

Anchor-aware Deep Metric Learning for Audio-visual Retrieval

Donghuo Zeng*
do-zeng@kddi-research.jp
KDDI Research, Inc.
Saitama, Japan

Kazushi Ikeda
kz-ikeda@kddi-research.jp
KDDI Research, Inc.
Saitama, Japan

Yanan Wang
wa-yanan@kddi-research.jp
KDDI Research, Inc.
Saitama, Japan

Yi Yu*
yiyu@hiroshima-u.ac.jp
Hiroshima University
Hiroshima, Japan

ABSTRACT

Metric learning minimizes the gap between similar (positive) pairs of data points and increases the separation of dissimilar (negative) pairs, aiming at capturing the underlying data structure and enhancing the performance of tasks like audio-visual cross-modal retrieval (AV-CMR). Recent works employ sampling methods to select impactful data points from the embedding space during training. However, the model training fails to fully explore the space due to the scarcity of training data points, resulting in an incomplete representation of the overall positive and negative distributions. In this paper, we propose an innovative Anchor-aware Deep Metric Learning (AADML) method to address this challenge by uncovering the underlying correlations among existing data points, which enhances the quality of the shared embedding space. Specifically, our method establishes a correlation graph-based manifold structure by considering the dependencies between each sample as the anchor and its semantically similar samples. Through dynamic weighting of the correlations within this underlying manifold structure using an attention-driven mechanism, Anchor Awareness (AA) scores are obtained for each anchor. These AA scores serve as data proxies to compute relative distances in metric learning approaches. Extensive experiments conducted on two audio-visual benchmark datasets demonstrate the effectiveness of our proposed AADML method, significantly surpassing state-of-the-art models. Furthermore, we investigate the integration of AA proxies with various metric learning methods, further highlighting the efficacy of our approach.

CCS CONCEPTS

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

*Corresponding authors.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ICMR '24, June 10–14, 2024, Phuket, Thailand
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0619-6/24/06
<https://doi.org/10.1145/3652583.3658067>

KEYWORDS

Audio-visual Retrieval, Anchor-aware, Deep Metric Learning, Triplet Loss

ACM Reference Format:

Donghuo Zeng, Yanan Wang, Kazushi Ikeda, and Yi Yu. 2024. Anchor-aware Deep Metric Learning for Audio-visual Retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24)*, June 10–14, 2024, Phuket, Thailand. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3652583.3658067>

1 INTRODUCTION

Metric learning endeavors to construct an embedded representation of the data, ensuring that similar pairs remain proximate while dissimilar pairs are distanced from each other within the embedding space [18, 23, 34, 35]. Metric learning serves as a cornerstone in various domains, such as image retrieval and clustering [20, 35], cross-modal learning [24], person re-identification [11]. In practice, as the objective of metric learning methods is to establish an embedding space where similar samples are drawn closer together while dissimilar samples are distinctly separated, it has been applied in AV-CMR task [40], by refining the similarity among data samples to enhance the efficacy of retrieval across different modalities.

Recent works [13, 18] on metric learning always incorporate deep learning approaches to leverage their capability of learning robust feature representations. As such deep learning-based metric learning approaches [13] often require large amounts of data to proficiently acquire intricate feature representations, the imbalanced distribution of similar and dissimilar data pairs in the massive training dataset leads to suboptimal results of metric learning. Existing works [17, 21, 30, 45] aim to generate synthetic hard negative samples where dissimilar data points are mistakenly identified, by employing adversarial learning [21, 45] and linear interpolation [17] or non-linear interpolation [30], which captures the potential of numerous easy negatives. However, these works face challenges in controlling the number of insertion points to obtain optimal models and often overlook the need to identify robust anchors to fully capture the positive and negative distribution to refine the embedding space. Overall, the scarcity of training data points will raise the problem of missing embeddings in existing methods. This problem leads to suboptimal learning of the embedding space and compromises the quality of sample representation, including positive and negative samples, consequently impairing the performance of deep metric learning.

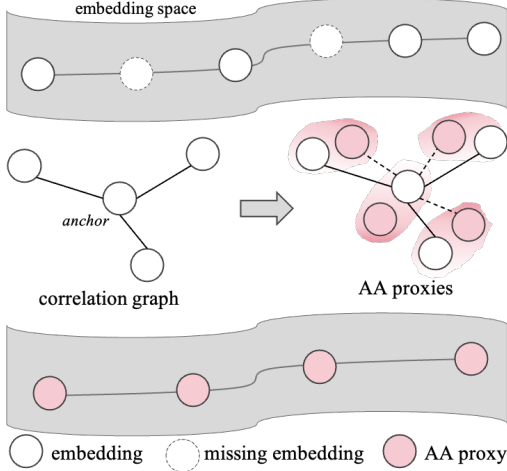


Figure 1: This diagram illustrates the role of the anchor-aware (AA) proxy in deep metric learning. Missing embeddings due to a lack of training data points leads to suboptimal learning of the embedding space. We introduce an AA proxy derived from the correlation graph for each embedding, facilitating the migration toward optimal embedding space learning.

To address the issue, we propose an innovative Anchor-aware Deep Metric Learning (AADML) method to uncover the intrinsic correlations among existing data points and compensate for the incomplete learning of the embedding space due to insufficient data points, as illustrated in Fig. 1. Our approach comprises three fundamental components: (1) as the manifold signifies the foundational structure and relationships within the data, we construct a correlation graph within each modality to facilitate the adept capture of the innate manifold structure. (2) To effectively capture correlations among similar samples and comprehensively consider each sample’s contribution to the embedding space learning process, we leverage manifold pairs from the correlation graph across modalities using an attention-driven mechanism. For example, if we select an audio sample as an anchor point and identify the K nearest audio samples relative to the anchor from the correlation graph, we create k tuples as inputs of an attention-driven model to generate an anchor-aware (AA) proxy for the anchor. (3) Leveraging AA proxies due to their ability to directly capture nuanced dependencies among similar pairs while indirectly mitigating the interference of dissimilar pairs, we employ these AA scores as sample proxies to calculate relative distances within a metric learning framework.

