Looking and Hearing into Details: Dual-enhanced Siamese Adversarial Network for Audio-Visual Matching

Jiaxiang Wang, Chenglong Li, Aihua Zheng*, Jin Tang, Bin Luo

***Abstract*—Audio-visual cross-modal matching aims to explore the intrinsic correspondence between face images and audio clips. Existing methods usually focus on the salient features of identities between visual images and voice clips, while neglecting their subtle differences, which are crucial to distinguishing cross-modal samples. To deal with this problem, we propose a novel Dual-enhanced Siamese Adversarial Network (DSANet), which pursues the adversarial dual enhancement to highlight both salient and subtle features for robust audio-visual cross-modal matching. First, we designed a dual enhancement mechanism to enhance potential subtle features by randomly selecting a region feature for salient feature suppression, while enhancing salient features in the corresponding region to ensure the global discriminative ability. Second, to establish the correlation of subtle features in the process of eliminating cross-modal heterogeneity, we design a siamese adversarial structure to perform modal heterogeneity elimination for both enhanced salient and subtle features in a parallel manner. Moreover, we propose an adaptive masked cross-entropy loss to force the network to focus on the feature differences among hard classes. Experiments on public benchmark datasets validate the effectiveness of the proposed algorithm.**

***Index Terms*—Audio-visual cross-modal matching, dual enhancement mechanism, Siamese adversarial network, adaptive masked cross-entropy.**

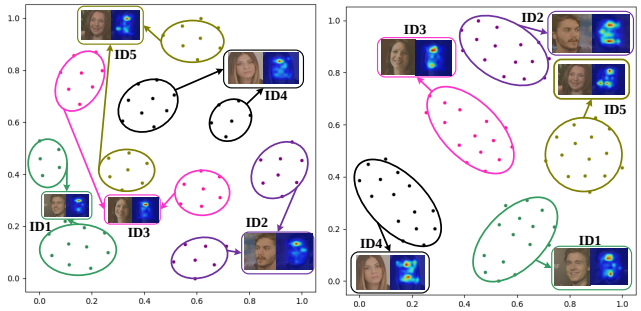
I. Introduction

The human brain can effectively link the perception between voice audio and facial information, as concluded by the renowned psychologists Bruce and Young [1]. After associative memory with a person's identity via face images and emitted audio, it is possible to access a specific person's face by means of audio information, and vice versa [2], [3]. Recently, there emerge research works on this frontier topic, named audio-visual learning, which aims to explore the connection between hearing and vision in artificial intelligence. It has the potential application to enhance conventional machine learning tasks, such as audio-visual speech separation [4], audio-video localization [5], [6], and audio-visual recognition [7]–[9].

This research is supported in part by the National Natural Science Foundation of China (61976002, and 62102344), the University Synergy Innovation Program of Anhui Province (GXXT-2021-038), and the Natural Science Foundation of Anhui Higher Education Institutions of China (No.KJ2021A0901).

A. Zheng and C. Li are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei, 230601, China (e-mail: ahzheng214@foxmail.com; lcl1314@foxmail.com).

J. Wang, J. Tang, and B. Luo are with Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China (e-mail: Netizen-wjx@foxmail.com; tangjin@ahu.edu.cn; ahu_lb@163.com)



(a) Feature distribution of one-layer adversarial network

(b) Feature distribution of dual enhancement Siamese adversarial network

Fig. 1. Comparison of the feature distribution between the dual-enhanced Siamese adversarial network and the one-layer adversarial network. Herein, five identities are selected to compare their feature distributions, where each identity has 15 samples indicated by the same color.

As a representative audio-visual learning task, audio-visual matching devotes to exploring the correlation between face images and speech clips with the same identity. Nagrani *et al.* [10] first propose an audio-visual cross-modal matching task and design a binary classification network to accomplish the classification task by identity supervision. However, the huge cross-modal heterogeneity restricts the performance considerably. Wang *et al.* [11] and Nawaz *et al.* [12] map two modal features through a common space and mitigate cross-modal differences by metric constraints. However, the metric only constrains the distance of features between classes, while cannot deceive the network's perception of modality. It is well known that generative adversarial networks (GAN) [13] can achieve Nash equilibrium via a minimax two-player game. Therefore, Zheng *et al.* [14] and Cheng *et al.* [15] employ GAN to eliminate audio-visual cross-modal heterogeneity. Despite the recent progress in audio-visual cross-modal matching, there are still three problems not well addressed.

The first problem is that the salient features learned by convolutional neural networks (CNNs) may neglect to learn subtle features of important information [16]–[18]. However, the hard samples are similar under different identities, and the same identity is multi-variant in appearance under different scenarios. For this reason, we should pay attention to subtle features to help narrow intra-class differences and enlarge inter-class distances. Therefore, Sun *et al.* [19] and Chen *et al.* [16] propose the salient feature suppression module,

which allows the network to learn more subtle features. Among them, suppression for randomly selected features is an important tool, which can make the network distracted to focusing on more feature regions. However, the salient features are significantly weakened which leads to a decrease in the inter-class distance. In this work, we propose a dual enhancement mechanism (DEM) to simultaneously learn enhancement salient features while retaining the enhancement subtle features learning. The enhancement of salient features is to maximize the discrimination keeping of salient features, while the enhancement of subtle features is to improve the learning of inter-class differences and intra-class compactness. As shown in Fig. 1, compared with the feature distribution of a one-layer adversarial network, the features extracted with dual enhancement siamese adversarial can both improve the intra-class compactness in different scenarios of the same identity and increase the inter-class distance for similar samples of different identities.

The second issue is the existence of data heterogeneity between audio and visual modalities. Despite the achievement of GANs [14], [15] to eliminate cross-modal data differences, a single GAN tends to focus on salient feature learning while ignoring the elimination of modal heterogeneity of subtle features, which leads to difficulties in distinguishing inter-class variations when identifying hard samples. To solve this problem, we design GANs as a dual-stream structure, called the siamese adversarial structure (SAS), which is composed of two GANs. One GAN is used to deal with enhanced salient features to eliminate modal heterogeneity to achieve cross-modal salient feature associations, while the other GAN learns discriminative subtle feature associations. The two GANs that share parameters learn different features to update the model parameters separately. Since SAS is a two-stream structure with shared parameters, the same feature is considered unlearned in SAS only if it is dropped twice, otherwise it can still learn cross-modal associations. As shown in Table I, the SAS has a higher feature loss rate and feature learning rate. The high feature dropout rate can suppress network overfitting [20] and the high feature learning rate learns diverse features generated from the dual enhancement mechanism to increase the generalization of the network [21].

TABLE I
FEATURE DROPOUT RATE AND LEARNING RATE COMPARISON OF GAN AND SAS, RESPECTIVELY, WHERE $\boldsymbol{p}$ IS THE FEATURE DROPOUT RATE OF EACH GENERATOR. THE GAN MODULE COMPUTES THE SUM OF THE FEATURE DROPOUT RATE AND THE FEATURE LEARNING RATE AS 1, WHILE THE SAS SUM IS NOT.

Methods	Feature dropout rate	Feature learning rate
GAN	$\boldsymbol{p}$	$1-\boldsymbol{p}$
SAS	$\boldsymbol{p}(2-\boldsymbol{p})$	$1-\boldsymbol{p}^2$

The third problem is that the existing methods do not purposely focus on hard samples, which may lead to slower convergence and low recall [19]. However, popular loss functions for classification tasks, such as cross-entropy, assign an equal probability of being misclassified in any of the hard negative classes, which does not prevent misclassification among them. Therefore, Sun *et al.* [19] proposed gradient-boosting cross-entropy (GBCE) loss to resolve the ambiguity between closely related hard negative classes. However, not every small batch has hard negative class samples, and GBCE performs gradient optimization by fixing the number of candidate negative classes, which increases the optimization computation. In this paper, we propose the adaptive masked cross-entropy loss ($\mathcal{L}_{AMCE}$), which allows the selection of hard negative classes by threshold. As shown in Fig. 2 (b), adaptive masked cross-entropy performs gradient optimization by calculating the loss of hard negative classes while ignoring the loss of easily distinguishable classes to further improve inter-class discrimination.

