# Bi-directional Image-Speech Retrieval Through Geometric Consistency

Xinyuan Qian[1], Wei Xue[2], Qiquan Zhang[3], Ruijie Tao[4], Yiming Wang[5], Kainan Chen[6], Haizhou Li[4]
[1] University of Science and Technology Beijing [2] Hong Kong University of Science and Technology
[3] University of New South Wales [4] National University of Singapore
[5] Fondazione Bruno Kessler [6] Eigenspace GmbH

## Abstract

*Cross-modal Retrieval (CMR) is formulated for scenarios where the queries and retrieval results are of different modalities. Existing CMR studies mainly focus on the common contextualized information between text transcripts and images, and the synchronized event information in audio-visual recordings. Differently, in this paper, we investigate the geometric correspondence between images and speech recordings captured in the same space and formulate a novel CMR task, called Spatial Image-Acoustic Retrieval (SIAR). We propose the Contrastive Speech Image Retrieval (CSIR) method which uses supervised contrastive learning to attract the same-space cross-modal features while repelling the ones from different spaces. Then, image and speech features are directly compared and we predict the SIAR result with the maximum similarity. Experiments demonstrate the effectiveness and feasibility of our proposal.*

## 1. Introduction

CMR performs flexible information retrieval among heterogenous modalities (*e.g.*, images, texts and audio signals) [24]. As a compelling research topic, it has been extensively studied with a broad range of applications, such as visual question and answering [14], video captioning [2], text-image retrieval [8] and action localization [17] *etc*. Specifically, CMR employs an uni-modal query to retrieve the counterpart from a different modality. It allows users to obtain comprehensive information about enclosed events or contextualized information from different modalities.

Most existing CMR works focus on the intersection of computer vision and natural language processing fields [6, 10, 13, 16, 26–28]. For example, the Contrastive Language–Image Pre-training (CLIP) model [16] learns highly generalizable representations by exploiting common contextualized content through massive image-text pairs, followed by other extensive explorations [6, 10, 28]. In [13], a spatial-temporal graph-based framework is proposed to facilitate multi-modal machine translation. In [26], the semantic consistency of the image and text was enhanced through a deep discrete cross-modal hashing network. Whereas CMR between audio-visual signals is still in infancy. As a representative work, a recent method [25] systematically distills audio representations from the pre-trained CLIP model [16] to facilitate downstream audio event classification and retrieval tasks. However, it only explores the same motive and semantics goal from synchronized audio-visual recordings, while neglecting their geometric dependence.

