DREAM: Decoupled Discriminative Learning with Bigraph-aware Alignment for Semi-supervised 2D-3D Cross-modal Retrieval

Fan Zhang[1], Changhu Wang[2], Zebang Cheng[3], Xiaojiang Peng[3], Dongjie Wang[4], Yijia Xiao[5], Chong Chen[6], Xian-Sheng Hua[6], Xiao Luo[5]*

[1]Department of Electrical and Computer Engineering, Georgia Institute of Technology, USA
[2]Department of Statistics, University of California, Los Angeles, USA
[3]College of Big Data and Internet, Shenzhen Technology University, China
[4]Department of Electrical Engineering and Computer Science, University of Kansas, USA
[5]Department of Computer Science, University of California, Los Angeles, USA
[6]Terminus Group, China
{fanzhang.karl, huaxiansheng, chenchong.cz, zebang.cheng}@gmail.com, wangch156@g.ucla.edu
pengxiaojiang@sztu.edu.cn, wangdongjie@ku.edu, {yijia.xiao, xiaoluo}@cs.ucla.edu

Abstract

With the burst of big data, 2D-3D cross-modal retrieval has received increasing attention, which targets at retrieving relevant data of one modality using the query of the other modality. In this paper, we study an underexplored yet applicable problem of semi-supervised 2D-3D cross-modal retrieval, which could suffer from serious label scarcity in real-world applications. Moreover, the huge heterogeneous gap could deteriorate the process of learning from unlabeled data. In this work, we propose a novel approach named Decoupled Discriminative Learning with Bigraph-aware Alignment (DREAM) for semi-supervised 2D-3D cross-modal retrieval. The core of our DREAM is to decouple the label prediction and reliability measurement processes to reduce overconfident samples in discriminative learning. In particular, we enhance a label prediction module with label propagation from labeled samples and additionally introduce a reliability measurement module to learn the scores of predicted labels. To reduce class-related bias, we compare reliability scores with class-specific adaptive thresholds to identify samples for additional learning. In addition, negative labels are estimated for unselected samples, which guides soft semantic learning to make the best use of all the information. To further minimize the heterogeneous gap, we build a bigraph graph that connects cross-modal similar examples and then conduct learning to cluster with most edges kept for alignment. Extensive experiments on several benchmark datasets validate the superiority of the proposed DREAM against many state-of-the-art baselines.

Introduction

3D visual understanding has demonstrated increasing significance as 3D data such as point clouds can effectively capture internal spatial architecture (Song et al. 2023). As a crucial challenge, 3D retrieval (Yang et al. 2019) aims to retrieve relevant 3D shape samples given a query sample. Early research primarily focused on single-modal retrieval, which maps 3D shapes into an embedding space while preserving the similarity structure. However, with the exponential growth of multimodal data from artificial intelligence generated content (AIGC) (Rombach et al. 2022; Peebles and Xie 2023; Zhu et al. 2018; Achlioptas et al. 2018; Vahdat et al. 2022; Luo and Hu 2021), 2D-3D cross-modal retrieval has captured rising attention, which seeks to retrieve samples from one modality using a query from another modality and thus opening up new opportunities for seamless integration of heterogeneous data sources (Lin et al. 2021; Nie et al. 2023). Recent 2D-3D cross-modal retrieval approaches (Li et al. 2015; Pham et al. 2020; Liu et al. 2022; Jing et al. 2021; Feng et al. 2023; Xu et al. 2022; Xue et al. 2023) typically involve mapping both 2D and 3D data into a shared hidden space and then calculating similarity scores for ranking and retrieval. Although they have achieved some progress, these retrieval approaches are generally data-hungry, requiring a large number of labeled samples for end-to-end training. However, obtaining labels can be prohibitively expensive and time-consuming, as it demands extensive human labor for annotation. Moreover, accurate annotation of 3D data becomes particularly challenging when related textual descriptions are absent (Jaritz et al. 2022; Jain et al. 2022). Note that AIGC has the capability to produce a vast amount of unlabeled 2D and 3D data without annotation, which is common in practice. In light of this, we focus on the applicable problem of semi-supervised 2D-3D cross-modal retrieval (Zhang et al. 2024b).

However, developing an effective framework for this problem is highly challenging, since it requires addressing two critical questions. (1) *How to effectively learn from 2D and 3D data under label scarcity?* Previous works typically (Hu et al. 2017; Sohn et al. 2020; Zhang et al. 2021) employ pseudo-labeling strategies to annotate unlabeled data using the model itself for discriminative learning. However, these approaches often utilize the same module for label generation and selection, which carries a high risk of overfitting and error accumulation. (2) *How to reduce the discrepancy in the embedding space across different modalities?* Recent works (Jing et al. 2021; Feng et al. 2023) commonly map 2D and 3D data into a shared hidden space by pairwise labels or class-wise anchors. However, these techniques cannot measure the intricate semantics gap of different modalities, resulting in

*Corresponding author.

significant discrepancies in multimodal representations.

To solve the aforementioned challenges, we propose a new approach named Decoupled Discriminative Learning with Bigraph-aware Alignment (DREAM) for semi-supervised 2D-3D cross-modal retrieval. The core of DREAM is to decouple the label prediction and reliability measurement processes from a learning-to-selectively-learn perspective, which could effectively filter unreliable examples. Specifically, our DREAM has a propagation-enhanced label prediction module, which combines labeled samples and the affinity across different representations for accurate predictions. More importantly, DREAM introduces a learnable reliability measurement module, which aggregates both samples and predictions to output the reliability scores. The measurement module is supervised by the similarity between predictions and ground truth in the training data and unlabeled data with scores over adaptive class-specific thresholds is identified as extra labeled data. To make the best of unselected data, we introduce multimodal negative learning, which identifies negative labels by comparing cross-modal pairs and then minimizing their log-likelihood for soft supervision. In addition, to promote the modality invariance, we introduce bigraph-aware modality alignment, which first constructs a bipartite graph with similar cross-modal examples connected, and then conducts learning to cluster with the maximum cross-modal edges kept. Extensive experiments on benchmark datasets validate the superiority of our DREAM in comparison to various state-of-the-art methods.

