



Pull It Together: Reducing the Modality Gap in Contrastive Learning

Amit Sofer¹, Yoav Goldman¹, Shlomo E. Chazan¹

¹OriginAI, Israel

amits@originai.co, yoavg@originai.co, shlomi@originai.co

Abstract

Contrastive learning has become a powerful strategy for aligning different modalities in a shared embedding space. Contrastive Language–Image Pre-training (CLIP) has achieved remarkable performance across various downstream tasks. This methodology has been extended to the audio-text domain through Contrastive Language–Audio Pre-training (CLAP), demonstrating strong performance in related tasks. However, recent work highlights a modality gap in CLIP’s embedding space, where embeddings from different modalities remain partially separated rather than fully integrated. In this paper, we begin by analyzing the CLAP embedding space and identify a similar modality gap. Furthermore, we propose a novel solution combining a modality classifier with a Gradient Reverse Layer (GRL) to reduce this gap. Our experiments on CLIP and CLAP confirm that our approach reduces the modality gap while improving performance, and even achieving new State Of The Art (SOTA) results in text-audio retrieval.

Index Terms: multi modal, CLAP, CLIP, modality gap

1. Introduction

CLIP [1] has become one of the most influential multimodal foundation models, particularly for vision and language tasks. By contrastively aligning images and text in a shared embedding space, CLIP enables strong zero-shot classification, cross-modal retrieval, and provides a robust starting point for further fine-tuning on downstream tasks.

At its core, CLIP employs contrastive learning to bridge the gap between visual and textual data. Each modality is processed by a dedicated encoder, mapping them into a unified semantic space. Training on a massive dataset of image-text pairs, the model learns to associate corresponding pairs while distinguishing unrelated ones, effectively capturing meaningful cross-modal relationships.

Following the success in the image-text domain, the research was expanded to the audio-text domain by [2–4]. However, performance was constrained by the relatively limited amount of training data, which was significantly smaller than the vast image-text pairs used in CLIP.

The C-CLAPA model [5] addressed this challenge by incorporating extensive data augmentation techniques, including text augmentation using Large Language Models (LLMs), audio augmentations, and pair augmentations, which artificially increased effective training data. In addition, a captioning decoder was added during training, providing a form of regularization and further boosting retrieval performance. These strategies allowed C-CLAPA to achieve SOTA results.

Contrastive learning in models like CLIP and CLAP is designed to align different modalities into a shared embedding

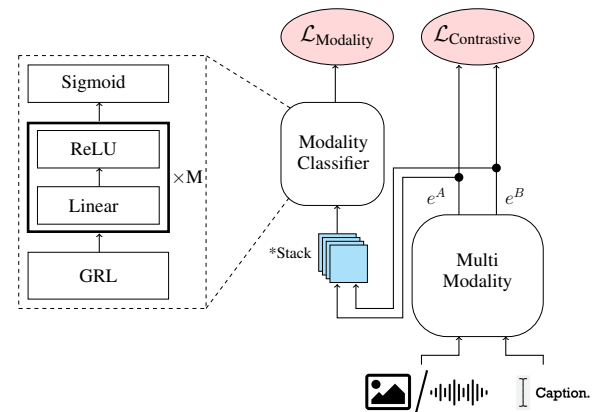


Figure 1: *Our proposed method. The text and audio embeddings (or text and image in the case of CLIP) are passed into a Modality classifier which predicts if the embedding is from text or an audio. The *Stack operation indicate a stacking of both modalities embeddings in the batch dimension.*

space. The primary goal is to ensure that semantically related pairs are mapped to nearby points in this space, while unrelated pairs are pushed apart. This shared embedding space enables cross-modal retrieval, zero-shot learning, and improved generalization across modalities. Ideally, the optimal structure of this embedding space should be such that the embeddings of different modalities, should occupy the same region, potentially forming clusters corresponding to specific topics (e.g. animals, household items, people, etc...).