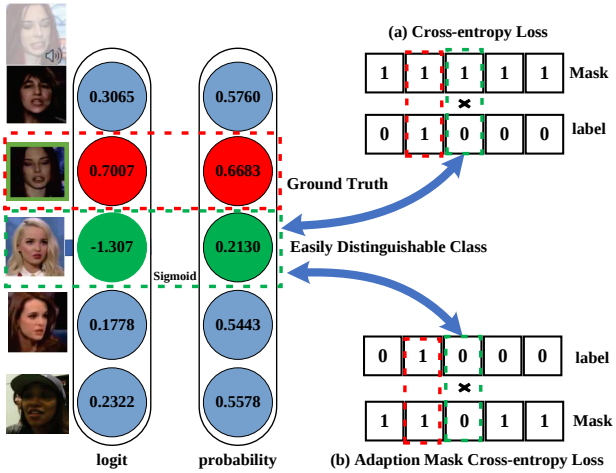


Fig. 2. Illustration of cross-entropy loss ($\mathcal{L}_{CE}$) and our proposed AMCE loss ($\mathcal{L}_{AMCE}$). The audio and image matching probabilities are given the same mask value on cross-entropy for all negative classes, while AMCE assigns the mask of the class with a lower matching probability to 0.

Overall, the main contributions of this work can be summed up as follows:

- We propose the dual enhancement mechanism that forces the network to focus on more feature regions to find subtle differences between identity categories.
- We propose the siamese adversarial structure which can learn enhanced salient and subtle features with pattern-independent audio-visual feature associations in a parallel manner.
- We propose adaptive masked cross-entropy which enables adaptive selection of hard negative classes to learn inter-class distinguishability.
- A comparison with the state-of-the-art algorithm to achieve optimal performance on the audio-visual cross-modal matching task illustrates the effectiveness of the proposed dual-enhanced siamese adversarial network. To verify the general applicability of the model, we extended the model to cross-modal audio-visual retrieval tasks as well.

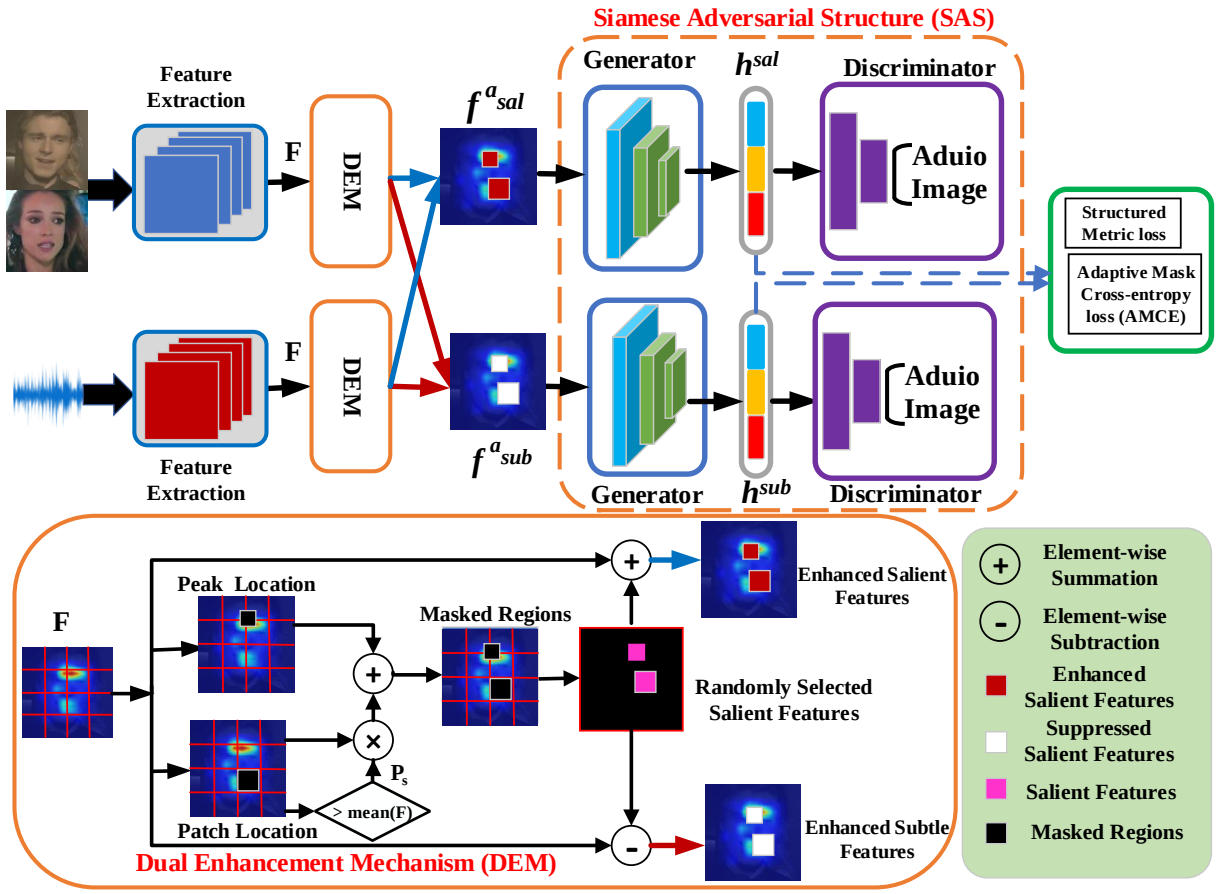


Fig. 3. Overview of our overall architecture. Our approach contains three new components: dual enhancement mechanism, siamese adversarial structure, and adaptive masked cross-entropy loss. The DEM is used to enhance salient features to maintain discriminative ability while enhancing the learned subtle features to help the model learn intra-class similarity and inter-class dissimilarity. The SAS removes modal heterogeneity by separating the enhanced salient features from the enhanced subtle features. The AMCE forces the network to focus on hard negative class samples to learn inter-class variation.

II. Related Work

A. Audio-visual Matching

Audio-visual association matching has attracted a lot of research attention in recent years. To the best of our knowledge, Nagrani *et al.* [10] first proposed sound (face) to face (voice) matching as a binary classification task. And by designing two-stream deep neural networks were able to achieve classification performance comparable to the human baseline. Nagrani *et al.* [22] extended the audio-visual matching task to verification and retrieval, which proposed a joint curriculum learning and contrast loss optimization embedding network. To learn a shared representation instead of directly associating audio clips and face images, Wen *et al.* [23] adopt more attribute information, such as nationality and gender, combined to co-supervise network training. Wang *et al.* [11] proposed an end-to-end joint embedding network for learning face-voice discriminative features with bi-directional ranking constraints, identity constraints, and centrality constraints in a small batch of data. To discriminate between classes by using salient and subtle features in both audio and visual data, we propose an adversarial dual enhancement-based audio-visual matching network, which is called DSANet in this paper, as shown in Fig. 3.

Despite the great progress of the above methods in audio-visual matching, there are still some unresolved issues. The first problem is that the contrast loss function can only learn local information in small batches of data, which may lead to slow convergence of the network. The second problem is that some hard identities are difficult to be learned effectively by the network which should be ignored for this part of the sample. Wen *et al.* [24] proposed a two-level modal alignment approach that is performed for global information. Hard but valuable identities are better learned by a dynamic re-weighting scheme, while identities that are difficult for the network to learn should be filtered out. Existing work can effectively handle the task of matching, verification, and retrieval between audio-visual identities, but fails to address the heterogeneity between the two modalities, which is an unavoidable problem. Zheng *et al.* [14] proposed an adversarial measurement learning model for audio-visual matching. The modality-independent feature representation is learned by generative adversarial networks and combined with similarity

measures to render the learned networks more robust.

B. Cross-modal Heterogeneity Elimination

To address the problem of cross-modal heterogeneity, Zhen *et al.* [25] and Li *et al.* [26], and Xu *et al.* [27] proposed to mitigate modal heterogeneity by feature embedding layers. In contrast to the feature embedding approach, previous studies [28]–[30] proposed the GAN-based model to eliminate modal heterogeneity, which is an effective way to bridge cross-modal feature associations. The method is widely applied to cross-modal retrieval, cross-spectrum face recognition, and audio-video matching.

