Adversarial-Metric Learning for Audio-Visual Cross-Modal Matching

Aihua Zheng, Menglan Hu, Bo Jiang, Yan Huang, Yan Yan, and Bin Luo

***Abstract*—Audio-visual matching aims to learn the intrinsic correspondence between image and audio clip. Existing works mainly concentrate on learning discriminative features, while ignore the cross-modal heterogeneous issue between audio and visual modalities. To deal with this issue, we propose a novel Adversarial-Metric Learning (AML) model for audio-visual matching. AML aims to generate a modality-independent representation for each person in each modality via adversarial learning, while simultaneously learns a robust similarity measure for cross-modality matching via metric learning. By integrating the discriminative modality-independent representation and robust cross-modality metric learning into an end-to-end trainable deep network, AML can overcome the heterogeneous issue with promising performance for audio-visual matching. Experiments on the various audio-visual learning tasks, including audio-visual matching, audio-visual verification and audio-visual retrieval on benchmark dataset demonstrate the effectiveness of the proposed AML model. The implementation codes are available on https://github.com/MLanHu/AML.**

***Index Terms*—Adversarial learning, audio-visual matching, cross-modal learning, metric learning.**

Manuscript received June 10, 2020; revised October 10, 2020 and December 1, 2020; accepted January 1, 2021. Date of publication January 12, 2021; date of current version January 21, 2022. This work was supported in part by the Major Project for New Generation of AI under Grant 2018AAA0100400, in part by the National Natural Science Foundation of China under Grants 61976002 and 62076004, in part by the Natural Science Foundation of Anhui Higher Education Institutions of China under Grant KJ2019A0033, and in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (201900046), and the Cooperative Research Project Program of Nanjing Artificial Intelligence Chip Research, Institute of Automation, Chinese Academy of Sciences. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Concetto Spampinato. *(Corresponding author: Bo Jiang.)*

Aihua Zheng, Menglan Hu, and Bin Luo are with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, and also with the School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: ahzheng214@foxmail.com; 2364244962@qq.com; luobin@ahu.edu.cn).

Bo Jiang is with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education and with the School of Computer Science and Technology, Anhui University, Hefei 230601, China, and also with the Institute of Physical Science, and Information Technology, Anhui University, Hefei 230601, China (e-mail: jiangbo@ahu.edu.cn).

Yan Huang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yan.huang@nlpr.ia.ac.cn).

Yan Yan is with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616 USA (e-mail: tom_yan@txstate.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3050089.

Digital Object Identifier 10.1109/TMM.2021.3050089

I. Introduction

VISION and audition are two ways that humans exploit the world in different modalities. The studies on both human perception and neurology [1], [2] reveal the ability of humans to relate the audio segment with corresponding visual image of the same identity. Recently, there emerges an interesting topic in audio-visual learning, which aims to recognize the identity between the audio and visual cross-modality data, e.g., visual facial images and speech audio clip, as shown in Fig. 1. We refer this topic as audio-visual data recognition in this paper to distinguish it from the other audio-visual learning topics. There are three common tasks in audio-visual data recognition, e.g., Audio-visual verification, Audio-visual matching, and Audio-visual retrieval, as shown in Fig. 1. Generally speaking, audio-visual data recognition falls into two challenging categories: A $\rightarrow$ V, which aims to response the given audio clip with corresponding visual image(s) with the same identity as the audio speech from the gallery, and vise versa in V $\rightarrow$ A challenge.

Audio-visual data recognition can be potentially used in many applications in modern smart society, such as criminal investigation, identity authentication, information retrieval, etc. However, it encounters one main challenge due to the heterogeneous issue existing between audio and visual data.

For audio-visual matching, Nagrani *et al.* [3] first present the audio-visual matching model by designing a two-stream deep neural network of Seeing Voice and Hearing Faces (SVHF-Net). They address the problem by first learning the voice and facial features respectively and then formulating it as a matching problem. Albanie *et al.* [4] propose to learn a joint embedding for audio and visual modalities. They combine curriculum learning and contrastive loss optimizing in a self-supervised way for audio-visual verification, matching, and retrieval. Wen *et al.* [5] consider more covariates of the attributes information such as gender, ethnicity, identity, and learn a shared representation among them instead of directly relating audio clips and their images for audio-visual matching. Nawaz *et al.* [6] propose a single stream network which takes less cost on parameter training to learn the joint representation of visual and audio information without pairwise or triplet supervision for the three tasks (matching, verification and retrieval).

Images usually contain more visual-specific information such as color, texture, etc., while audio segments include more audio-specific information such as tone, amplitude. Despite the great progress of the existing works on audio-visual learning,

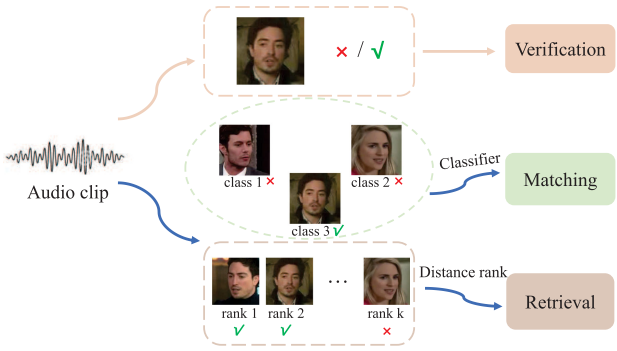


Fig. 1. Audio-visual matching, verification and retrieval in the form of A → V challenge. Audio-visual verification aims to verify whether the paired audio clip and visual image are from the same identity. Audio-visual matching tries to match the correct visual image against certain number of images for the given audio clip, normally via a classifier. Audio-visual retrieval ranks the similarity of the query against the k samples in the gallery via the distance learning scheme. The V → A challenge of matching, verification and retrieval tasks can be defined in the same manner.

the intrinsic heterogeneous issue between different modalities has not been fully studied. This modality-specific information poses a great challenge for learning discriminative feature representation for cross-modal learning. Recent works [7]–[12] indicate that adversarial learning provides an effective technique to reduce the modality gap in heterogeneous data representation. Herein, we propose to tackle the heterogeneous challenge by learning discriminative representation between audio and visual modalities via adversarial learning. To our best knowledge, it is the first work to exploit and emphasize adversarial learning in audio-visual data recognition.

Different from the original GANs [13] which aims to generate fake images that have the same distribution as the true images, we aim to learn a kind of distribution between audio and image modality, to bridge the cross-modality gap between audio and visual domains. Specifically, for a modality-specific feature such as audio-specific feature, we generate its modality-independent feature via the generator to fool the discriminator, and simultaneously train the discriminator to distinguish whether the generated feature is audio or visual modality feature. The modality-independent feature can be obtained when both generator and discriminator reach optima.

Based on the modality-independent feature, one can achieve audio-visual learning via some specific strategies for the certain tasks. For instance, we utilize a fully connected classification layer for matching task, a standard multi-layer perceptron (MLP) layer for verification task, or a distance learning scheme for retrieval. However, these strategies obviously neglect the intrinsic relationship among samples. Herein, we further propose to integrate deep metric learning [14]–[20] into the adversarial learning framework, to obtain a more robust similarity measure for audio-visual matching. Specifically, we pursue to enhance the similarity between the positive pair (the anchor/query audio feature and the corresponding image feature), while enlarge the distance between the negative set and the positive pair. This metric learning method has been demonstrated to obtain a better metric embedding in training data [19].

Based on the above observations, we propose a novel Adversarial-Metric Learning (AML) framework for audio-visual data recognition as shown in Fig. 2. AML aims to generate discriminative modality-independent representation via adversarial learning, while simultaneously learns a robust metric for audio-visual cross-modality matching and retrieval in an end-to-end manner. Specifically, the proposed AML consists of four modules: 1) Audio-visual sub-networks to extract the voice and facial features respectively. 2) Adversarial learning module which contains a generator and a discriminator to learn a modality-independent representation. 3) Metric learning module to simultaneously learn the feature embedding for audio-visual matching and retrieval. 4) Task-specific module for specific task, such as matching, verification and retrieval in audio-visual data recognition problem.

To our best knowledge, this is the first work to jointly employ adversarial and metric learning for audio-visual cross-modal matching although they have been used in some other cross-modal learning problems [21]–[24]. The main contributions are summarized as follows:

- We propose to employ an adversarial learning mechanism to explore the discriminative feature representation in audio-visual learning. The proposed method can obtain the desired modality-independent feature representation.
- We propose to incorporate metric learning into adversarial learning to learn a robust feature embedding among audio-visual modality-independent features.
- Comprehensive experiments on three audio-visual learning tasks, including matching, verification and retrieval yield to a new state-of-the-art comparing to prevalent methods on the benchmark dataset constructed from the overlap of VoxCeleb [25] and VGGFace [26].

II. Related Works

Audio-visual learning aims to discover the intrinsic relationship between audio and visual data. Audio-visual matching is a new branch of audio-visual learning, and can be regarded as a special cross-modal learning task. Herein, we will briefly review the literature on general audio-visual learning, together with specific audio-visual matching and the advances in common cross-modal learning as the related works.