However, recent observations on the shared embedding space showed that a gap exists and the modalities are not perfectly aligned. This phenomenon, where embeddings from different modalities occupy distinct, non-overlapping regions, referred to as the modality gap, was first identified in [6]. Previous work has been conducted to investigate and bridge the modality gap in CLIP [7, 8] by analyzing the structural differences between text and image embeddings. Different strategies for improving multi-modal alignment were explored, including alignment-based regularization techniques and architectural modifications. While eliminating the gap entirely did not always lead to performance gains, addressing it often resulted in smoother embedding spaces and improved downstream task performance.

The GRL [9] has been widely used in domain adaptation [10–12] to address distribution shifts between a source and target domain. In this setting, GRL helps align feature representations across domains by inverting the gradients of a domain

classifier, encouraging the feature extractor to produce domain-invariant features.

In this study, we extend the analysis of the modality gap to the CLAP model, demonstrating that this gap is even more pronounced in the audio-text embedding space. To address this issue, we propose a novel approach that leverages a GRL to enhance cross-modal alignment. Specifically, we use GRL to reduce the modality gap. By applying GRL to a modality classifier, we actively force the encoders to generate embeddings that are indistinguishable with respect to their originating modality, leading to improved cross-modal alignment. Unlike domain adaptation scenarios, where GRL aligns different distributions of the same modality, our approach aligns fundamentally different data types, such as text and audio, or text and images. Our method effectively reduces the modality gap, leading to consistent improvements in embedding alignment across modalities. Through extensive experiments on CLAP (audio-text) and CLIP (image-text), we validate our approach and achieve SOTA performance in text-audio retrieval tasks. Notably, our method introduces minimal computational overhead during training and imposes no additional cost at inference time.

Our key contributions are as follows:

- We extend the analysis of the modality gap to the CLAP embedding space and demonstrate that this gap is not only present but is even more pronounced in the audio-text embedding space compared to the CLIP embedding space.
- We propose a novel approach leveraging a GRL to mitigate the modality gap, introducing minimal computational overhead during training and no additional cost during inference.
- We show that our method effectively enhances cross-modal alignment and improves performance across different multimodal settings.

2. Contrastive learning

While contrastive learning predates CLIP, the contrastive learning paradigm as popularized today—particularly through the use of the InfoNCE loss—was first established in its current form by CLIP, introduced in [1], and has since become the foundation for models like CLIP and CLAP. Given a batch of N pairs (A_n, B_n) , $n = 0, \dots, N-1$, where A_n, B_n represent the n -th sample pair from modality A and modality B, respectively, the goal is to learn which of the $N \times N$ possible pairs are correctly aligned. Essentially, \mathbf{f}_A and \mathbf{f}_B are the embedding functions (encoders) that map each modality into a shared representation space, and $e_n^A = \mathbf{f}_A(A_n)$ and $e_n^B = \mathbf{f}_B(B_n)$ are the resulting embeddings of (A_n, B_n) , respectively. The objective of contrastive learning is to maximize the similarity (commonly cosine similarity) between the two modalities embeddings for the matching N pairs in the batch while minimizing the similarity for the other unrelated $N^2 - N$ pairs.

The contrastive learning objective consists of two symmetric parts: (1) aligning modality A with its corresponding modality B, and (2) aligning the modality B with its correct modality A. This bidirectional objective is captured by the following loss function:

$$\mathcal{L}_{\text{Contrastive}} = \frac{1}{2}\mathcal{L}_{A \rightarrow B} + \frac{1}{2}\mathcal{L}_{B \rightarrow A} \quad (1)$$

where, $\mathcal{L}_{A \rightarrow B}$ is defined as:

$$\mathcal{L}_{A \rightarrow B} = -\frac{1}{N} \sum_{n=0}^{N-1} \log \frac{\exp(e_n^A \cdot e_n^B / \tau)}{\sum_{\tilde{n}=0}^{N-1} \exp(e_n^A \cdot e_{\tilde{n}}^B / \tau)}, \quad (2)$$

