A NOVEL DISTANCE LEARNING FOR ELASTIC CROSS-MODAL AUDIO-VISUAL MATCHING

Wangrui[1], Huaibo Huang[2,3], Xufeng Zhang[1], Jixin Ma[4], Aihua Zheng[1,]*

[1]School of Computer Science and Technology, Anhui University, Hefei, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Center for Research on Intelligent Perception and Computing, CASIA, Beijing, China
[4]Department of Computing and Information Systems, University of Greenwich, London, UK

ABSTRACT

In this work we propose a novel network formulation for joint representation of cross-modal audio and visual information base on metric learning. We employ a distance learning framework as a training procedure. For this purpose we introduce an elastic matching network (EmNet) and a novel loss function to learn the shared latent space representation of multi-modal information. The elastic matching network is capable of matching given face image (or audio voice clip) from diverse number of audio clips (or face images). We quantitatively and qualitatively evaluate the purposed approach on the standard audio-visual matching evaluation dataset, the overlap of VoxCeleb and VGGFace by both multi-way and binary audio-visual matching tasks. The promising performance comparing to the existing methods verifies the effectiveness of the proposed approach, which yields to a new state-of-the-art for cross-modal audio-visual matching.

Index Terms— Cross-modality, Audio-visual matching, Elastic multi-way matching, Distance learning

1. INTRODUCTION

Audio-visual matching aims to match the given query audio voice clip to the corresponding person from the gallery face images (F2V) or vice versa (V2F), as shown in Fig. 1. It has potential applications such as criminal investigation, face detection, identity determination, etc. One of the challenging in audio-visual matching is to measure the similarity or the distance between the cross-modal information, which appears heterogeneously.

Comparing to the conventional face recognition [1] [2], cross-modal audio-visual matching is a recently emerged research topic. The pioneer work was Nagrani et al. [3] which proposed a Network of Seeing Voice and Hearing Face (SVHF-Net) based on a two-stream architecture to learn the audio and visual features respectively, then the spliced feature were fed into the Softmax layer to obtain the probability of classification. Wen et al. [4] proposed a DIsjoint Mapping Network (DIMNet) which made full use of the covariates of the attributes among different people as a pivotal condition to improve the accuracy of cross-modal matching task. However, it can only handle multi-way F2V task. Furthermore, it required to introduce more branches when the number of face images increasing in multi-way matching. Recently, Chung et. al. [5] proposed a new training scheme by increasing the number of negative samples based on the baseline models [6] [7] to improve the discrimination ability of the network. Albanie et al. [8] proposed a cross-modal embedding for Person Identity Nodes (PINs) by curriculum mining and contrastive loss for diverse cross-modal audio-visual tasks, including retrieval, matching, ect. However, the above methods mainly focused on the feature representation learnt from corresponding networks while ignoring the inter-modal difference and the intra-modal similarity. Herein, we pursue to enhance the intra-modal similarity and expand the inter-modal difference inspired by the metric learning mechanism in this paper. We name our proposed method as Elastic match-

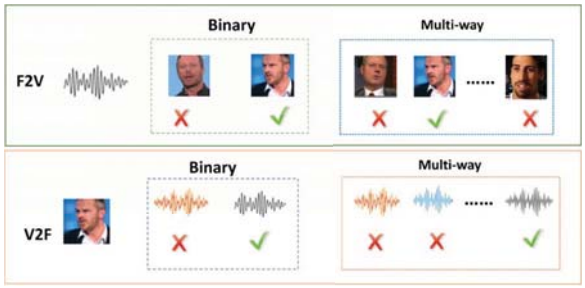

Fig. 1. Binary and multi-way cross-modal audio-visual matching. F2V aims to match the given query audio voice clip to the corresponding identity from the gallery face images. V2F aims to match the given query face image to the corresponding identity from the gallery audio voice clips. The binary matching task indicates only two samples in gallery, which can be regarded as the special case of multi-way task.