GAN is an effective method to mitigate modal heterogeneity. However, there are many differences in the final results of selecting different adversarial models, which are determined by the convergence of GAN. For the design of GAN, here GANs [31], LSGANs [32] and WGANs [33] are representative network models. These GANs network models are applied to audio-visual matching to verify the effectiveness of the models in eliminating modal heterogeneity. Therefore, Zheng *et al.* [14] proposed the Wasserstein generative adversarial network (WGAN) for learning modality-aligned embedding to eliminate modal heterogeneity. Cheng *et al.* [15] proposed a new adversarial deep semantic matching network to learn the interaction of face and voice to build associations. And triplet loss and multimodal center loss are jointly used to explicitly regularize the correspondence between them. To learn the correlation of effective features, we adopt the Self-Attention Generative Adversarial Network (SAGAN) proposed by Zhang *et al.* [34] as the backbone of our adversarial network. Spectral parameter regularization and large gradient truncation operations are imposed on the adversarial network to achieve more stable convergence. More detailed and effective methods can be read in the audio-visual review proposed by Zhu *et al.* [30].

III. Method

We propose the Dual-enhanced Siamese Adversarial Network (DSANet) to learn the intrinsic association between audio and visual cross-modal data in the audio-visual matching task. In particular, to maximize the discrimination of salient and subtle features simultaneously, we propose a dual enhancement mechanism (DEM) to maintain the discriminative ability for salient features while learning subtle features to help expand the inter-class feature distance and increase the intra-class compactness. Then, we propose a siamese adversarial structure (SAS) to handle the heterogeneity of salient and subtle features between audio and visual data. Finally, we design the adaptive mask cross-entropy (AMCE) loss, which enables the network to focus on hard negative class learning distinctions.

A. Dual Enhancement Mechanism

First, audio and facial image features are extracted by respective convolutional networks to obtain activation maps corresponding to the features, such that the extracted audio and face images activation mappings are $\boldsymbol{f}^a \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{f}^v \in \mathbb{R}^{C\times H\times W}$, respectively. Here, C is the number of channels of activation maps, and H and W are the height and width of the activation maps, respectively. Then, the DEM can find the mask region $\mathbf{M} \in \mathbb{R}^{C\times HW}$ based on the activation maps, so that the network can enhance and suppress the activation maps in the selected region. To make the activation maps correspond to the mask dimension, we represent the activation maps of the previous audio-visual features uniformly with the $\mathbf{F} \in \mathbb{R}^{C\times HW}$ of the transform matrix dimension. As can be seen from Fig. 3, the mask localization is determined by both peak and patch.

Peak location: The peak mask is selected as the location of the response value of the maximum activation map because it is the most discriminative activation map for classification. Let $\mathbf{P}_{max}$ be the location of the peak maps from the activation maps $\mathbf{F}$ and denoted as:

$$\mathbf{P}_{max}(i,j) = \begin{cases} 1, & if \quad \mathbf{F}(i,j) = max(\mathbf{F}) \\ 0, & otherwise. \end{cases}, \tag{1}$$

where $max(\mathbf{F})$ denotes the maximum of activation maps matrix $\mathbf{F}$. The i and j correspond to the rows and columns of the index matrix.

Patch location: To clarify salient and subtle features, we use average features as thresholds for definition. For subsequent processing, we determine the position $\mathbf{P}_s$ representation corresponding to the salient features on the activation map as follows:

$$\mathbf{P}_s(i,j) = \begin{cases} 1, & if \quad \mathbf{F}(i,j) > mean(\mathbf{F}) \\ 0, & otherwise. \end{cases}. \tag{2}$$

Next, we describe how to select the patch localization on the activation maps $\mathbf{F}$. We divide each $\mathbf{F}$ into a grid of patches, where each patch $\mathbf{M}_g(l,m)$ is set to a fixed size $r \times c$ and indexed by row l and column m. The indexes of all such patches on the grid are represented as follows:

$$\mathbf{M}_g(l,m) \in \mathbb{R}^{\frac{C}{r}\times\frac{HW}{c}}, l \in [1, \frac{C}{r}], m \in [1, \frac{HW}{c}], \tag{3}$$

where r and c are fixed-length values for dividing activation maps. The positions corresponding to the patch mask $\mathbf{M}_g$ are set to 1 by randomly selecting l and m. Otherwise, the elements of $\mathbf{M}_g$ are set to 0.

The final position of the activation maps corresponding to the selected patch is determined by $\mathbf{P}_s$ and $\mathbf{M}_g$ together. $\mathbf{P}_{sg}$ is the corresponding position of the salient feature in the selected patch, which is denoted as:

$$\mathbf{P}_{sg} = \mathbf{P}_s \odot \mathbf{M}_g, \tag{4}$$

where $\odot$ refers to the element-wise product.

Considering that the peak position may overlap with the selected patch position, we set the position corresponding to the peak in the patch to 0. The formula is as follows:

$$\mathbf{P}_{sg}(x,y) = 0, if \quad \mathbf{M}_g(x,y) = max(\mathbf{F}). \tag{5}$$

The final mask position corresponding to the selected activation maps is derived as:

$$\mathbf{M} = \mathbf{P}_{sg} + \mathbf{P}_{max}. \tag{6}$$

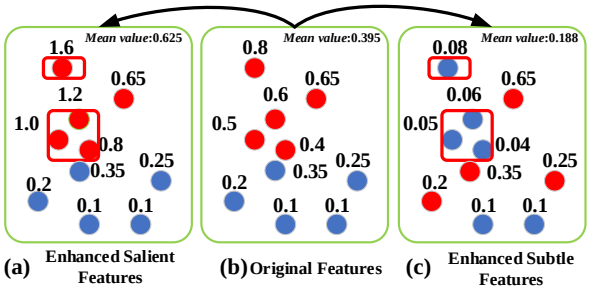


Fig. 4. The feature activation mapping values, where the red and blue dots represent the values higher and lower than the corresponding mean value respectively. The red boxes in (a) and (c) locate the selected salient and subtle features respectively.

Dual enhancement: The salient features corresponding to the patch mask can be expressed as:

$$\mathbf{F}^s = \mathbf{M} \odot \mathbf{F}. \tag{7}$$

The salient features can identify easily distinguishable classes while learning more subtle features can help the network further distinguish hard classes. Therefore, we implement the enhancement of the salient features corresponding to the patch mask by the activation factor, which is calculated as follows:

$$\boldsymbol{f}^{sal} = \mathbf{F} + \alpha * \mathbf{F}^s, \tag{8}$$

where α denotes the enhancing factor. In the same way, enhancing the subtle features is done by suppressing the features $\mathbf{F}^s$ corresponding to the activation maps mask region, which is calculated as follows:

$$\boldsymbol{f}^{sub} = \mathbf{F} - \beta * \mathbf{F}^s, \tag{9}$$

where β denotes the suppressing factor. In general, setting both α and β to higher numbers achieves better performance. Based on our experimental setup, we set α to 1 and β to 0.9.

As shown in Fig. 4 (b), the red features are above-average features, which the network focuses on learning to match the corresponding candidate identities. As shown in Fig. 4 (a) and (c), both enhancement techniques force the network to learn different feature activation mappings, which leads to a more robust feature representation. The results are discussed in the experimental section.

B. Siamese Adversarial Structure

Given a matching data tuple is composed of an audio clip $\boldsymbol{f}_{i0}^{a}$ and k visual face images $\boldsymbol{f}_i^v = \{\boldsymbol{f}_{i1}^v, \cdots, \boldsymbol{f}_{ik}^v\}$. The features of the audio clip and visual face image after the dual enhancement mechanism the features become $\boldsymbol{f}_{i0}^{a_{sal}}$,$\boldsymbol{f}_{i0}^{a_{sub}}$, $\boldsymbol{f}_{ik}^{v_{sal}}$ and $\boldsymbol{f}_{ik}^{v_{sub}}$ respectively. The dimension of each feature $\boldsymbol{f}_{ik}$ is $C \times H \times W$. The purpose of audio-visual cross-modal matching is to find the face images in the corresponding candidate identities by audio information and vice versa, where i is denoted as the i-th matching data tuple. The corresponding identity is represented by the label $L_i \in [1, k]$. Note that where $k = 2$ denotes binary matching and $k > 2$ is multi-way matching.