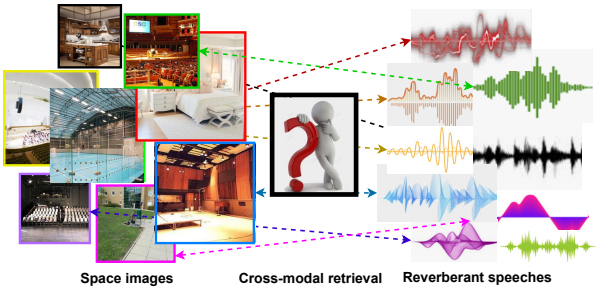


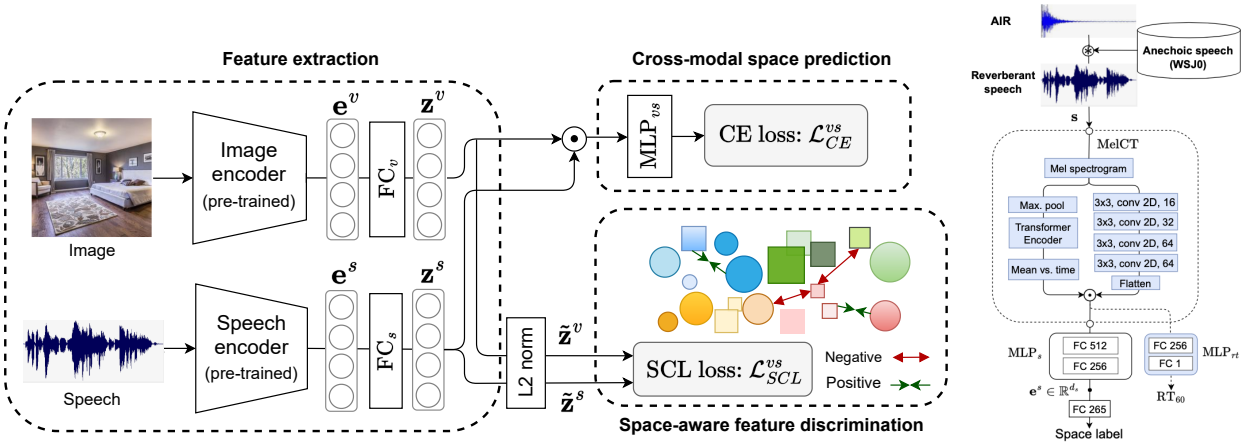

Figure 1. Our proposed SIAR task, which performs bi-directional cross-modal retrieval between images and reverberant speeches captured in the same space (data from different spaces are indexed by different colors).

While exciting as proof of concept and potential applications, there exists no work that correlates visual images and speeches via the common spatial properties. How to bridge their *heterogeneity gap* remains under-explored. In this work, we introduce the novel SIAR task which relies on the image-speech geometric correspondence. As shown in Fig. 1, it retrieves speeches captured in a geometric space depicted by the query image and vice versa.

## 2. Proposed Method

Let us start by giving a few definitions. We denote the training set as $\mathbf{O} = \{\mathbf{o}_i = (\mathbf{v}_i, \mathbf{s}_i, c_i), i = 1, 2, ..., n\}$, includ-

This work is sponsored by Young Scientists Fund of the National Natural Science Foundation of China (NSFC) Grant No. 62306029; CCF-Tencent Rhino-Bird Open Research Fund; ECS-HKUST22201322 from Research Grant Council of Hong Kong; NSFCY-62206234 from China; EU ISFP PRECRISIS (ISFP-2022-TFI-AG-PROTECT-02-101100539).




(a) proposed CSIR approach (b) speech encoder

Figure 2. (a) The block diagram of our proposed CSIR approach for the novel bi-directional SIAR task. The extracted image features $\mathbf{e}^v$ and speech features $\mathbf{e}^s$ are projected into a common representation space, denoted as $\mathbf{z}^v$ and $\mathbf{z}^s$, respectively. (1) CE loss: cross-modal features are concatenated for space prediction; (2) SCL loss: The $L2$-normalized cross-modal features $\tilde{\mathbf{z}}^v$ and $\tilde{\mathbf{z}}^s$ are marked with rounds and rectangles where colors index distinct space labels. In particular, the positive feature pairs are depicted by red bi-directional arrows while the negative ones are marked with green arrows ($\odot$ denotes pairwise concatenation); (b) The architecture of the speech encoder for space classification ($\circledast$ denotes convolution).

ing the set of visual space images $\mathbf{V}=\{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_n}\}$, reverberant speech waveforms $\mathbf{S}=\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$, and space labels $\mathbf{C} = \{c_1, c_2, ..., c_n\}$. Then the cross-modal correspondence can be learnt from paired image-speech samples $(\mathbf{v}_i, \mathbf{s}_i)$. Let us denote the test set as $\mathbf{T} = \{(\mathbf{v}_j, \mathbf{s}_j), j = 1, ..., n\}$ where $\mathbf{O} \cap \mathbf{T} = \varnothing$. SIAR aims to learn a similarity measure that given a query speech waveform $\mathbf{s}_q \in \mathbf{T}$ or an image $\mathbf{v}_q \in \mathbf{T}$, retrieve the counterpart from the same space. These bi-directional Speech-to-image Retrieval (SIR) process is formulated as:

$$\hat{\mathbf{v}} = \underset{\mathbf{v}_j \in \mathbf{T}}{\arg\max}\, \mathcal{F}_{sim}(\mathbf{s}_q, \mathbf{v}_j; \Upsilon^{sim}) \tag{1}$$

where $\mathcal{F}_{sim}(\cdot)$ denotes the similarity computation derived from our proposed approach and $\Upsilon^{sim}$ are the trainable parameters. The Image-to-speech Retrieval (ISR) process is formulated in a similar way. We consider a retrieval result is correct if it has the same space label as the query.