* Corresponding author. E-mail: ahzheng214@foxmail.com

978-1-5386-9214-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMEW.2019.00-70

ing Network (EmNet) since it is capable of matching the given audio clip (or face image) for arbitrary number of face images (or audio clips) in order to make full use of similarity between different modalities. The capability of the state-of-the-art audio visual matching methods on various tasks is summarized in Table. 1.

Table 1. The capability of the state-of-the-art audio visual matching methods on various tasks, where '$\times$' denotes not capable, while '-' indicates not available.

	Binary task		Multi-way task	
	F2V	**V2F**	**F2V**	**V2F**
SVHF-Net [3]	$\surd$	$\surd$	$\surd$	$-$
DIMNet [4]	$\surd$	$\surd$	$\surd$	$\times$
PINs [8]	$\surd$	$\surd$	$\surd$	$-$
EmNet (Ours)	$\surd$	$\surd$	$\surd$	$\surd$

Based on above discussion, we propose an elastic network (EmNet) for cross-modal audio-visual matching by introducing a novel distance loss function base on metric learning [9], which maps each modality into the same joint-embedding space [10] [8]. By minimizing the distance between positive samples while maximizing the distance between multiple negative samples, the matching result is achieved by calculating the distance of corresponding features. Therefore, it can take advantage of the similarity between features and tolerate elastic number of the samples in gallery without altering the architecture of the network. The main contribution of this paper can be summarized as:

- We propose a novel distance measurement inspired by metric learning for cross-modal audio-visual matching, which maps the cross-modal information into the joint-embedding space and learn the representations in the shared latent space.
- We propose an elastic matching network (EmNet) based on the proposed distance function which can tolerate the diverse number of sample in gallery for both F2V and V2F tasks in audio-visual matching with fixed architecture of the network.
- Quantitative and qualitative evaluations on benchmark dataset VoxCeleb [11] demonstrate the effectiveness of the proposed model, which yields a new state-of-the-art for audio-visual matching, comparing to the other methods.

2. RELATED WORKS

2.1. Audio-visual Retrieval

Audio-visual retrieval aims to exploit the correlation between audio and visual information. Representative works include hash transformation [12, 13], subspace learning [14, 15] and metric learning [16, 17]. Yang et al. [12] proposed a novel end-to-end network base on deep cross-modal hashing method and added decorrelation constraints to improve the discrimination of each hash bit. Zhen et al. [13] proposed a hashing-based method based on spectral analysis of different modal correlation matrices. Wang et al. [14] addressed the problem of the measuring the relevance and coupled feature selection by mapping different modalities into same subspace. Xu et al. [15] verified the quality of various cross-modal retrieval algorithms on sketch-based image retrieval problem. Xu et al. [16] proposed a deep adversarial metric learning to map data from different modalities into a shared latent subspace. Zhai et al. [17] proposed joint graph regularized heterogeneous metric learning which mapped different modalities into a joint graph regularization and learned a high-level semantic metric based on label propagation.

2.2. Audio-visual Generating

With the blossoming of GANs [18], audio-visual generating task has become more and more popular. Owens et al. [19] introduced a model which predicts the subband envelopes of the audio waveform. Jalalifar et al. [20] produced a sequence of realistic faces that synchronized with the input audio by two networks. Chen et al. [21] proposed a model that exploit speech to generate lip movement. Rithesh et al. [22] proposed ObamaNet to generate video condition on key point rather that generate directly. In [23], Zhou et al. design an end-to-end model to solve the task of generating sounds from in-the-wild videos. Hao et al [24]. proposed a CMCGAN to tackle cross-modal visual-audio mutual generation by organizing all subnetworks in a cycle architecture.

3. ARCHITECTURE OF NETWORK

At the first place, we use FFMPEG[1] to divide the raw video into audio clips and visual images while leaving the interference in audio such as noise untouched. Similar as the existing works [3] [6], we employ the dual branch CNN architecture to process the audio and visual information respectively.

