



Audio-text Retrieval with Transformer-based Hierarchical Alignment and Disentangled Cross-modal Representation

Yifei Xin, Zhihong Zhu, Xuxin Cheng, Xusheng Yang, Yuexian Zou*

School of ECE, Peking University, Shenzhen, China

xinyifei@stu.pku.edu.cn

Abstract

Most existing audio-text retrieval (ATR) approaches typically rely on a single-level interaction to associate audio and text, limiting their ability to align different modalities and leading to suboptimal matches. To tackle this issue, we present a novel ATR framework that leverages two-stream Transformers in conjunction with a Hierarchical Alignment (THA) module to identify multi-level correspondences of different Transformer blocks between audio and text. Moreover, current ATR methods mainly focus on learning a global-level representation, missing out on intricate details to capture audio occurrences that correspond to textual semantics. To bridge this gap, we introduce a Disentangled Cross-modal Representation (DCR) approach that disentangles high-dimensional features into compact latent factors to grasp fine-grained audio-text semantic correlations. Additionally, we develop a confidence-aware (CA) module to estimate the confidence of each latent factor pair and adaptively aggregate cross-modal latent factors to achieve local semantic alignment. Experiments show that our THA effectively boosts ATR performance, with the DCR approach further contributing to consistent performance gains.

Index Terms: audio-text retrieval, hierarchical alignment, disentangled representation learning

1. Introduction

The audio-text retrieval (ATR) task is to retrieve items from one modality when provided with a caption or audio clip from another. To align audio and text modalities, deep neural networks [1–3] have been leveraged to extract audio and text representations, which are subsequently mapped to a joint latent space to measure the relevance of audio-text pairs. However, many existing ATR methods adopt two-stream encoders with different architectures: for instance, CNNs for audio and Transformers for text [4, 5]. Such architectural disparities may lead to varied semantic distribution spaces, potentially weakening the audio-text alignment. Moreover, most existing ATR strategies predominantly emphasize single-level alignment across two independent streams [6, 7], neglecting multi-level correspondences of different modalities. It's notable that semantics across different modalities are multifaceted. Incorporating information from another modality can lead to better semantic representations.

To fill the above research gaps, we present a novel ATR framework based on dual-stream Transformers, composed of an audio Transformer, a text Transformer, and a Transformer-based hierarchical alignment (THA) module. With such unified

architecture, the dual-stream encoders could yield features with more similar characteristics for audio and text, facilitating more straightforward audio-text interactions and alignments. As prior studies have pointed out that shallow blocks in Transformers tend to capture more static and local semantics while deeper blocks encompass more contextual information [8, 9], our THA module enables the ATR network to learn to grasp the rich correspondences between audio and text across different semantic levels.

Besides, current ATR methods mainly concentrate on establishing a global alignment across modalities [10, 11], which treats discriminative information equally and curtails the potential to capture fine-grained audio-text correlations. A more refined strategy would involve aligning text and audio based on local semantic clues, emulating human cognition and reasoning process. In pursuit of extracting salient cross-modal information, we employ disentangled representation learning [12–14] to separately encode semantic information about each latent variable in only a few dimensions.

Building upon the above understanding, we introduce a disentangled cross-modal representation (DCR) approach for ATR, which disentangles high-dimensional audio and text representations into compact latent factors that purposefully capture fine-grained audio-text semantic relationships. Our optimization of latent factors is twofold: inter-factor and intra-factor. For the former, we minimize mutual information among different latent factors to identify representation subspaces with minimal relevance to each other for disentangling representation. For the latter, we maximize mutual information for each latent factor pair separately, facilitating local alignment between corresponding audio and text pair. In order to further enhance the semantic matching of cross-modal information contained in audio-text pairs, we introduce a confidence-aware (CA) module to estimate the confidence of each cross-modal latent factor pair. Lastly, we utilize the confidence as the weight to adaptively aggregate all factor pairs to compute the text-audio similarity.

The primary contributions of this paper include:

- We present an ATR framework that leverages two-stream Transformers combined with a Hierarchical Alignment (THA) module which comprehensively explores multi-level correspondences of different Transformer blocks between audio and text.
- We introduce a disentangled cross-modal representation (DCR) approach integrated with a confidence-aware (CA) module for ATR, which disentangles high-dimensional audio and text features into compact latent factors to capture fine-grained audio-text relationships and achieve local semantic alignment.
- Experiments demonstrate that our THA significantly im-

This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No:GXWD20201231165807007-20200814115301001).

* Yuexian Zou is the corresponding author.

proves the ATR performance. Moreover, our DCR also consistently boosts the performance to a great extent.

2. Proposed Methods

2.1. Dual-stream Transformers

Text Transformer. Given a text sentence T comprising N words, we leverage BERT [15] to extract word-level representations $T^l = \{t_n^l \mid t_1^l, t_2^l, \dots, t_N^l\}$, where l corresponds to the l -th Transformer block of BERT. In our approach, we utilize BERT-base with 12 Transformer [16–18] blocks.

Audio Transformer. Instead of the common practice of leveraging CNN-based architectures like ResNet38 [2, 19] for audio feature extraction, we utilize HTSAT [20] as the audio encoder, employing Swin Transformer [21] as the backbone, which demonstrates great potential and surpasses CNN models in various audio-related tasks [20, 22, 23]. Given an audio spectrogram $A \in \mathbb{R}^{H \times W \times 1}$, we first split the audio spectrogram into several non-overlapping patches, each considered a token to feed into the Transformer blocks. We take the output features of multiple stages as hierarchical audio representations to convey richer semantics, and represent the output frame-level features of each stage as A^i , where i denotes the i -th stage. Due to the computational intensity of stage 1’s large token numbers, we discard the output feature of stage 1. Outputs from stages 2-4 are labeled as low-, middle-, and high-level semantic audio representations. Thus, the overall audio hierarchical representation A_h is denoted as: $A_h = \text{HTSAT}(A) = \{A^{s2}, A^{s3}, A^{s4}\}$. To be consistent with BERT’s block representation, we map stages in A^{s2}, A^{s3} , and A^{s4} to their corresponding blocks, i.e., 4, 10, 12-th block in BERT, given that HTSAT includes 4 stages with 2, 2, 6, 2 swin-transformer blocks.

2.2. Hierarchical Alignment Module

After obtaining the frame-level and word-level representations, our Transformer-based hierarchical alignment (THA) integrates the three-level semantics to learn correlations between audio and text. As illustrated in Figure 1, we calculate the cosine similarities for each level and combine the hierarchical similarities by addition. Specifically, we implement text-enhanced and audio-enhanced stacked cross attention [24, 25] respectively. Supposing the output of the l -th block of Swin Transformer $A^l = \{a_m^l \mid a_1^l, a_2^l, \dots, a_M^l\}$ with M audio patch tokens, and the output of the l -th block of BERT $T^l = \{t_n^l \mid t_1^l, t_2^l, \dots, t_N^l\}$ with N words, the text-enhanced stacked cross attention first calculates the similarities between all token-word pairs: $s_{mn}^l = \frac{a_m^{lT} t_n^l}{\|a_m^l\| \|t_n^l\|}$, where s_{mn}^l denotes the similarity between m -th token and n -th word in the l -th block of both HTSAT and BERT. Based on the token-word similarity, we aggregate the weighted words representations for each audio token to obtain the cross-modal fused text representations:

$$\hat{a}_m^l = \sum_{n=1}^N \alpha_{mn}^l t_n^l, \quad \alpha_{mn}^l = \frac{\exp(\lambda * s_{mn}^l)}{\sum_{n=1}^N \exp(\lambda * \bar{s}_{mn}^l)}, \quad (1)$$

where λ denotes the temperature hyper-parameter. \bar{s}_{mn}^l is normalized similarity: $\bar{s}_{mn}^l = \frac{[s_{mn}^l]_+}{\sqrt{\sum_{m=1}^M [s_{mn}^l]_+^2}}$, where $[x]_+ = \max(x, 0)$ denotes the hinge function to keep the similarity with positive value. Lastly, we compute the cosine similarity for each token-word pair (a_m^l, \hat{a}_m^l) , and aggregate all the pairs

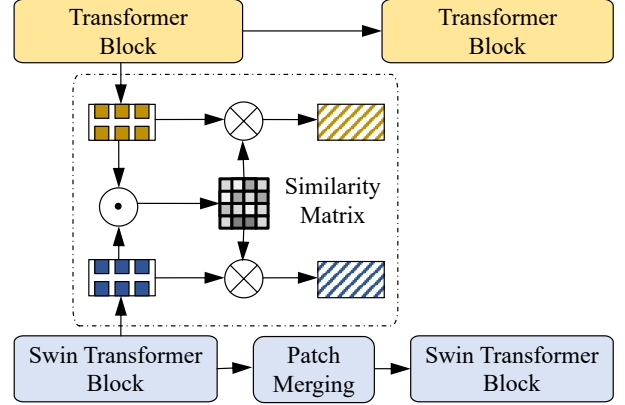


Figure 1: Illustration of our Transformer-based hierarchical alignment module. We take one-stage interaction as an example.

for the overall similarity: $S_{THA}(A, T) = \sum_l \sum_m \frac{a_m^{lT} \hat{a}_m^l}{\|a_m^l\| \|\hat{a}_m^l\|}$. Likewise, the audio-enhanced stacked cross attention follows similar patterns.

After passing through all blocks of BERT and HTSAT, we also follow previous works [2, 6] that take the [CLS] token output as the global text feature \bar{t} , and leverages an average pooling layer to aggregate frame-level embeddings into a global representation \bar{a} , which is followed by a dot-product operation between the two holistic representations to measure feature similarity: $S_{DP}(\bar{a}, \bar{t}) = \frac{\bar{a} \cdot \bar{t}}{\|\bar{a}\|_2 \|\bar{t}\|_2}$. Next, we will introduce a disentangled cross-modal representation (DCR) module to further improve the discriminative power of our ATR network for fine-grained cross-modal matching.

2.3. Disentangled Cross-modal Representation from both Inter-factor and Intra-factor Perspectives

Currently, most ATR methods [10, 11] yield holistic text representations and audio representations for similarity calculation. However, such global representations characterize semantic information of the input text and audio in an entangled state. As a result, direct similarity matching of these representations cannot capture fine-grained cross-modal correspondences between a set of diverse phrases in textual descriptions and a set of various audio events within audio clips. To address this, we learn discriminative disentangled latent factors from text and audio representations to explicitly measure local cross-modal correlations.

To initiate, we simplify by considering the sentence-level representation $\bar{t} \in \mathbb{R}^D$. According to the setting of disentangled representation learning [26, 27], we assume that the sentence-level representation can be decomposed into K independent latent factors, i.e., $E^t = [e_1^t, e_2^t, \dots, e_K^t]$. Each latent factor $e_k^t \in \mathbb{R}^{\frac{D}{K}}$ symbolizes a distinct textual semantic element. The independence of the latent factors ensures little semantic overlap among them. In this case, we project the text feature \bar{t} into K factors separately, and get the k -th latent factor e_k^t : $e_k^t = W_k^t \bar{t}$, where $W_k^t \in \mathbb{R}^{\frac{D}{K} \times D}$ denotes a trainable parameter. Similarly, the latent factor e_k^a of audio representation is calculated as: $e_k^a = W_k^a \bar{a}$. By combining the inter-factor decoupling and intra-factor alignment approaches introduced below, we explicitly map features to distinct representational subspaces corresponding to different semantic elements, empowering the model

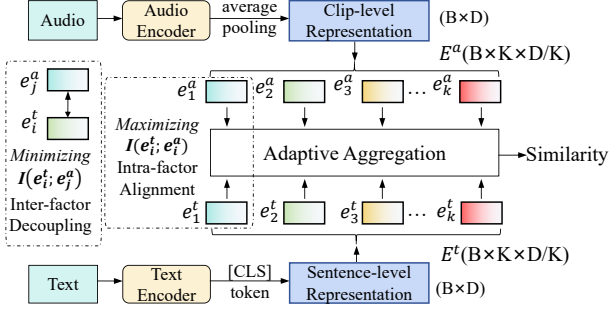


Figure 2: Overview of the disentangled cross-modal representation (DCR) framework.

to optimize and reason cross-modal semantic information separately from these representation subspaces.

Inter-factor Decoupling. To identify representation subspaces with minimal relevance to each other and thus enhance the discriminative power for cross-modal semantic matching, we minimize the inter-factor mutual information. Taking two latent factors e_i^t and e_j^a , their mutual information is captured via their probabilistic density functions as:

$$I(e_i^t; e_j^a) = \mathbb{E}_{t,a} \left[p(e_i^t, e_j^a) \log \frac{p(e_i^t, e_j^a)}{p(e_i^t) p(e_j^a)} \right]. \quad (2)$$

Given the difficulty of directly quantifying mutual information, we follow previous work [28, 29] to measure the mutual information implicitly. Specifically, for the latent factors $e^t \in \mathbb{R}^{\frac{D}{K}}$ and $e^a \in \mathbb{R}^{\frac{D}{K}}$, we first normalize them: $z^t = \frac{e^t - \mathbb{E}[e^t]}{\sqrt{\text{Var}[e^t]}}$, $z^a = \frac{e^a - \mathbb{E}[e^a]}{\sqrt{\text{Var}[e^a]}}$, where z^t, z^a have the same mean and standard deviation. Then, we compute the covariance of z^t and z^a : $C_{i,j} = \mathbb{E}_{t,a} \left[(z_i^t)^\top z_j^a \right]$, where z_i^t and z_j^a are the normalized features of e_i^t and e_j^a , respectively. As the meaning of $p(e_i, e_j)$ is the probability that e_i matches e_j , it can be expressed as:

$$p(e_i, e_j) = \frac{\mathbb{E} \left[(z_i)^\top z_j \right]}{\sum_{k=1}^K \mathbb{E} \left[(z_k)^\top z_j \right]} = \frac{C_{i,j}}{\sum_{k=1}^K C_{k,j}}. \quad (3)$$

In this case, minimizing $p(e_i, e_j)$ is equivalent to the minimization of $C_{i,j}$. Therefore, our approach minimizes $I(e_i^t; e_j^a)$ by minimizing the inter-factor decoupling loss \mathcal{L}_D defined as: $\mathcal{L}_D = \sum_i \sum_{j \neq i} (C_{i,j})^2$. This loss minimizes the mutual information for the negative pairs (e_i^t, e_j^a) , achieving the decoupling of latent factors.

Intra-factor Alignment. To effectively disentangle the K latent factors from both text and audio representations, it's crucial to discern cross-modal semantics within text-audio correspondences. In this case, we optimize the representation subspace for each latent factor pair independently. Central to our strategy is treating each representation subspace separately to powerfully capture the latent factors' correlation within text-audio pairs. In detail, we maximize the mutual information between the matching text and audio latent factors within the identical subspace. The intra-factor alignment loss \mathcal{L}_A is denoted as: $\mathcal{L}_A = \sum_i (1 - C_{i,i})^2$, which maximizes the mutual information of each positive pair (e_i^t, e_i^a) independently.

2.4. Adaptive Aggregation for Similarity Calculation

To further enhance cross-modal semantic alignment for audio-text latent factors, we first present a confidence-aware (CA) module to estimate the confidence for each audio-text latent factor pair. Concretely, for the i_{th} subspace, we concatenate the latent factor $e_i^t \in \mathbb{R}^{\frac{D}{K}}$ with the latent factor $e_i^a \in \mathbb{R}^{\frac{D}{K}}$ to form the input $\hat{e}_i = [e_i^t, e_i^a] \in \mathbb{R}^{\frac{2D}{K}}$ for our CA module. The confidence score for the i_{th} subspace is then derived using the function: $g_i = \text{MLP}(\hat{e}_i)$, where MLP consists of two linear layers and a ReLU activation function. A smaller value of g_i implies that the semantic element is less likely to match the i_{th} subspace. Leveraging these confidence scores as weights, we adaptively aggregate all factor pairs to compute the text-audio similarity, defined as:

$$S_{DCR} = \sum_{i=1}^K g_i \frac{(e_i^t)^\top e_i^a}{\|e_i^t\| \|e_i^a\|}. \quad (4)$$

Following common practice, we adopt a contrastive learning framework [2, 30, 31], specifically utilizing the NT-Xent loss [32], to optimize the cross-modal similarity:

$$\mathcal{L}_S = -\frac{1}{B} \left(\sum_{i=1}^B \log \frac{\exp(S_{i,i}/\tau)}{\sum_{j=1}^B \exp(S_{i,j}/\tau)} + \sum_{i=1}^B \log \frac{\exp(S_{i,i}/\tau)}{\sum_{j=1}^B \exp(S_{j,i}/\tau)} \right), \quad (5)$$

where B is the batch size, τ serves as a temperature hyper-parameter, i and j indicate the sample index within a batch. Notably, the overall similarity S can be only S_{DCR} , or it can be a combination of multiple similarities (e.g., $S_{THA} + S_{DCR}$). Combining the aforementioned objectives for text-audio similarity \mathcal{L}_S , inter-factor decoupling \mathcal{L}_D and intra-factor alignment \mathcal{L}_A , the whole training loss is $\mathcal{L} = \mathcal{L}_S + \alpha \mathcal{L}_D + \beta \mathcal{L}_A$, where α and β are optimization hyper-parameters guiding the balance between individual loss components.

3. Experiments

3.1. Dataset and Experiment Details

We evaluate our methods on the AudioCaps [33] and Clotho [34] datasets. AudioCaps contains around 50,000 audio samples; its training set consists of 49,274 samples each with one caption, while the validation and evaluation sets contain 494 and 957 samples, each with five captions. Clotho v2 includes 3,839 training, 1,045 validation, and 1,045 testing samples, each associated with five descriptive sentences. In this study, we follow the same pipeline in [2] to train our network. We perform experiments by fine-tuning the audio and text encoders. We set $K = 8$, $\alpha = 0.01$, $\beta = 0.005$. We repeat experiments three times with different training seeds, presenting the mean and standard deviation of the metrics on the AudioCaps dataset. Recall at rank k ($R@k$) is utilized as the evaluation metric. $R@1$, $R@5$, and $R@10$ are presented in our results.

3.2. Experimental Results

As shown in Table 1 and Table 2, we compare our proposed methods with previous baselines on the AudioCaps and Clotho datasets. "DP (ResNet38)" denotes the existing mainstream ATR approach leveraging ResNet38 [19] as the audio encoder and employing the NT-Xent loss [32] that per-

Table 1: ATR performance on the AudioCaps dataset.

Methods	Text-to-Audio			Audio-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
DP (ResNet38) [2]	33.9±0.4	69.7±0.2	82.6±0.3	39.4±1.0	72.0±1.0	83.9±0.6
DP (HTSAT)	38.3±0.2	72.2±0.1	83.2±0.5	45.1±0.6	77.7±0.8	88.1±0.6
THA	39.9±0.3	73.6±0.7	84.9±0.1	46.8±0.4	79.1±0.2	89.4±0.5
THA+DP	40.1±0.4	73.9±0.2	85.2±0.3	47.1±0.1	79.4±0.4	89.6±0.4
DCR	40.6±0.3	74.1±0.8	84.3±0.5	46.6±0.2	79.3±1.0	89.9±0.9
THA+DCR	41.1±0.6	74.9±0.8	85.7±0.1	47.8±0.1	80.6±0.4	90.6±0.3

Table 2: ATR performance on the Clotho dataset.

Methods	Text-to-Audio			Audio-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
DP(HTSAT)	11.4	31.9	43.1	14.0	32.4	45.4
THA	13.2	33.4	44.9	15.1	33.7	46.6
THA+DP	13.7	33.9	45.2	15.4	33.8	46.7
DCR	14.1	34.2	45.5	15.9	34.2	46.9
THA+DCR	14.8	34.9	46.7	16.8	35.6	47.8

Table 3: Effect of each part of our DCR method.

Method	Text-to-Audio			Audio-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
DP(HTSAT)	38.3	72.2	83.2	45.1	77.7	88.1
+ \mathcal{L}_D	39.1	72.8	83.6	45.9	78.2	88.7
+ \mathcal{L}_A	39.4	73.3	83.9	46.2	78.6	89.1
+ CA	40.6	74.1	84.3	46.6	79.3	89.9

forms the dot product operation to compute the global similarity between audio-text pairs, while “DP (HTSAT)” indicates that we adopt HTSAT [20] as the audio encoder, based on the Swin Transformer backbone. “THA” represents that we only use our Transformer-based hierarchical alignment method, while “THA+DP” refers to an integration of our THA module with the prevalent dot product strategy to enhance feature similarity computation. “DCR” denotes that we only leverage our disentangled cross-modal representation method, while “THA+DCR” combines our THA and DCR approaches.

It is clear that our methods consistently achieve better performance over the baselines both text-to-audio and audio-to-text retrieval tasks. Particularly, our THA+DCR combination stands out, delivering the highest scores across all evaluation metrics. This underlines the efficacy of combining the hierarchical alignment with our representation disentanglement approaches. When applied individually, both THA and DCR manifest commendable performance, which not only demonstrates the potency of each method but also underscores their potential as standalone techniques. Moreover, the marginal improvements observed in THA+DP over THA suggest that integrating existing techniques like dot product operations can further refine our methods, indicating the compatibility of our approaches.

3.3. Ablation Study

In this part, we discuss the influences of each part of our DCR method, the number of latent factors, and the selection of hyper-parameters α and β on the AudioCaps dataset.

Each part of our DCR method. Table 3 studies the importance of each part of our DCR method. We find that compared with the intra-factor alignment (\mathcal{L}_A) of latent factors, the improvements of the inter-factor decoupling (\mathcal{L}_D) is more notable. This is likely ascribed to the efficacy of the inter-factor decoupling strategy in mitigating the interference of mismatched semantic elements. Additionally, the integration of our

Table 4: Effect of the number of latent factors.

Method	Text-to-Audio			Audio-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
6	40.3	73.7	84.0	46.4	78.9	89.5
8	40.6	74.1	84.3	46.6	79.3	89.9
10	40.2	73.8	83.8	46.3	78.9	89.8

Table 5: Effect of the trade-off hyper-parameters α and β .

Method	Text-to-Audio			Audio-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10
$\alpha = 0.005$	40.3	74.0	84.2	46.6	78.9	89.5
$\alpha = 0.01$	40.6	74.1	84.3	46.6	79.3	89.9
$\alpha = 0.02$	40.4	73.7	84.2	46.3	79.1	89.7
$\beta = 0.001$	40.2	73.9	83.9	46.2	79.3	89.6
$\beta = 0.005$	40.6	74.1	84.3	46.6	79.3	89.9
$\beta = 0.01$	40.5	74.1	84.1	46.4	79.2	89.8

confidence-aware (CA) module yields marked improvement, attributed to its adaptive mechanism in determining the text-audio similarity for local cross-modal matching. Finally, our full method obtains the best performance, thus strongly showing that the three components are all beneficial for aligning audio and text representations.

The number of latent factors. In Table 4, we show the results of different numbers of latent factors $E = [e_1, e_2, \dots, e_K]$, where $e_k \in \mathbb{R}^{\frac{D}{K}}$. We find that $K = 8$ achieves the optimal performance. A smaller set of latent factors potentially constrains the network’s ability to exploit fine-grained information. Conversely, an excessive number of latent factors diminishes the dimension of each latent factor e_k , thus compromising the discriminative ability of the factors.

The selection of hyper-parameters α and β . The hyper-parameters α and β indicate the importance of \mathcal{L}_D and \mathcal{L}_A respectively. As shown in Table 5, we find that the optimal performance is achieved when α is set to 0.01. This suggests a slight but notable improvement over the settings where α equals 0.005 and 0.02. A similar pattern is observed for the β parameter, where setting β to 0.005 yields the best results.

4. Conclusions

In this work, we present an ATR framework that leverages two-stream Transformers combined with a hierarchical alignment (THA) module to unify the architectures of audio and text encoders, and enhance multi-level correspondences between audio and text. Furthermore, we introduce a disentangled cross-modal representation (DCR) approach that disentangles high-dimensional features into compact latent factors and adaptively aggregates cross-modal latent factors to achieve fine-grained local semantic alignment. Experiments show that our THA effectively boosts the ATR performance, with the DCR approach further enhancing these improvements.

5. References

- [1] A.-M. Oncescu, A. Koepke, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries," *arXiv preprint arXiv:2105.02192*, 2021.
- [2] X. Mei, X. Liu, J. Sun, M. Plumbley, and W. Wang, "On Metric Learning for Audio-Text Cross-Modal Retrieval," in *Proc. Interspeech 2022*, 2022, pp. 4142–4146.
- [3] Y. Xin, B. Wang, and L. Shang, "Cooperative game modeling with weighted token-level alignment for audio-text retrieval," *IEEE Signal Processing Letters*, 2023.
- [4] S. Lou, X. Xu, M. Wu, and K. Yu, "Audio-text retrieval in context," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4793–4797.
- [5] Y. Xin and Y. Zou, "Improving audio-text retrieval via hierarchical cross-modal interaction and auxiliary captions," in *Proc. INTERSPEECH 2023*, 2023, pp. 341–345.
- [6] Y. Wu, T. Zhang, and K. Chen, "Text-to-audio retrieval via large-scale contrastive training," DCASE2022 Challenge, Tech. Rep., 2022.
- [7] Y. Xin, D. Yang, and Y. Zou, "Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [8] W. de Vries, A. van Cranenburgh, and M. Nissim, "What's so special about bert's layers? a closer look at the nlp pipeline in monolingual and multilingual models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4339–4350.
- [9] B. Van Aken, B. Winter, A. Löser, and F. A. Gers, "How does bert answer questions? a layer-wise analysis of transformer representations," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1823–1832.
- [10] X. Mei, X. Liu, H. Liu, J. Sun, M. D. Plumbley, and W. Wang, "Language-based audio retrieval with pre-trained models," Tech. rep., DCASE2022 Challenge, Tech. Rep., 2022.
- [11] S. Deshmukh, B. Elizalde, and H. Wang, "Audio Retrieval with WavText5K and CLAP Training," in *Proc. INTERSPEECH 2023*, 2023, pp. 2948–2952.
- [12] T. Wang, Z. Yue, J. Huang, Q. Sun, and H. Zhang, "Self-supervised learning disentangled group representation as feature," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 225–18 240, 2021.
- [13] F. Liu, H. Chen, Z. Cheng, A. Liu, L. Nie, and M. Kankanhalli, "Disentangled multimodal representation learning for recommendation," *IEEE Transactions on Multimedia*, 2022.
- [14] P. Jin, H. Li, Z. Cheng, J. Huang, Z. Wang, L. Yuan, C. Liu, and J. Chen, "Text-video retrieval with disentangled conceptualization and set-to-set alignment," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. International Joint Conferences on Artificial Intelligence Organization, 8 2023, pp. 938–946.
- [15] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacl-HLT*, 2019, pp. 4171–4186.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [17] X. Cheng, Z. Zhu, H. Li, Y. Li, X. Zhuang, and Y. Zou, "Towards multi-intent spoken language understanding via hierarchical attention and optimal transport," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 844–17 852.
- [18] X. Cheng, Q. Dong, F. Yue, T. Ko, M. Wang, and Y. Zou, "M 3 st: Mix at three levels for speech translation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [19] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [20] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.
- [21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.
- [22] Y. Xin, D. Yang, F. Cui, Y. Wang, and Y. Zou, "Improving weakly supervised sound event detection with causal intervention," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [23] Y. Xin, D. Yang, and Y. Zou, "Background-aware modeling for weakly supervised sound event detection," in *Proc. INTERSPEECH*, vol. 2023, 2023, pp. 1199–1203.
- [24] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 201–216.
- [25] Y. Bin, H. Li, Y. Xu, X. Xu, Y. Yang, and H. T. Shen, "Unifying two-stream encoders with transformers for cross-modal retrieval," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3041–3050.
- [26] C.-C. Yang, W.-C. Fan, C.-F. Yang, and Y.-C. F. Wang, "Cross-modal mutual learning for audio-visual speech recognition and manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3036–3044.
- [27] J. Ma, P. Cui, K. Kuang, X. Wang, and W. Zhu, "Disentangled graph convolutional networks," in *International conference on machine learning*. PMLR, 2019, pp. 4212–4221.
- [28] P. A. Sreekar, U. Tiwari, and A. Namboodiri, "Mutual information based method for unsupervised disentanglement of video representation," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 6396–6403.
- [29] K. Do and T. Tran, "Theory and evaluation metrics for learning disentangled representations," in *International Conference on Learning Representations*, 2020.
- [30] X. Cheng, B. Cao, Q. Ye, Z. Zhu, H. Li, and Y. Zou, "MI-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 6492–6505.
- [31] X. Cheng, W. Xu, Z. Zhu, H. Li, and Y. Zou, "Towards spoken language understanding via multi-level multi-grained contrastive learning," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 326–336.
- [32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [33] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.
- [34] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.