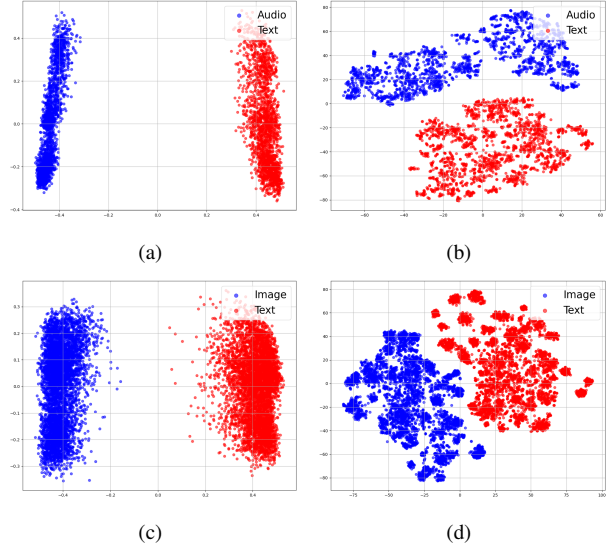


Figure 2: Visualization of PCA and t-SNE projections for audio and image datasets. For CLAP: (a) PCA and (b) t-SNE projections of the AudioCaps and Clotho validation sets. For CLIP: (c) PCA and (d) t-SNE projections of the MSCOCO set.

where the parameter τ is a learned temperature that scales the similarities, controlling the sharpness of the alignment scores between embeddings.

Similarly, $\mathcal{L}_{B \rightarrow A}$, is given by:

$$\mathcal{L}_{B \rightarrow A} = -\frac{1}{N} \sum_{n=0}^{N-1} \log \frac{\exp(e_n^A \cdot e_n^B / \tau)}{\sum_{\tilde{n}=0}^{N-1} \exp(e_{\tilde{n}}^B \cdot e_n^A / \tau)}. \quad (3)$$

This method ensures that corresponding pairs move closer while non-matching pairs move apart, and it can be applied to any two modalities.

3. Modality Gap

It is commonly assumed that multimodal embeddings form a homogeneous space where data from different modalities naturally cluster according to semantic content. In an ideal scenario, embeddings for related concepts (e.g., animals or household items) from different modalities would overlap significantly. However, an analysis of the latent space using metrics such as the Silhouette Score [13] and Euclidean distance from the modality cluster centroid reveals a presence of a modality bias. This bias is further evident when visualizing the embeddings in a lower-dimensional space using Principal Component Analysis (PCA), T-distributed Stochastic Neighbor Embedding (TSNE), or a Uniform Manifold Approximation and Projection (UMAP), where it becomes clear that the intra-modality distance exceeds the inter-modality distance. A PCA and TSNE visualization of the CLIP embedding space on the MSCOCO [14] validation-set are depicted in Figure 2c and 2d, respectively. Evidently, the different modalities are noticeably separated in the embedding space.

Following [6], we define the modality gap Δ_{gap} as the euclidean distance between the mean of each modality cluster:

$$\Delta_{\text{gap}} = \frac{1}{D} \sum_{d=0}^{D-1} e_d^A - \frac{1}{D} \sum_{d=0}^{D-1} e_d^B \quad (4)$$

where D is the number all pairs in the dataset.

According to [6], the two main reasons that contribute to this phenomenon are related to model initialization and the contrastive learning optimization. First, the models often initialize with representations confined to a narrow cone in the embedding space. This inherent bias leads to distinct initial separations between modalities, as each encoder (e.g., for images and text) starts with embeddings concentrated in different sub-regions. Second, during training, the contrastive learning objective, which aims to maximize the similarity between matched pairs and minimize it between unmatched pairs, tends to maintain and even reinforce the initial separation. The degree of this separation is influenced by factors such as the learned temperature parameter in the loss function.

These combined factors result in a persistent modality gap, where embeddings from different modalities remain distinct within the shared representation space.

In a previous work, [7] initially attempted to directly optimize the modality gap function to achieve exact alignment between modalities in the latent space. However, this approach did not lead to improved results. Consequently, they argued that such alignment is suboptimal for downstream tasks and shifted their focus toward constructing better latent modality structures.