The contribution of this work can be summarized as follows: (1) *New Perspective.* We pioneer a new learning-to-selectively-learn perspective that decouples the label prediction and reliability measurement processes for semi-supervised 2D-3D cross-modal retrieval. (2) *Novel Methodology.* Our DREAM not only introduces a learnable reliability measurement module that fuses deep representations and label embeddings to output the reliability scores, but also conducts learning to cluster on cross-modal bigraph for semantics-aware modality alignment. (3) *Comprehensive Experiments.* Extensive experiments across multiple datasets validate that DREAM is superior to various baselines.

Related Work

Cross-modal Retrieval. To build the connection across diverse modalities, cross-modal retrieval has received extensive attention (Zhen et al. 2019; Hu et al. 2023; Qian et al. 2022; Zhu et al. 2022; Kim et al. 2023; Jing et al. 2021). Existing methods can be roughly divided into two groups: local approaches and global approaches. Local approaches (Lee et al. 2018; Li et al. 2019; Dong et al. 2019; Hu et al. 2023) explore fine-grained semantics at the token and patch levels. In contrast, global approaches (Liu et al. 2020; Diao et al. 2021; Li et al. 2023a; Zhang et al. 2024a) map data from both modalities into the same embedding space, which is more efficient in practice. However, most of the previous approaches are data-hungry with the requirement of extensive label annotation, which could be unavailable in real-world applications (Jing et al. 2021; Feng et al. 2023). Towards this end, we explore the task of semi-supervised cross-modal retrieval to ease the practical label scarcity issue.

Semi-supervised Learning. Semi-supervised learning aims to leverage abundant unlabeled samples to reduce the cost of label annotation (Wu et al. 2023; Assran et al. 2021; Li et al. 2023b; Chen et al. 2023b; Li et al. 2022). It has achieved extensive progress in different areas including image segmentation (Qiao et al. 2023; Chen et al. 2023a) and object detection (Hua et al. 2023; Zhang et al. 2023). Pseudo-labeling (Berthelot et al. 2019; Hu et al. 2021b) and consistency regularization (Miyato et al. 2018; Xie et al. 2020) are two mainstream strategies in semi-supervised learning. Pseudo-labeling assigns unlabeled samples with soft or hard pseudo labels for dataset extension while consistency regularization minimizes the discrepancy when adding various perturbations. In contrast to previous works (Sohn et al. 2020; Zhang et al. 2021; Garg et al. 2022) which often utilize the model itself to annotate extensive unlabeled samples, DREAM decouples the label prediction and reliability measurement processes, which can effectively filter overconfident samples for reliable discriminative learning.

The Proposed DREAM

Problem Definition. Let $\mathcal{D}^{2d} = \mathcal{D}^{2d,l} \cup \mathcal{D}^{2d,u}$ denote a dataset containing 2D labeled examples $\mathcal{D}^{2d,l} = \{(\boldsymbol{x}_i^{2d,l}, y_i^{2d,l})\}_{i=1}^{N^{2d,l}}$ and unlabeled examples $\mathcal{D}^{2d,u} = \{(\boldsymbol{x}_i^{2d,u})\}_{i=1}^{N^{2d,u}}$. Let $\mathcal{D}^{3d} = \mathcal{D}^{3d,l} \cup \mathcal{D}^{3d,u}$ denote a dataset containing 3D labeled examples $\mathcal{D}^{3d,l} = \{(\boldsymbol{x}_j^{3d,l}, y_j^{3d,l})\}_{j=1}^{N^{3d,l}}$ and unlabeled examples $\mathcal{D}^{3d,u} = \{(\boldsymbol{x}_j^{3d,u})\}_{j=1}^{N^{3d,u}}\}$. y_i^{2d} and y_j^{3d} are semantic labels for 2D and 3D samples from C common classes. N is the total sample number and N^{2d} and N^{3d} denote the total number of 2D and 3D samples, respectively. To bridge 2D and 3D modalities, our target is to map multimodal samples into a shared embedding space with the heterogeneous gap minimized.

Framework Overview. In this paper, we study the problem of semi-supervised 2D-3D cross-modal retrieval, which is highly complicated because of serious label scarcity and the heterogeneous gap. To solve this problem, we propose a novel approach DREAM, which maps multimodal data into a shared embedding space as $\boldsymbol{h}_i^{2d} = \phi^{2d}(\boldsymbol{x}_i^{2d}), \boldsymbol{h}_j^{3d} = \phi^{3d}(\boldsymbol{x}_j^{3d})$ where $\phi^{2d}(\cdot)$ and $\phi^{3d}(\cdot)$ are two separate encoders for representation learning, respectively. $\boldsymbol{x}_i^{2d}$ and $\boldsymbol{x}_j^{3d}$ denotes 2D and 3D inputs with output $\boldsymbol{h}_i^{2d}$ and $\boldsymbol{h}_j^{3d}$. Our DREAM mainly consists of two parts: (1) *Decoupled Discriminative Learning*, which consists of a label prediction module and a reliability measurement module to provide reliable samples with accurate labeling for additional training. Furthermore, DREAM utilizes multimodal negative learning to make use of unselected data in a soft manner. (2) *Bigraph-aware Modality Alignment*, which constructs a bigraph connecting similar cross-modal graphs and then conducts learning to cluster with the maximum edges kept effective cross-modal retrieval. More details can be found in Figure 1.