A. Audio-Visual Learning

Audio-visual learning is a board research area and contains various topics [27], including audio-visual separation and localization, audio-visual correspondence learning, audio-visual generation and audio-visual representation learning. We briefly highlight several representative tasks, such as vision aided speech recognition [28], [29], object sound localization [30]–[32], audio-visual generation [33]–[36], etc. Although conventional speech recognition [37], [38] achieve remarkable performance on isolating a single speaker from a noisy environment in solely audio modality, recent efforts [28], [29] take the advantage of the complementary visual information and audio information to boost the performance of conventional speech recognition. Object sound localization [30], [31] aims to

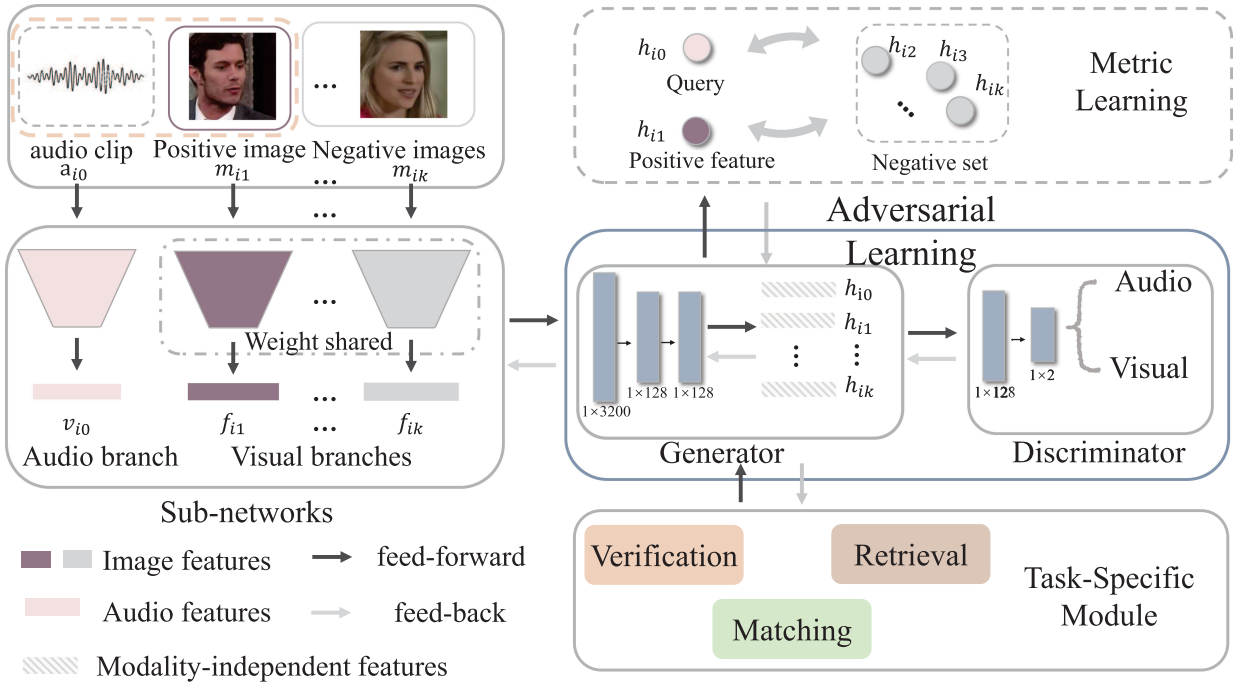


Fig. 2. Pipeline of the proposed AML in the term of A → V challenge. We take a data tuple consists of a query audio clip a_{i0}, one positive image m_{i1} and $k-1$ negative images as gallery. After obtaining the audio feature v_{i0} and the k image features $\{f_{i1}, .., f_{ik}\}$, we propose to learn the modality-independent feature representation $\{h_{i0}, h_{i1}, .., h_{ik}\}$ via adversarial learning. Meanwhile, we propose to learn a robust feature embedding for similarity measure via metric learning. Note that adversarial learning and metric learning are jointly learned in a unified framework. The task-specific module is followed for the specific audio-visual learning tasks, including matching, verification and retrieval. In practise, the metric learning module is wiped for verification task since only one (negative or positive) sample in the gallery.

localize the sound source in the visual context. With the increasing number of videos, various object sound localization methods develop from supervised manner [30], [31] to unsupervised learning [39]. Moreover, attention mechanism [40] has been noted as an effective technique to emphasize the sound-emitting objects in object sound localization. With the blossom of the Generative Adversarial Networks (GANs), there emerges an interesting topic named audio-visual generation, which aims to generate visual context from sound [41], or vise versa [42], or both [43], [44]. In generating audio speeches from videos, related works mainly focus on recovering audio from lip movements [45], [46] or silent videos [47]. While in generating images/videos from audio clips, talking face generation [33], [34], [48], body motion generation [35], [36] and image generation from audio data [49], [50] have attracted increasing attention recently. More details and works on audio-visual learning can be referred in the recent survey [27].

B. Audio-Visual Matching

To recognize the cross-modality audio and visual data, Hoover *et al.* [51] present a method to associate faces with audio segments in a video by detecting and clustering. They first detect and cluster images and audio segments via VLAD (vector of locally aggregated descriptors) [52] followed by frame cluster assignment to the given voice cluster based on majority principle. With the blossom of deep neural networks, Arsha Nagrani *et al.* [3] first announce the audio-visual matching mission based on a two-stream deep neural network, which learns audio and visual features respectively followed by a softmax layer for classification. Albanie *et al.* [4] propose a self-supervised method to learn a joint embedding of audio clips and facial images for cross-modal retrieval through contrastive loss. Furthermore, they design a novel curriculum learning strategy to obtain more information during training. Kim *et al.* [53] propose a feature representation learning method by computing the common information between audio and visual modality to learn a co-embedding. Furthermore, Wen *et al.* [5] consider more covariates of the attributes information such as gender, ethnicity, identity, etc., and learn a shared representation among them instead of directly relating audio clips and images. Different from above two-stream network based methods, Nawaz *et al.* [6] suggest a single stream network to learn the joint representation of visual and audio information with fewer parameter during training and without pairwise or triplet information supervision. However, these methods mainly concentrate on the audio and visual feature representation, while neglecting the heterogeneous issue between audio and visual modalities, which brings a huge challenge in audio-visual cross-modal data.

C. Cross-Modal Learning

Cross-modal learning is a general research topic which mainly focuses on learning the correspondence among multimedia data, such as text, image, audio, video, etc. Representative works include hashing transformation [54], [55], adversarial

learning [10], [21], [22] and both [56], [57]. Chen *et al.* [55] propose a dual cross-modal hash method for image-text retrieval in three stages, which supervised the second stage visual hash code learning by the learned text hash code in the first stage, and optimized them together via reconstruction in the last stage. Su *et al.* [58] propose to explore the latent semantic relations among input images and texts by computing a joint-semantics affinity matrix from the neighborhood information and reconstruct the joint-semantics structure maximally by the learned binary codes. Jiang *et al.* [23] propose a novel adversarial learning network to learn the view-invariant and consistent pixel-level representation for RGB and depth images in salient object detection. Xu *et al.* [22] propose to minimize the gap between modalities and intra-class variation for image-text retrieval. Specifically, they project labeled data pairs into a shared nonlinear latent space via adversarial learning. Li *et al.* [57] propose an unsupervised hashing network to learn a shared representation and generated robust hash codes simultaneously via a coupled cycle network. It aggregates the advantage of generative adversarial network and hashing transformation for image-text retrieval. Zhang *et al.* [59] propose to utilize the generative adversarial model to exploit the underlying manifold structure across modality in an unsupervised way. Yu *et al.* [60] propose a semi-supervised method to exploit semantic information of unlabeled data. To obtain the optimal matrices across modalities, they utilize the label graph to maintain the geometric structure among cross-modal features and employ $l_{2,1}$-norm for feature selection. Song *et al.* [61] propose a memory network to store discriminative features, followed by the learning network for the common representation with the help of the memory mechanism and adversarial learning. Xu *et al.* [12] propose to solve the zero-shot image-text retrieval problem via adversarial learning to maximize the consistency and correlation between different modality features. Chi *et al.* [11] propose to utilize two generative adversarial networks to learn image and text common embedding respectively via generation and reconstruction. Despite the widely applied adversarial learning in other retrieval tasks, we first try to settle the audio-visual learning via adversarial learning in this paper.

III. Proposed Method

In the sake of generality and simplicity, we only elaborate the A $\rightarrow$ V challenge in audio-visual matching for methodology description. The V $\rightarrow$ A challenge can be defined in the same manner.