In [8], the authors addressed the modality gap in CLIP’s embedding space by introducing parameter sharing between modality-specific encoders and implementing intra-modality separation. These architectural modifications enhanced cross-modal alignment, leading to improved performance in tasks such as zero-shot image classification and multi-modal retrieval.

The goal of this work is to close the gap in an elegant way, without changing the architecture or the contrastive loss, and still improve the bidirectional retrieval performances.

4. Proposed method

While previous work has primarily examined the modality gap in image-text embeddings, our first objective is to investigate whether a similar gap exists in the audio-text domain using the CLAP model. Subsequently, we introduce a novel approach to address this challenge.

4.1. Audio-Text modality gap

Our first objective is to examine whether a similar modality gap exists in this domain. To this end, we utilize the C-CLAPA model and construct the embedding space using the AudioCaps [15] and Clotho [16] test sets. To visualize the modality gap, we employ PCA and t-SNE techniques. Figures 2a and 2b illustrate the PCA and t-SNE visualizations of the C-CLAPA model on these test sets, revealing that CLAP also exhibits a modality gap. The modality gap in the CLIP embedding space is measured at 0.822, while in the C-CLAPA embedding space, it is even larger at 0.871. Furthermore, PCA visualizations for both CLIP and CLAP reveal that the first principal component, which captures the greatest variance, primarily serves to separate the modalities. In CLAP’s PCA visualization, this first principal component is almost entirely devoted to distinguishing the two modalities, with minimal variation within each modality. In contrast, CLIP’s first component retains some degree of inter-modality variation. These results suggest that CLAP exhibits an even more pronounced modality gap than CLIP. Motivated by these findings, we aim to develop a solution to mitigate this gap, which may also improve the model’s downstream perfor-

mance.

4.2. The GRL-based gap reducer

Inspired by work in domain adaptation, where the use of a GRL has led to numerous breakthroughs [9–12] [17], we adopt a similar strategy for addressing the modality gap. In domain adaptation, GRLs are employed to close the distribution shift between source and target domains by making their latent embeddings indistinguishable to a classifier.

Similarly, we propose using a GRL to make the embeddings of the two modalities (e.g., text and audio) indistinguishable in the shared embedding space. By employing a simple Multi-Layer Perceptron (MLP) as a modality classifier to predict which modality an embedding comes from, and then applying the GRL, the encoders of both modalities are encouraged to produce embeddings that are closer to each other in the latent space. The Binary Cross Entropy (BCE) loss is used as the modality classification loss, and the overall loss of the model is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{Contrastive}} + \lambda \cdot \mathcal{L}_{\text{Modality}} \quad (5)$$

Here, $\mathcal{L}_{\text{Modality}}$ represents the BCE loss applied to the modality classifier, with gradients reversed to the encoders via the GRL. The parameter λ controls the weight of the modality loss in the overall objective. $\mathcal{L}_{\text{Contrastive}}$ can represent the contrastive loss for CLIP ($\mathcal{L}_{\text{CLIP}}$), CLAP ($\mathcal{L}_{\text{CLAP}}$), or C-CLAPA ($\mathcal{L}_{\text{C-CLAPA}}$), with the latter incorporating both contrastive loss and the captioning decoder loss as described in [5]. In practice, we take the N text embeddings, and N audio embeddings and stack them at the batch dimension. This is the input to the modality classifier, as shown at figure 1. This classifier purpose is to take the embeddings of both the audio and text, and classify them according to the modality type. the first layer of this classifier is the GRL, which does nothing in the forward pass of the model, but reverses the gradients entering both encoders, text and audio (or image in the case of CLIP).

5. Experimental set up

For the modality classifier, we use a $M = 5$ layer MLP, with a ReLU activation between each linear layer, a GRL at the start of the MLP and a sigmoid at the end. We test 4 different values of λ , 0, 0.5, 1 and 3. with 0 meaning without $\mathcal{L}_{\text{Modality}}$.

5.1. Audio-Text

As a CLAP model we used the C-CLAPA [5] (approximately 220M parameters), with a learning rate of 5×10^{-3} , SGD [18] optimizer, using a batch size of 110 per device on 8 A100 GPUs, yielding a total batch size of 880. The only modification was adding the modality classifier and the GRL layer. We trained the model on the training datasets described in [5].