Decoupled Discriminative Learning for Multimodal Representations

The major challenge is label scarcity in real-world applications. Previous methods (Sohn et al. 2020; Yang et al. 2023;

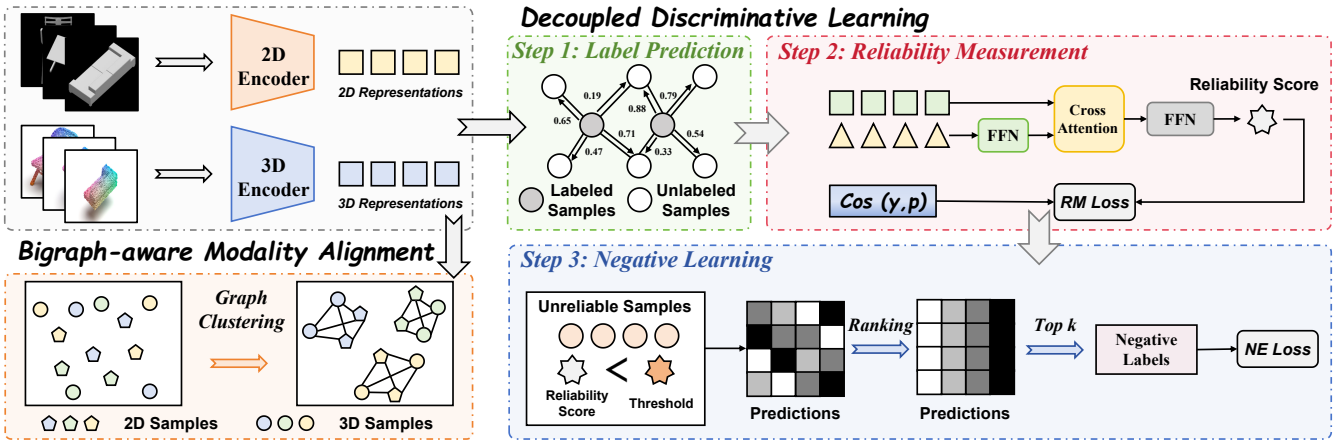


Figure 1: An overview of our DREAM. DREAM first employs separate encoders to obtain multimodal representations, and then adopts decoupled discriminative learning to obtain reliability scores for optimization in both semantic learning and negative learning. Moreover, DREAM constructs a bipartite graph for multimodal semantics alignment with discriminability preserved.

Saberi et al. 2024) usually adopt pseudo-labeling, which first uses the model itself to select data with high confidence and then adds them into labeled datasets. Unfortunately, a single model for both labeling and selection could easily produce overconfident results, resulting in serious error accumulation. To tackle this issue, we provide a new perspective of learning to selectively learn. On this basis, we decouple the label prediction and reliability measurement using separate modules and then compare the reliability scores with adaptive class-specific thresholds for high-quality unlabeled samples.

To be specific, we first leverage labeled data for discriminative learning by mapping labels into the embedding space, i.e., $\{\boldsymbol{f}_1, \cdots, \boldsymbol{f}_C\}$ (Wang et al. 2023). Then, we encourage each labeled sample to approach their corresponding anchors:

$$\mathcal{L}_S = \sum_{i=1}^{N^{2d,l}} \frac{\exp\left({\boldsymbol{h}_i^{2d,l}}^\top \boldsymbol{f}_{y_i^{2d,l}}\right)}{\sum_{c=1}^{C} \exp\left({\boldsymbol{h}_i^{2d,l}}^\top \boldsymbol{f}_c\right)} + \sum_{j=1}^{N^{3d,l}} \frac{\exp\left({\boldsymbol{h}_j^{3d,l}}^\top \boldsymbol{f}_{y_j^{3d,l}}\right)}{\sum_{c=1}^{C} \exp\left({\boldsymbol{h}_j^{3d,l}}^\top \boldsymbol{f}_c\right)}, \tag{1}$$

where $\boldsymbol{h}_i^{2d,l}$ and $\boldsymbol{h}_j^{3d,l}$ are representations of $\boldsymbol{x}_i^{2d,l}$ and $\boldsymbol{x}_j^{3d,l}$.

Label Prediction Module. In the scenario of label scarcity, we aim to make use of abundant unlabeled data for discriminative multimodal representations. An intuitive solution is to assign each unlabeled sample $\boldsymbol{x}_i$[1] with a pseudo-label:

$$\hat{y}_i = argmax_c[\boldsymbol{h}_i^T \boldsymbol{f}_c], \tag{2}$$

where $\boldsymbol{h}_i$ denotes the representation for an 2D or 3D sample. However, when labeled data is scarce, these pseudo-labels could be inaccurate due to insufficient training. To tackle this, we utilize label propagation (Iscen et al. 2019; Sun et al. 2024; Zhu and Ghahramani 2002) to transfer the label information from labeled data toward unlabeled data from a global view. Here, we compare the distance between different samples to generate a transition matrix $\boldsymbol{T} \in [0,1]^{N^u \times N}$ where N^u denotes the number of unlabeled samples:

$$\boldsymbol{T}_{im} = \frac{\exp(||\boldsymbol{h}_i - \boldsymbol{h}_m||/\sigma^2)}{\sum_{m=1}^{N_u} \exp(||\boldsymbol{h}_i - \boldsymbol{h}_m||/\sigma^2)}, \tag{3}$$

where σ^2 is a parameter to regularize the distance. Moreover, we construct the label matrix $\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}^l \\ \boldsymbol{P}^u \end{bmatrix} \in [0,1]^{N \times C}$ by stacking both one-hot label embeddings $\boldsymbol{p}_i^l$ and predicted distributions $\boldsymbol{p}_i^u = [\boldsymbol{h}_i^T \boldsymbol{f}_1, \cdots, \boldsymbol{h}_i^T \boldsymbol{f}_C]$for labeled and unlabeled samples, respectively. Then, we update the label matrix using the transition matrix till the convergence:

$$\boldsymbol{P}^u \leftarrow \boldsymbol{T}\boldsymbol{P}. \tag{4}$$

In this way, we smoothly transform semantic information from labeled data to unlabeled data. Compared with pointwise pseudo-labeling, our method can incorporate all these labeled samples explicitly to adjust label predictions.