Problem Formulation: Given a data tuple consisting of an audio modality clip a_{i0} as the query and k visual modality images $m_i = \{m_{i1}, .., m_{ik}\}$ as the gallery, audio-visual matching aims to discover the corresponding facial image(s) from the gallery for the query voice audio. Specifically, i denotes the i-th data tuple, and $l_i \in \{1, k\}$ is the label, indicating the current query a_{i0} matches the the l_i-th image m_{il_i}. Note that $k = 1$ indicates the verification task to verify whether the audio-visual pair is consistent with the same identity. A certain fixed number of $k > 1$ involves the matching task, where $k = 2$ indicates the binary matching case, otherwise the multi-way matching case. Retrieval task can be regarded as a general case which queries the correct hit from a certain number of k samples in the gallery based on the distance learning scheme.

A. Overall Framework

As shown in Fig. 2, our framework consists of four modules. (1) Audio-visual sub-networks, with an audio branch and k visual branches, to extract the audio-specific and visual-specific features respectively. (2) Adversarial learning based representation learning, with a generator and a discriminator, to learn a modality-independent feature representation via the min-max game between feature generator and modality discriminator. (3) Metric learning based feature embedding, to learn a robust distance metric for audio-visual learning. (4) Task-specific module, to fulfill different tasks in audio-visual learning, such as matching, verification and retrieval.

B. Audio-Visual Sub-Networks

Let $\mathcal{A}$ and $\mathcal{V}$ denote the audio and visual feature spaces respectively. First, we obtain the audio-specific feature $\boldsymbol{f_{i0}^a} \in \mathcal{A}$ and visual-specific features $\{\boldsymbol{f_{i1}^v}, .., \boldsymbol{f_{ik}^v}\} \in \mathcal{V}$ via the audio-visual sub-networks, containing one audio branch and k visual branches. The audio branch and visual branch consists of different five-convolutional-layers respectively in our framework. All the images are normalized to the size of 224*224*3. The image feature of each face image is extracted via the corresponding visual branch with parameter θ_V. The input of the audio branch is an audio clip that has been resized to 224*125*1 after converting to a single-channel audio spectrogram. The audio feature of each audio clip is extracted by the audio branch with parameter θ_A.

C. Adversarial Learning Based Representation

To eliminate the undesired impact of the heterogeneous issue between audio and visual modalities, we propose to learn modality-independent representation to mitigate the cross-modality gap via adversarial learning (AL). AL consists of a generator G and a discriminator D to beat each other with a min-max game to find the latent feature space $\mathcal{H}$. The G takes $\boldsymbol{f_{i0}^a}$ and $\{\boldsymbol{f_{i1}^v}, .., \boldsymbol{f_{ik}^v}\}$ as input, and aims to generate modality-independent features $\{\boldsymbol{h_{i0}}, .., \boldsymbol{h_{ik}}\} \in \mathcal{H}$, while the D provides a modality classifier to discriminate the modality of audio and visual features.

Generator: The generator G with parameter θ_G is constructed with a standard MLP to learn two mapping functions $\{\phi : \mathcal{V} \rightarrow \mathcal{H}\}$ and $\{\psi : \mathcal{A} \rightarrow \mathcal{H}\}$,

$$\boldsymbol{h_{i0}} = \phi(\boldsymbol{f_{i0}^a}; \theta_G), \tag{1}$$

$$\boldsymbol{h_{ij}} = \psi(\boldsymbol{f_{ij}^v}; \theta_G), j \in [1, k]. \tag{2}$$

The ϕ and ψ aims to map audio-specific feature $\boldsymbol{f_{i0}^a}$ and visual-specific features $\{\boldsymbol{f_{i1}^v}, .., \boldsymbol{f_{ik}^v}\}$ to modality-independent features respectively.

Discriminator: The discriminator D is designed as a FC network, which is defined as a modality classifier with parameter θ_D to discriminate the original modality of feature representation $\boldsymbol{h_{ij}}$ obtained from the generator. In order to avoid the

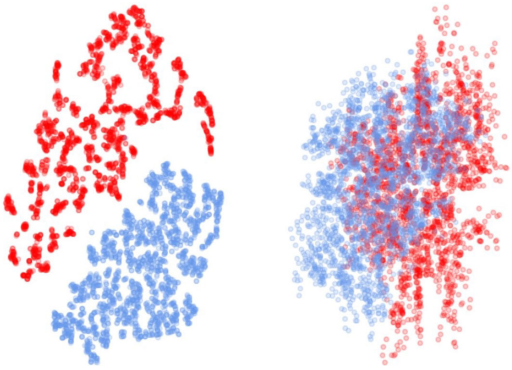


Fig. 3. Visualization of the feature representation of 50 identities before and after the adversarial learning. Red and blue points represent the audio and image features respectively.

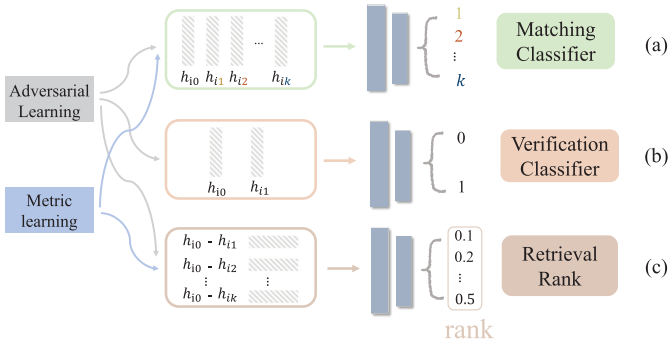


Fig. 4. Architecture of audio-visual matching, verification and retrieval tasks. (a) Audio-visual matching, takes the features of AL and ML as input and aims to predict the class of the given audio clip in the k class gallery. $k = 2$ for the binary case. (b) Audio-visual verification, is achieved via a binary classifier to verify whether the input audio and image features after adversarial learning are with the same identity. (c) Audio-visual retrieval, utilizes the MLP layer to compute the distance between the query sample and each gallery sample and then ranks the query sample based on the distance.

mode collapse and instability problem, we adopt the weight clipping scheme[1] during the discriminator training as mentioned in WGAN [62]. The discriminator is trained by minimizing,

$$\mathcal{L}_{dis} = -\frac{1}{M_{train}} \sum_{i=1}^{M_{train}} \sum_{j=0}^{k} y_{ij} \log D(\boldsymbol{h_{ij}}; \theta_D), \tag{3}$$

where y_{ij} represents the modality label of the j-th sample in the i-th data tuple, $D(\boldsymbol{h_{ij}}; \theta_D)$ is the modality probability of the output of generator. M_{train} denotes the number of training data tuples.

After the adversarial learning, we can obtain modality-independent representations $\{\boldsymbol{h_{i0}}, .., \boldsymbol{h_{ik}}\}$ for both audio and visual modalities.

To better demonstrate the advantage of adversarial learning in the proposed AML, we use a 2D t-SNE [63] to visualize the features of 50 identities before and after adversarial learning as shown in Fig. 3. Note that the metric learning is not into consideration here. One can note that the distances between audio and visual features become closer and better mixed together after adversarial learning, which demonstrates that adversarial learning can alleviate the undesired impact of the heterogeneous issue existing between audio and visual modalities.

D. Deep Metric Learning Based Embedding

After obtaining the modality-independent features from generator, we further propose to employ deep metric learning (ML) method for better similarity measure learning. Inspired by [19] which preserves the intra-class structure and inter-class variation during learning, we propose to preserve the structure between positive pair and the variation among negative set. Specifically, we propose to employ the structured loss function to pull the distance between the positive pair, $i.e.$ $\boldsymbol{h_{i0}}$ and $\boldsymbol{h_{i1}}$, while pushing the negative set $\{\boldsymbol{h_{i2}}, .., \boldsymbol{h_{ik}}\}$ away from the positive pair as shown in Fig. 2. The optimization objective of metric learning is defined as [19],

$$\mathcal{L}_{metric} = \frac{1}{2M_{train}} \sum_{i=1}^{M_{train}} \max(0, \mathcal{J}_i)^2, \tag{4}$$

$$\mathcal{J}_i = \log\left(\sum_{j\in[2,k]} e^{\lambda - d_{i0,ij}} + \sum_{q\in[2,k]} e^{\lambda - d_{i1,iq}} \right) + d_{i0,i1}, \tag{5}$$

where $d_{i0,i1}$ measures the Euclidean distance between $\boldsymbol{h_{i0}}$ and $\boldsymbol{h_{i1}}$, $d_{i1,iq}$ measures the Euclidean distance between $\boldsymbol{h_{i1}}$ and $\boldsymbol{h_{iq}}$, and λ is a hyper-parameter which controls the margin of the distance between the negative set and positive pair. $\mathcal{J}_i$ is the distance that contains the similarity between positive pairs and dissimilarities between negative sets and positive pairs. From the Eq. (4) and Eq. (5), we encourage to minimize $\mathcal{J}_i$ by enlarging the inter-class distances $d_{i0,ij}$ and $d_{i1,iq}$ while minimizing the intra-class distance $d_{i0,i1}$.