To validate the effectiveness of the proposed AADML approach, extensive experiments are conducted on two cross-validated audio-visual benchmark datasets. These datasets encompass annotation labels present in both audio and visual modalities and offer a comprehensive evaluation setup. Experimental results exhibit the superiority of our anchor-aware approach, achieving substantial performance enhancements, and outperforming state-of-the-art models in AV-CMR tasks by 3.0% on large dataset VEGAS and by 45.6%

on small dataset AVE in terms of mean average precision (MAP), respectively. Additionally, we demonstrate the applicability of the AA proxy across various metric learning methods, resulting in performance surpassing existing state-of-the-art losses in AV-CMR tasks.

The remainder of this paper unfolds as follows: Section 2 provides an overview of related works. Section 3 elaborates on the proposed Anchor-Aware Deep Metric Learning (AADML) approach and Section 4 outlines the experimental setup and results. Lastly, in Section 6, we summarize our work and discuss potential future directions in the AV-CMR task.

2 RELATED WORK

2.1 Audio-visual Cross-modal Retrieval

Audio-visual cross-modal retrieval is a challenging task that aims to retrieve relevant content across different modalities such as audio and visual modalities [28, 38, 40]. Canonical Correlation Analysis (CCA) [4] is a classical linear technique used to find correlated patterns between two sets of variables. In the context of AV-CMR, CCA-based methods have been extensively studied to learn joint representations that capture the correlations between audio and visual data. Representative approaches include traditional CCA [4], K-CCA [19], Cluster-CCA [26], Deep CCA (DCCA) [1], C-DCCA [36], Triplet with Cluster-CCA (TNN-C-CCA) [41], and VAE-CCA [43]. While CCA-based methods offer interpretable embeddings, they struggle to capture complex and non-linear relationships present in real-world audio-visual data.

Deep learning is a powerful approach for representation learning in AV-CMR, which employs deep neural networks to learn joint embeddings by minimizing the distance between corresponding samples. Notably, the CLIP [25], ACMR [32], DSCMR [44], BiC-Net [10], DCIL [46], CCTL [38], EICS [39], and MSNSCA [42]. These models have exhibited exceptional performance in cross-modal retrieval between audio-visual [38, 39, 42], positioning them as valuable contributions to the field of representation learning. While prior efforts emphasize training intricate neural networks to acquire representations for AV-CMR, our approach introduces a novel direction by focusing on advanced metric learning techniques.

2.2 Metric Learning for Retrieval

Metric learning is of paramount importance in retrieval tasks, particularly when tailored to the specific requirements of AV-CMR. The objective is to develop a suitable distance or similarity metric that preserves the underlying structure of data across different modalities to enhance retrieval accuracy.

Traditional metric learning methods such as Mahalanobis distances [5], focus on preserving the intra-class similarity while increasing the inter-class distance. However, they may struggle with capturing non-linear relationships within complex audio-visual data. The emergence of deep metric learning has gained significant traction across various retrieval tasks, including AV-CMR. Prominent techniques, such as contrastive loss [9] and triplet loss [11], aim to pull similar samples closer together and push dissimilar samples further apart in the learned embedding space. These methods excel in generating discriminative embeddings by analyzing relative distances between samples. Additionally, an array of deep metric

learning variations [12, 13, 17, 21, 30, 45] has been introduced. These novel deep metric learning methods offer the potential to further enhance AV-CMR. For instance, the CCTL [38] model adapts triplet loss to audio-visual cross-modal data. However, it has limitations in terms of model training difficulty and sensitivity to data scarcity situations.

The works [7, 15, 31] have proposed advanced metric learning in some clues. Notably, approaches such as harnessing hyperbolic geometry to capture the intricate correlations present in natural data have emerged [7, 15, 31], which exists its potential complexity and computational overhead. Similar problems will happen in the generator model involved in data augmentation [17, 22, 45] methods. These methods are an effective strategy to address data scarcity by generating synthetic samples through various transformations. However, the effectiveness of data augmentation is sensitive to the quality of synthetic points and the specific generation methods employed, raising concerns about model reliability.

Instead of relying on data augmentation or mining effective data settings to train by complex selection methods, our proposed approach efficiently computes global attention-driven dependencies between an anchor and other similar samples in parallel to dynamically weigh the correlations of their manifold structure, the correlations scores as proxies for data samples will improve the performance of AV-CMR while using such proxies for the deep metric learning.

3 PROBLEM FORMULATION

In audio and visual cross-modal retrieval tasks, the primary challenge arises from the intricate features of audio and visual data, coupled with their different feature distributions. This disparity makes direct comparisons using basic metrics such as distance or similarity metrics unavailable. Hence, there arises a need to devise cross-modal retrieval methods. These methods aim to map audio and visual features into a shared embedding space, facilitating the generation of new features for each modality. These features can then be utilized for comparison computations directly.

Assume an audio-visual dataset comprising n videos, symbolized as $\Sigma = \{\sigma_i\}_{i=0}^{n-1}$, where each video consists an audio and visual pair, represented as $\sigma_i = (a_i, v_i)$. Here, $a_i \in R^{128}$ denotes audio features extracted by VGGish pre-trained model¹, and $v_i \in R^{1024}$ signifies visual features extracted by Inception V3 pre-trained model². Each σ_i instance is associated with a semantic vector label $Y_i = [y_i^0, y_i^1, \dots, y_i^{c-1}]$, where c represents the number of semantic categories. In this context, y_i^j is a binary value; it equals 1 if the sample belongs to the j -th category ($j = 0, 1, 2, \dots, c-1$), and 0 otherwise.

Our objective is to feed the a_i and v_i pair into a shared label space, where the feature dimension aligns with the number of pre-defined labels. To enhance the learning of shared features, we introduce a novel metric learning approach aligned with triplet loss. This method creates a transformative proxy for each anchor. For instance, if we designate an audio sample a_i as an anchor, its transformative proxy is denoted as $AA(a_i)$. This proxy captures intricate dependencies and generates associations through attention-driven mechanisms. Empowered by $AA(a_i)$, samples belonging to

the same class collaborate to recalibrate and refine their embeddings. This collective effort culminates in the creation of a more nuanced, comprehensive, and accurate semantic embedding space. Consequently, it enhances the performance of fundamental metric learning techniques.

4 APPROACH

In this section, we comprehensively present an Anchor-Aware Deep Metric Learning (AADML) model, which is a cohesive framework that combines: 4.1. correlation graph-based manifold structure, 4.2. anchor-aware (AA) proxy, and 4.3. leveraging the AA proxy in metric learning to enhance the accuracy of AV-CMR, as seen in Fig. 2.