Reliability Measurement Module. Still, our label prediction module could make mistakes, which should be carefully checked (Li et al. 2023b; Zheng et al. 2022; Wu et al. 2023; Chen et al. 2023b; Li et al. 2022). To effectively select high-quality unlabeled data, we introduce a reliability measurement module, which provides unbiased feedback on labeling. We begin by consolidating a sample along with its predicted label distribution into a unified pair. Subsequently, this pair undergoes evaluation within the measurement module to generate a reliability score. This process can be formulated as $\hat{s}_i = \mathcal{M}(\boldsymbol{x}_i, \boldsymbol{p}_i)$. In particular, we first extract the representation $\boldsymbol{h}_i$ and its estimated label predictions $\boldsymbol{w}_i = \phi^l(\boldsymbol{p}_i)$ using two separate encoders. Then, these embeddings are fused using the cross-attention mechanism:

$$\tilde{\boldsymbol{w}}_i = \text{softmax}\left(\frac{\boldsymbol{W}^Q \boldsymbol{h}_i \cdot [\boldsymbol{W}^K \boldsymbol{w}_i]^T}{\sqrt{d}}\right) \cdot \boldsymbol{W}^V \boldsymbol{w}_i, \tag{5}$$

$$\tilde{\boldsymbol{h}}_i = \text{softmax}\left(\frac{\boldsymbol{W}^{Q'} \boldsymbol{w}_i \cdot [\boldsymbol{W}^{K'} \boldsymbol{h}_i]^T}{\sqrt{d}}\right) \cdot \boldsymbol{W}^{V'} \boldsymbol{h}_i, \tag{6}$$

[1]It can be a 2D or 3D sample with the superscript omitted.

where $\boldsymbol{W}^*$ denotes learnable matrix for feature transformation and d is the hidden dimension of embedding space. Then, we concatenate the fused embeddings and use a feed-forwarding network to output the reliability scores $\hat{s}_i = \psi([\tilde{\boldsymbol{h}}_i, \tilde{\boldsymbol{w}}_i])$ where a large $\hat{s}_i \in [0, 1]$ indicates a more reliable pair $(\boldsymbol{x}_i, \boldsymbol{p}_i)$. Since we do not have any ground truth information of unlabeled data, labeled data is used to train our measurement module. The ground truth reliability score is the cosine similarity between ground truth and the prediction, i.e., $cos(\boldsymbol{y}_i, \boldsymbol{p}_i)$ on the labeled dataset. The loss objective for training the measurement module is written as:

$$\mathcal{L}_{RM} = \sum_{i=1}^{N^l} ||\hat{s}_i - cos(\boldsymbol{y}_i, \boldsymbol{p}_i)||_2^2, \tag{7}$$

where $\boldsymbol{y}_i$ is the one-hot embedding vector of y_i. After generating the reliability scores for unlabeled data, we introduce a class-specific threshold τ_c to promise that $N^u \cdot r/C$ samples have the scores over the threshold for each class in which N^u denotes the number of unlabeled examples and r denotes the ratio of correct predictions. The selected unlabeled data can be collected into the set $\mathcal{S}$ as:

$$\mathcal{S} = \cup_{c=1}^{C} \left\{ \boldsymbol{x}_i \mid c = p_i, \hat{s}_i > \tau_c \right\}. \tag{8}$$

Our reliable sample selection strategy has two features. Firstly, we decouple the labeling and reliability measurement process, which reduces the potential overconfident pseudo-labels and error accumulation by their interaction. Secondly, class-specific thresholds are adopted to further reduce the class imbalance in pseudo-labeling. These two features guarantee the high quality of pseudo-labels.

Optimization. After identifying reliable samples, we utilize the predicted distribution $\boldsymbol{p}_i$ to supervise the optimization along with labeled data. Let p_i denote the label inferred from $\boldsymbol{p}_i$, and the training objective $\mathcal{L}_S$ for these reliable samples and labeled samples can be reformulated as:

$$\mathcal{L}_S = \sum_{\boldsymbol{x}_i \in \mathcal{D}^l \cup \mathcal{S}} \frac{\exp\left(\boldsymbol{h}_i{}^\top \boldsymbol{f}_{p_i}\right)}{\sum_{c=1}^{C} \exp\left(\boldsymbol{h}_i{}^\top \boldsymbol{f}_c\right)}. \tag{9}$$

Negative Learning Module. However, there are still a large number of unreliable samples with abundant semantic information (Chen et al. 2023b; Kim et al. 2019; Cole et al. 2021). They cannot directly provide guidance with their predicted labels due to the potential misleading. Therefore, we introduce multimodal negative learning to learn reliable information softly from these samples, which first estimates the thresholds of negative labels using cross-modal information and minimizes their log-likelihood for semantic learning.

In detail, we first feed each unlabeled cross-modal pair $(\boldsymbol{x}_{i_1}^{2d}, \boldsymbol{x}_{i_2}^{3d})$ into the 2D and 3D encoder respectively, and then generate the predicted distributions, i.e., $\boldsymbol{p}_{i_1}^{2d}$ and $\boldsymbol{p}_{i_2}^{3d}$. Due to the heterogeneous gap, there could be huge differences between them, which are utilized to measure negative labels. To achieve this, we measure the maximal positive ranking for each predicted label (the index of maximal values in the predicted distribution) in their corresponding ranking as the maximal error and then consider the ranking over the threshold as a negative ranking. In formulation, the threshold can be written as:

$$\begin{aligned} t_n = \min_k \{k | \operatorname{argmax}_c(\boldsymbol{p}_{i_1}^{2d}[c]) \in \operatorname{top}_{k_{c'}}(\boldsymbol{p}_{i_2}^{3d}[c']) \\ \wedge \operatorname{argmax}_c(\boldsymbol{p}_{i_2}^{3d}[c]) \in \operatorname{top}_{k_{c'}}(\boldsymbol{p}_{i_1}^{2d}[c'])\}, \end{aligned} \tag{10}$$

where $\operatorname{argmax}(\cdot)$ returns the index of the maximum value in the vector and $\operatorname{top}_k(\cdot)$ returns the index set of the top k values in the vector. Then, we consider all the labels with ranking over the threshold as negative ones and minimize their log-likelihood. In other words, the loss objective for negative learning is written as:

$$\begin{aligned} \mathcal{L}_{NE} = \sum_{c=1}^{C} (\sum_{\boldsymbol{x}_{i_1}^{2d} \in \mathcal{D}^u / \mathcal{S}} 1_{c \notin \operatorname{top}_{t_n} \boldsymbol{p}_{i_1}^{2d}[c']} \log \boldsymbol{p}_{i_1}^{2d}[c] \\ + \sum_{\boldsymbol{x}_{i_2}^{3d} \in \mathcal{D}^u / \mathcal{S}} 1_{c \notin \operatorname{top}_{t_n} \boldsymbol{p}_{i_2}^{3d}[c']} \log \boldsymbol{p}_{i_2}^{3d}[c]), \end{aligned} \tag{11}$$

where $\mathcal{D}^u$ collects all the unlabeled data and $\operatorname{top}_{t_n}(\cdot)$ returns the index set of the top t_n values in the vector. Through multimodal negative learning, we minimize the likelihood of these improbable labels inferred from cross-modal information, which further enhances semantic learning and provides additional weak supervision with extensive uninformative unlabeled data in a safe manner.

Bigraph-aware Modality Alignment

Another crucial task in cross-modal retrieval is to generate modality-invariant representations by minimizing the heterogeneous gap across two modalities (Hwang et al. 2024; Xu et al. 2020; Wang et al. 2017; Zhang et al. 2024c). To solve this task with discriminability considered, we introduce bigraph-aware modality alignment, which first builds a bipartite graph with similar cross-modal pairs connected by edges and then encourages their alignment by learning to cluster.

In particular, each sample is a node in our bipartite graph. Given a unified label $\tilde{y}_i$ from labels or predicted labels ($\tilde{y}_i = y_i$ as a labeled sample and $\tilde{y}_i = p_i$ as an unlabeled sample, we add edges between any cross-modal pair $(\boldsymbol{x}_i^{2d}, \boldsymbol{x}_j^{3d})$ with the adjacency matrix as:

$$\boldsymbol{A}_{i,j} = \begin{cases} 1, & \tilde{y}_i^{2d} = \tilde{y}_j^{3d} \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

In our bigraph, similar cross-modal pairs are connected. Then, we conduct graph clustering on label predictions to keep most of these edges within clusters, which can not only minimize the modality discrepancy, but also enhance the discriminative learning of multimodal representations. On the basis of graph clustering, we minimize the following equation:

$$\mathcal{L}_{BMA} = \left\| (\boldsymbol{I} - \boldsymbol{L}) - \boldsymbol{P}\boldsymbol{P}^T \right\|_F^2, \tag{13}$$

in which $\boldsymbol{I}$ denotes the identity matrix, $\boldsymbol{L}$ denotes the normalized Laplacian matrix of $\boldsymbol{A}$, and $\boldsymbol{P}$ is the label prediction matrix. Eqn. 13 can be rewritten as:

$$\mathcal{L}_{BMA} = -2 \sum_{i,j} \frac{\boldsymbol{A}_{ij}}{\sqrt{d_i}\sqrt{d_j}} \boldsymbol{p}_i^T \boldsymbol{p}_j + \sum_{i,j} \left(\boldsymbol{p}_i^T \boldsymbol{p}_j\right)^2 + CONST, \tag{14}$$

where d_i denotes the degree of $\boldsymbol{x}_i$ in the bigraph. Assuming the class balance, we have $d_i \approx N/C$ and the final alignment loss is written as:

$$\mathcal{L}_{BMA} = -\sum_{\boldsymbol{A}_{ij}=1} \frac{2C}{N} \boldsymbol{p}_i^T \boldsymbol{p}_j + \sum_{i,j} \left(\boldsymbol{p}_i^T \boldsymbol{p}_j\right)^2. \quad (15)$$

Summarization

In a nutshell, the final loss objective is summarized by combining all these losses:

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_{RM} + \alpha \mathcal{L}_{NE} + \beta \mathcal{L}_{BMA}, \quad (16)$$

where α and β are two coefficients to control the weight of the negative learning module and bigraph-aware modality alignment, respectively. During optimization, we first warm up the whole framework by minimizing the supervised objective and then minimize Eqn. 16 gradually.

Theoretical Analysis

In this part, we provide a theoretical analysis to demonstrate how DREAM works. For each unlabeled sample $\boldsymbol{x}_i$, we define $\boldsymbol{y}_i^*$ and p_i as the unknown one-hot embedding vector of the true label y_i^* and the predicted label from $\boldsymbol{p}_i$, respectively. We first provide the following theorem, which shows that with high probability, every reliable sample $\boldsymbol{x}_i \in \mathcal{S}$ possesses the accurate label.

Theorem 1 (Reliability Measurement Module I)**.** *Assume that there exists a universal constant $0 < c_0 < 1/4$, such that for each unlabeled sample $\boldsymbol{x}_i$, the corresponding predicted reliability scores $\hat{s}_i$ satisfy*

$$\lim_{N\to\infty} \mathbb{P}\left(\bigcup_{\boldsymbol{x}_i \in \mathcal{D}^u} \{|\hat{s}_i - \cos(\boldsymbol{p}_i, \boldsymbol{y}_i^*)| \geq c_0\}\right) = 0, \quad (17)$$

where $\mathbb{P}(A)$ represents the probability of the event A. Then, there exists a threshold τ^* in Eqn. 8 such that

$$\lim_{N\to\infty} \mathbb{P}(p_i = y_i^*, \textit{ for all } \boldsymbol{x}_i \in \mathcal{S}) = 1, \quad (18)$$

where y_i^* and p_i as the unknown true label and the predicted label from $\boldsymbol{p}_i$, respectively.

Theorem 1 ensures the validity of Eqn. 9. The condition in Theorem 1 puts the restriction on the reliability measurement module and is very mild because we allow a constant gap c_0 between the ground truth and the predicted reliability scores, indicating that the module does not need to be a perfect predictor. A smaller c_0 indicates a higher accuracy of the reliability measurement module.