3.1. Dual Branch Architecture

3.1.1. Audio branch

The input to the audio branch are 3 seconds audio clips, we first convert them to the single channel audio spectrogram. Since the spectrum features from different identities vary widely in frequency, amplitude, and striations. Then we resize them into the same resolution of 224$\times$125, following the protocol in [25] [7] [3]. Each branch consists of 5 convolution layers and 3 pooling layers (kernel-sizes in first and

[1] URL: https://sourceforge.net/projects/ffmpeg/

second convolution layers are (2, 1), (3, 6) respectively and 3 for the rest).

3.1.2. Visual branch

The RGB face images fed into visual branch are resized to the resolution of 224×224. For each 3-second video segment, we sample the face image at the rate of 10 fps. We call visual data as 'Certain Face' due to it is a single RGB image at a certain moment without any temporal information. Similar to audio branch, each visual branch consists of 5 convolution layers and 3 pooling layers (kernel-sizes are 7, 5, 2, 3, 3 for each convolution layer). Note that all visual branches share parameters.

3.2. Cross-modal Audio-visual Matching

We conduct both binary and multi-way audio-visual matching tasks in this paper, where the binary task can be regarded as the special case of the multi-way task.

3.2.1. Elastic Multi-way Network

We render the multi-way audio-visual matching as the multi-way classification task, where the multi-way F2V aims to match the given audio voice clip from N face images (one positive face image and $N-1$ negative face images), as shown in Fig. 2. After obtaining the feature vectors of the query audio clip and the face images in gallery by one audio branch and N visual branches. The key issue is to measure the distance between them.

Distance learning The traditional cross-modal matching tasks usually use Softmax function to the last layers of the network to achieve the binary or multi-classification task on the feature maps [4] [3]. The main limitation is they cannot change the number of the face images (or audio voice clips) in the gallery. Inspired by the metric learning [9] [26] [27], which can directly calculate the distance between features to accomplish multi-classification task, we propose a novel distance-based loss function for multi-way audio-visual matching. The main idea is to minimize the distance between the anchor and the positive sample while maximizing the distance between the anchor to the multiple negative samples. The distance loss function can be written as:

$$Loss_i = \sum_{i=1}^{N-1} \varphi([D(a, Pos_{face}) - D(a, Neg_i) + \beta_i]_+) \quad (1)$$

where a is the anchor, $D(a, Pos_{face})$ indicates the distance between anchor and the positive sample, while $D(a, Neg_i)$ indicates the distance between anchor and the i-th negative sample, β_i represents the margin value of anchor and the i-th negative sample. φ represents the $Relu$ function.

$$D(a, Pos_{face}) = ||Aud(a) - Vis(Pos_{face})||_p \quad (2)$$

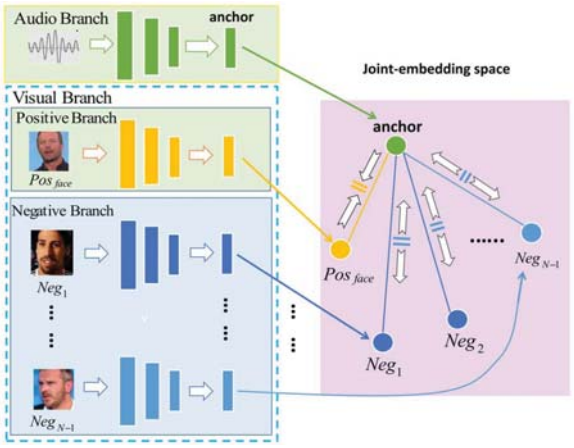

Fig. 2. The architecture of the proposed elastic matching network in the case of multi-way F2V task, while the multi-way V2F task can be constructed in the same manner. Pos_{face} is positive face, Neg_i indicates the i-th negative face and anchor symbolizes audio (in F2V tasks) or face (in V2F task). The goal of EmNet is to enlarge the distance between anchor and Neg_i while shrinking the distance between anchor and Pos_{face}.

$$D(a, Neg_i) = ||Aud(a) - Vis(Neg_i)||_p \quad (3)$$

where Aud(·), Vis(·) denote features obtained from audio and visual sub-network. It's worth noting that our network is elastic for diverse number in gallery. For the multiple samples in gallery, we only care about the distance between them which avoids complicated operations on features and enhances the applicability of the network. The multi-way V2F task can be achieved in the same manner.