4.1 Correlation Graph-based Manifold Structure

The utilization of manifold structure is crucial for facilitating our Anchor-Aware mechanism to accurately capture dependencies of similar pairs. We employ an approach based on correlation graphs to capture the underlying manifold structure inherent in data samples. The objective of this approach is to ensure that the similarity distance between data samples within the same manifold remains small, thereby promoting retrieval accuracy.

We commence by constructing undirected graphs, $Graph = (V_m, S_m)$, m individually for the audio or visual modality. Here, V_m can be audio set $\{a_0, a_1, \dots, a_{n-1}\}$ or visual set $\{v_0, v_1, \dots, v_{n-1}\}$, which represents the vertices, and S_m constitutes the cosine similarity matrix of m modality. The matrix S_m is meticulously defined as follows.

$$S_m(p, q) = \begin{cases} 1 & \text{if } x_p \in k\text{-NN}(x_q) \text{ and } (p, q) \in C_i \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

where p and q are pointers ranging from 0 to $n-1$, $x \in \{a, v\}$, x_q and x_p indicate q -th and p -th of x , respectively. $(p, q) \in C_i$ represents x_p and x_q belong to the same category C_i , $i = 0, 1, \dots, c-1$. The $k\text{-NN}(x_q)$ is the k -nearest neighbor of x_q within the training set. The k -nearest neighbors algorithm consists of three main steps for identifying the nearest neighbors to a given reference vector x_q .

Firstly, we compute the cosine similarity between the x_q for one modality and each vector x_p from another modality in the dataset. The cosine similarity between two vectors x_q and x_p is defined in Eq. 2:

$$\text{Similarity}(x_q, x_p) = \frac{x_q \cdot x_p}{\|x_q\| \cdot \|x_p\|} \quad (2)$$

Secondly, after computing the cosine similarity for each pair (x_q, x_p) , we rank the computed similarities for all pairs by decreasing order as a rank list, which is computed by Eq. 3.

$$\text{rank}(x_q, x_p) = \text{argsort}_{p=0}^{n-1}(\text{Similarity}(x_q, x_p)) \quad (3)$$

where given a query x_q and fixed the indicator q , we rank the similarity between x_q and all the x_p , $p = 0, 1, \dots, n-1$.

Thirdly, we select the top k similarities from the ranked list $\text{rank}(x_q, x_p)$ to obtain the k -nearest neighbors. These selected neighbors, denoted as the set $k\text{-NN}(x_q)$, can be represented in Eq. 4

$$k\text{-NN}(x_q) = \{x_{p1}, x_{p2}, \dots, x_{pk}\} \quad (4)$$

¹<https://github.com/tensorflow/models/blob/master/research/audioset/vggish/README.md>

²<https://huggingface.co/docs/timm/en/models/inception-v3>

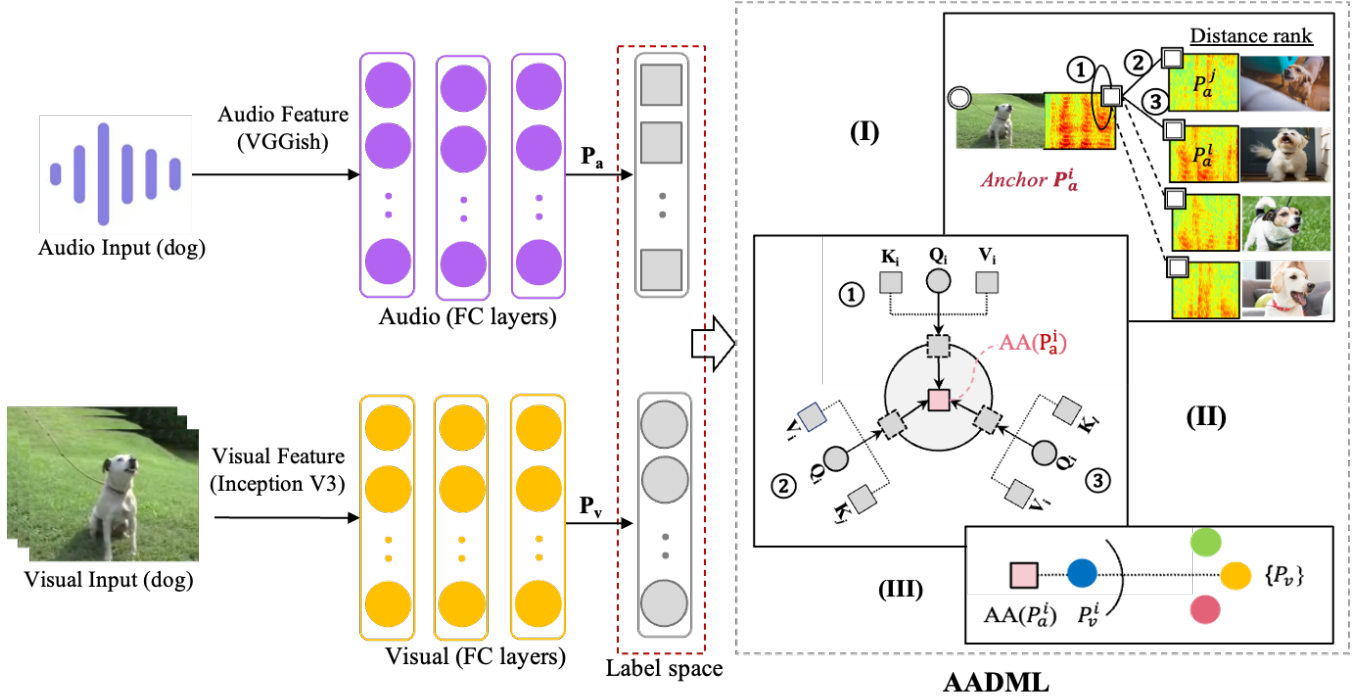


Figure 2: The framework of our proposed model. The audio and visual features extracted by pre-trained models VGGish and Inception V3, respectively, are projected into label space as predicted label embeddings. AADML approach operates within the label space and comprises three distinct components: (I) Choosing an audio sample P_a^i as the anchor, we traverse the correlation graph to discern the k ($k=3$) nearest audio samples (P_a^k vs. P_a^l) relative to the anchor, thus forming three manifold pairs as key-value pairs for (II), to compute the attention score $A(\cdot)$ while the anchor as query (Q_i) with each pair: $P_a^{i,j,k}$ as key ($K_{i,j,k}$), P_a^i as value (V_i). The anchor-aware $AA(\cdot)$ score (pink box) is then obtained as the average of this $A(\cdot)$ across the three key-value pairs. (III) This score is subsequently utilized as an anchor proxy for foundational metric learning methods like contrastive and triplet loss.

where p_1, p_2, \dots, p_k are the top k indices of rank list.