Theorem 2 (Reliability Measurement Module II)**.** *Under the same condition as in Theorem 1, if there exists a positive constant c_1 such that $\#\{\boldsymbol{x}_i : \cos(\boldsymbol{p}_i, \boldsymbol{y}_i^*) \geq \tau^* + c_0\} \geq c_1 N^u$, then we have*

$$\lim_{N\to\infty} \mathbb{P}(\#\mathcal{S} \geq c_1 N^u) = 1, \quad (19)$$

where $\#A$ represents the cardinality of the set A.

Theorem 2 ensures a sufficient amount of data in $\mathcal{S}$ and further ensures an ample supply of accurately labeled data to enhance our model's performance. The inequality $\#\{i : \cos(\boldsymbol{p}_i, \boldsymbol{y}_i^*) \geq \tau^* + c_0\} \geq c_1 N^u$ assumes that the true reliability scores exhibit a significant number of high values, suggesting the efficacy of the label prediction module in forecasting a subset of the labels.

Theorem 3 (Negative Learning Module)**.** *Assume that there exists an integer k^* such that*

$$\lim_{N\to\infty} \mathbb{P}\left(\bigcup_{\boldsymbol{x}_i \in \mathcal{D}^u} \{y_i^* \notin \text{top}_{k^*}(\boldsymbol{p}_i)\}\right) = 0$$
$$\textit{and } \lim_{N\to\infty} \mathbb{P}\left(\bigcap_{\boldsymbol{x}_i \in \mathcal{S}} \left\{y_i^* \in \text{top}_{k^*-1}(\boldsymbol{p}_i)\right\}\right) = 0. \quad (20)$$

Then, we have

$$\lim_{N\to\infty} \mathbb{P}(t_n < k^*) = 0$$
$$\textit{and } \lim_{N\to\infty} \mathbb{P}\left(\bigcup_{\boldsymbol{x}_i \in \mathcal{D}^u} \left\{y_i^* \notin \text{top}_{t_n}(\boldsymbol{p}_i)\right\}\right) = 0, \quad (21)$$

where t_n is defined as in Eqn. 10.

Theorem 3 ensures that all the labels with ranking over the threshold t_n are true negative ones. It further guarantees the validity of Eqn. 11. Condition in Theorem 3 is quite lenient as we do not necessitate the label prediction module to precisely predict the labels of unlabeled data. Rather, we simply assume that the true label falls within the top k^* index set of $\boldsymbol{p}_i$. A smaller k^* indicates a higher accuracy of the label prediction module. When $k^* = 1$, the label prediction module is optimal. Conversely, when $k^* = C$, then the label prediction module becomes ineffective. By combining Theorem 1–3, we provide theoretical guarantees to support the effectiveness of the proposed decoupled discriminative learning process. The proofs can be found in the supplementary materials.

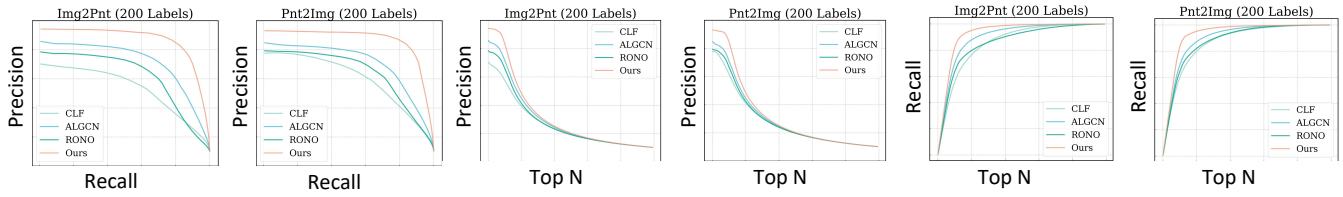
Experiments

Experimental Setup

Datasets and Baselines. We conduct evaluation on three popular datasets, including 3D MNIST (Xu et al. 2016), ModelNet10 (Wu et al. 2015), and ModelNet40 (Wu et al. 2015). For comparison, we include three state-of-the-art 2D-3D cross-modal retrieval methods (CLF (Jing et al. 2021) , RONO (Feng et al. 2023), and HOPE (Zhang et al. 2024b)) as baselines. Due to the scarcity of methods tailored specifically for the 2D-3D retrieval task, we also extend some image-text retrieval methods to the 2D-3D retrieval scenario, including DSCMR (Zhen et al. 2019), MRL (Hu et al. 2021a), ALGCN (Qian et al. 2021), DA-I-GCN (Qian et al. 2022), DA-I-GAT (Qian et al. 2022), DA-P-GCN (Qian et al. 2022), and DA-P-GAT (Qian et al. 2022). The widely recognized mean average precision (MAP) score is employed as our evaluation metric. More details about datasets and baselines are provided in the supplementary materials.