The key of SIAR is how to project the heterogenous features into a common representation space for direct similarity measure. We first extract features from the separately *pre-trained* image and speech encoders and then use individual Fully Connected (FC) layers to make the projections:

$$\mathbf{z}^v = \text{FC}_v(\mathbf{e}^v; \Upsilon^{zv}) \in \mathbb{R}^d \tag{2}$$

$$\mathbf{z}^s = \text{FC}_s(\mathbf{e}^s; \Upsilon^{zs}) \in \mathbb{R}^d \tag{3}$$

where $\mathbf{z}^v$ and $\mathbf{z}^s$ are the normalized features, $\Upsilon^{zv}$ and $\Upsilon^{zs}$ are FC parameters, $d$ is the common feature dimension.

## 2.1. Speech Encoder

There is no pre-trained model of space recognition from reverberant human speech. Thus, we design and pre-train a novel speech encoder, to extract the space-aware acoustic signatures, illustrated in Fig. 2(b) and formulated as:

$$\mathbf{e}^s = \mathcal{F}^s(\mathbf{s}; \Upsilon^s) \in \mathbb{R}^{d_s} \tag{4}$$

where $\mathcal{F}^s$ denotes the speech encoder parameterized by $\Upsilon^s$, and $d_s$ represents the speech feature dimension.

Specifically, the speech encoder takes Mel spectrogram as the input and ensembles both Convolutional Neural Network (CNN) and transformer [23] architectures by taking their complementary characteristics: CNNs are effective in processing contextual information of spectrograms [19, 22] but limited at modeling time series signals, while transformer excels in capturing the long-distance correlation because of the self-attention mechanism [1, 23].

Outputs of the two branches are concatenated after an average (over time) and a flatten operation to predict the *posterior probability* of the space label where we use Cross Entropy (CE) as the optimization criterion. Since $\text{RT}_{60}$ characterizes the environmental reverberation effect by computing the time taken for a sound to decay to one millionth of its original intensity (*i.e.*, 60 dB of decay), it is also used as a supervision. Moreover, since speech encoder can be enhanced by maximizing the intra-class feature similarities while minimizing the inter-class similarities, we also adopt the SCL as a loss.

### 2.2. Image Encoder

CNNs are useful in estimating late-reverberation statistics from images [11, 12]. Compared to CNN, ResNet [9] uses identity mapping to tackle the vanishing gradient problem, resulting in remarkable success in computer vision applications. Thus, we believe that the ResNet-based structure will also be beneficial to image-acoustics correspondence learning. We select a *pre-trained* space recognition model on Place365 dataset [30] to extract visual features: $\mathbf{e}^v = \mathcal{F}^v(\mathbf{v}; \Upsilon^v) \in \mathbb{R}^{d_v}$ where $\mathcal{F}^v$ denotes the image encoder parameterized by $\Upsilon^v$ and $d_v$ represents the visual feature dimension.

### 2.3. Cross-modal Retrieval.

Given the supervision of space labels, our proposed CSIR network, illustrated in Fig. 2(a), incorporates feature extraction, cross-modal space prediction and space-aware contrastive feature discrimination into a unified framework to facilitate bi-directional SIAR.

**Cross-modal space prediction.** Since both images and speeches are captured in the same environment, to leverage the space-specific supervision, we fuse the cross-modal features to predict the *posterior probability* of each semantic space label:

$$p_{vs}(c) = \text{MLP}_{vs}(\mathbf{z}^v \odot \mathbf{z}^s; \Upsilon^{vs}) \in \mathbb{R}^C \tag{5}$$

where $\text{MLP}_{vs}$ denotes the space classifier parameterized by $\Upsilon^{vs}$ and $\odot$ indicates pairwise feature concatenation. Then, we use CE loss as the training objective:

$$\mathcal{L}_{CE}^{vs} = -\sum_{i=1}^{C} p(c_i) log(p_m(c_i)) \tag{6}$$

**Space-aware feature discrimination.** We use SCL [5] to bridge the cross-modal *heterogeneity gap* by learning common representations with maximized intra-class distance and minimized inter-class distance. In particular, image and speech features are intrinsically different and can be considered as augmented views to each other. Thus, in the resulting *multiviewed* batch, cross-modal features from the same space class are considered as positive pairs while those from different spaces are negative pairs. Then, the cross-modal SCL loss is defined as:

$$\text{L}_{SCL}^{vs} = -\sum_{i=1}^{2N} \frac{1}{2N_{c_i}-1} \sum_{j=1}^{2N} \delta \cdot \log\left(\frac{\exp(\tilde{\mathbf{z}}_i^s \cdot \tilde{\mathbf{z}}_j^v/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{i\neq k} \cdot \exp(\tilde{\mathbf{z}}_i^s \cdot \tilde{\mathbf{z}}_k^v/\tau)}\right) \tag{7}$$

where $\delta = \mathbb{1}_{i\neq j, c_i=c_j}$ is a binary indicator and $\tilde{z}_i^s$ denotes the $L2$ normalized version of $z_i^s$.

The overall optimization function is defined as:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{CE}^{vs} + (1-\alpha)\mathcal{L}_{SCL}^{vs} \tag{8}$$

where $\alpha$ is a pre-defined parameter to scale the contributions of different loss items which is set to 0.6 empirically.

Table 1. Specifications of the Image2Reveb dataset [18].

| | | | |
|---|---|---|---|
| AIR sampling rate | 22,050 Hz | | |
| Image size | 224 × 224 pixels | | |
| Total No. of spaces | 265 | | |
| # in/outdoor spaces | 235/30 | | |
| | Train | Test | Total |
| # image-AIR pairs (indoor/outdoor) | 10,207 (8,963/1,244) | 1,647 (1,515/132) | 11,854 (10,478/1,376) |
| Speech generation | Randomly sample 5 recordings from the WSJ dataset to convolve with each Acoustic Impulse Response (AIR). | | |

## 3. Dataset

We use the publicly available Image2Reverb dataset [18] for SIAR experiments. It enables learning of late-field reverberation characteristics by providing paired AIRs and images in high variability of spaces (*e.g.*, musical halls, bedrooms, and cathedrals). Apart from Image2Reverb, we also investigate other alternative datasets. Sound-space[1] [4] provides acoustically realistic audio renderings on the vision-based Replica [20] and Matterport3D [3] datasets. However, both the acoustic signals and visual scenes are simulated in similar indoor smart home scenarios, thus being less realistic and varied than Image2Reverb. For VEGAS [31] and AVE [21], the most explored audio-visual datasets, they only provide category labels (*e.g.*, helicopter and chainsaw) which do not fit our space-specific research objective. Therefore, Image2Reverb is our best-fit dataset.

We re-arrange the Image2Reverb dataset with augmented reverberant speech samples for the SIAR purpose. Table 1 lists the specifications. The dataset involves 265 scenes with 235 indoor and 30 outdoor scenarios. There are totally 11,854 paired image-AIR examples and we split them into non-overlapping train-test partitions with 10,207 and 1,647 sequences, respectively. To augment reverberant speeches, we randomly select five anechoic speech samples from the WSJ0 dataset [15] to convolve with each AIR. This setting is sufficient for characterization of the environment. Moreover, we resample AIRs and speech waveforms at 22.05 kHz and truncate them to the same duration of 6 seconds. The space images are normalized and center-cropped at the size of 224 × 224 pixels.