E. Task-Specific Module for Audio-Visual Learning

Audio-Visual Matching Classifier: After the adversarial learning and metric learning, one can directly compute the distance between the given audio clip to each visual sample for matching evaluation. Considering the nonlinear fitting ability of neural networks, we design a fully connected network as the position classifier on the modality-independent features for the audio-visual matching. Specifically, we concatenate the modality-independent features and feed them to the matching classifier with parameter θ_{C_M} as shown in Fig. 4. We can use the commonly used cross-entropy loss [24] which is defined as

$$\mathcal{L}_{cls} = -\frac{1}{M_{train}} \sum_{i=1}^{M_{train}} l_i \log C_M(\{\boldsymbol{h_{i0}}, .., \boldsymbol{h_{ik}}\}; \theta_{C_M}), \tag{6}$$

where C_M denotes the audio-visual matching classifier to compute the probability that the audio clip belongs to each image. Note that, $k = 2$ for the binary matching case.

[1]URL: [Online]. Available: https://github.com/YadiraF/GAN

Audio-visual Verification Classifier: Verification task can be regarded as the situation with only one sample in the gallery. Therefore, we only conduct the adversarial learning procedure while omitting the metric learning in our AML framework for verification task. After the adversarial learning, the verification task is fulfilled via a binary classifier with parameter θ_{C_V} to discriminate whether the paired modality-independent features are from the same identity. SSNet [6] verifies the paired data by adjusting whether the distance between the two features is smaller than an adaptive threshold. Different from this linear transformation, we employ a nonlinear verification classifier for verification, shown in Fig. 4. Specifically, the paired modality-independent features are fed into the binary classifier after concatenated. The commonly used cross-entropy loss [24] can be defined as

$$\mathcal{L}_{cls} = -\frac{1}{M_{train}} \sum_{i=1}^{M_{train}} l_i \log C_V(\{\boldsymbol{h_{i0}}, \boldsymbol{h_{i1}}\}; \theta_{C_V}), \quad (7)$$

where C_V denotes the audio-visual verification classifier to determine whether the paired data is consistent.

Audio-visual Retrieval Distance Learning: Audio-visual retrieval task aims to search the correct visual sample(s) from the certain number of samples in the gallery for each query.

As the gallery size varies in the audio-visual retrieval problem, a classifier such as in matching or verification is not suitable after AL and ML. Herein, we design a distance learning method to learn the similarity between query and gallery. Different from the existing method [6] which directly ranks the features in the gallery via Euclidean distance, we design a MLP layer to compute the distance between the query sample $\boldsymbol{h_{i0}}$ and each gallery sample $\boldsymbol{h_{ij}}$ because of its nonlinear fitting ability. following the commonly used cross-entropy loss [24], we define the loss function as

$$\mathcal{L}_{cls} = -\frac{1}{M_{train}} \sum_{i=1}^{M_{train}} \sum_{j=1}^{k} l_i \log C_R(\{\boldsymbol{h_{i0}} - \boldsymbol{h_{ij}}\}; \theta_{C_R}), \quad (8)$$

where $C_R(\{\boldsymbol{h_{i0}} - \boldsymbol{h_{ij}}\}; \theta_{C_R})$ outputs the distance of $\boldsymbol{h_{i0}}$ and $\boldsymbol{h_{ij}}$ and k is the number of the samples in the gallery.

F. Joint Learning Process

Our model integrates metric learning into adversarial learning to joint learn the modality-independent feature representation and feature embedding. Specifically, we update the generator by minimizing,

$$\mathcal{L}_{total} = \beta \mathcal{L}_{metric} + \gamma \mathcal{L}_{cls}, \quad (9)$$

where β and γ are hyper-parameters. For simplicity, we let $\theta_C = \{\theta_{C_M}, \theta_{C_V}, \theta_{C_R}\}$. The objective is optimized by alternatively solving,

$$\min_{\theta_V, \theta_A, \theta_G, \theta_C} (\mathcal{L}_{total}(\theta_V, \theta_A, \theta_G, \theta_C) - \mathcal{L}_{dis}(\hat{\theta}_D)) \quad (10)$$

$$\max_{\theta_D} (\mathcal{L}_{total}(\hat{\theta}_V, \hat{\theta}_A, \hat{\theta}_G, \hat{\theta}_C) - \mathcal{L}_{dis}(\theta_D)) \quad (11)$$

Algorithm 1: Optimization Process of AML

Require: Audio branch θ_A, visual branch θ_V, generator θ_G, classifier θ_C, discriminator θ_D, hyper-parameters: λ, β, γ, mini-batch: N, learning rate: r_D, r_G, the number of training steps of the discriminator: T

1: **for** $i < M_{train}$ **do**
2: Randomly select training pairs$\{a_{i0}, m_i\}$
3: (a_{i0} is the audio clip, $m_i = \{m_{i1}, .., m_{ik}\}$ are images)
4: **end for**
5: **while** not convergence**do**
6: Calculate the loss $\mathcal{L}_{dis}$ and $\mathcal{L}_{total}$
7: **for** T steps**do**
8: Update parameter θ_D by ascending their stochastic gradients
9: $\theta_D \leftarrow \theta_D + r_D \nabla_{\theta_D} \frac{1}{N}(\mathcal{L}_{total} - \mathcal{L}_{Dis})$
10: **end for**
11: Update parameter $\theta_V, \theta_A, \theta_G, \theta_C$ by descending their stochastic gradients
12: $\theta_V \leftarrow \theta_V - r_G \nabla_{\theta_V} \frac{1}{N}(\mathcal{L}_{total} - \mathcal{L}_{Dis})$
13: $\theta_A \leftarrow \theta_A - r_G \nabla_{\theta_A} \frac{1}{N}(\mathcal{L}_{total} - \mathcal{L}_{Dis})$
14: $\theta_C \leftarrow \theta_C - r_G \nabla_{\theta_C} \frac{1}{N}(\mathcal{L}_{total} - \mathcal{L}_{Dis})$
15: $\theta_G \leftarrow \theta_G - r_G \nabla_{\theta_G} \frac{1}{N}(\mathcal{L}_{total} - \mathcal{L}_{Dis})$
16: **end while**
17: **return** Feature representation mapping function $\{\phi : \mathcal{V} \rightarrow \mathcal{H}\}$ and $\{\psi : \mathcal{A} \rightarrow \mathcal{H}\}$

where $\hat{\theta}$ indicates fixing the parameter. In each epoch, we first update our discriminator for T steps by $\mathcal{L}_{dis}$, then update our generator by $\mathcal{L}_{total}$. Similar to works [13], [21], we use the following algorithm to optimize $\mathcal{L}_{total}$ and $\mathcal{L}_{dis}$ as shown in Algorithm 1.

G. Implementation Details

We conduct all our experiments on NVIDIA GeForce 1080Ti graphic card. We first extract the audio and image features with the same size of 10*10*32 via audio and visual branches respectively, as the input to the generator. During the adversarial learning, the generator applies a standard multi-layer perceptron, which transforms the 3200 dimensional features to 128 dimensional ones for both audio and visual modality. Then the discriminator further transforms them to 2 dimensional ones, indicating the probabilities belonging to audio and visual modalities respectively. As for the metric learning, we reproduce the method Lifted Struct [19] due to its ability to preserve the intra-class structure and inter-class variation, which can be referred in our released codes on Github.[2] All the matching and verification classifiers, and the retrieval distance learning scheme apply a single layer fully connected network but with different sizes. Specifically, the matching classifier receives $(k+1)*128$ dimension features and output k dimensional ones in the form of vectors indicating the probability of probe belongs to each

[2]URL: [Online]. Available: https://github.com/MLanHu/AML

TABLE I
DATASET INFORMATION

VGGFace	Identities	Male	690
		Female	561
	Face images		995,705
VoxCeleb	Identities	Male	690
		Female	561
	Audio segments		153,486

data in the gallery. The verification classifier receives two 128 dimensional audio and visual features respectively and outputs a 2 dimensional feature, determining whether the input audio and visual features derive from the same identity or not. The distance learning scheme in retrieval transforms a set of 128 dimensional features into 1 dimensional distances representing their distances to the query. we initialize the weights of the network by Glorot initialization [64]. During the training, the batch size set to 50. And we adopt batch normalization by Adam [65] to optimize the network. We set distinguishing learning rates for audio-visual sub-networks, generator, task-specific module and discriminator in AML, due to their diverse convergence speeds. Specifically, we use a logarithmically decaying learning rate strategy for audio-visual sub-networks, generator and task-specific module, which decaying from 5e-3 to 5e-5, and set the learning rate as 5e-3 for the discriminator. The iteration is fixed as $50, 20, 100$ for matching,verification and retrieval respectively for the best performance. In addition, we empirically set the hyper-parameters $\{\lambda, \beta, \gamma\}$ to {1.0, 3.0, 2.0} for cross-modal matching and retrieval. In cross-modal verification task, we set $\{\lambda, \beta, \gamma\}$ to {0.0, 0.0, 2.0} as the metric learning method demands more than one sample in the gallery while there is only one (negative or positive) sample in the gallery for verification problem.

IV. EXPERIMENTS

We have implemented the the proposed method on audio-visual learning on three tasks, including audio-visual matching, verification and retrieval to verify the performance of the proposed AML comparing with the state-of-the-art methods.