To alleviate the modal heterogeneity between audio and face images, we propose to learn modality-independent representations by the SAS. The SAS consists of two generators G and two discriminators D that share parameters during the training process. In the following, we use the audio and visual image features with enhanced salient features as examples to illustrate the generative adversarial process. Audio features $\boldsymbol{f}_{i0}^{a_{sal}}$ and face images $\{\boldsymbol{f}_{i1}^{v_{sal}}, \cdots, \boldsymbol{f}_{ik}^{v_{sal}}\}$ are used as inputs to G, which generates modality-independent features $\{\boldsymbol{h}_{i0}^{sal}, \cdots, \boldsymbol{h}_{ik}^{sal}\} \in \mathcal{H}$. The features $\{\boldsymbol{h}_{i0}^{sal}, \cdots, \boldsymbol{h}_{ik}^{sal}\} \in \mathcal{H}$ as D inputs are adopted by a modal classifier to distinguish the modality of audio and face features. Each modal feature finds the modality-independent feature space $\mathcal{H}$ by a min-max game.

Generator. We use SAGAN [15] as the backbone of the adversarial network. The generator G with parameters θ_G is constructed using self-attention and fully connected (FC) layers. The feature function after mapping through G is represented as:

$$h_{i0} = \psi(\phi(\boldsymbol{f}_{i0}^{a_{sal}}; \theta_G)), \tag{10}$$

$$h_{ij} = \psi(\phi(\boldsymbol{f}_{ij}^{v_{sal}}; \theta_G)), j \in [1, k], \tag{11}$$

where ψ and ϕ are the two FC layers and self-attention, respectively, which are used to map the audio $\boldsymbol{f}_{i0}^{a_{sal}}$ and face image features $\{\boldsymbol{f}_{i1}^{v_{sal}}, \cdots, \boldsymbol{f}_{ik}^{v_{sal}}\}$ to a modality-independent feature space. Similarly, the process is used to deal with the enhanced subtle features in the generative network.

Discriminator. The discriminator D is a binary FC network with training parameter θ_D, which is used to discriminate $\boldsymbol{h}_{ij}^{sal}$ features from the original audio-visual modality. The discriminator is trained by minimizing:

$$\mathcal{L}_{disc}^{sal} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{k} Y_{ij} \log D(\boldsymbol{h}_{ij}^{sal}; \theta_D), \tag{12}$$

where Y_{ij} represents the modality label of the j-th sample in the i-th data tuple, $D(\boldsymbol{h}_{ij}^{sal}; \theta_D)$ is the modality probability of the output of D. N denotes the number of training data tuples. Similarly, the process is used to deal with the enhanced subtle features in the discriminator network. The loss computed for the subtle features enhanced under the discriminator is $\mathcal{L}_{disc}^{sub}$, which has a similar form of computation to $\mathcal{L}_{disc}^{sal}$. The total discriminative loss can be represented as:

$$\mathcal{L}_{disc}^{total} = \mathcal{L}_{disc}^{sal} + \mathcal{L}_{disc}^{sub}. \tag{13}$$

C. Adaptive Mask Cross-entropy Loss

To learn the distinguishability of hard samples, we propose the adaptive mask cross-entropy (AMCE) loss, which can focus on hard negative classes to learn inter-class variance.

Adaptive Mask Cross-entropy. As the features extracted by the network are divided into two enhanced feature outputs via a dual enhancement mechanism, we propose the adaptive mask cross-entropy by calculating the loss of the enhanced salient

features. For audio $\boldsymbol{f}_{i0}^{a_{sal}}$ matching face $\boldsymbol{f}_{ik}^{v_{sal}}$, the AMCE can be written as follows:

$$\mathcal{L}_{AMCE}^{sal} = -\sum_{j=1}^{k} \omega_j \log(p_j^{sal}), \tag{14}$$

where binary mask term ω_j is a non-zero i.e. one. If ω_j is all ones, all negative samples are taken into account in the loss.

Given a probe is an instance from one modality and is utilized to find a match among k candidates in another modality. m_j^{sal} denotes the j-th value of the output k candidates of the classification network. By applying the sigmoid activation function, the output probability p_j^{sal} is represented by the formula shown below:

$$p_j^{sal} = \frac{1}{1 + e^{-m_j^{sal}}}, \tag{15}$$

where the maximum probability value for each match is p_j^{max}. By setting the hyperparameters η, the corresponding masks ω_j^{sal} are defined as follows:

$$\omega_j^{sal} = \begin{cases} 1, & if \;\; L_i = j \\ 1, & if \;\; L_i \neq j \; and \;\; p_j^{sal} - P_j^{max} + \eta > 0 \\ 0, & if \;\; L_i \neq j \; and \;\; p_j^{sal} - P_j^{max} + \eta < 0. \end{cases}, \tag{16}$$

where $L_i = j$ denotes the same identity for different modalities. Based on our experiments, we set η to 0.1. The gradient for m_i^{sal}, is derived as:

$$\frac{\partial \mathcal{L}_{AMCE}^{sal}}{m_i^{sal}} = \begin{cases} p_j^{sal} - 1, & if \;\; L_i = j \\ \omega_j^{sal} p_j^{sal}, & if \;\; L_i \neq j \end{cases}. \tag{17}$$

Compared to the gradient of the binary cross-entropy loss,

$$\frac{\partial \mathcal{L}_{BCE}^{sal}}{m_i^{sal}} = \begin{cases} p_j^{sal} - 1, & if \;\; L_i = j \\ 0, & if \;\; L_i \neq j \end{cases}, \tag{18}$$

clearly, we have,

$$\frac{\partial \mathcal{L}_{AMCE}^{sal}}{m_i^{sal}} > \frac{\partial \mathcal{L}_{BCE}^{sal}}{m_i^{sal}}. \tag{19}$$

The larger gradient forces the network to learn to distinguish between hard sample classes and ground truth classes.

Similarly, in the adaptive mask cross-entropy, the loss of enhanced subtle features is $\mathcal{L}_{AMCE}^{sub}$ which has the same computational form as $\mathcal{L}_{AMCE}^{sal}$. The total classification loss is summed as:

$$\mathcal{L}_{AMCE}^{total} = \mathcal{L}_{AMCE}^{sal} + \mathcal{L}_{AMCE}^{sub}. \tag{20}$$

D. Joint Learning Algorithm

Inspired by Peng *et al.* [35], we propose a structured metric to constrain the intra-class compactness and inter-class variability of audio-visual data, which is formulated as:

$$\mathcal{L}_{metric}^{sal} = \frac{1}{2N} \sum_{i=1}^{N} max(0, E_i^{sal}), \tag{21}$$

$$E_i^{sal} = log(\max_{j \in [2,k]} e^{\theta - d_{i0,ij}^{sal}} + \max_{q \in [2,k]} e^{\theta - d_{i1,iq}^{sal}}) + d_{i0,i1}^{sal}, \tag{22}$$

where $d_{i0,i1}^{sal}$ measures the Euclidean distance between the cross-modal anchor data h_{i0}^{sal} and the positive sample h_{i1}^{sal}, and $d_{i1,iq}^{sal}$ measures the Euclidean distance between the same-modal anchor data h_{i1}^{sal} and the negative sample h_{iq}^{sal}. Considering different numbers of $d_{i0,i1}^{sal}$ and $d_{i1,iq}^{sal}$ may lead to optimization imbalance problems. Only use a negative instance that reaches the maximum value is activated so that the imbalance problem can be naturally avoided. θ is a hyper-parameter that controls the margin of the distance between the negative set and positive pair. E_i^{sal} is used to measure the distance between the anchor and the positive sample and the negative sample.