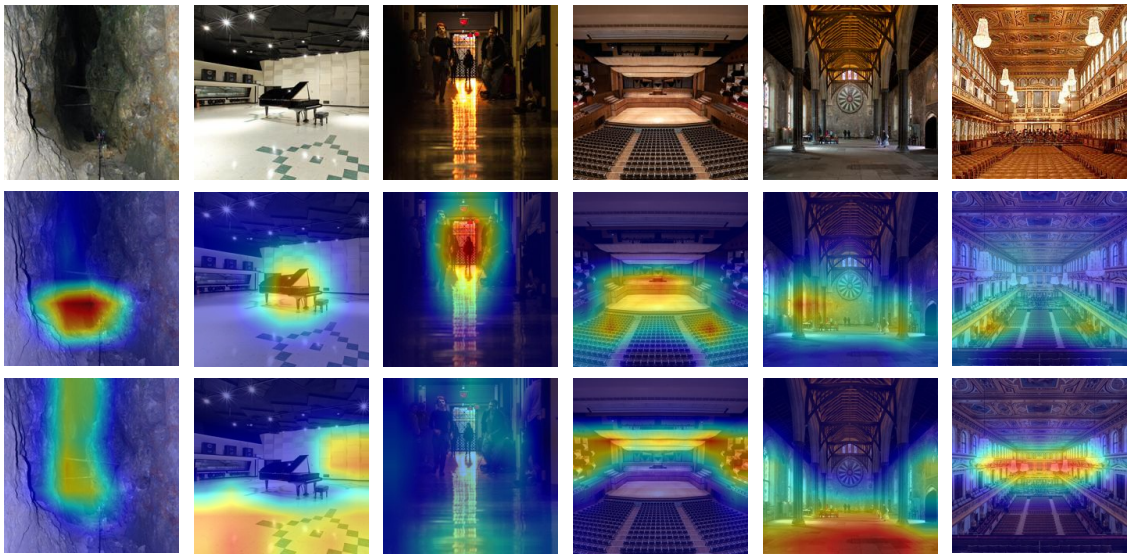
## 4. Experiments

We directly use a pre-trained visual space recognition model (*i.e.*, ResNet50 model [9] pre-trained on the Places365 dataset [30]) to extract visual features ($d_v = 365$) in the Image2Reverb dataset. To characterize space acoustics, we compute Short-time Fourier Transform (STFT) with a 2048-

[1] https://github.com/facebookresearch/sound-spaces

Table 2. The results of bi-directional cross-modal retrieval between space images and reverberant speeches.

| Model | Speech-to-image retrieval (SIR) | | | | | | Image-to-speech retrieval (ISR) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Indoor | | Outdoor | | Total | | Indoor | | Outdoor | | Total | |
| | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| DCMR [29] | 33.49 | 33.53 | 64.55 | 64.55 | 35.98 | 36.02 | 39.65 | 50.29 | 59.85 | 71.21 | 41.27 | 51.97 |
| CSIR (ours) | **68.20** | **68.34** | **88.16** | **88.16** | **69.80** | **69.93** | **90.68** | **95.70** | **96.97** | **100.00** | **91.18** | **96.05** |



(a) (b) (c) (d) (e) (f)

Figure 3. The Grad-CAMs for visualizing what the image encoder is looking at. The RGB images (first row) are fed into the pre-trained image encoder (second row) and our fine-tuned one (bottom row) for SIAR. The varying colors from red to blue correspond to the higher and lower activation values.

point Hanning window and a hop size of 441. Then, Mel filter banks are applied to get a 128-D Mel spectrogram. For Cross-modal Learning (CML), both audio and visual features are projected into a common space where $d = 64$ in Eq. 2 and Eq. 3. For evaluation, we use recall at Rank K as the metric, denoted as R@K (%), which describes the percentage of queries where the desired label is within the top K retrieval answers (K=1 or 5). Table 2 gives the bi-directional retrieval results between space images and reverberant speeches. We can see that our proposed CSIR model can successfully perform SIR where R@1 and R@5 equal 69.80% and 69.93%, respectively. On the contrary, our proposal can retrieve a speech transmitted in the space specified by a query image *i.e.*, ISR where R@1 and R@5 equal 91.18% and 96.05%, respectively. It achieves superior results than DCMR [29] in all conditions.

