overall framework performs favourably against the existing SOTA methods.

6. REFERENCES

[1] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8427–8436.

[2] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 71–88.

[3] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[4] Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu, "Face-voice matching using cross-modal embeddings," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1011–1019.

[5] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik, "On learning associations of faces and voices," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 276–292.

[6] Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood, and Alessandro Calefati, "Deep latent space learning for cross-modal mapping of audio and visual signals," in *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2019, pp. 1–7.

[7] Peisong Wen, Qianqian Xu, Yangbangyan Jiang, Zhiyong Yang, Yuan He, and Qingming Huang, "Seeking the shape of sound: An adaptive framework for learning voice-face association," *arXiv preprint arXiv:2103.07293*, 2021.

[8] Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, and Rita Singh, "Disjoint mapping network for cross-modal matching of voices and faces," in *7th International Conference on Learning Representations, ICLR 2019, USA, May 6-9, 2019*, 2019.

[9] Shah Nawaz, Muhammad Saad Saeed, Pietro Morerio, Arif Mahmood, Ignazio Gallo, Muhammad Haroon Yousaf, and Alessio Del Bue, "Cross-modal speaker verification and recognition: A multilingual perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1682–1691.

[10] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[11] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[12] Zhengyang Chen, Shuai Wang, and Yanmin Qian, "Multi-modality matters: A performance leap on voxceleb," *Proc. Interspeech 2020*, pp. 2252–2256, 2020.

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[14] Munawar Hayat, Salman Khan, Syed Waqas Zamir, Jianbing Shen, and Ling Shao, "Gaussian affinity for max-margin class imbalanced learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6469–6479.

[15] Alessandro Calefati, Muhammad Kamran Janjua, Shah Nawaz, and Ignazio Gallo, "Git loss for deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[16] Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan, "Orthogonal projection loss," *arXiv preprint arXiv:2103.14021*, 2021.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[18] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," 2015.

[19] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.