Equivalently, the loss of the enhanced subtle features in the structure metric is calculated as $\mathcal{L}_{metric}^{sub}$ which is calculated in the same form as the distance metric of $\mathcal{L}_{metric}^{sal}$. The total structural metric loss is denoted as:

$$\mathcal{L}_{metric}^{m} = \mathcal{L}_{metric}^{sal} + \mathcal{L}_{metric}^{sub}. \tag{23}$$

Our model uses the learned modality-independent feature representations for classification loss and integrates structured metric loss into adversarial learning. The total loss is calculated as follows:

$$\mathcal{L}_{total}^{m} = \mathcal{L}_{disc}^{total} + \lambda \mathcal{L}_{metric}^{m} + \mu \mathcal{L}_{AMCE}^{total}, \tag{24}$$

where λ and μ are hyper-parameters.

E. Extension to Audio-visual Cross-model Retrieval

To verify the generality of the proposed DSANet, we extend it to the more challenging task, audio-visual retrieval, which aims to retrieve one or more matching samples from the entire gallery for each cross-modal probe. It is a more challenging task since the retrieved candidates are variable in appearance and background, which leads to learning difficulty in distinguishing a relative number of hard samples. In the retrieval training, the original matching network is kept unchanged, and the data input is changed to the audio-visual pair data.

In contrast to the total loss in the matching task as shown in Eq. (24), the retrieval task learns modality-independent feature representations by generating adversarial networks G, D and combining global classification C_R and metric learning $\mathcal{L}_{metric}^{r}$. The total loss is represented as follows:

$$\mathcal{L}_{total}^{r} = \mathcal{L}_{disc}^{total} + \lambda_1 \mathcal{L}_{metric}^{r} + \mu_1 \mathcal{L}_{cls}, \tag{25}$$

$$\begin{aligned} \mathcal{L}_{cls} = &-\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} l_i \log C_R(\{\boldsymbol{h}_{i0}^{sub} - \boldsymbol{h}_{ij}^{sub}\}; \theta_{C_R}) \\ &-\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} l_i \log C_R(\{\boldsymbol{h}_{i0}^{sal} - \boldsymbol{h}_{ij}^{sal}\}; \theta_{C_R}), \end{aligned} \tag{26}$$

$$\begin{aligned} \mathcal{L}_{metric}^{r} = \mathcal{L}_{metric}^{m} &+ \frac{1}{N} \sum_{i=1}^{N} ||\boldsymbol{h}_i^{sub} - c_{L_i}^{sub}|| \\ &+ \frac{1}{N} \sum_{i=1}^{N} ||\boldsymbol{h}_i^{sal} - c_{L_i}^{sal}||, \end{aligned} \tag{27}$$

where θ_{C_R} denotes network parameters of classification network C_R. $C_R(\{\boldsymbol{h}_{i0}^{sal} - \boldsymbol{h}_{ij}^{sal}\}; \theta_{C_R})$ and $C_R(\{\boldsymbol{h}_{i0}^{sub} -$

$\boldsymbol{h}_{ij}^{sub}\}; \theta_{C_R})$ are the distance between the enhanced audio and face modalities of salient features and subtle features respectively. L_i is the label of the i-th image in a mini-batch. c_i denotes the L_i-th class center of deep features [36]. λ_1 and μ_1 are hyper-parameters.

IV. Experiments

A. Implementation Details

Network architecture. We conduct our experiments on NVIDIA GeForce RTX 3090 graphic card. The PyTorch architecture with python version 3.7 is used to execute the commands. We employ ResNet18 [37] as the feature extractor backbone. The inputs are face image with the shape of $224 * 224 * 3$ and audio clip spectrogram with $224 * 125 * 1$. We transfer the audio and image separately to the corresponding feature extractors to obtain features with the same size of $12 * 12 * 32$. In the adversarial learning process, the generation module contains a self-attention and FC layers with spectral parametric regularization, and transforms the 4068-dimensional features of audio and visual modalities into 256-dimensions, and then to 128 dimensions. Then, the discriminator is a binary classification network that outputs a 2-dimensional representation of the probability belonging to the corresponding modality. For the matching task, the feature combination of the anchor sample and k candidate matching samples is $(k + 1) * 128$ dimension, which is classified as k-dimensional output to represent the probability of matching between them. For the retrieval task, the distance between the anchor samples and each candidate sample is represented as a 128-dimensional feature that is categorized as a 1-dimensional output to denote the matching relationship between the samples.

Evaluation protocol. The cross-modal matching performance is measured in terms of accuracy (ACC). During the matching training process, we set the batch size to 50 and use Adam [38] with a momentum of 0.9 and weight decay of 0.0005 to fine-tune the network. The initial learning rate of the siamese generator, the siamese discriminator, and the classification are $5e-3$, $5e-3$, and $5e-2$ respectively. The learning rates of the siamese-generator and siamese-discriminator are both delayed from $5e-3$ to $5e-5$, while from $5e-2$ to $5e-4$ for the classifiers. The retrieval results are reported in terms of mean average precision (mAP). During the retrieval training process, we used Adam [38] as the optimizer, where the batch size and momentum were set to 128 and 0.9, respectively. Siamese generator, Siamese discriminator, and classifier learning rates were initialized to $5e-3$, $5e-3$, and $5e-2$ respectively, and decayed by 0.1 in 600 and 1k iterations with a maximum iteration T_{max} of 1.2k.

Dataset. The performance of the proposed algorithm is evaluated on Voxceleb [39] and VGGFace [40] public datasets, which contain 149354 speech segments for 1225 speaker identities and 137060 face images for 1225 face identities, respectively, where the speech identities are aligned with the face identities. For a fair comparison, the number of sampled data and the segmentation scheme is referred to [23], [24], where the validation set is composed of data with the names of people starting with ['**A**', '**B**'], and the test set has consisted of data with the names of people starting with ['**C**', '**D**', '**E**'], and the rest of people were selected as the training set. Table II shows the information on the data splitting.

TABLE II
THE DATA SPLITTING TO TRAINING, VALIDATION, AND TESTING AFTER SAMPLING.

Item	Train	Validation	Test	Total
Identities	924	112	189	1225
Face Images	104724	12260	20076	137060
Audio Clips	113322	14182	21850	149354

B. Comparison Results

In order to evaluate the effectiveness of the proposed method (DSANet), Table IV reports the comparison results with five state-of-the-art algorithms, including AML [14], Wen *et al.* [24], SVHF [10], DIMNet [23] and Wang *et al.* [11], followed by the data splitting scheme in Wen *et al.* [24]. We can see that our model outperforms the other algorithms in both audio-visual matching and retrieval tasks in both V-F and F-V scenarios, with an average improvement of about 1.5%, which evidences the effectiveness of the proposed DSANet for correlating cross-modal audio and visual information. Note that the performance of F-V is generally inferior to the V-F scenario in both binary and multi-way cases. The main reason is, as the visual information, audio signals are more sensitive to the environment and present higher inter-class similarity compared with the facial images [41]. This leads to relative difficulty to distinguish the voice audio features in the F-V scenario. Furthermore, for a fair comparison, we conduct the experiments in the audio-visual matching task following the data splitting scheme in PINs [22] as shown in Talbe III. Consistently, our DSANet significantly outperforms the state-of-the-art methods on all the metrics.

TABLE III
COMPARISON RESULTS OF AUDIO-VISUAL MATCHING WITH THE STATE-OF-THE-ART METHOD IN THE BINARY (k=2) AND MULTI-BINARY (k=10) CASES. WHERE " -" MEANS "NOT AVAILABLE". THE EXPERIMENTAL RESULTS IN THE TABLE ARE OBTAINED FOLLOWING THE DATA SETTINGS PROPOSED BY PINS [22]

Methods	Venue	Binary (ACC)		Multi-way (ACC)	
		V-F	F-V	V-F	F-V
DIMNet [23]	ICLR2019	84.12	84.03	39.75	-
PINs [22]	ECCV2018	84.00	-	31.00	-
SSNet [12]	DIC2019	78.00	78.50	30.00	30.05
AML [14]	TMM2021	92.72	93.3	43.45	39.35
DSANet	Ours	**95.25**	**94.28**	**46.83**	**43.36**

To further validate the superiority of our method, we compare the $1:k$ multi-way matching results in Fig. 5. In multi-way matching, the accuracy decreases as the number of matching candidates increases, but our method maintains a better performance overall, especially in V-F multi-way cases, where

TABLE IV
THE QUALITATIVE RESULTS OF MATCHING AND RETRIEVAL TASKS. BINARY DENOTES THE 1:2 MATCHING WHILE MULTI-WAY DENOTES THE 1:k (k = 10) MATCHING. V-F REPRESENTS MATCHING THE AUDIO ANCHOR TO THE GALLERY FACES, AND VISA VERSA FOR F-V. THE EXPERIMENTAL RESULTS IN THE TABLE ARE OBTAINED FOLLOWING THE DATA SETTINGS PROPOSED BY WEN *et al.* [24].