A. Dataset

Following the protocol in [3]–[5], we evaluate the proposed AML on the overlap of two large-scale benchmark datasets, the speaker identification dataset VoxCeleb [25] and face recognition dataset VGGFace [26]. VoxCeleb dataset [25] consists of 153486 audio segments of 1251 speakers, and the VGGFace [26] consists of 995705 aligned face images of 1251 identities. More information can be observed in Table I.

B. Evaluation on Audio-Visual Matching

We compare our method on audio-visual matching task to the prevalent methods SVHF-Net [3], DIMNet [5], PINs [4] and SSNet [6] in both binary and multi-way cases. We construct the training and testing sets in the same manner as [3]–[5] for fair comparison. Specifically, we select 153 identities whose names start with 'C,' 'D,' and 'E' for testing, while the rest 942 identities whose names starting with 'F' - 'Z' for training. The gender and age distributions are balanced in both training and testing sets. Table II illustrates the number of data tuples we selected during training and testing. Taking A $\rightarrow$ V challenge as an example, we randomly select fifteen negative facial images for every paired data in binary while fifty sets of data consisting of $k-1$ negative facial images in multi-way case. We randomly select data tuples before training and fixed them throughout the training process.

Table III reports the matching accuracies in both V $\rightarrow$ A and A $\rightarrow$ V challenges in both binary and 10-way cases. It is clearly to see that, 1) AML yields a new state-of-the-art on both binary and 10-way in either V $\rightarrow$ A or A $\rightarrow$ V challenge. 2) Comparing to the second-best method DIMNet [5] which utilizes the attribute information, our AML still dominate by large margins, which verifies the competency of AML in more challenge scenarios.

TABLE II
NUMBERS OF TUPLES DURING TRAINING AND TESTING FOR AUDIO-VISUAL MATCHING

Training		Testing	
Binary	Multi-way	Binary	Multi-way
14130	2295	47100	7650

TABLE III
COMPARISON RESULTS OF AUDIO-VISUAL MATCHING AGAINST STATE-OF-THE-ART METHODS ON BOTH BINARY ($k=2$) AND MULTI-WAY ($k=10$) CASES, WHERE '-' INDICATES 'NOT AVAILABLE,' '$\times$' INDICATES 'NOT CAPABLE'. (IN %)

Method \ Task	Binary		Multi-way	
	A $\rightarrow$ V	V $\rightarrow$ A	A $\rightarrow$ V	V $\rightarrow$ A
SVHF-Net [3]	81.00	79.50	34.50*	$\times$
DIMNet [5]	84.12	84.03	39.75	-
PINs [4]	84.00*	-	31.00*	-
SSNet [6]	78.00	78.50*	30.00	30.05*
AML	**92.72**	**93.33**	**43.45**	**39.35**

*indicates the approximate values we estimated from the limited results provided by corresponding publications.

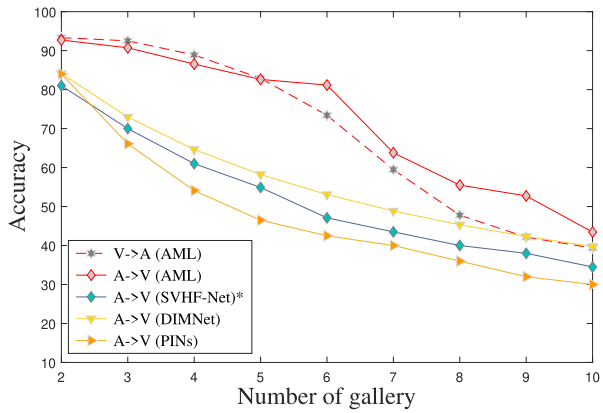


Fig. 5. Matching accuracy of different number of samples in gallery from binary ($k=2$) to multi-way ($k=10$) cases in both V $\rightarrow$ A and A $\rightarrow$ V challenges.

Fig. 6. Qualitative results of audio-visual cross-modal matching of the proposed AML comparing to DIMNet [5], SVHF-Net [3] in A → V challenge with $k = 2$. The shadowed images with the volume icons in the lower right corners represent the query audio clips with corresponding identities. The facial images with the green ticked and the red crossed bounding boxes indicate the correct and wrong matching results respectively.

Fig. 5 demonstrates the performance of the proposed AML against other methods with the number of samples in the gallery k from 2 to 10. As can be seen that, 1) with the number of the samples in the gallery (k) increasing, the challenge of the matching consistently increases in both V → A and A→ V challenges, which leads to the decreasing matching accuracy. 2) The proposed AML outperforms the state-of-the-art methods in all the different numbers of gallery in A → V challenge.> Note that the compared state-of-the-art methods are either not available or not capable in V → A challenge, we only demonstrate our results in V → A challenge.

Furthermore, we qualitatively demonstrate two matching results of our proposed AML comparing with the state-of-the-art methods DIMNet [5] and SVHF-Net [3] in A → V challenge with gallery number $k = 2$, which is the only scenario in the released SVHF-Net [3] codes. As shown in Fig. 6, due to the large inter-class similarity, both DIMNet [5] and SVHF-Net [3] produce the wrong matching for query (b) and (a) respectively, while our AML successfully hits the correct matchings for both queries. This further indicate the robustness of our AML on handling audio-visual cross-modality matching.

C. Evaluation on Audio-Visual Verification

Audio-visual cross-modal verification aims to verify whether the input pair of audio clip and visual image belongs to the same identity. Following the experimental protocol in [4], we split the data in Voxceleb dataset [25] into three sets: training set, seen-heard set and unseen-unheard set. First, we select 901 identities and choose a part of samples (in the same protocol as [4]) of each identity for training, while the rest part of samples for testing. Since each identity in the testing set appears in the training set, we name it as "seen-heard" set. Then, we further select 250 identities not existed in the training set as the unseen-unheard testing set. Table IV elaborates the data structure for verification in details.

We quantify the verification results on the standard metric, the AUC (Area Under Curve) of the ROC (Receiver Operating Characteristic) curve, which reflects the true positive and false positive rates. The higher AUC, the better verification accuracy. Note that AUC $\subseteq [0.5, 1.0]$.

Table V reports the verification results comparing to the state-of-the-art methods. One can note that AML outperforms the prevalent methods SSNet [6] and PINs [4] especially on the more challenging unseen-unheard testing set, which verifies the effectiveness of the proposed adversarial learning for audio-visual cross-modal learning.

TABLE IV
DATA STRUCTURE OF THE TRAINING SET, SEEN-HEARD, AND UNSEEN-UNHEARD TESTING SETS RESPECTIVELY

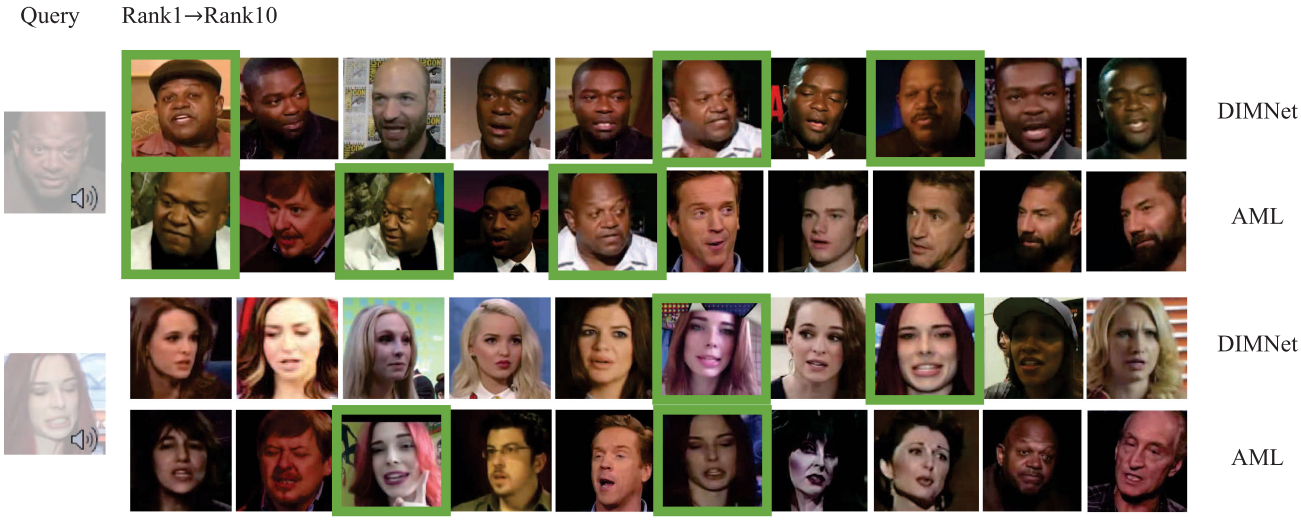
	Training	Testing	
		seen-heard	unseen-unheard
Identity	901	901	250
Data pair	423,004	18,020	30,496

TABLE V
COMPARISON RESULTS OF AUDIO-VISUAL VERIFICATION AGAINST STATE-OF-THE-ART METHODS ON BOTH SEEN-HEARD AND UNSEEN-UNHEARD CASES ON METRIC AUC (IN %)