3.2.2. Binary Network:

The binary network can be regarded as the special case of the multi-way network where N=2 as shown in Fig. 3, which in turn means we need one audio branch and two visual branches to obtain the audio and visual features. In this case, the loss function can be rewritten as:

$$Loss = [D(a, Pos_{face}) - D(a, Neg_{face}) + \beta]_+ \quad (4)$$

where $D(a, Pos_{face})$ denote the distance between anchor and the positive sample while $D(a, Neg_{face})$ denote the distance between the anchor and the negative sample. Eq. (4) is exactly the tripletloss. The binary V2F task can be also achieved in the same manner.

4. EXPERIMENTS

In this section, we shall introduce the details of our experiments and the results comparing to the state-of-the-arts.

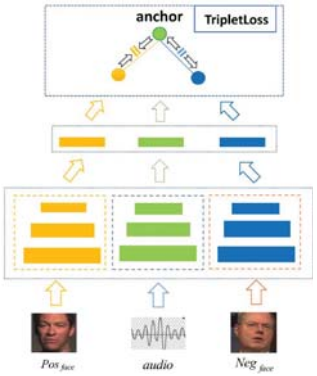

Fig. 3. Demonstration of binary F2V matching task. Where the proposed distance learning can evolved into tripletloss. The binary V2F matching task can be achieved in the same manner.

4.1. Dataset

We evaluate the proposed model on the large-scale benchmark datasets VoxCeleb [11] and VGGFace [28], which contain rich face and audio information and have been widely used in audio-visual research. Following the protocol in [3], we evaluate our EmNet on the overlap part of these two datasets, which contains about 1,000 identities.

Split of training and testing: We split the evaluation data into 604 identities for training and 189 identities for testing. Each identity contains several face images. For multi-way task, we construct the data as tuple$\{audio, Pos_{face}, Neg_1, Neg_2, ..., Neg_{N-1}\}$ for F2V task. The data construction for V2F task can be achieved in the same manner. Therefore, for binary task, we construct the data into tuples as$\{audio, Pos_{face}, Neg_{face}\}$ in F2V task while $\{face, Pos_{audio}, Neg_{audio}\}$ in V2F task. The detailed number of tuples of each task is shown in Table. 2.

Table 2. Numbers of binary task and multi-way task.

	train	**test**
identities	604	189
tuples	241,600	75,600

4.2. Implementation details

The multi-way matching task can be regarded as $N:1$ classification task where N is elastic for various integers. We set the maximum number of N = 5 in our paper and the margin in loss function is determined by p in Eq. (2). The binary matching is the special case of multi-way task when N = 2 and the loss function becomes tripletloss. In Binary task, we set β in Eq. (4) of tripletloss to be 0.6 while β_i in Eq. (1) is $\{0.1, 0.2, 0.3, 0.4\}$ in multi-way task.

4.3. Experimental results

In this section we compared the results of our experiments with the recent state-of-the-art methods SVHF-Net [3], DIMNet [4] and PINs [8] in both binary and multi-way matching tasks. Following the protocol in SVHF-Net [3], we use identification accuracy to measure the performance of the proposed method. The number of the norm (p) in Eq. (2) is set as 2-norm for multi-way and binary task.

Table 3. Results of multi-way matching task($N = 5, p = 2$) comparing to the state-of-the-arts, '$\times$' denotes not capable while '-' indicates not available. (in %)

task / model	**F2V**	**V2F**	**reference**
SVHF-Net [3]	54.92	-	CVPR 2018
DIMNet [4]	58.91	$\times$	Arxiv 2018
PINs [8]	46.13	-	ECCV 2018
EmNet	**69.37**	**68.02**	Ours

Table. 3 reports the comparison results of multi-way matching task. The matching accuracy of our model improves approximately 15% comparing to the second best method SVHF-Net [3] in F2V task which demonstrate the effectiveness of our model.