Methods	Venue	Binary (ACC)		Multi-way (ACC)		Retrieval (mAP)	
		V-F	F-V	V-F	F-V	V-F	F-V
SVHF [10]	CVPR2018	81.0	79.5	34.5	×	-	-
DIMNet [23]	ICLR2019	81.3	81.9	38.4	36.2	4.3	3.8
Wang *et al.* [11]	ACM2020	83.4	84.2	39.7	36.4	4.4	3.4
Wen *et al.* [24]	CVPR2021	87.2	86.5	48.2	44.8	5.5	5.8
AML [14]	TMM2021	89.4	86.3	46.2	43.7	4.8	4.5
DSANet	Ours	**92.5**	**88.4**	**49.1**	**46.8**	**6.3**	**6.1**

we have a smaller drop than any other algorithms, which also validates the robustness of our proposed model. The retrieval results present really poor performance as shown in Table IV since the retrieved candidates are variable in appearance and context, which leads to difficulties in learning robust intrinsic associations for audio-visual data. Even though, we are still able to achieve the most superior performance with the dual-enhanced siamese adversarial network, which demonstrates that it can learn relatively robust associations between audio-visual data. Due to the extremely low performance of the retrieval, we evaluate the following experiments only on matching tasks.

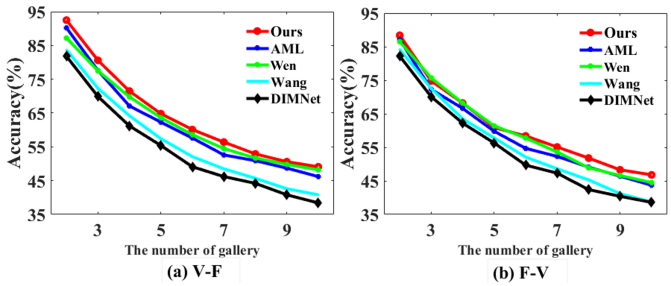


Fig. 5. The quantitative results of 1 : k matching task in V-F and F-V scenario, where k indicates the number of matching candidates in the gallery.

C. Ablation Study

Effects of different components. To verify the effectiveness of our proposed three components, we perform ablation experiments on the proposed dual enhancement mechanism (DEM), siamese adversarial structure (SAS), and adaptive mask cross-entropy (AMCE) loss respectively in the binary-way matching task. The performance of V-F and F-V matching on the baseline model is 90.2% and 86.6%, respectively in Table V (a). Integrating siamese adversarial structure and adaptive mask cross-entropy loss effectively improves the performance of the baseline, as shown in Table V (c) and (d). As shown in Table V (b), direct integrating the dual enhancement mechanism (DEM) slightly decreases the matching performance of the baseline. The main reason is that the subtle features mined by DEM are not guaranteed to eliminate pattern heterogeneity in a single GAN. Meanwhile, the two enhanced features have relatively abundant common information, which also degrades the performance due to information redundancy. However, based on the siamese adversarial structure, the dual enhancement mechanism significantly improves the performance, comparing Table V (e) with (c), or comparing Table V (g) with (f). This demonstrates that each of our proposed components makes a positive contribution. Utilizing all three components achieves the best performance.

TABLE V
ABLATION STUDY OF PROPOSED DSANET ON THE AUDIO-VISUAL MATCHING TASK IN BINARY (WHEN $k = 2$) CASES. '✓' MEANS THE CORRESPONDING COMPONENT IS INCLUDED.

	Component			Binary ($k = 2$)	
	DEM	SAS	$\mathcal{L}_{AMCE}$	V-F	F-V
a				90.2	86.6
b	✓			89.9	87.9
c		✓		91.3	87.0
d			✓	91.6	87.5
e	✓	✓		92.1	88.0
f		✓	✓	91.6	87.9
g	✓	✓	✓	**92.5**	**88.4**

Evaluation on Dual Enhancement Mechanism. To further evaluate the effectiveness of the proposed dual enhancement mechanism, we design three dual enhancement structures as shown in Fig. 6 (a), (b), and (c). Note that cascading multiple dual enhancement mechanisms in a network framework does not significantly improve network performance as shown in Table VI. Considering the complexity of the model, we use the 1-layer dual enhancement mechanism in the training network if not specified. Furthermore, our proposed DSANet model uses only a 1-layer dual enhancement mechanism. We demonstrate the features extracted by DSANet for visualization as shown in Fig. 7. Our DSANet with the dual enhancement mechanism of 1-layer can focus on relatively sufficient feature regions. The dual enhancement mechanism is too diverse for the extracted features, which forces the network to focus on different feature regions. However, the multi-layer dual enhancement mechanism only increases the number of diverse features which may not learn more associations between audio and visual feature regions. As can be seen from Fig. 7, the network tends to focus on the forehead and mouth more compared to the eyes. This is because the face motion of the

same identity varies greatly in the data, while the information on the forehead and mouth is relatively stable, which enables the features of this region to be easily learned as well as associated with the audio clip features.

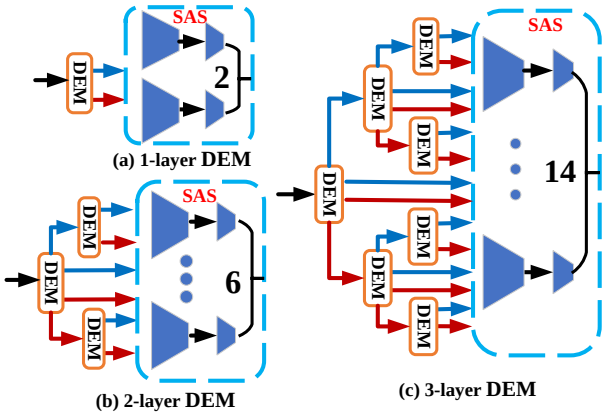


Fig. 6. Demonstration of the three dual enhancement mechanism schemes, where the numbers 2, 6, and 14 denote the number of GANs [34] in the siamese adversarial structure with shared parameters, respectively.

TABLE VI
THE NEW NETWORK STRUCTURE IS FORMED BY CASCADING MULTIPLE LAYERS OF DEM MODULES.

Demo	Methods	V-F	F-V
Fig. 6 (a)	1-layer DEM	92.5	88.4
Fig. 6 (b)	2-layer DEM	92.6	**88.6**
Fig. 6 (c)	3-layer DEM	**92.7**	88.0

Fig. 7. Class activation maps (CAM) generated by the proposed DSANet compared with the baseline.

Evaluation on Siamese Adversarial Structure. To further validate the effectiveness of the proposed siamese adversarial structure, we compare the performance of the enhanced features by handling enhanced features in two adversarial forms. To demonstrate each scheme, we combine the feature enhancement approach and the adversarial forms into five structures, as shown in Fig. 8 (a), (b), (c), (d), and (e). Among them, Fig. 8 (a), (b), and (c) are only subtle features, only salient features and salient and subtle features combined are enhanced, respectively, and the matching performance is obtained by the single-stream adversarial structure. While Fig. 8 (d) and (e) are enhanced with only subtle features and only salient features, respectively, and the matching performance is obtained by the siamese adversarial structure. As shown in Table VII, the single-stream adversarial structure obtains lower performance than the siamese adversarial structure in all cases. Our proposed DSANet is a parallel input of enhanced salient features and enhanced subtle features into the siamese adversarial structure, which can capture more regions of features for model robustness, thus it is outperformed by the baseline network. This further validates that the designed siamese adversarial structure is effective.