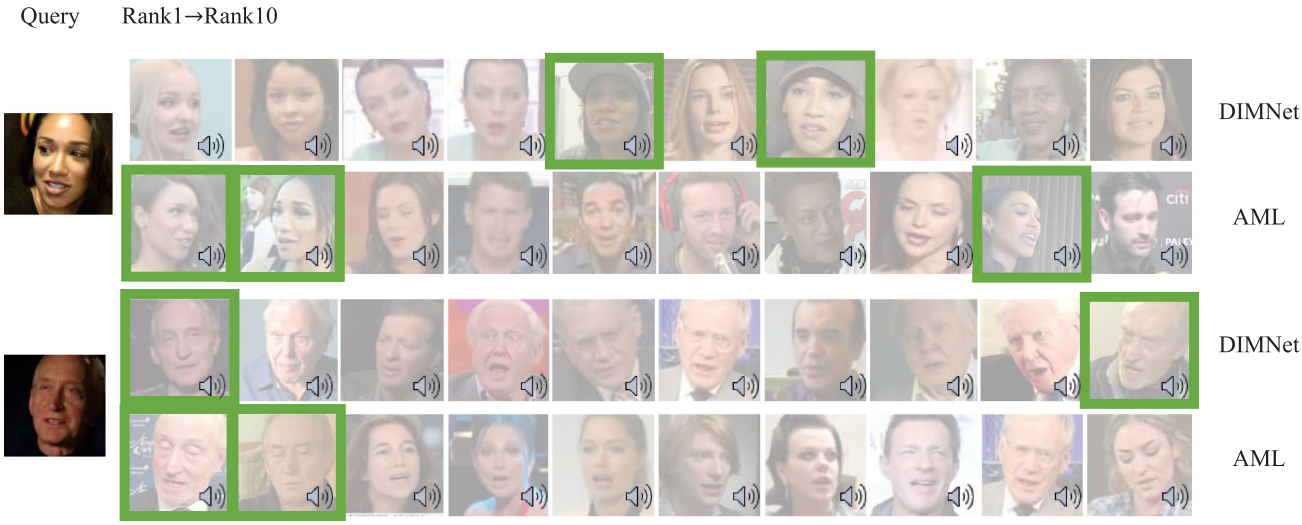
	seen-heard	unseen-unheard
PINs [4]	73.80	63.50
SSNet [6]	91.20	78.80
AML	**92.30**	**80.60**

TABLE VI
COMPARISON RESULTS OF AUDIO-VISUAL RETRIEVAL AGAINST SSNET [6] AND DIMNET [5] IN BOTH A → V AND V → A CHALLENGES ON METRIC R@10 (IN %)

Challenge \ Method		SSNet	AML	DIMNet	AML
V→A	seen-heard	50.00	**55.75**	64.05	**65.35**
	unseen-unheard	13.20	**16.03**		
A→V	seen-heard	36.27	**47.83**	87.58	**88.23**
	unseen-unheard	8.70	**11.39**		



(a) A → V



(b) V → A

Fig. 7. Qualitative results of audio-visual cross-modal retrieval of the proposed AML comparing with DIMNet [5] in both A → V and V → A challenges. We demonstrate the gallery rankings from $R@1$ to $R@10$, where the results in green boxes indicate the right hits.

TABLE VII
ABLATION STUDY ON OF PROPOSED AML ON AUDIO-VISUAL MATCHING TASK IN BOTH BINARY (WHEN $k = 2$) AND MULTI-WAY (WHEN $k = 10$) CASES IN V → A CHALLENGE. '✓' MEANS THE CORRESPONDING COMPONENT IS INCLUDED. (IN %)

Component \ Task	Binary ($k = 2$)				Multi-way ($k = 10$)			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
Adversarial Learning		✓		✓		✓		✓
Metric Learning			✓	✓			✓	✓
Accuracy	83.13	86.23	85.18	**93.33**	30.36	33.68	33.74	**39.35**

D. Evaluation on Audio-Visual Retrieval

Audio-visual cross-modal retrieval aims to retrieve the same identity with the given query from the gallery in the other modality. Compared to audio-visual matching task, audio-visual retrieval is more general yet more challenging due to the larger number of gallery. Different from the existing methods [5], [6] which ranks the distance between the query and each cross-modal samples in the gallery according to Euclidean distance, we design a MLP layer to evaluate the distance between the query sample and the gallery samples.

We evaluate the performance of AML, SSNet [6] and DIMNet [5] in both A → V and V → A challenges on audio-visual

retrieval task on the metric $R@10$. Table VI reports the comparison results. Note that the data splitting protocols are different in SSNet [6] and DIMNet [5] on retrieval task. We follow the two protocols respectively while comparing with corresponding method in Table VI.

To compare with SSNet [6], we follow the same the protocol as mentioned in [4], where the data is split into training, seen-heard and unseen-unheard testing sets in the same manner as in verification as in Section IV-C. Specifically, we select each audio clip in the testing set as the query and remain all the visual images (901 identities for seen-heard testing set while 250 for unseen-unheard) as the gallery in A $\rightarrow$ V challenge, vise versa in V $\rightarrow$ A challenge. Since there is no seen-heard or unseen-unheard split in DIMNet [5], we only compare the retrieval results in both A $\rightarrow$ V and V $\rightarrow$ A challenges. Specifically, we select each audio clip in the testing set (153 identities) as the query and take all the visual images as gallery in A $\rightarrow$ V challenge, vise versa in V $\rightarrow$ A challenge.

As shown in Table VI, generally speaking, the audio-visual cross-modal retrieval task is much more challenging comparing to both verification and matching tasks due to the larger number of samples in the gallery, which results in much lower accuracies with both our AML and SSNet [6] especially in unseen-unheard scenarios. It is clear that AML outperforms SSNet [6] in both A $\rightarrow$ V and V $\rightarrow$ A challenges in either seen-heard or unseen-unheard case. Noted that the improvement of our AML comparing to SSNet [6] on seen-heard case in much higher than unseen-unheard case, the main reason is that the evaluation on unseen-unheard set is a zero-shot problem which is more challenging than seen-heard case. DIMNet [5] achieves impressive performance on the retrieval task due to the utilizing of attribute information. However, it is still overshadowed by AML in both A $\rightarrow$ V and V $\rightarrow$ A challenges, which indicates the contribution of the adversarial and metric learning of AML on handling audio-visual cross-modal retrieval task.

Fig. 7 further demonstrates some quantitative results of our AML on audio-visual retrieval task in both A $\rightarrow$ V and V $\rightarrow$ A challenges comparing with DIMNet [5], which is the only method that has released the codes on retrieval task. From which we can see, even without any attribute information, AML still can produce more correct hits in foregoing ranks in the gallery than DIMNet [5], which verifies the effectiveness of our method on audio-visual cross-modal retrieval task.

E. Ablation Study

To evaluate the contribution of the two crucial components in AML, e.g., adversarial learning (AL) and metric learning (ML). We conduct the ablation study on cross-modal matching task with four variants in both binary and multi-way cases in V $\rightarrow$ A challenge, as shown in Table VII. Note that, we conduct experiments on three variants. Specifically, we conduct Table VII (a) by removing both discriminator and metric learning loss, while Table VII (b) and (c) by removing discriminator and metric learning loss respectively. From which we can see, 1) Both AL and ML play crucial role in audio-visual task, comparing Table VII (b) and (c) by introducing AL and ML respectively to the baseline as shown in Table VII (a). 2) Without both adversarial and metric learning, Table VII (a) achieves impressive result (83.1%), which is even higher than the state-of-the-art method SVHF-Net [3] as shown in Table III. The main reason is the deeper network architecture in our method. Specifically, our network further includes three more fully connected layers which can thus extract more discriminative features for classification. 3) By collaborating AL and ML, our AML (as shown in Table VII (d)) significantly boost the performance, especially for the more challenging multi-way scenario, which verifies the contribution of joint adversarial learning and metric learning.