Table 4. Results of binary matching task ($N = 2, p = 2$) comparing to the state-of-the-arts, where '-' indicates not available. (in %)

task / model	**F2V**	**V2F**	**reference**
SVHF-Net [3]	79.50	81.01	CVPR 2018
DIMNet [4]	84.12	84.03	Arxiv 2018
PINs [8]	83.80	-	ECCV 2018
EmNet	**89.84**	**93.06**	Ours

Table. 4 reports the comparison results in binary matching task. Consistently with the results of multi-way matching task, our EmNet significantly beats the state-of-the-art methods SVHF-Net [3], PINs [8] and DIMNet [4] in both F2V and V2F tasks.

4.4. Ablation study

4.4.1. Evaluation on the number of norms (p)

The number of the norm in Eq. (2) is one of the key factor for the distance learning. We evaluate the number of the norms varying from 1 to 4 on both binary matching and multi-way tasks in Fig. 5 and Fig. 4.

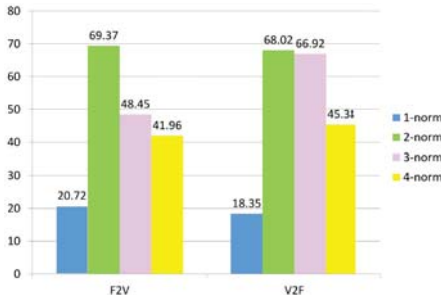

Fig. 4. Accuracy of F2V and V2F tasks of multi-way matching (N=5) against varying number of norms.

From which we can see that, in both multi-way and binary matching tasks, our method consistently achieves the best performance with 2-norm for both F2V and V2F tasks. In order to keep the parameter consistency during the evaluation, we set 2-norm for multi-way and binary matching.

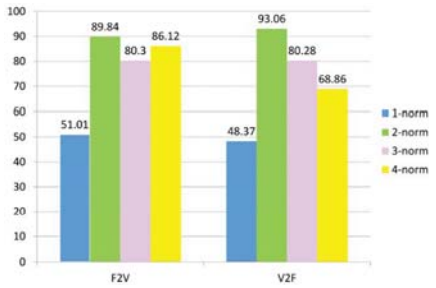

Fig. 5. Accuracy of F2V and V2F tasks of binary matching against varying number of norms.

4.4.2. Evaluation on the number of samples (N) in the gallery

To demonstrate the performance of the proposed elastic matching network, we further evaluate our method with various number of face image (or audio clips) in both F2V and V2F tasks with the fixed number of norm as 2. Fig. 6 illustrates the evaluation results where N=2 indicates the binary matching and other values for elastic multi-way matching.

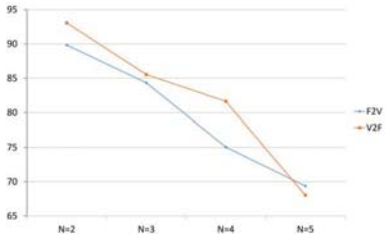

Fig. 6. Matching accuracy against the number of samples in gallery (N) with 2-norm.

As we predicted, due to the challenge of the intra-modal similarities among face images (or audio clips) and the inter-modal heterogeneity between audio and visual information, the matching accuracy decreases while N increasing. Even though, our method outperforms the state-of-the-art method with the same number of ways during the matching as mentioned in Table. 3 and Table. 4.

4.5. Training Protocol and analysis

Our approach is an end-to-end task which is trained with batch normalization by stochastic gradient descent. We set the min-batch size of 64 and use Adam optimizer with β_1 = 0.5, β_2 = 0.999 and learning-rate = $1e$-4 during the training. The weight of audio and visual branches are initialized from a Gaussian distribution. In the test, we first compute the distance between the anchor and each negative sample, and compare it with its distance between the positive samples. If $D(a, Pos_{face})$ is the smallest, or the difference between $D(a, Pos_{face})$ and $D_{min}(a, Neg_i)$ is less than β, the network stops updating.