Notably, pairwise information inherently exists in cross-modal data, if the corresponding audio a_k is within the same manifold as audio query q_j , the paired visual v_k is also within the same manifold as q_j , and vice versa, seen in component (I) of the Fig. 2. With this definition, we aim to leverage the underlying data manifold of different modalities to construct the anchor-aware proxy.

4.2 Anchor-Aware (AA) Proxy

After obtaining the generated manifold pairs, the method proceeds to predict the relevance score of the anchor by leveraging its corresponding manifold pairs using the attention model. The relevance score is represented as an AA proxy for each anchor.

4.2.1 Correlation Graph for AA Proxy. In this context, we select a sample from one modality as the anchor and then identify the k nearest audio samples relative to the anchor from the k -NN(\mathbf{x}) (Eq. 4) in the correlation graph. This process forms three manifold pairs as key-value pairs for the attention-driven mechanism, which computes the AA proxy score for the anchor.

For instance, suppose we select the audio sample P_a^i as the designated anchor point and set the hyperparameter k as 3, we derive three manifold pairs as key-value pairs: (P_a^i, P_a^j) , (P_a^i, P_a^l) , and (P_a^l, P_a^j) . These pairs are then used as inputs for the attention-driven mechanism to compute the AA proxy score for the anchor P_a^i . The resulting tuples are expressed as $(Q, K, V) = \{(P_a^i, P_a^j, P_a^j), (P_a^i, P_a^l, P_a^l), (P_a^l, P_a^j, P_a^j)\}$.

4.2.2 Anchor-Aware (AA) Proxy Computation Method. Drawing inspiration from established concepts in scaled dot-product attention and multi-head attention models, our goal is to capture intricate dependencies within modalities and interdependencies across modalities, while simultaneously extracting adaptive weights for each data point. To achieve this, we utilize scaled dot-product attention to compute attention scores. Then, we employ a multi-head self-attention mechanism in a parallel manner.

Firstly, we aim to compute the attention score $A(Q, K, V)$ for each tuple (Q, K, V) by using the scaled dot-product attention mechanism, which is a fundamental aspect of AA proxy computation. The $A(Q, K, V)$ score quantifies the degree of focus to allocate to other samples within the input sequence, defined by Eq. 5

$$A(Q, K, V) = \text{softmax} \left(\frac{QW^Q K^T W^K}{\sqrt{d_k}} \right) V W^V \quad (5)$$

where W^Q , W^K , and W^V are the parameter matrices of Q , K , and V , respectively.

Secondly, multi-head attention focuses on different parts of the input simultaneously by splitting the input into multiple "heads" and computing separate attention scores by projecting the queries, keys, and values h times.

$$\hat{A}(Q, K, V) = \text{Concat}(A_1, A_2, \dots, A_h) W^O \quad (6)$$

where h is the number of attention heads, W^O represents the output projection matrix, which is applied to the concatenated attention heads to compute the final output of the multi-head attention layer.

Finally, we compute the $AA(\cdot)$ proxy, using a_i as the input example.

$$AA(a_i) = \frac{1}{k} \sum_{\theta=1}^k \hat{A}(v_i, a_\theta, a_i) \quad (7)$$

where k is the hyperparameter of Eq. 4.

The essence of AA proxy lies in establishing global dependencies between anchors (queries) and their semantic counterparts across both intra- and inter-modal relationships. Computed in parallel, these relationships yield a dynamic AA proxy score that encapsulates the degree of association between anchor points and analogous samples.

4.3 Leveraging the AA Proxy in Metric Learning

We explore the pragmatic applications of the AA proxy within the realm of metric learning. With $AA(\cdot)$ positioned as an innovative anchor proxy, we redefine conventional metric learning techniques to align with it. Specifically, we demonstrate the seamless integration of $AA(\cdot)$ in computing relative distances for both triplet loss [11] and contrastive loss [9].

The AA proxy effectively replaces the conventional anchor, while the positive and negative components remain unaltered. This reformulated paradigm infuses the metric learning process with heightened sensitivity to the intricate dependencies captured by $AA(\cdot)$, and harnesses the inherent parallelism within the $AA(\cdot)$, thereby yielding more accurate and robust embeddings. The $AA(\cdot)$, acting as a conduit for similarity measurement, are adeptly utilized as proxies for computing relative distances. For example, Triplet loss with AA proxy can be defined as:

$$\mathcal{L}_{AA+triplet} = \sum_{i=1}^N [\|AA(a_i) - AA(p_i)\|_2^2 - \|AA(a_i) - AA(n_i)\|_2^2 + \alpha]_+ \quad (8)$$

where N is the number of sample with one batch, p_i and n_i are the positive and negative sets of original anchor sample a_i . The hyperparameter α serves as the margin in triplet loss, defining a threshold that regulates the degree of dissimilarity permissible between the anchor-positive and anchor-negative pairs. The $\|\cdot\|_2$ represents the Frobenius norm. The formula for contrastive loss

can be expressed as:

$$L_{AA+contrastive} = (1-y) \frac{1}{2} D^2 + y \frac{1}{2} \max(0, m-D)^2 \quad (9)$$

$$D = \|AA(x_i) - AA(x_j)\|_2^2$$

where y indicate the given two data points x_i and x_j is similar ($y=1$) or dissimilar ($y=0$). m is the margin hyperparameter that defines the minimum distance between dissimilar pairs.

Our final objective loss of AA in metric learning, such as triplet loss (TL) or contrastive loss (CL), is defined as follows:

$$\begin{aligned} Loss &= Loss_{label} + Loss_{AA+TL/CL} \\ Loss_{label} &= \frac{1}{n} \|f_a(a_i) - Y(a_i)\|_F + \frac{1}{n} \|f_v(v_i) - Y(v_i)\|_F \end{aligned} \quad (10)$$

where $\|\cdot\|_F$ signifies the Frobenius norm, $f(x)$ represents the projected feature in the shared label space and $Y(\cdot)$ denotes the label representations. In the end, the optimization of the final objective loss is carried out using the stochastic gradient descent (SGD) algorithm.

5 EXPERIMENT

In this section, we conducted experiments to evaluate our proposed AADML model in AV-CMR tasks, we compared our method with the existing state-of-the-art methods. In order to identify whether our method can be useful for the improvement of exiting deep metric learning methods, we combine our method with them to further show the power of our method. We also perform some ablation studies to analyze the effectiveness of our method.

5.1 Dataset and Metrics

Our model achieved success in the AV-CMR task based on the assumption that audio and visual modalities share identical semantic information. As a result, we select video datasets containing audio-visual tracks and ensure that both tracks are labeled identically on the time series. We select two special video datasets: VEGAS [47] and AVE [29]. Both datasets underwent a thorough process of label double-checking, ensuring uniform labeling across all frames in both modalities. The VEGAS dataset consists of 28,103 videos annotated with 10 labels, while the AVE dataset includes 1,955 videos annotated with 15 labels. We adhere to the same data partitioning strategy for training and testing sets, as well as the identical approach for feature extraction as described in the referenced work [41].

For model evaluation metrics, we adopt the same measures as those works [38, 39]: mean average precision (MAP) and Precision-scope@K. The MAP functions as an evaluative metric to gauge the effectiveness of models in AV-CMR. This entails calculating the average precision (AP) for each individual query and subsequently deriving the mean value from these AP scores, providing a comprehensive measure of overall performance. The Precision-scope@K quantifies the proportion of relevant items retrieved out of the total relevant items, up to a specified rank "K" in the ranked list. The model performs better as the values of both metrics increase.

5.2 Implementation Details

In the subnetwork of our model, we develop three fully connected (FC) layers and one predicted layer as a decoder for both audio

Models	VEGAS Dataset			AVE Dataset		
	A→V	V→A	Avg.	A→V	V→A	Avg.
Random case	0.110	0.109	0.109	0.127	0.124	0.126
CCA [4]	0.332	0.327	0.330	0.190	0.189	0.190
KCCA [19]	0.288	0.273	0.281	0.133	0.135	0.134
DCCA [1]	0.478	0.457	0.468	0.221	0.223	0.222
C-CCA [26]	0.711	0.704	0.708	0.228	0.226	0.227
C-DCCA [36]	0.722	0.716	0.719	0.230	0.227	0.229
TNN-C-CCA [41]	0.751	0.738	0.745	0.253	0.258	0.256
VAE-CCA [43]	0.821	0.824	0.822	0.328	0.302	0.315
AGAH [8]	0.578	0.568	0.573	0.200	0.196	0.198
ACMR [32]	0.465	0.442	0.454	0.162	0.159	0.161
DSCMR [44]	0.732	0.721	0.727	0.314	0.256	0.285
CLIP [25]	0.473	0.617	0.545	0.129	0.161	0.145
BiC-Net [10]	0.680	0.653	0.667	0.188	0.187	0.188
DCIL [46]	0.726	0.722	0.724	0.244	0.213	0.228
CCTL [38]	0.766	0.765	0.766	0.328	0.267	0.298
VideoAdviser [33]	0.825	0.819	0.822	-	-	-
EICS [39]	0.797	0.779	0.788	0.337	0.279	0.308
TLCA [37]	0.822	0.838	0.830	<u>0.410</u>	<u>0.451</u>	<u>0.431</u>
MSNSCA [42]	<u>0.866</u>	<u>0.865</u>	<u>0.866</u>	0.323	0.343	0.333
AADML (Ours)	0.901	0.891	0.896	0.890	0.883	0.887

Table 1: Model comparison: MAP scores of our approach versus state-of-the-art methods. The best MAP scores are presented in bold, and the second-best MAP scores are indicated with underlining.

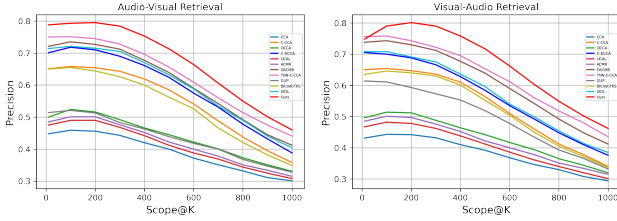


Figure 3: Precision-scope@K curves on the VEGAS dataset for audio \rightarrow visual and visual \rightarrow audio retrieval experiments, spanning different values of K from 10 to 1000.

and visual inputs, each fully connected layer was configured with 1024 hidden units, and set dropout as 0.1. To project audio and visual features into a shared label space, we ensure that the feature dimension of the projected features in the label space matches the number of labels. Specifically, we maintain both the predicted and pre-defined categories at 10 for VEGAS and 15 for AVE. We achieved optimal results with batch sizes of 400 and 200, respectively, and conducted training over 400 epochs. For the margin in triplet loss and metric learning, the best results were obtained with values of 1.2 and 1.0, respectively.

We implement all the experiments of our methods with the machine learning framework PyTorch³, and based on Ubuntu Linux 22.04.2 with NVIDIA GeForce 3080 (10G). We utilize the Adam optimizer [16] with its default parameter configuration to train our model, while the learning rate is designated as 0.0001.

³<https://pytorch.org/>

5.3 Comparison with Existing AV-CMR Approaches

For the purpose of evaluating the efficacy of our proposed method, a comprehensive comparison is conducted against a repertoire of 17 distinct algorithms. This ensemble encompasses seven methods grounded in Canonical Correlation Analysis (CCA), as well as eleven leading-edge techniques rooted in deep learning for cross-modal models. CCA [4] is employed to derive linear transformations for two data sets to optimize their correlation. KCCA [19] and DCCA [1] try to learn non-linear transformations by using the kernel method and deep learning, respectively. C-CCA [26] and C-DCCA [36] aim to discover transformations for both modalities by respectively clustering cross-modal data points into classes to enhance intra-cluster correlation. TNN-C-CCA [41] and VAE-CCA [43] improve C-CCA to enhance the correlation of C-CCA by triplet loss and VAE methods. ACMR [32], and AGAH [8] utilize adversarial learning to enhance cross-modal representation discrimination. DSCMR [44] and EICS [39] achieve discriminative features by introducing representation and label spaces. CLIP [25] and VideoAdviser [33] learn transferable visual models from natural language supervision by using a simple pre-training task. BiC-Net [10] and MSNSCA [42] utilize transformer architecture to effectively bridge dual modalities. DCIL [46] introduces an instance loss to improve rank loss by finding appropriate triplets. Conversely, the CCTL [38] considers all the triplets across two modalities. TLCA [37] employs a progressive training approach, prioritizing easy-to-hard triplet learning over a single-pass model training strategy.

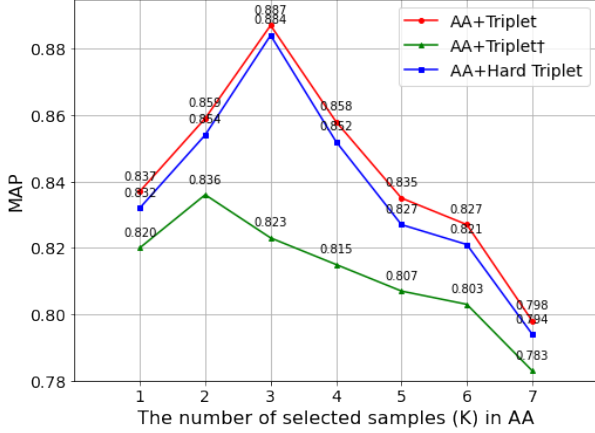


Figure 4: MAP trends on AVE dataset: AA proxy combined with three distinct triplet methods, varying with sample selections (1 to 7) in AA.

The AV-CMR performances of our method and comparison approaches on two audio-visual datasets are shown in Table 1. From the results, we can observe that our proposed AADML model can achieve the best results on both datasets all over the MAP metrics, with gains of 3.50%, 2.6%, and 3.0% on VEGAS, 48.0%, 43.2%, and 45.6% on AVE in terms of A2V, V2A, and Average, which indicate the effectiveness of our proposed model. Fig. 3 illustrates the variation in average precision across scope@K, where K ranges from 10 to 1000. The curve demonstrates the superior performance of our method across all ranges. This consistency aligns with the MAP comparison presented in Table 1 when employing audio and visual as queries.

5.4 Ablation Study

5.4.1 Impact of the number of selected samples on AA. The number of selected samples in AA serves as the sole hyper-parameter in our proposed method. Fig. 4 presents an experimental analysis of the effect of varying the number of selected samples on AA with three different triplet selection strategies. This exploration aims to prove the significance of this hyperparameter.

For AA combined with either *triplet* loss or *hard triplet* loss, the system’s performance displays a consistent upward track until approximately 3 samples are selected. Beyond this point, a decrement in performance is observed. In contrast, in the scenario of AA combined with *triplet†* loss, the performance attains its peak when 2 samples are chosen, following which a decline sets in. These trends result from the interaction between selected samples and AA methodology. In conclusion, the most optimal outcome achieved by AADML occurs through the combination of AA and Triplet techniques, particularly when the AA approach involves the selection of three samples.

5.4.2 Impact of Triplet Selection Strategy. The triplet selection strategy directly impacts model performance by determining the quality and diversity of training triplets. Well-chosen triplets facilitate effective learning of data distribution and improve the model’s ability to

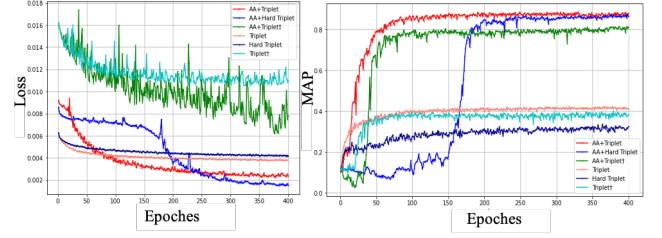


Figure 5: Loss value and MAP performance of training and test set from AVE dataset. Comparative analysis of triplet loss variants: exploring Triplet†, Triplet, and Hard Triplet losses with and without AA.

Models	VEGAS Dataset		
	A→V	V→A	Avg.
Triplet	0.750	0.718	0.734
Triplet†	0.766	0.765	0.766
Hard Triplet	0.761	0.747	0.754
DAML (Triplet)	0.792	0.783	0.788
HDML (Triplet)	0.779	0.784	0.781
AA+Triplet†	0.867	0.856	0.862
AA+Hard Triplet	0.865	0.851	0.858
AA+Triplet (ours)	0.901	0.890	0.896
Contrastive	0.701	0.702	0.704
DSML (Contrastive)	0.745	0.736	0.41
SupCon (Contrastive)	0.732	0.727	0.730
CrossCLR (Contrastive)	0.683	0.675	0.679
AA+Contrastive	0.783	0.807	0.795
N-pair	0.563	0.548	0.555
DSML (N-pair)	0.606	0.578	0.592
SupCon (N-pair)	0.621	0.589	0.605
AA+N-pair	0.647	0.612	0.629
Angular	0.581	0.533	0.557
AA+Angular	0.613	0.549	0.581
Hinge	0.704	0.698	0.702
AA+Hinge	0.761	0.764	0.762
DSL	0.694	0.692	0.693
AA+DSL	0.758	0.752	0.755

Table 2: VEGAS Dataset: The impact of anchor-aware proxy on different metric learning methods. the best MAP scores are presented in bold.

discriminate between categories. In this subsection, we explore the influence of different triplet selection strategies on model training and evaluation using the AVE dataset.

The first strategy involves selecting all possible triplets within a batch, where a triplet consists of an anchor (1st modality), a positive (2nd modality), and a negative (2nd modality), denoted as *Triplet*. The second strategy revolves around choosing a single triplet from each batch, wherein the triplet comprises an anchor (1st modality), a positive (2nd modality), and the hardest negative (2nd modality), designated as *Hard Triplet*. The third strategy derived from this work [38] entails selecting all available triplets within a batch,

Models	AVE Dataset		
	A→V	V→A	Avg.
Triplet	0.404	0.431	0.418
Triplet†	0.333	0.351	0.342
Hard Triplet	0.259	0.355	0.307
DAML (Triplet)	0.370	0.379	0.375
HDML (Triplet)	0.339	0.293	0.316
AA+Triplet†	0.839	0.833	0.836
AA+ Hard Triplet	0.883	0.885	0.884
AA+Triplet (ours)	0.890	0.883	0.887
Contrastive	0.371	0.375	0.373
DSML (Contrastive)	0.482	0.463	0.473
SupCon (Contrastive)	0.478	0.455	0.467
CrossCLR (Contrastive)	0.402	0.376	0.389
AA+Contrastive	0.798	0.854	0.826
N-pair	0.321	0.335	0.328
DSML (N-pair)	0.467	0.448	0.458
SupCon (N-pair)	0.474	0.443	0.459
AA+N-pair	0.828	0.833	0.830
Angular	0.326	0.333	0.330
AA+Angular	0.833	0.838	0.836
Hinge	0.358	0.381	0.369
AA+Hinge	0.786	0.822	0.804
DSL	0.371	0.369	0.370
AA+DSL	0.731	0.753	0.742

Table 3: AVE Dataset: The impact of anchor-aware proxy on different metric learning methods. the best MAP scores are presented in bold.

where a triplet is composed of an anchor (1st or 2nd modality), a positive (1st or 2nd modality), and a negative (1st or 2nd modality), denoted as *Triplet†*.

To investigate the impact of the triplet selection strategy with AA, we conduct experiments involving these three types of triplets with and without the integration of AA, as illustrated in Fig. 5. It is evident that incorporating AA significantly enhances the performance of the triplets compared to scenarios without AA overall. Remarkably, the AA+Triplet configuration achieves exceptional results during both the training and testing phases. However, the AA+Hard Triplet strategy exhibits suboptimal performance prior to approximately 150 epochs. Interestingly, before this threshold, the performance of AA+Hard Triplet even lags behind that of the baseline without AA. Notably, this phenomenon is accompanied by considerable fluctuations in the model’s loss value, possibly due to the varying detection of similar pairs within different batches. The performance begins to exhibit improvement around the epoch before 150, indicating the model’s convergence on the training set.

5.4.3 Impact of AA Proxy on Metric Learning. In this subsection, we leverage the assessment of the AA Proxy’s impact on existing metric learning techniques, seen in Table 2 and Table 3. We integrate our AA proxy with existing novel deep metric learning methods including contrastive loss and triplet loss, and compare them against the state-of-the-art advanced methods designed to enhance these deep metric learning techniques. This comparative

analysis aims to determine whether the combination of our AA Proxy with deep metric learning techniques can yield improvements in the performance of AV-CMR tasks.

The experimental results notably demonstrate the substantial potential of our AA proxy in enhancing the capabilities of deep metric learning methods for AV-CMR. Furthermore, our approach achieves a significant performance advantage over DSML [6], HDML [45], SupCon [14], and CrossCLR [48] methods when combined with deep metric learning techniques.

To highlight the efficacy of AA as a complementary enhancement for established deep metric learning methodologies, we extend the application of AA proxies to other advanced metric learning methods, including N-pair loss [27], Angular loss [27], Hinge loss [2], and DSL loss [3]. The results conclusively reveal that employing AA proxies in conjunction with these advanced deep metric learning methods consistently yields superior results compared to scenarios where AA is not applied.

6 CONCLUSION

In this work, we address the challenges of metric learning, which seeks to narrow the gap between similar pairs and enhance the separation of dissimilar pairs for the AV-CMR task. While recent methods show promise by selecting impactful data points during training, the limited number of training samples restricts the full exploration of the embedding space, resulting in an incomplete representation of data distributions. To overcome this, we propose an innovative Anchor-aware Deep Metric Learning approach that adeptly navigates the embedding space even with limited data. Our method simultaneously calculates attention-driven dependencies, considering each sample as an anchor alongside its semantically similar samples. This dynamic correlation weighing within the underlying manifold structure yields Anchor-Aware (AA) scores. Leveraging the parallel computation of AA scores, we employ them as proxies to compute relative distances within the metric learning framework. Extensive experiments on benchmark datasets affirm the effectiveness of AADML, outperforming state-of-the-art models. Furthermore, our exploration of combining AA proxies with various metric learning methods underscores the potency of our approach in advancing the field. In the future, we aim to leverage our methodology for cross-modal retrieval across various modalities and to further expand its applicability into the realm of unsupervised learning.

REFERENCES

- [1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep Canonical Correlation Analysis. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. Vol.28)*. PMLR, Atlanta, Georgia, USA, pp.1247–1255.
- [2] Christian Bailer, Kiran Varanasi, and Didier Stricker. 2017. CNN-based patch matching for optical flow with thresholded hinge embedding loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3250–3259.
- [3] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. 2021. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290* (2021).
- [4] Hardoon David, R., Szedmák Sándor, and Shawe-Taylor John. 2004. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*. Vol.16, no.12 (2004), pp.2639–2664. <https://doi.org/10.1162/0899766042321814>
- [5] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems* 50, 1 (2000), 1–18.

- [6] Yueqi Duan, Wenzhao Zheng, Xudong Lin, Jiwen Lu, and Jie Zhou. 2018. Deep adversarial metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2780–2789.
- [7] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khruikov, Nicu Sebe, and Ivan Oseledets. 2022. Hyperbolic vision transformers: Combining improvements in metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7409–7419.
- [8] Wen Gu, Xiaoyan Gu, Jingzi Gu, Bo Li, Zhi Xiong, and Weiping Wang. 2019. Adversary Guided Asymmetric Hashing for Cross-Modal Retrieval. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. ACM, New York, NY, USA, pp.159–167. <https://doi.org/10.1145/3323873.3325045>
- [9] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.
- [10] Ning Han, Jingjing Chen, Guangyi Xiao, Yawen Zeng, Chuhan Shi, and Hao Chen. 2021. Visual Spatio-temporal Relation-enhanced Network for Cross-modal Text-Video Retrieval. <https://arxiv.org/abs/2110.15609>
- [11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).
- [12] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*. Springer, 84–92.
- [13] Mahmut Kaya and Hasan Şakir Bilge. 2019. Deep metric learning: A survey. *Symmetry* 11, 9 (2019), 1066.
- [14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.
- [15] Sungyeon Kim, Boseung Jeong, and Suha Kwak. 2023. HIER: Metric Learning Beyond Class Labels via Hierarchical Regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19903–19912.
- [16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [17] Byungsoo Ko and Geonmo Gu. 2020. Embedding expansion: Augmentation in embedding space for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7255–7264.
- [18] Brian Kulis et al. 2013. Metric learning: A survey. *Foundations and Trends® in Machine Learning* 5, 4 (2013), 287–364.
- [19] Pei Ling Lai and Colin Fyfe. 2000. Kernel and Nonlinear Canonical Correlation Analysis. *Int. J. Neural Syst.* Vol.10, no.5 (2000), pp.365–377. <https://doi.org/10.1142/S012906570000034X>
- [20] Jung-Eun Lee, Rong Jin, and Anil K Jain. 2008. Rank-based distance metric learning: An application to image retrieval. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.
- [21] Pandeng Li, Yan Li, Hongtao Xie, and Lei Zhang. 2022. Neighborhood-adaptive structure augmented metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1367–1375.
- [22] Lizhao Liu, Shangxin Huang, Zhuangwei Zhuang, Ran Yang, Minghui Tan, and Yaowei Wang. 2022. Das: Densely-anchored sampling for deep metric learning. In *European Conference on Computer Vision*. Springer, 399–417.
- [23] David G Lowe. 1995. Similarity metric learning for a variable-kernel classifier. *Neural computation* 7, 1 (1995), 72–85.
- [24] Alexis Mignon and Frédéric Jurie. 2012. CMML: A new metric learning approach for cross modal matching. In *Asian Conference on Computer Vision*. 14–pages.
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. Vol.139)*. PMLR, Virtual Event, pp.8748–8763.
- [26] Nikhil Rasiwasia, Dhruv Mahajan, Vijay Mahadevan, and Gaurav Aggarwal. 2014. Cluster Canonical Correlation Analysis. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. JMLR.org, Reykjavik, Iceland, pp.823–831. <https://doi.org/10.1201/b18358-8>
- [27] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems* 29 (2016).
- [28] Didac Suris, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i Nieto. 2018. Cross-modal embeddings for video and audio retrieval. In *Proceedings of the european conference on computer vision (eccv) workshops*. 0–0.
- [29] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-Visual Event Localization in Unconstrained Videos. In *Computer Vision - ECCV 2018 - 15th European Conference (Lecture Notes in Computer Science, Vol. Vol.11206)*. Springer, Munich, Germany, pp.252–268. https://doi.org/10.1007/978-3-030-01216-8_16
- [30] Shashanka Venkataramanan, Bill Psomas, Ewa Kijak, Laurent Amsaleg, Konstantinos Karantzas, and Yannis Avrithis. 2021. It takes two to tango: Mixup for deep metric learning. *arXiv preprint arXiv:2106.04990* (2021).
- [31] Lucas Vinh Tran, Yi Tay, Shuai Zhang, Gao Cong, and Xiaoli Li. 2020. Hyperml: A boosting metric learning approach in hyperbolic space for recommender systems. In *Proceedings of the 13th international conference on web search and data mining*. 609–617.
- [32] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial Cross-Modal Retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia (Mountain View, California, USA) (MM '17)*. ACM, New York, NY, USA, pp.154–162. <https://doi.org/10.1145/3123266.3123326>
- [33] Yanan Wang, Donghuo Zeng, Shinya Wada, and Satoshi Kurihara. 2023. VideoAdviser: Video Knowledge Distillation for Multimodal Transfer Learning. *IEEE Access* (2023).
- [34] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. 2005. Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems* 18 (2005).
- [35] Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. 2002. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems* 15 (2002).
- [36] Yi Yu, Suhua Tang, Kiyoharu Aizawa, and Akiko Aizawa. 2019. Category-based deep CCA for fine-grained venue discovery from multimodal data. *IEEE transactions on neural networks and learning systems* Vol.30, no.4 (2019), pp.1250–1258.
- [37] Donghuo Zeng and Kazushi Ikeda. 2023. Two-Stage Triplet Loss Training with Curriculum Augmentation for Audio-Visual Retrieval. *arXiv preprint arXiv:2310.13451* (2023).
- [38] Donghuo Zeng, Yanan Wang, Jianming Wu, and Kazushi Ikeda. 2022. Complete Cross-triplet Loss in Label Space for Audio-visual Cross-modal Retrieval. In *2022 IEEE International Symposium on Multimedia (ISM)*. IEEE, 1–9.
- [39] Donghuo Zeng, Jianming Wu, Gen Hattori, Rong Xu, and Yi Yu. 2023. Learning Explicit and Implicit Dual Common Subspaces for Audio-visual Cross-modal Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* vol.19, no.2s (2023), pp.1–23.
- [40] Donghuo Zeng, Yi Yu, and Keizo Oyama. 2018. Audio-visual embedding for cross-modal music video retrieval through supervised deep CCA. In *2018 IEEE International Symposium on Multimedia (ISM)*. IEEE, 143–150.
- [41] Donghuo Zeng, Yi Yu, and Keizo Oyama. 2020. Deep Triplet Neural Networks with Cluster-CCA for Audio-Visual Cross-Modal Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* vol.16, no.3 (2020), pp.1–23. <https://doi.org/10.1145/3387164>
- [42] Jiwei Zhang, Yi Yu, Suhua Tang, Wei Li, and Jianming Wu. 2023. Multi-scale network with shared cross-attention for audio-visual correlation learning. *Neural Computing and Applications* (2023), 1–15.
- [43] Jiwei Zhang, Yi Yu, Suhua Tang, Jianming Wu, and Wei Li. 2023. Variational Autoencoder with CCA for Audio-Visual Cross-Modal Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 3s, Article 130 (feb 2023), 21 pages. <https://doi.org/10.1145/3575658>
- [44] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep Supervised Cross-Modal Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, CA, USA, pp.10394–10403. <https://doi.org/10.1109/CVPR.2019.01064>
- [45] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. 2019. Hardness-aware deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 72–81.
- [46] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications* Vol.16, no.2 (2020), pp.1–23.
- [47] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. 2018. Visual to Sound: Generating Natural Sound for Videos in the Wild. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, Utah, USA, pp.3550–3558. <https://doi.org/10.1109/CVPR.2018.00374>
- [48] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. 2021. Cross-clr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1450–1459.