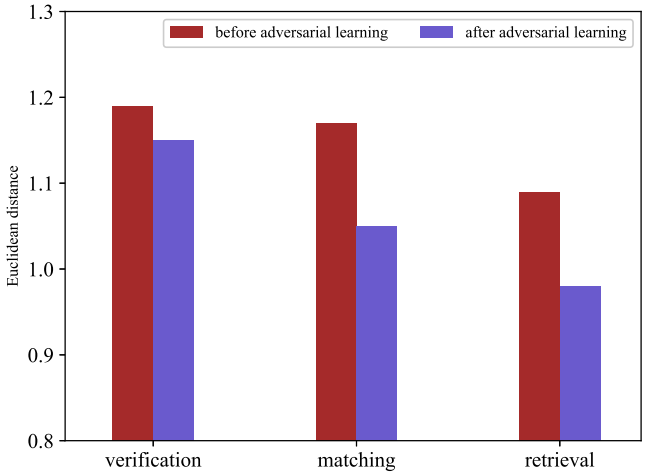


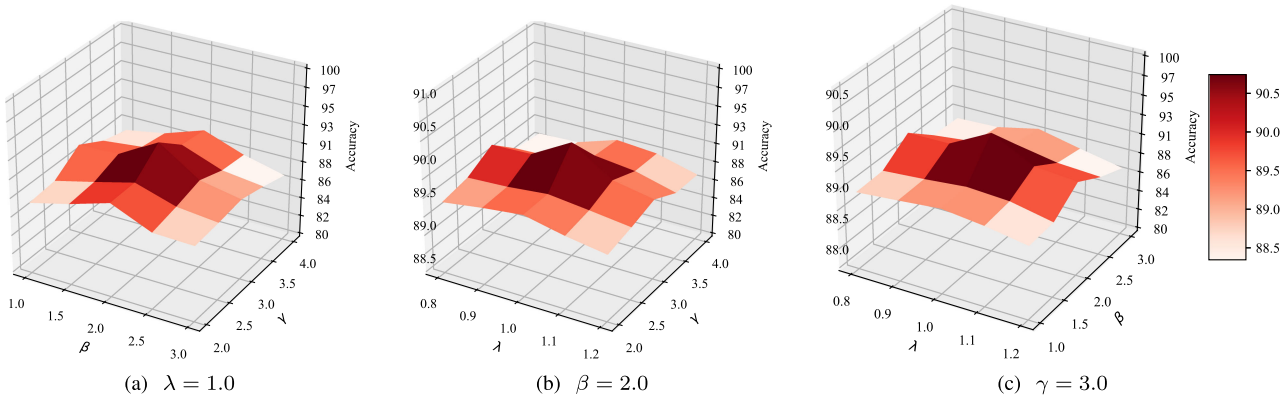
Fig. 8. Average Euclidean distance between audio and visual features before and after adversarial learning of all identities in the dataset in V $\rightarrow$ A challenge.

TABLE VIII
COMPARISON TO DIFFERENT ADVERSARIAL LEARNING METHODS ON AUDIO-VISUAL MATCHING TASK IN V $\rightarrow$ A CHALLENGE IN BOTH BINARY ($k = 2$) AND MULTI-WAY ($k = 10$) CASES

Method \ Accuracy	Binary	Multi-way ($k = 10$)
GANs [13]	91.76	36.23
LSGANs [66]	90.45	35.79
WGAN [62]	**93.33**	**39.35**

Evaluation on Adversarial Learning: In this section, we further investigate the effectiveness of the proposed AML with different adversarial learning methods on audio-visual matching task. Different from our adversarial learning, which is trained based on WGAN [62] as presented in Section III-C, we substitutively train the generator and discriminator by the original GANs [13] and LSGANs [66] for comparison. Table VIII compares the matching results on different adversarial learning methods. From which we can see, the proposed AML offers promising performance regardless the fashions of the generator and discriminator, which verifies the significance of the adversarial learning in audio-visual learning.

Evaluation on Metric Learning: To further evaluate the dependency of the proposed AML method on metric learning, we compare the metric learning method, Lifted Struct [19], to



(a) $\lambda = 1.0$ (b) $\beta = 2.0$ (c) $\gamma = 3.0$

Fig. 9. Parameter analysis of $\{\lambda, \beta, \gamma\}$ on audio-visual matching task in binary ($k = 2$) case in V $\rightarrow$ A challenge.

TABLE IX
COMPARISON ON DIFFERENT METRIC LEARNING METHODS ON AUDIO-VISUAL MATCHING TASK IN V $\rightarrow$ A CHALLENGE IN BOTH BINARY ($k = 2$) AND MULTI-WAY ($k = 10$) CASES

Method \ Accuracy	Binary	Multi-way ($k = 10$)
RankList [14]	88.14	34.18
Triplet [67]	91.02	36.46
Lifted Struct [19]	**93.33**	**39.35**

two state-of-the-art metric learning methods, RankList [14] and Triplet loss [67]. Specifically, we construct $k - 1$ triplet-tuples for every negative sample when using triplet loss. Table IX reports the comparison results with different metric learning methods. From which we can see, all the three metric learning methods perform comparably which verifies the effectiveness of introducing the deep metric learning in the audio-visual learning framework.

F. Analysis on Heterogeneous Issue

To analyze the effectiveness of adversarial learning via solving the heterogeneous issue, we further evaluate the distance between the cross-modal audio and visual features after the adversarial learning. Specifically, we calculate the average Euclidean distance between audio and visual features of the same identity in the whole dataset before and after adversarial learning respectively on all the three tasks, including verification, matching, and retrieval in V $\rightarrow$ A challenge. As visualized in Fig. 8, after adversarial learning, the cross-modality heterogeneous gap between audio and visual features significantly drops, which demonstrates the effectiveness of the proposed method on handling the heterogeneous issue for audio-visual learning.

G. Parameter Analysis

There are three important hyper-parameters in our model, the distance margin λ in Eq. (5) and two balance hyper-parameters β and γ in Eq. (9). We empirically set $\{\lambda, \beta, \gamma\}$ as $\{1.0, 2.0, 3.0\}$ for the best performance. In order to evaluate the impact of these hyper-parameters, we analyze the performance of our model on audio-visual matching task by fixing λ, β, and γ respectively, while varying the other two parameters. As shown in Fig. 9, generally speaking, the accuracy slightly varies with diverse combinations of the hyper-parameters, which indicates that our model is not sensitive to these hyper-parameters.

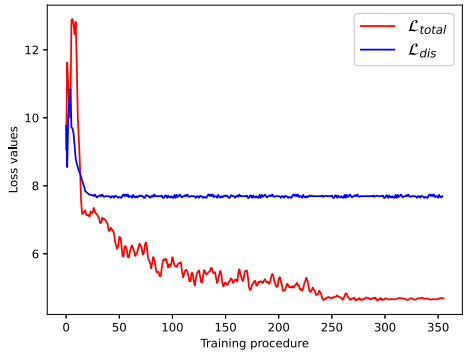
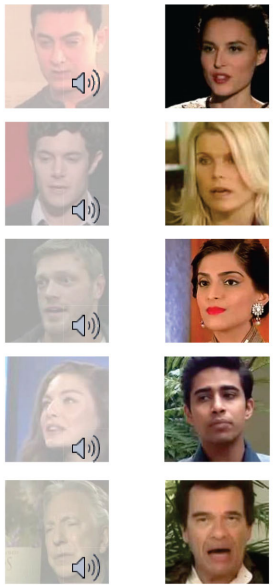


Fig. 10. Demonstration of $\mathcal{L}_{total}$ and $\mathcal{L}_{dis}$ values during training procedure.

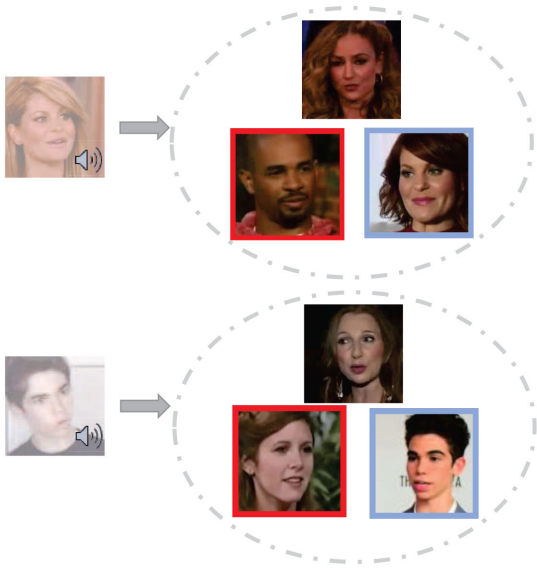
H. Analysis on Model Convergence and Complexity

In order to show the convergence of AML, we record the values of the final loss $\mathcal{L}_{total}$ and the discriminator loss $\mathcal{L}_{dis}$ across the every three batches in matching task when $k = 2$ as shown in Fig. 10. We can see that the final loss $\mathcal{L}_{total}$ decreases vibrationally until reaching the convergence. The discriminator loss $\mathcal{L}_{dis}$ first decreases since we first train the discriminator and then it stabilizes in the following training procedure.

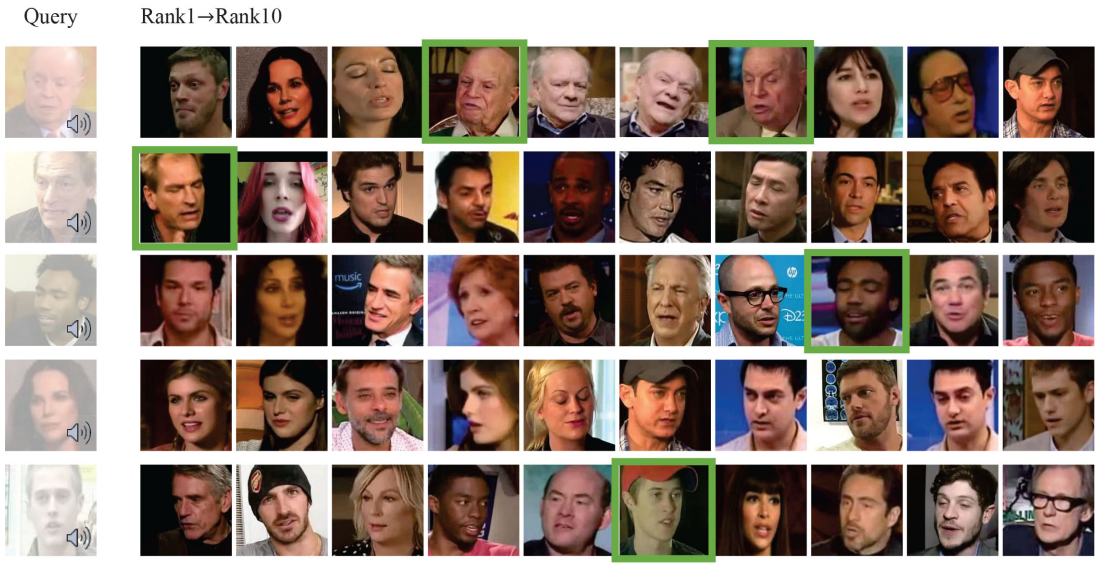
Furthermore, we analysis the complexity of our model, by calculating the average running time of AML in A $\rightarrow$ V challenge with $k = 2$. The average training and testing time are 0.016 s and 0.003 s respectively, while 0.005 s and 0.012 s in DIMNet [5] whose codes are released. Note that, we only compare the training time of ID classifier without the gender and nationality classifiers due to the limited codes releasing in DIMNet [5], which in practice takes longer than 0.005 s during the training. Due to the existence of adversarial and metric learning, the training of AML needs more time cost than testing. Due to the lighter structure of our model with only five convolution layers for face and voice feature extraction, while thirteen in DIMNet [5], AML performs faster than DIMNet [5] in the testing.



(a) verification



(b) matching ($k = 3$)



(c) retrieval

Fig. 11. Limitations of AML in audio-visual learning in A → V challenge on (a) verification, (b) matching, and (c) retrieval tasks. (a) The wrong verification results. (b) The wrong matching results in red boxes while the images in blue boxes indicate the ground truth. (c) The retrieval results from $R@1$ to $R@10$ where right hits are highlighted in green boxes.

I. Limitation

We have also encountered a key limitation of the proposed AML on audio-visual learning during the evaluation. As shown in Fig. 11, the AML may produce apparently incorrect results across the ages, genders or nationalities. For instance, As shown in the fifth row in Fig. 11 (a) on verification task, AML assigns a young man's image to an old man's audio clip. Similar cross gender/nationality/age mistakes also occur in Fig. 11 (b) and (c) on matching and retrieval tasks. The main reason is the audio-visual sub-networks in AML only focus on the speech and appearance features while ignoring the high-level semantic attribute information. The lack of high-level attribute information leads these intuitive mistakes. This limitation inspires us to explore the attribute-driven audio-visual learning framework in the future to better bridge the heterogeneous issue among cross-modal data.

V. Conclusion

In this paper, we have presented a novel adversarial-metric learning (AML) method for cross-modal audio-visual matching. Considering the heterogeneous issue between audio and visual data, we first propose to learn the modality-independent

representations for different modalities via adversarial learning. Meanwhile, we propose to learn a robust feature embedding for cross-modal similarity measure. AML tackles the challenge by generating discriminative feature representations and learning a robust feature metric for cross-modal audio-visual matching simultaneously. Comprehensive experiments on three audio-visual learning tasks, including audio-visual matching, verification and retrieval comparing with the state-of-the-art methods verify the effectiveness of the proposed AML. Our future work will concentrate on exploiting the common attributes between different modalities for audio-visual learning.

REFERENCES

[1] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "Putting the face to the voice': Matching identity across modality," *Curr. Biol.*, vol. 13, no. 19, pp. 1709–1714, 2003.

[2] H. M. J. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey, "Matching novel face and voice identity using static and dynamic facial images," *Attention Perception Psychophys.*, vol. 78, no. 3, pp. 868–879, 2016.

[3] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8427–8436.

[4] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable PINs: Cross-modal embeddings for person identity," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 71–88.

[5] 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. Int. Conf. Learn. Representations*, 2019.

[6] 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 *Proc. Int. Conf. Digit. Image Comput.: Techn. Appl.*, 2019, pp. 1–7.

[7] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7181–7189.

[8] H. Yi, W. Nannan, L. Jie, and G. Xinbo, "HSME: Hypersphere manifold embedding for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8385–8392.

[9] Y. Peng and J. Qi, "CM-GANs: Cross-modal generative adversarial networks for common representation learning," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 15, no. 1, pp. 22:1–22:24, 2019.

[10] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R2GAN: Cross-modal recipe retrieval with generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11 477–11486.

[11] J. Chi and Y. Peng, "Zero-shot cross-media embedding learning with dual adversarial distribution network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1173–1187, Apr. 2020.

[12] X. Xu *et al.*, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2400–2413, Jun. 2020.

[13] I. J. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[14] X. Wang *et al.*, "Ranked list loss for deep metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5207–5216.

[15] J. Lu, J. Hu, and Y. P. Tan, "Discriminative deep metric learning for face and kinship verification," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4269–4282, Sep. 2017.

[16] Z. Liu, W. Dong, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2429–2438.

[17] C. Yin, Z. Feng, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1153–1162.

[18] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1857–1865.

[19] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4004–4012.

[20] V. E. Liong, J. Lu, Y. P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1234–1244, Jun. 2017.

[21] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia*, pp. 154–162, 2017.

[22] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, 2019.

[23] B. Jiang, Z. Zhou, X. Wang, J. Tang, and B. Luo, "cmSalGAN: RGB-D salient object detection with cross-view generative adversarial networks," *IEEE Trans. Multimedia*, to be published.

[24] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 677–683.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," *Conf. Int. Speech Comm. Assoc.*, pp. 2616–2620, 2017.

[26] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.

[27] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, "Deep audio-visual learning: A survey," 2020, *arXiv:2001.04758*.

[28] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, 2018.

[29] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," 2018, *arXiv:1804.04121*.

[30] J. R. Hershey and J. R. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 813–819.

[31] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–53.

[32] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 631–648.

[33] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9299–9306.

[34] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 538–553.

[35] T. Tang, J. Jia, and H. Mao, "Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1598–1606.

[36] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman, "Audio to body dynamics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7574–7583.

[37] J. W. FisherIII, T. Darrell, W. T. Freeman, and P. A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 772–778.

[38] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 3051–3055.

[39] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon, "Learning to localize sound source in visual scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4358–4366.

[40] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 247–263.

[41] A. Davis *et al.*, "The visual microphone: Passive recovery of sound from video," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–10, 2014.

[42] C.-H. Wan, S.-P. Chuang, and H.-Y. Lee, "Towards audio to scene image synthesis using generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 496–500.

[43] L. Chen, S Srivastava. , Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 349–357.

[44] W. Hao, Z. Zhang, and H. Guan, "CMCGAN: A. uniform framework for cross-modal visual-audio mutual generation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 6886–6893.

[45] T. L. Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3355–3359.

[46] T. L. Cornu and B. Milner, "Generating intelligible audio speech from visual speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 9, pp. 1751–1761, Sep. 2017.

[47] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3550–3558.
[48] S. A. Jalalifar, H. Hasani, and H. Aghajan, "Speech-driven facial reenactment using conditional generative adversarial networks," 2018, *arXiv:1803.07461*.
[49] Y. Qiu and H. Kataoka, "Image generation associated with music data." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2510–2513.
[50] A. Duarte *et al.*, "Wav2Pix: Speech-conditioned face generation using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 8633–8637.
[51] 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," 2017, *arXiv:1706.00079*.
[52] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3304–3311.
[53] C. Kim *et al.* "On learning associations of faces and voices," 2018, arXiv:1805.05553.
[54] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1618–1625.
[55] Z.-D. Chen, W.-J. Yu, C.-X. Li, L. Nie, and X.-S. Xu, "Dual deep neural networks cross-modal hashing," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 274–281.
[56] X. Zhao, G. Ding, Y. Guo, J. Han, and Y. Gao, "TUCH: Turning cross-view hashing into single-view hashing via generative adversarial nets," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3511–3517.
[57] C. Li, C. Deng, L. Wang, D. Xie, and X. Liu, "Coupled cycleGAN: Unsupervised hashing network for cross-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 176–183.
[58] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3027–3035.
[59] J. Zhang and Y. Peng, "Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 174–187, Jan. 2020.
[60] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1276–1288, May 2019.
[61] G. Song, D. Wang, and X. Tan, "Deep memory network for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1261–1275, May 2019.
[62] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
[63] V. D. M. Laurens and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 2605, pp. 2579–2605, 2008.
[64] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2010, pp. 249–256.
[65] D. P. Kingma and J. Ba, "Adam: A. method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[66] X. Mao *et al.*, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2794–2802.
[67] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A. unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.