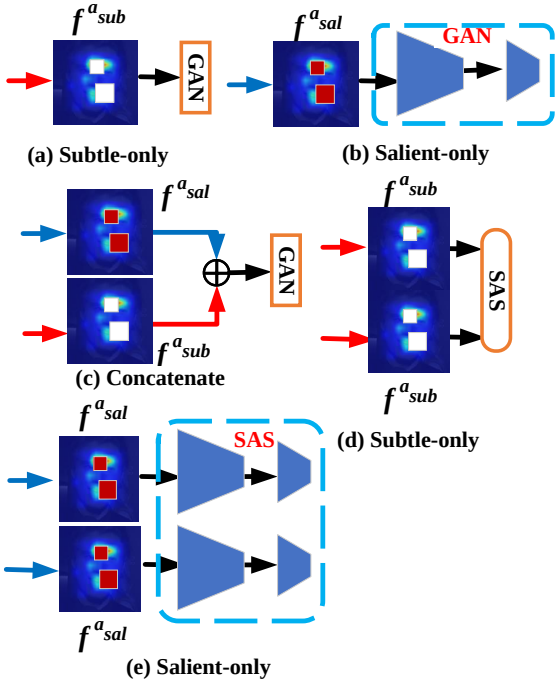


Fig. 8. Five combinations of feature enhancement methods and adversarial approaches.

Evaluation on Adaptive Mask Cross-entropy loss. To learn to distinguish hard samples, we introduce $\mathcal{L}_{AMCE}$ loss to focus on hard negative class samples for learning inter-class variability. To verify the effectiveness of AMCE, we compared it with three state-of-the-art cross-entropy methods, binary cross-entropy ($\mathcal{L}_{BCE}$) [14], gradient-boosting cross-entropy ($\mathcal{L}_{GBCE}$) [19] and focal loss ($\mathcal{L}_{Focal}$) [42], on binary and multi-way matching tasks. $\mathcal{L}_{GBCE}$ has low performance in the binary matching case due to a smaller number of negative sample comparisons, while $\mathcal{L}_{Focal}$ does not eliminate simple samples compared to $\mathcal{L}_{AMCE}$ to learn distinction for hard samples. The other cross entropy directly replaces the adaptive

TABLE VII
THE IMPACT OF THE PRESENCE AND ABSENCE OF SAS ON THE MODEL IS COMPARED IN THE BINARY AUDIO-VISUAL MATCHING TASK. WHERE CONCATENATE IS THE MERGING OF FEATURES FROM THE OUTPUT OF THE DE MECHANISM.

Structure	Enhancement	V-F	F-V
GAN	Baseline	90.2	86.6
	Subtle_only (Fig. 8 (a))	91.8	87.7
	Salient_only (Fig. 8 (b))	91.9	86.3
	Concatenate (Fig. 8 (c))	89.9	87.9
SAS	Subtle_only (Fig. 8 (d))	91.8	87.9
	Salient_only (Fig. 8 (e))	92.0	88.0
	Subtle + Salient (DSANet)	**92.5**	**88.4**

mask cross entropy to calculate the matching probability loss of the classification. As shown in Table VIII, $\mathcal{L}_{AMCE}$ outperforms the other two losses in the binary task and can also perform comparably in the multi-way case.

TABLE VIII
COMPARISON OF THE PERFORMANCE IMPACT OF FOUR CROSS-ENTROPY LOSSES ON BINARY AND MULTI-WAY 1:k (k=10) CASES FOR AUDIO-VISUAL MATCHING.

Methods	Binary (ACC)		Multi-way (ACC)	
	V-F	F-V	V-F	F-V
$\mathcal{L}_{BCE}$ [14]	92.1	88.0	47.2	47.0
$\mathcal{L}_{GBCE}$ [19]	91.8	88.2	49.0	**47.5**
$\mathcal{L}_{Focal}$ [42]	92.2	87.7	48.5	46.9
$\mathcal{L}_{AMCE}$	**92.5**	**88.4**	**49.1**	46.8

Evaluation on Adversarial Network. To verify the impact of modal heterogeneity elimination on network performance, we use multiple adversarial networks on DSANet to evaluate the ability to handle modal heterogeneity. Specifically, all three original GAN [13], Wasserstein GAN (WGAN) [33] and self-attention GAN (SAGAN) [34] were designed to compare the output performance of the siamese structure. It can be seen from Table IX that superior performance can be achieved when the SAGAN is used. This also validates that generative adversarial networks can play an important role in audio-visual cross-modal matching tasks.

TABLE IX
COMPARISON TO DIFFERENT ADVERSARIAL NETWORK METHODS ON AUDIO-VISUAL MATCHING TASK.

Methods	Binary (ACC)		Multi-way (ACC)	
	V-F	F-V	V-F	F-V
GAN [13]	90.9	86.2	45.8	43.9
WGAN [33]	92.1	87.5	47.6	45.4
SAGAN [34]	**92.5**	**88.4**	**49.1**	**46.8**

Evaluation on Metric Learning. To evaluate the dependence of the proposed DSANet method on metric loss, our proposed structural metric is compared with the commonly used the triplet loss [43] and the lifted struct loss [44]. In the audio-visual cross-modal matching task, we propose a structural metric with both intra-modal metric (positive-negative) and inter-modal (anchor-positive and anchor-negative) metric constraints. The triplet loss [43] is simply a comparison of the inter-modal distance (anchor-positive) with the intra-modal distance (positive-negative). Compared to the triple loss, the structural metric can better constrain the identity feature distribution to reach a more robust performance. The lifted structure loss [44] and structure metric loss have the same distance constraint. And there is only one negative class sample in the binary matching such that the results are consistent under the metric constraint. In the case of multi-way matching, the former sums over all negative class sample distances while the latter selects the minimum negative class sample distance, which naturally avoids the positive and negative sample distance statistical imbalance problem. The comparison results are shown in Table X, which verifies that the structure metric is more effective in the cross-modal matching of audio and visuals.

TABLE X
COMPARISON TO DIFFERENT METRIC LOSS ON AUDIO-VISUAL MATCHING TASK.

Methods	Binary (ACC)		Multi-way (ACC)	
	V-F	F-V	V-F	F-V
Triplet [43]	91.7	84.5	46.5	44.6
Lifted Structure [44]	**92.5**	**88.4**	48.9	45.5
Structure Metric	**92.5**	**88.4**	**49.1**	**46.8**

Evaluation on Feature Selection Strategy. To verify the effects of different feature selection strategies, we evaluate multiple different feature selection strategies and two feature manipulation techniques in our network. The feature suppression technique helps the network focus on more subtle feature regions [16], [19]. However, with only one feature selection strategy, the model falls slightly below the superior performance, as shown in Table XI (a-b). The feature enhancement technique also helps to improve the performance of the model, which is less affected by the feature selection strategy, as shown in Table XI (c-d). Therefore, we operate both feature enhancement and feature suppression techniques under the peak region, the randomly selected region, and the combined region of the two regions. As can be seen in Table XI (e-g), we achieve the best performance by simultaneously performing the dual enhancement mechanism on the features selected from both selection strategies, which indicates the effectiveness of both feature selection strategies.

D. Hyper-parameters Analysis

We analyze the hyperparameters of λ, μ, α, β, c, r, λ_1 and μ_1 in this paper. Specifically, λ and μ in Eq. (24) indicate the weights of structured metric and adaptive masked cross-entropy in the matching loss, α and β in Eq. (8) and (9) indicate the factors of salient features and subtle feature enhancement, respectively. c and r in Eq. (3) determine the random patch locations, while λ_1 and μ_1 in Eq. (25) indicate the weights in the retrieval loss. As shown in Table XII, the matching task is not sensitive to the hyperparameters λ, μ,

TABLE XI
COMPARISON OF DIFFERENT FEATURE SELECTION STRATEGIES IN THE BINARY AUDIO-VISUAL MATCHING TASK.