Task	Dataset	3D MNIST					ModelNet10					ModelNet40				
	Label	200	400	600	800	Avg	200	400	600	800	Avg	800	1600	2400	3200	Avg
2D → 3D	DSCMR	68.26	86.26	91.10	92.34	84.49	49.26	72.69	74.23	80.90	69.27	46.56	56.21	62.13	67.24	58.04
	MRL	46.06	65.05	71.66	78.71	65.37	45.23	53.38	57.92	62.89	54.86	26.30	26.96	32.62	35.94	30.46
	ALGCN	83.30	89.55	91.13	92.40	89.10	70.92	76.21	79.97	81.73	77.21	53.77	58.53	60.41	63.28	59.00
	DA-I-GCN	75.75	86.45	88.40	90.08	85.17	44.51	52.52	55.59	60.74	53.34	30.65	44.19	49.09	58.19	45.53
	DA-I-GAT	77.43	86.23	86.92	90.34	85.23	44.86	50.04	55.72	59.99	52.65	31.22	44.31	45.22	54.46	43.80
	DA-P-GCN	79.02	86.47	89.10	90.49	86.27	49.08	61.74	63.94	68.09	60.71	34.63	40.13	54.10	58.37	46.81
	DA-P-GAT	77.14	86.63	88.74	89.77	85.57	50.19	61.09	65.03	69.13	61.36	34.39	48.37	51.63	60.17	48.64
	CLF	67.44	86.54	89.77	91.12	83.72	50.60	71.65	76.78	82.33	70.34	50.42	59.74	66.57	70.72	61.86
	RONO	60.40	81.08	85.66	88.85	79.00	57.45	72.80	79.66	81.09	72.75	46.39	62.13	65.07	72.93	61.63
	HOPE	92.05	92.77	93.63	93.96	93.10	81.94	83.86	84.82	86.64	84.32	72.48	75.03	75.60	76.05	74.79
	Ours	**92.20**	**93.28**	**93.79**	**94.54**	**93.45**	**84.89**	**86.08**	**86.12**	**87.58**	**86.17**	**79.68**	**80.17**	**80.72**	**80.84**	**80.35**
3D → 2D	DSCMR	67.74	82.51	88.17	89.11	81.88	46.01	68.28	72.22	79.16	66.42	28.05	46.24	54.96	59.70	47.24
	MRL	46.46	64.45	70.88	78.69	65.12	44.90	52.21	55.50	60.99	53.40	25.96	27.29	32.80	36.94	30.75
	ALGCN	82.57	88.19	89.67	90.83	87.82	64.14	70.66	73.94	78.82	71.89	35.03	48.25	51.12	53.67	47.02
	DA-I-GCN	75.61	85.89	87.18	88.92	84.40	42.87	51.86	55.91	60.64	52.82	32.08	42.92	47.96	56.94	44.98
	DA-I-GAT	77.81	85.61	86.45	89.39	84.82	43.09	50.77	57.96	59.92	52.94	32.14	42.25	44.88	53.64	43.23
	DA-P-GCN	80.14	86.51	88.03	89.39	86.02	47.47	60.02	63.37	67.40	59.57	35.03	39.82	52.56	56.08	45.87
	DA-P-GAT	79.51	86.59	87.66	88.91	85.67	48.12	61.30	63.98	68.22	60.41	34.11	46.85	50.17	58.53	47.42
	CLF	65.65	86.26	88.53	89.61	82.51	46.50	69.11	72.58	81.23	67.36	41.93	54.22	63.21	67.03	56.60
	RONO	65.75	82.24	84.95	88.57	80.38	49.56	68.86	75.33	78.98	68.18	35.08	54.58	59.37	69.40	54.61
	HOPE	90.85	91.42	92.52	92.74	91.88	81.89	83.15	84.90	86.26	84.05	72.11	74.61	74.83	74.94	74.12
	Ours	**91.35**	**93.01**	**93.70**	**94.15**	**93.05**	**84.88**	**86.32**	**86.92**	**88.16**	**86.57**	**79.49**	**80.13**	**80.28**	**80.79**	**80.17**

Table 1: MAP score comparison (%) with various amounts of labeled samples. The best results are shown in **boldface**.



(a) Precision-Recall (b) Precision-Top N (c) Recall-Top N

Figure 2: Precision-Recall, Precision-Top N, and Recall-Top N curves with 200 labeled samples on 3D MNIST.

Implementation Details. The experiments are conducted using the Pytorch on NVIDIA Tesla A100 GPUs. We adopt ResNet-18 and DGCNN as the 2D and 3D backbone, respectively. Subsequently, two different fully connected layers are employed to obtain 256-dimensional 2D and 3D features. For a fair comparison, we reproduce the baseline approaches according to the settings in their papers and carefully tune them to the best results. For our DREAM, we optimize the model for 50 epochs with the batch size 50. The learning rates for 2D and 3D networks are set to 5e-5 and 1e-4, respectively.

Experimental Results

Quantitative Comparison. We make extensive quantitative comparisons on three benchmark datasets, with results shown in Table 1. From the results, it can be observed that DREAM consistently outperforms all the compared methods. We attribute this to the following reasons: Firstly, traditional methods heavily rely on the vast amount of labeled training samples for discriminative learning, thus they suffer from performance degradation under label scarcity conditions. Although methods like HOPE tackle label scarcity with pseudo-labeling, they still leverage a single model for both labeling and selection, thus suffering from overconfident results. In contrast, DREAM presents a decoupled discriminative learning perspective to make full use of labeled and unlabeled data for joint training. Secondly, the compared methods typically align multimodal representations by pairwise labels or classwise anchors, but most of them couldn't result in high-quality representations with insufficient supervision. In contrast, DREAM proposes to align the semantics of multimodal representations by constructing a bipartite graph, and subsequently push intra-class multimodal representations to cluster in the common space. The proposed bigraph-aware modality alignment takes both labels and pseudo-labels into account, which is more effective than existing techniques.

Qualitative Comparison. Besides quantitative comparisons, we make further qualitative comparisons by plotting precision-recall, precision-top N, and recall-top N curves in Figure 2, and by conducting t-SNE visualization (Van der Maaten and Hinton 2008) in Figure 3 (a). From the observations in Figure 2, we can conclude that DREAM outperforms the compared baselines in both precision and recall scores. Moreover, the performance of DREAM is robust to the number of returned samples. The results depicted in Figure 3 vividly demonstrate the superiority of the representations learned by our DREAM over the compared methods. This superiority stems from the enhanced alignment of representations across different modalities, as well as the more cohesive clustering of representations within the same class.

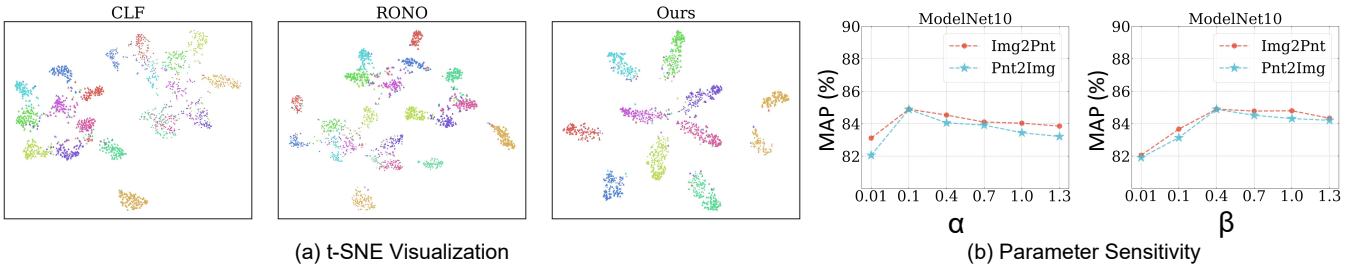


Figure 3: The (a) t-SNE visualization on 3D MNIST and (b) sensitivity analysis on ModelNet10 with 200 labeled samples.

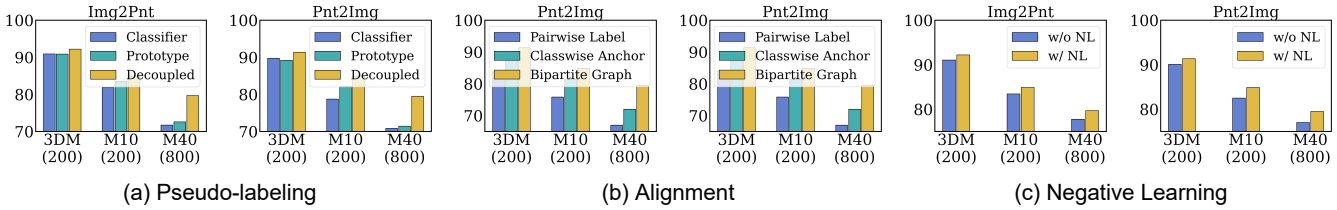


Figure 4: Ablation study on (a) pseudo-labeling techniques, (b) alignment strategies, (c) the negative learning module.

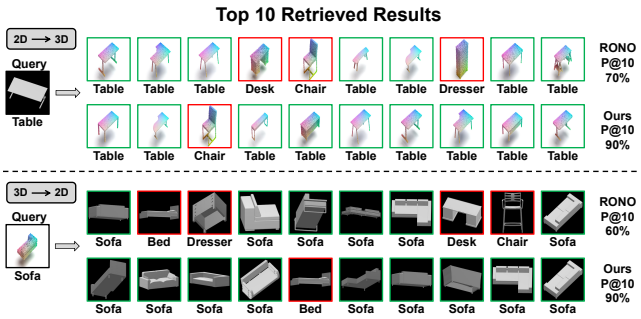


Figure 5: The top 10 returned samples on ModelNet10.

Ablation Study. We conduct comprehensive experiments to answer the following questions. (1) *Is decoupled discriminative learning superior to other pseudo-labeling techniques?* In Figure 4 (a), we provide some model variants to compare our proposed decoupled discriminative learning with two mainstream pseudo-labeling (classifier-based and prototype-based) techniques. From the results, we can find that the proposed decoupled discriminative learning can let the model learn to selectively learn, which can prevent overconfident results and thus result in better retrieval performance. (2) *Is bigraph-aware modality alignment better than other alignment strategies?* In Figure 4 (b), we provide some model variants to validate the effectiveness of our bigraph-aware modality alignment strategy. The results indicate that other strategies using pairwise labels or classwise anchors do not yield results as promising as the proposed bigraph-aware alignment strategy. This demonstrates that our bigraph-aware alignment could result in high-quality modality-invariant representations. (3) *Is negative learning effective for semi-supervised 2D-3D cross-modal retrieval?* To validate whether unreliable samples can provide supervision for cross-modal retrieval, we examine the function of the negative learning module in Figure 4 (c). The results indicate that when the negative learning module is removed, the model discards unreliable samples, leading to performance degradation. This validates the effectiveness of the negative learning module, which enables the model to leverage unreliable unlabeled samples for additional supervision.

Sensitivity Analysis. In Figure 3 (b), we provide the sensitivity analysis of two hyper-parameters. Firstly, we vary α within the interval of $\{0.01, 0.1, 0.4, 0.7, 1.0, 1.3\}$ while keeping other parameters constant. We can observe that the model performs best when $\alpha = 0.1$. This phenomenon is reasonable since the negative learning module only provides weak supervision and α should be less than 1 to prioritize other losses. Next, we fix α at 0.1 and vary β from 0.01 to 1.3. Similarly, the performance first increases and then saturates, with optimal performance observed when $\beta = 0.4$. Therefore, we obtain the recommended values for these two hyper-parameters are $\alpha = 0.1$ and $\beta = 0.4$.

Case Study. As depicted in Figure 5, we make a case study to test the top 10 returned samples by RONO and our method with 800 labeled samples. The samples marked in green represent those belonging to the same category as the query sample, while those marked in red represent the opposite. From the results, it can be observed that DREAM is clearly superior to the compared RONO, which is consistent with previous quantitative and qualitative comparisons.

Conclusion

In this paper, we design a novel approach DREAM for the problem of semi-supervised 2D-3D cross-modal retrieval. The main idea of DREAM is to decouple the discriminative learning process to make the model learn to selectively learn, and construct a bipartite graph for modality alignment with discriminablity preserved. Through extensive experiments on three benchmark datasets, we verify the superiority of DREAM, as well as the effectiveness of each proposed module. In the future, we plan to extend the proposed DREAM to more advanced scenarios including fine-grained cross-modal retrieval and multimodal foundation models.

References

Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. 2018. Learning representations and generative models for 3d point clouds. In *ICML*, 40–49. PMLR.

Assran, M.; Caron, M.; Misra, I.; Bojanowski, P.; Joulin, A.; Ballas, N.; and Rabbat, M. 2021. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *ICCV*, 8443–8452.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 32.

Chen, D.; Bai, Y.; Shen, W.; Li, Q.; Yu, L.; and Wang, Y. 2023a. MagicNet: Semi-Supervised Multi-Organ Segmentation via Magic-Cube Partition and Recovery. In *CVPR*, 23869–23878.

Chen, Y.; Tan, X.; Zhao, B.; Chen, Z.; