LISTEN AND LOOK: MULTI-MODAL AGGREGATION AND CO-ATTENTION NETWORK FOR VIDEO-AUDIO RETRIEVAL

Xiaoshuai Hao[1,2] *Wanqian Zhang*[2*] *Dayan Wu*[2] *Fei Zhu*[1,2] *Bo Li*[2]

[1] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, 100049, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100093, China
{haoxiaoshuai, zhangwanqian, wudayan, zhufei, libo}@iie.ac.cn;

ABSTRACT

Video is a natural source of multi-modal data with intrinsic correlations between different modalities, such as objects, motions and captions. Though intuitive, such inherent supervision has not been well explored in previous video-audio retrieval works. Besides, existing methods exploit the video stream and the audio stream seperately, whereas ignoring the mutual interactions between them. In this paper, we propose a two-stream model named Multi-modal Aggregation and Co-attention network (MAC), which processes video and audio inputs with co-attentional interactions. Specifically, our method takes raw videos as inputs and extracts aggregated features from multiple modalities to benefit the video representation learning. Then, we introduce the self-attention mechanism to make videos adaptively assign higher weights to the representative modalities. Moreover, we introduce a co-attention transformer module to better capture the relations among videos and audios. By exchanging key-value pairs in the multi-headed attention, this module enables video-attended audio features to be incorporated into video representations and vice versa. Experiments show that our method significantly outperform other state-of-the-arts.

Index Terms— video-audio retrieval, multi-modal aggregation, co-attention transformer

1. INTRODUCTION

With the rapid growth of user-generated multimedia data, cross-modal retrieval between videos and audios, known as video-audio retrieval, has attracted much attention [1, 2]. Contrastive learning, as the dominant paradigm for video-audio retrieval, has delivered impressive retrieval performances[3, 4]. It maps audio queries and the videos in database into a joint embedding space, where the semantically-similar audios and videos are much closer to each other and vice versa. While producing satisfactory results, these methods often ignore other modalities in the videos, such as objects, motions and captions, which are informative and effective for video representation learning [5]. How to fully exploit knowledge from these experts and combine these heterogeneous features in one video is still an open problem.

Moreover, The differences between audio and video are obvious, such as input representations, network architectures and benchmarks. On one hand, audio classification methods often rely on short term Fourier analysis to produce log-mel spectrograms, which are thereafter fed into CNNs originally designed for images [6, 4]. Contrarily, the video stream is usually three-dimensional, and there exists the unique challenge, i.e., the high redundancy across multiple frames. This is to some extent a counter-intuitive schema. On the other hand, humans perceive the world by concurrently processing and fusing multiple feelings, which can be intepreted as high dimensional inputs such as vision and audio. While machine learning models, in stark contrast, are typically modality-specific and trained on unimodal benchmarks. How to bridge the gap between video and audio modalities and further utilize their interactions is the primary concern of this work.

In this paper, we propose a novel two-stream model named Multi-modal Aggregation and Co-attention network (MAC), which processes both video and audio inputs in separate streams with co-attentional interactions. As shown in Fig. 1, we first take raw videos as inputs and extracts aggregated features from multiple modalities to benefit the video representation learning. Then, we introduce the self-attention mechanism to make videos adaptively assign higher weights to the representative modalities. Video and audio modalities are highly correlated and complementary, which inspires us to explore whether it is possible to utilize the pairwise relations between them. Thus, we further introduce a co-attention transformer module to better capture the semantic relations among video and audio features. By exchanging key-value pairs in multi-headed attention, this module can not only allow for the independent processing in each modality, but also enable the interactions between two modalities. Experiments show that our method improves the mutual interactions of the learned video-audio representations, and significantly outperforms other state-of-the-arts. The main contributions of this work can be summarized as follows:

∗ Corresponding author:zhangwanqian@iie.ac.cn

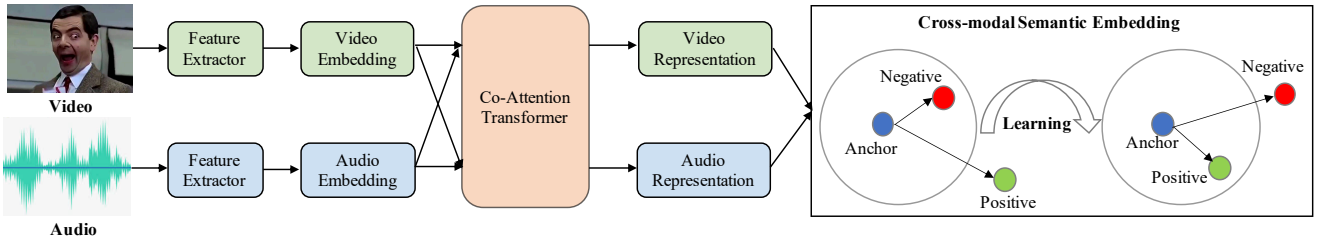


Fig. 1: The Framework of Multi-modal Aggregation and Co-attention network(MAC). Aggregated features from multiple modalities in video stream are extracted with a self-attention mechanism to benefit the video representation learning. Then, the co-attention transformer module is introduced to better exploit the semantic relations among videos and audios, which not only allows for the independent processing in each modality, but also enable the interactions between them.

- We propose a novel method named Multi-modal Aggregation and Co-attention network (MAC) for video-audio retrieval, which fully utilizes the mutual interactions between vidos and audios.
- We extracts aggregated features from multiple modalities with the self-attention mechanism to benefit the video representation learning. We also introduce a co-attention transformer module to enable video-attended audio features to be incorporated into video representations and vice versa.
- Extensive experiments show promising retrieval performance compared with the state-of-the-art baselines on three datasets, indicating the superiority of our method.

2. RELATED WORK

Cross-modal video-text retrieval aims to find relevant videos given text queries [5, 7, 8]. CE [8] adopts video features extracted from all modalities to encode a video. HGR [5] proposes a Hierarchical Graph Reasoning (HGR) model, which decomposes video-text pairs into global-to-local levels. MMT [7] presents a multi-modal transformer to jointly encode the different modalities in videos. Recently, the Contrastive Language-Image Pretraining (CLIP) [9] model is widely used in video-text retrieval. CLIP4Clip [10] investigates three mechanisms of similarity calculation based on the pre-trained CLIP. Similarly, CLIP2video [11] focuses on the spatial semantics captured by the CLIP model. However, these works focus on learning a joint multi-modal embedding space between texts and videos, which do not incorporate the audio stream.

Audio-Visual Representation Learning focuses on connecting the visual and sound data of different modalities [1, 4, 2]. AVSlowFast [1] encourages the model to capture fine-grained temporal information by utilizing different temporal scales of the audio and visual data. [3] learns patch-level audio-visual correspondence by drawing positive/negative patches iteratively along with audio-visual feature correlation. VATT [4] takes raw signals as inputs and extracts multimodal representations with a modality-agnostic, single-backbone and weight-sharing transformer. CoMVT [2] proposes a visually conditioned Future Utterance Prediction (FUP) learning task, where the goal is to predict the next utterance in an instructional video using both visual frames and transcribed speech. Unlike the above works, we focus on the downstream task, i.e., video-audio retrieval, which finds and retrieves the relevant videos according to given audio queries [6].

3. METHODOLOGY

Let $\{(v_i, a_i)\,|v_i \in V, a_i \in A\}$ be a set of videos with v_i being the visual representation of the i^{th} video and a_i the corresponding audio description. Our goal of video-audio retrieval is to learn a pair of functions $\varphi\,(v)$ and $\psi\,(a)$ to map videos and audio descriptions into a joint embedding space, in which embeddings for matched audio descriptions and videos should lie close together and vice versa. Next, we elaborate details of the video and audio embeddings, the co-attention transformer module and the loss function.

3.1. Video Embedding

In order to make full use of the information in one video, we draw on a collection of pre-trained models to extract video features from different modalities. These operations project the video to a collection of N modality features $\left\{I_{var}^{(1)}, ..., I_{var}^{(i)}, ..., I_{var}^{(N)}\right\}$, where $I_{var}^{(i)}$ represents the i^{th} modality feature and subscript *var* denotes a variable-length output. Each element is then aggregated along its temporal dimension, producing fixed-length video feature embeddings per video $\left\{I^{(1)}, ..., I^{(i)}, ..., I^{(N)}\right\}$. Next, we apply linear projections to transform these time-aggregated embeddings into a common dimensionality. Thus, extracted video embeddings can be written as $V = \{I_i\}_{i=1}^{N}$. To further combine multiple modality features, we adopt two successive fully-connected layers as the attention weighting function, which is a straightforward way considering the relations between multiple modalities. By assigning different weights to different modality features, we can finally obtain discriminative video embeddings.

3.2. Audio Embedding

Audios can exert influences on videos in various cases. For example, 'playing piano' is the scenario where sound dominates, while 'humming a tune' is the scenario where the action itself is difficult to detect in videos. Inspired by [6], we adopt the QuerYD dataset to obtain the audio embeddings. QuerYD [6] is gathered from user-contributed descriptions provided by the YouDescribe community [1], which contribute audio descriptions to videos hosted on YouTube to assist the visually-impaired persons. Then, a portion of these audio descriptions is further accompanied by user-provided transcriptions. To handle cases in which such a transcription is not provided, we use the Google Speech-to-Text API [2] to transcribe audio descriptions. These word embeddings are then aggregated into a single audio vector to obtain the entire audio description using the NetVLAD aggregation module [12]. After the aggregation, we project the aggregated audio vector to the separated subspaces for each video feature using Gated Embedding Module (GEM). The audio representation consisting of N embeddings can thus be written as $A = \{\psi_i\}_{i=1}^{N}$.

3.3. Co-Attention Transformer Module

We argue intuitively that lower layers are involved in processing low-level features, while higher layers are focused on learning semantic concepts. Note that humans learn to understand audios, recognize visions, and identify their correspondences by learning patterns in what they see and hear [1, 4, 2]. However, previous methods have developed video-audio models only capable of generating unimodal features seperately on specific benchmarks.

Thus, we introduce a Co-Attention Transformer module (CAT) to learn the interactive information between audio features and video features. CAT encourages the whole model to optimize the representations with respect to the mutual interactions [13]. Specifically, CAT consists of two streams, i.e., video embeddings $V = \{I_i\}_{i=1}^{N} \in \mathcal{R}^{N\times D}$ and audio embeddings $A = \{\psi_i\}_{i=1}^{N} \in \mathcal{R}^{N\times D}$. As in Fig. 2, the whole CAT module computes query, key and value matrices as in a standard transformer block. Note that the keys and values from each modality are passed as inputs to the other modality's multi-headed attention block. Consequently, the attention block produces attention-pooled features for one modality conditioned on the other, i.e., performs video-conditioned audio attention in the visual stream and audio-conditioned video attention in the audio stream.

For the video stream, we compute key $(\bar{K})$-value$(\bar{V})$ pairs based on audio features and query $(\bar{Q})$ based on video features. The scaled dot-product attention is then calculated by:

$$V_{attn} = softmax\left(\frac{\bar{Q}\bar{K}^T}{\sqrt{d}}\right)\bar{V}, \quad (1)$$

where V_{attn} is a weighted sum of values (audio features), and the weight of each value is calculated based on its interaction with the video features $\bar{Q}$. The final video representation can be written as $\phi(V) = V_{attn} + V$.

For the audio stream, we compute key $\left(\widetilde{K}\right)$-value $\left(\widetilde{V}\right)$ pairs based on video features and query $\left(\widetilde{Q}\right)$ based on the audio features. Similarly, the scaled dot-product attention is then calculated by:

$$A_{attn} = softmax\left(\frac{\widetilde{Q}\widetilde{K}^T}{\sqrt{d}}\right)\widetilde{V}, \quad (2)$$

where A_{attn} is a weighted sum of values (video features), and the weight of each value is calculated based on its interaction with the audio features $\widetilde{Q}$. The final audio representation can be written as $\psi(A) = A_{attn} + A$.

Note that we restrict early layers of the network to focus on unimodal processing, while only introduce cross-modal connections at later layers. The intuition is that low-level visual features such as edges and corners may not have a particular audio signature, therefore fail to benefit from early fusion with audios. To this end, the proposed CAT module can not only allow for the independent processing in each modality, but also enable the interactions between two modalities.

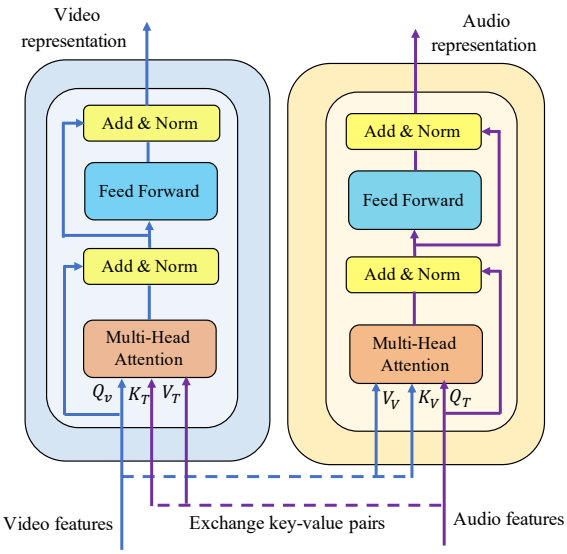


Fig. 2: Illustration of the introduced co-attention transformer module.

3.4. Loss Function

To train the model, we adopt the simple yet effective bi-directional hard-negatives ranking loss for cross-modal se-

[1] https://youdescribe.org/

[2] https://cloud.google.com/speech-to-text

Table 1: Comparison of video-audio retrieval methods trained with paragraph-level information on the QuerYD dataset.

Method	Audio-to-Video Retrieval				Video-to-Audio Retrieval				rsum
	R@1	R@5	R@10	Med R	R@1	R@5	R@10	Med R	
E2EWS [14]	13.5	27.5	34.5	35	12.4	23.8	30.8	33	142.5
MoEE [15]	11.6	30.2	43.2	14.2	13.0	30.9	43.0	14.5	171.9
CE [8]	13.9	37.6	48.3	11.3	13.7	35.2	46.9	12.3	195.6
MAC	**16.6**	**39.8**	**52.5**	**9**	**17.3**	**39.8**	**50.7**	**10**	**216.7**

Table 2: Comparison of localisation methods trained with oracle temporal proposals information on the QuerYD dataset.

Method	Audio-to-Video Retrieval				Video-to-Audio Retrieval				rsum
	R@1	R@5	R@10	Med R	R@1	R@5	R@10	Med R	
E2EWS [14]	6.7	14.7	20.4	133	8.4	15.4	19.8	154.5	85.4
MoEE [15]	19.0	38.9	47.9	12	19.8	39.6	47.6	13	212.8
CE [8]	18.2	38.1	46.8	13.3	18.1	37.3	45.9	14	204.4
MAC	**19.9**	**41.6**	**50.3**	**10**	**20.1**	**40.7**	**50.2**	**10**	**222.8**

mantic embedding [16, 17], which can be formulated as:

$$L = \frac{1}{B}\sum_{i=1}^{B}\sum_{j\neq i}[max\left(0,\ m+\widehat{s_{i,j}}-s_{i,i}\right) + max\left(0,\ m+\widehat{s_{j,i}}-s_{i,i}\right)], \tag{3}$$

where B is the batch size, m is the margin value and $s_{i,j} = s(\psi\left(A_i\right), \phi\left(V_j\right))$ is the similarity score of audio description A_i and video V_j. $\widehat{s_{i,j}}$ and $\widehat{s_{j,i}}$ indicate a negative text pair for V and a negative video pair for A, respectively.

4. EXPERIMENTS

4.1. Experimental Settings

Datasets. We carry out experiments on QuerYD [6], AUDIOCAPS [18] and CLOTHO [19] datasets. QUERYD [6] is a dataset of described videos sourced from YouTube and the YouDescribe platform. It is accompanied by audio descriptions that are provided with the explicit aim of conveying the video content to visually impaired users. Therefore, the provided descriptions focus heavily on the visual modality. The training, validation and test partitions of the dataset are the same as in prior works [6]. AUDIOCAPS [18] is a dataset of sounds with event descriptions that was introduced for the task of audio captioning, where sounds are sourced from the AudioSet dataset [20]. We follow the standard split with 49,291 training, 428 validation, and 816 testing samples in the official split. CLOTHO [19] is a dataset of described sounds introduced for the task of audio captioning, with sounds sourced from the Freesound platform [1]. We use the training and validation set with 2,314 and 579 samples, respectively.

Evaluation Criteria We adopt standard retrieval metrics (following [6]) to evaluate the performance of video-audio retrieval. We measure rank-based performance by R@K (higher is better) and Median Rank, i.e., MR, (lower is better). We also report the sum of R@1, R@5 and R@10 as Sum of Recalls.

[1] https://freesound.org/

4.2. Implementation Details

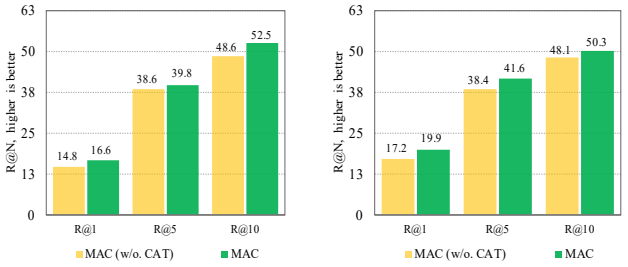
We adopt the Adam optimizer for all our experiments, and set the margin of the bi-directional hard-negatives ranking loss to 0.3. Inspired by [6], we also freeze our pre-trained models for video feature extraction. We set N=4 and extract video features of scene, sound, object and action, which are publicly released by [8]. To be specific, for scene, object and action, we average frame-level features along the temporal dimension to produce a single feature vector per video. For sound features, we adopt the NetVLAD mechanism, which has proven effective for the video-text retrieval [8]. All aggregated video features are projected to the same size as 768 before fed into the co-attention transformer module (i.e., D=768). Moreover, we utilize one self-attention layer and four attention heads for the co-attention transformer module. For QuerYD paragraph-level video-audio retrieval task and clip localisation task, we set the batch size to 128, learning rate to 0.01, and weight decay to 1e-3.

4.3. Comparision with state-of-the-arts

In this section we demonstrate the application of QuerYD to two video understanding tasks: paragraph-level video-audio retrieval and clip localisation. We consider three models:

The E2EWS (End-to-end Weakly Supervised) model proposed by [14] is a cross-modal retrieval model. We use the video and text encoders without any form of fine-tuning on QuerYD, providing a calibration of task difficulty.

The MoEE (Mixture of Embedded Experts) model proposed by [15] comprises a multi-modal video model in combination with a system of context gates that learn to fuse together different pretrained 'experts' to form a robust cross-modal text-video embedding.



(a) Paragraph-level video retrieval task (b) Clip localisation task

Fig. 3: Ablation study on the proposed CAT module.

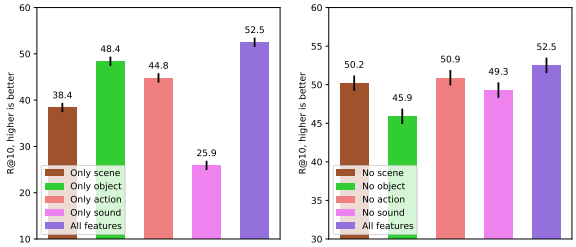


Fig. 4: Illustration of the differences between multiple experts in videos.

The CE (Collaborative Experts) model proposed by [8] similarly learns a cross-modal embedding by fusing together a collection of pretrained experts to form a video encoder. It uses a relation network sub-architecture to combine together different modalities, and represents the state-of-the-art on several retrieval benchmarks.

Table 1 compares the proposed MAC model with SOTA methods on the paragraph-level video retrieval task. For fairness, all methods adopt the same four pretrained experts for scene classification, action recognition, sound classification and image classification described in [8]. We can find that our method performs best and consistently outperforms state-of-the-art methods. For instance, it outperforms the SOTA model CE [8], and sum of recalls is increased from 195.6 to 216.7 on the paragraph-level video retrieval task.

Moreover, to demonstrate the robustness of our approach on different tasks, we further provide results on clip localisation task in Table 2. Event localisation is a task that aims to retrieve a specific temporal segment from a video given a natural language text description. We can see that our MAC model achieves consistent improvements, with 9.3%, 9.1% and 7.5% relative improvements compared with CE in R@1, R@5 and R@10, respectively. In a nutshell, MAC verifies the effectiveness of interactive information between video and audio modalities.