	Feature selection strategy	V-F	F-V
	Baseline	90.2	86.6
a	Only Peak Suppression	92.2	87.7
b	Only Patch Suppression	92.1	88.2
c	Only Peak Enhancement	92.0	88.1
d	Only Patch Enhancement	91.9	88.3
e	Only Peak DSANet	92.3	87.7
f	Only Patch DSANet	92.1	**88.5**
g	Subtle + Salient (DSANet)	**92.5**	**88.4**

α, and β, while the superior performance is achieved when $\lambda = 2$, $\mu = 3$, $\alpha = 1$, $\beta = 0.9$, $c = 4$, and $r = 4$. By contrast, the retrieval task is more sensitive to the hyperparameters λ_1 and μ_1 since the overall performance of the retrieval task is extremely poor due to the huge challenge of retrieving the cross-modality audio-visual data. Therefore, how to balance the global classification loss and metric loss is very important for the model to obtain robust performance.

TABLE XII
THE EFFECTS OF HYPERPARAMETERS OF λ, μ, α, β, c, AND r ON BINARY MATCHING TASK AND THE EFFECTS OF λ_1 AND μ_1 ON RETRIEVAL TASK.

Param	Settings	V-F	F-V	Param	Settings	V-F	F-V
λ	1	91.6	86.8	c	2	92.2	87.9
	1.5	92.0	87.5		3	92.3	88.1
	2	**92.5**	**88.4**		**4**	**92.5**	**88.4**
	2.5	91.9	87.3		5	92.2	88.0
μ	1	91.6	87.4	r	2	92.3	**88.4**
	2	92.0	87.8		3	91.9	88.0
	3	**92.5**	**88.4**		**4**	**92.5**	**88.4**
	4	92.1	88.1		5	92.0	88.3
α	0.5	92.0	**88.5**	β	1	92.2	**88.4**
	1	**92.5**	88.4		**0.9**	**92.5**	**88.4**
	1.5	91.9	88.3		0.7	92.3	87.8
	2	92.1	88.2		0.5	92.1	87.1
λ_1	0.05	6.1	6.0	μ_1	0.1	5.5	5.2
	0.1	**6.3**	**6.1**		**0.2**	**6.3**	**6.1**
	0.2	5.3	5.1		0.4	6.1	5.7
	0.4	4.8	4.5		0.6	5.6	5.5

V. CONCLUSION

The association present between the visual and audio information has attracted the attention of researchers. To find a highly matching relationship between the audio-visual cross-modal data, we developed a novel Dual-enhanced Siamese Adversarial Network (DSANet). Specifically, we first randomly select a region in which salient features are enhanced to maintain inter-class discriminability while salient features are suppressed for enhancing subtle features to help improve intra-class compactness and inter-class discriminability. Then, to uncouple the mutual impact between enhanced subtle and salient features, we eliminate the modal heterogeneity between enhanced salient and subtle features by siamese adversarial networks in a parallel manner. In addition, the network is further forced to focus on learning feature variances between hard classes by the adaptive masked cross-entropy loss. Experimental results on audio and face image data validate that the proposed DSANet compares favorably with state-of-the-art audio-visual cross-modal matching algorithms. In the future, we will apply the proposed DSANet model to other cross-modal data to implement tasks such as recognition and localization to prove the generalization of the model.

REFERENCES

[1] V. Bruce and A. Young, "Understanding face recognition," *British journal of psychology*, vol. 77, no. 3, pp. 305–327, 1986.

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

[3] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "Putting the face to the voice': Matching identity across modality," *Current Biology*, vol. 13, pp. 1709–1714, 2003.

[4] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15495–15505, 2021.

[5] D. Hu, R. Qian, M. Jiang, X. Tan, S. Wen, E. Ding, W. Lin, and D. Dou, "Discriminative sounding objects localization via self-supervised audio-visual matching," *Advances in Neural Information Processing Systems*, vol. 33, no. 19, pp. 10077–10087, 2020.

[6] C. Xue, X. Zhong, M. Cai, H. Chen, and W. Wang, "Audio-visual event localization by learning spatial and semantic co-attention," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[7] M. Kim, J. Hong, S. J. Park, and Y. M. Ro, "Cromm-vsr: Cross-modal memory augmented visual speech recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 4342–4355, 2021.

[8] F. Tao and C. Busso, "End-to-end audiovisual speech recognition system with multitask learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 1–11, 2021.

[9] Z. Li, J. Tang, L. Zhang, and J. Yang, "Weakly-supervised semantic guided hashing for social image retrieval," *International Journal of Computer Vision*, vol. 128, no. 8, pp. 2265–2278, 2020.

[10] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8427–8436, 2018.

[11] R. Wang, X. Liu, Y. Cheung, K. Cheng, N. Wang, and W. Fan, "Learning discriminative joint embeddings for efficient face and voice association," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1881–1884, 2020.

[12] 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 *Proceedings of the Digital Image Computing: Techniques and Applications*, pp. 1–7, 2019.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.

[14] A. Zheng, M. Hu, B. Jiang, Y. Huang, Y. Yan, and B. Luo, "Adversarial-metric learning for audio-visual cross-modal matching," *IEEE Transactions on Multimedia*, vol. 24, pp. 338–351, 2021.

[15] K. Cheng, X. Liu, Y.-m. Cheung, R. Wang, X. Xu, and B. Zhong, "Hearing like seeing: Improving voice-face interactions and associations via adversarial deep semantic matching network," in *Proceedings of the ACM International Conference on Multimedia*, pp. 448–455, 2020.

[16] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3300–3310, 2020.

[17] Z. Li and J. Tang, "Semi-supervised local feature selection for data classification," *Science China Information Sciences*, vol. 64, no. 9, pp. 1–12, 2021.

[18] Z. Li, Y. Sun, L. Zhang, and J. Tang, "Ctnet: Context-based tandem network for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[19] G. Sun, H. Cholakkal, S. Khan, F. Khan, and L. Shao, "Fine-grained recognition: Accounting for subtle differences between similar classes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12047–12054, 2020.

[20] X. Shen, X. Tian, T. Liu, F. Xu, and D. Tao, "Continuous dropout," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 3926–3937, 2017.

[21] B. O. Ayinde, T. Inanc, and J. M. Zurada, "Regularizing deep neural networks by enhancing diversity in feature extraction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2650–2661, 2019.

[22] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proceedings of the European Conference on Computer Vision*, pp. 71–88, 2018.

[23] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," *Proceedings of the International Conference on Learning Representations*, 2019.

[24] P. Wen, Q. Xu, Y. Jiang, Z. Yang, Y. He, and Q. Huang, "Seeking the shape of sound: An adaptive framework for learning voice-face association," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16347–16356, 2021.

[25] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10394–10403, 2019.

[26] R. Xu, C. Li, J. Yan, C. Deng, and X. Liu, "Graph convolutional network hashing for cross-modal retrieval.," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 982–988, 2019.

[27] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2070–2083, 2018.

[28] P. Hu, D. Peng, X. Wang, and Y. Xiang, "Multimodal adversarial network for cross-modal retrieval," *Knowledge-Based Systems*, vol. 180, pp. 38–50, 2019.

[29] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Adversarial cross-spectral face completion for nir-vis face recognition," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1025–1037, 2019.

[30] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *International Journal of Automation and Computing*, vol. 18, no. 3, pp. 351–376, 2021.

[31] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7181–7189, 2018.

[32] Y. Hao, N. Wang, J. Li, and X. Gao, "Hsme: Hypersphere manifold embedding for visible thermal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8385–8392, 2019.

[33] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R2gan: Cross-modal recipe retrieval with generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11477–11486, 2019.

[34] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the International Conference on Machine Learning*, pp. 7354–7363, 2019.

[35] Y. Peng and J. Qi, "Cm-gans: Cross-modal generative adversarial networks for common representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1, pp. 1–24, 2019.

[36] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1487–1495, 2019.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of the International Conference on Learning Representations*, 2015.

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

[40] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference*, pp. 41.1–41.12, 2015.

[41] Y. Wei, D. Hu, Y. Tian, and X. Li, "Learning in audio-visual context: A review, analysis, and new perspective," *arXiv preprint arXiv:2208.09579*, 2022.

[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[43] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[44] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016.