We test our model on the text to audio and audio to text retrieval on both the AudioCaps and Clotho test sets.

5.2. Image-Text

Using the OpenCLIP¹ framework, we trained a CLIP model which consists of a ResNet50-based image encoder and a transformer-based text encoder (total of 102M parameters) on the CC12M [19] dataset for 60 epochs. The training was done with a learning rate of 5×10^{-4} , AdamW [20] optimizer, and a weight decay of 0.2, using a batch size of 512

¹https://github.com/mlfoundations/open_clip

Table 1: Cross Domain Retrieval (CDR) results on both text to audio(T-A) and audio to text(A-T) retrieval on the Clotho and AudioCaps test sets.

Model	AudioCaps Eval						Clotho Eval						T-A Avg	A-T Avg
	T-A Retrieval			A-T Retrieval			T-A Retrieval			A-T Retrieval				
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10		
CLAP [3]	34.6	70.2	82.0	41.9	73.1	84.6	16.7	41.1	54.1	20.0	44.9	58.7	49.8	53.9
CLAP [4]	35.1	71.9	83.7	44.2	80.8	90.3	16.9	41.6	54.4	24.4	49.3	65.7	50.6	59.1
CLAP [5]	41.6	76.7	87.7	52.8	82.9	91.1	21.9	48.0	62.2	25.4	52.6	64.7	56.4	61.6
CLAP-GRL ($\lambda = 1$)	42.7	77.7	87.9	56.6	84.4	92.6	22.3	49.5	62.9	26.6	53.0	67.7	57.2	63.5

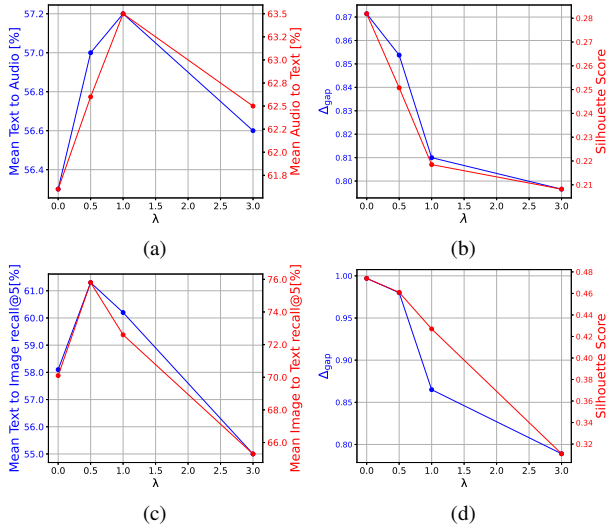


Figure 3: Mean retrieval results, modality gap Δ_{gap} , and Silhouette Score as a function of λ (a) CLAP mean retrieval results. (b) CLAP modality gap and Silhouette Score. (c) CLIP mean retrieval results. (d) CLIP modality gap and Silhouette Score.

per device on 8 A100 GPUs, yielding a total batch size of 4096. We tested the checkpoint with the lowest validation loss on the CC12M validation set. The model was tested on MSCOCO [14], Flickr8K [21] and Flickr30K [22] datasets.

6. Results

6.1. Audio-Text

Table 1 presents the retrieval results on the Clotho and AudioCaps datasets. Our method improves performance on both text-to-audio and audio-to-text retrieval tasks and achieves new SOTA results.

To investigate the impact of the λ parameter, the modality classifier loss weight during training, we conducted additional experiments. Figure 3a shows the mean retrieval scores (text-to-audio and audio-to-text) as a function of λ , where we observe that the best performance is achieved at $\lambda = 1$.

In Figure 3b, we plot the modality gap and the Silhouette Score² as a function of λ . It is clear that increasing λ reduces the modality gap, although this does not always guarantee improved performance.