5. CONCLUSION

We have proposed a novel distance learning method for elastic cross-modal audio-visual matching. Benefit from the metric learning, which minimizes the distance between the anchor and the positive sample while maximizing the distance between the anchor to the multiple negative samples, our approach outperform the state-of-the-art audio visual matching in both F2V and V2F tasks. Furthermore, our model allows elastic number of gallery which improves the matching capability of the cross-modal audio-visual information. Our future interest will focus on exploring the coherence between difference modalities for audio-visual matching.

6. ACKNOWLEDGEMENT

This work was partially supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (201900046), the National Natural Science Foundation of China (61602006, 61860206004, 61872005), and the Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2017A017).

7. REFERENCES

[1] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "A regularized correntropy framework for robust pattern recognition," *Neural Comput.*, pp. 2074–2100, 2011.

[2] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1561–1576, 2011.

[3] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," *IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 8427–8436, 2018.

[4] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," *arXiv preprint:abs/1807.04836*, 2018.

[5] S. Chung, J. S. Chung, and H. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," *arXiv preprint:abs/1809.08001*, 2018.

[6] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, Asian Conference on Computer Vision (ACCV)*, pp. 251–263, 2016.

[7] R. Arandjelovi and A. Zisserman, "Objects that sound," *European Conference on Computer Vision (ECCV)*, pp. 435–451, 2018.

[8] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," *European Conference on Computer Vision (ECCV)*, pp. 71–88, 2018.

[9] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*, pp. 84–92, 2015.

[10] A. S. J. T. Didac Surís, Amanda Duarte and X. Giró i Nieto, "Cross-modal embeddings for video and audio retrieval," *Workshop at European Conference on Computer Vision (ECCV)*, pp. 711–716, 2018.

[11] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint:abs/1706.08612*, 2017.

[12] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *AAAI*, 2017.

[13] Y. Zhen, Y. Gao, D.-Y. Yeung, H. Zha, and X. Li, "Spectral multimodal hashing and its application to multimedia retrieval," *IEEE Transactions on Cybernetics*, pp. 27–38, 2016.

[14] L. W. W. W. T. T. Kaiye Wang, Ran He, "Joint feature selection and subspace learning for cross-modal retrieval," in *IEEE*, pp. 2010 – 2023, 2015.

[15] P. Xu, K. Li, Z. Ma, Y.-Z. Song, L. Wang, and J. Guo, "Cross-modal subspace learning for sketch-based image retrieval: A comparative study," in *IEEE*, pp. 500–504, 2016.

[16] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, pp. 1–16, 2018.

[17] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," pp. 1198–1204, 2013.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems (NIPS)*, pp. 2672–2680, 2014.

[19] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2405–2413, 2016.

[20] S. A. Jalalifar, H. Hasani, and H. Aghajan, "Speech-driven facial reenactment using conditional generative adversarial networks," *arXiv preprint:abs/1803.07461*, 2018.

[21] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," *European Conference on Computer Vision (ECCV)*, pp. 520–535, 2018.

[22] R. Kumar, J. Sotelo, K. Kumar, A. de Brbisson, and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," *CoRR*, 2018.

[23] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3550–3558, 2018.

[24] W.-L. Hao, Z. Zhang, and H. Guan, "Cmcgan: A uniform framework for cross-modal visual-audio mutual generation," in *AAAI*, 2018.

[25] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *IEEE International Conference on Computer Vision (ICCV)*, vol. pages, pp. 609–617, 2017.

[26] J. Suárez, S. García, and F. Herrera, "A tutorial on distance metric learning: Mathematical foundations, algorithms and software," *arXiv preprint:abs/1812.05944*, 2018.

[27] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 521–528, 2003.

[28] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," pp. 41.1–41.12, BMVA Press, 2015.