4.4. Ablation Study

Effect of co-attention transformer module. To evaluate the contribution of co-attention transformer module, we removes the co-attention transformer module from the full MAC and present the results in Fig. 3. We can find that the full MAC method achieves 12.16% relative improvements compared with MAC(w/o.CAT) method in R@1 on the paragraph-level video retrieval task. Similarly, full MAC method achieves 15.69% relative improvements compared with MAC(w/o.CAT) method in R@1 on the clip localisation task. We argue that MAC(w/o.CAT) is inferior to the full MAC method, indicating the proposed module contributes to generating modality-agnostic and discriminative features in the video-audio retrieval task.

Comparison of the different experts. In Fig. 4, we show an ablation study when training our model on paragraph-level video retrieval task using only one expert (left), and using all experts but one (right). In the case of using only one expert, we note that the object expert provides the best results. We owe the poor performance of sounds to the fact that they are often absent, thus resulting in a zero vector input to our video encoder. While the scene expert shows a decent performance, if used alone, it does not contribute when combined with others, which might due to the semantics it encodes have already been captured by other experts like appearance or motion. Though the sound expert alone does not provide a good performance, it contributes the most when used in conjunction with the others. We owe this to the complementary cues it provides when compared to the other experts.

Table 3: Generalization to text-audio retrieval tasks.

Benchmark	Text→Audio		Audio→Text	
	R@1	R@10	R@1	R@10
AUDIOCAPS				
MoEE [15]	22.5	70.0	25.7	73.0
CE [8]	22.9	70.2	26.1	72.7
MAC(ours)	**23.9**	**74.2**	**28.8**	**76.0**
CLOTHO				
MoEE [15]	5.1	30.1	6.3	29.9
CE [8]	5.8	31.3	7.3	32.8
MAC(ours)	**6.4**	**33.0**	**7.7**	**33.9**

Generalization to different tasks. To verify that our method can be easily generalized to different tasks, as shown in Table 3, we generalized MAC to text-audio retrieval task on two datasets, i.e., AUDIOCAPS [18] and CLOTHO [19] datasets. We can find that our method performs best and consistently outperforms state-of-the-art methods on two text-audio datasets.

5. CONCLUSION

In this paper, we have proposed Multi-modal Aggregation and Co-attention network (MAC) to process video and audio inputs with co-attentional interactions. To benefit the video representation learning, we extract aggregated video fea-

tures from multiple modalities with the self-attention mechanism. Moreover, to better capture the relations among videos and audios, we introduce a co-attention transformer module, which can not only allow for the independent processing in each modality, but also enable the interactions between two modalities. Experiments on three video-audio benchmarks have demonstrated that our method achieves significant improvements compared to the state-of-the-arts.

6. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grants 62106258 and 62006242.

7. REFERENCES

[1] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[2] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid, "Look before you speak: Visually contextualized utterances," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[3] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang, "Unsupervised sound localization via iterative contrastive learning," *arXiv preprint arXiv:2104.00315*, 2021.

[4] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *arXiv preprint arXiv:2104.11178*, 2021.

[5] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[6] Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie, "QUERYD: A video dataset with high-quality text and audio narrations," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 2265–2269.

[7] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid, "Multi-modal transformer for video retrieval," in *European Conference on Computer Vision*, 2020.

[8] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," in *British Machine Vision Conference*, 2019.

[9] 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, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021.

[10] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li, "Clip4clip: An empirical study of CLIP for end to end video clip retrieval," *arXiv preprint arXiv:2104.08860*, 2021.

[11] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen, "Clip2video: Mastering video-text retrieval via image CLIP," *arXiv preprint arXiv:2106.11097*, 2021.

[12] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2019.

[14] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9876–9886.

[15] Antoine Miech, Ivan Laptev, and Josef Sivic, "Learning a text-video embedding from incomplete and heterogeneous data," *arXiv preprint arXiv:1804.02516*, 2018.

[16] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang, "Dual encoding for zero-example video retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9346–9355.

[17] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," in *British Machine Vision Conference*, 2018.

[18] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "Audiocaps: Generating captions for audios in the wild," in *NAACL*, 2019, pp. 119–132.

[19] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen, "Clotho: an audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 736–740.

[20] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 776–780.