By analyzing these two curves, we observe that reducing the modality gap using the λ parameter doesn't behaves linearly,

²Silhouette Score measures the clustering quality by evaluating how similar a data point is to its own cluster compared to other clusters, ranging from -1 (poor clustering) to 1 (well-clustered)

so while $\lambda=1$ reduces the gap by 0.06, increasing the value to $\lambda=3$ becomes more challenging and only reduces the gap by 0.08 while also leading to a decline in downstream task performance.

Table 2: CDR results on both image-to-text (I-T) and text-to-image (T-I) recall@5 retrieval on the MSCOCO, Flickr8K and Flickr30K test sets.

Model	MSCOCO		Flickr8K		Flickr30K		T-I Avg	I-T Avg
	T-I	I-T	T-I	I-T	T-I	I-T		
CLIP	43.7	56.9	64.3	77.1	66.5	76.5	58.1	70.1
CLIP GRL ($\lambda = 0.5$)	45.2	61.8	67.4	82.9	71.3	82.9	61.3	75.8

6.2. Image-Text

It is important to note that due to computational constraints we applied our method on small-scale CLIP model which was trained on smaller amount of data compared to SOTA CLIP model.

We trained both the baseline CLIP model and a CLIP model using our GRL modality classifier. The results are summarized in table 2. Similar ablation on the effect of the λ parameter in the Image-Text domain was made. As can be seen in figure 3c, the mean cross-modal retrieval over all test datasets (MSCOCO, Flickr8K and Flickr30K) was improved, getting optimal results when $\lambda = 0.5$. In figure 3d we plot the modality gap and the Silhouette accordingly.

7. Conclusions

In this paper, we introduced a novel approach for mitigating the modality gap in multi-modal embedding space using a GRL and a modality classifier. By applying the GRL, we significantly reduced this gap without adding any computational overhead during inference. Furthermore, in the Audio-Text domain our method improved cross-modal retrieval performance, achieving new SOTA results for both text-to-audio and audio-to-text retrieval on the AudioCaps and Clotho datasets. Furthermore, in Image-Text domain our method improves cross-modal retrieval performance while reducing the modality gap when evaluated on MSCOCO, Flickr8K and Flickr30K. The ablation study further confirmed the positive impact of the GRL and modality classifier on both retrieval performance and modality alignment. Future work could explore refining the modality alignment technique, possibly integrating it with other domain adaptation methods. Additionally, applying this method to other multi-modal models could provide broader insights on the generalization of the method. Ultimately, our approach represents a step forward in improving the robustness and effectiveness of multi-modal contrastive learning.

8. References

- [1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [2] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, “Clap learning audio concepts from natural language supervision,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [3] S. Deshmukh, B. Elizalde, and H. Wang, “Audio retrieval with wavtext5k and clap training,” *arXiv preprint arXiv:2209.14275*, 2022.
- [4] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [5] A. Sofer and S. E. Chazan, “C-clapa: Improving text-audio cross domain retrieval with captioning and augmentations,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8040–8044.
- [6] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 612–17 625, 2022.
- [7] Q. Jiang, C. Chen, H. Zhao, L. Chen, Q. Ping, S. D. Tran, Y. Xu, B. Zeng, and T. Chilimbi, “Understanding and constructing latent modality structures in multi-modal representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7661–7671.
- [8] S. Eslami and G. de Melo, “Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip,” *arXiv preprint arXiv:2406.17639*, 2024.
- [9] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.
- [10] C. S. Anoop, A. Prathosh, and A. Ramakrishnan, “Unsupervised domain adaptation schemes for building asr in low-resource languages,” in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 342–349.
- [11] K. Osumi, T. Yamashita, and H. Fujiyoshi, “Domain adaptation using a gradient reversal layer with instance weighting,” in *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE, 2019, pp. 1–5.
- [12] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, “Unsupervised domain adaptation via domain adversarial training for speaker recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4889–4893.
- [13] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
- [14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
- [15] 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.
- [16] 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.
- [17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, “Domain-adversarial training of neural networks,” *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.
- [18] H. Robbins and S. Monro, “A stochastic approximation method,” *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.
- [19] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, “Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3558–3568.
- [20] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2019.
- [21] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.
- [22] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.