To understand the effect of the image encoder, we use Gradient-weighted Class Activation Mapping (Grad-CAM) [7] to analyze which image regions are more related to environmental acoustic characteristics. Specifically, Grad-CAM is a widely used strategy for visually interpreting CNNs by assigning higher values to more important regions. By default[2], we extract outputs from layer 4 of the original and our fine-tuned ResNet50 image encoder to generate heat maps in test images. As illustrated in Fig. 3, high-value features are observed to be mostly associated with activations of visual regions related to large reflective regions. For example, in Fig. 3(b), instead of focusing on the building, the image encoder tends to pay more attention to the large reflective meadow. In Fig. 3(e), the ceiling light is highlighted instead of the stage and the seats, thus is more spatial-related to the transmitted speech. In summary, Fig. 3 shows that the fine-tuned image encoder in our proposed CSIR model can successfully emphasize reflective surfaces, which are more correlated with environmental acoustics.

## 5. Conclusion

We proposed a novel task, SIAR, that tackles the bi-directional retrieval problem between images and speeches captured in the same space by considering their geometrical correspondence. We proposed the CSIR approach which adopts cross-modal contrastive learning to achieve space-aware feature discrimination. The experimental results demonstrate the efficiency of our proposed speech encoder and the feasibility of the SIAR task.

[2]GradCAM: https://github.com/jacobgil/pytorch-grad-cam

## References


[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *Proc. of Int. Conf. on Learning Representations*, 2015. 2

[2] Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and Dimosthenis Karatzas. Good news, Everyone! context driven entity-aware captioning for news images. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 12466–12475, 2019. 1

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, pages 374–382, 2017. 3

[4] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigaton in 3D environments. In *Proc. of European Conf. on Computer Vision*, 2020. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. of Int. Conf. on Machine Learning*, pages 1597–1607, 2020. 3

[6] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021. 1

[7] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 4

[8] Chunbin Gu, Jiajun Bu, Xixi Zhou, Chengwei Yao, Dongfang Ma, Zhi Yu, and Xifeng Yan. Cross-modal image retrieval with deep mutual information maximization. *Neurocomputing*, pages 166–177, 2022. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. of Int. Conf. on Machine Learning*, pages 4904–4916, 2021. 1

[11] Homare Kon and Hideki Koike. Estimation of late reverberation characteristics from a single two-dimensional environmental image using convolutional neural networks. *Journal of the Audio Engineering Society*, 67(7/8):540–548, 2019. 3

[12] Homare Kon and Hideki Koike. An auditory scaling method for reverb synthesis from a single two-dimensional image. *Acoustical Science and Technology*, 41(4):675–685, 2020. 3

[13] Mingjie Li, Po-Yao Huang, Xiaojun Chang, Junjie Hu, Yi Yang, and Alex Hauptmann. Video pivoting unsupervised multi-modal machine translation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 3918–3932, 2022. 1

[14] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 299–307, 2017. 1

[15] Douglas B Paul and Janet Baker. The design for the wall street journal-based csr corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992. 3

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of Int. Conf. on Machine Learning*, pages 8748–8763, 2021. 1

[17] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022. 1

[18] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proc. of Int. Conf. on Computer Vision*, pages 286–295, 2021. 3

[19] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 6447–6456, 2017. 2

[20] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[21] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proc. of European Conf. on Computer Vision*, pages 247–263, 2018. 3

[22] J-M Valin, François Michaud, and Jean Rouat. Robust 3D localization and tracking of sound sources using beamforming and particle filtering. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, pages 134–138, May 2006. 2

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of Int. Conf. on Neural Information Proc. Systems*, 2017. 2

[24] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proc. of ACM Int. Conf. on Multimedia*, pages 154–162, 2017. 1

[25] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2CLIP: Learning robust audio representations from clip. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, pages 4563–4567, 2022. 1

[26] En Yu, Jianhua Ma, Jiande Sun, Xiaojun Chang, Huaxiang Zhang, and Alexander G Hauptmann. Deep discrete cross-modal hashing with multiple supervision. *Neurocomputing*, 486:215–224, 2022. 1

[27] En Yu, Jiande Sun, Jing Li, Xiaojun Chang, Xian-Hua Han, and Alexander G Hauptmann. Adaptive semi-supervised feature selection for cross-modal retrieval. *IEEE Trans. on Multimedia*, 21(5):1276–1288, 2018. 1

[28] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 1

[29] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019. 4

[30] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014. 3

[31] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 3550–3558, 2018. 3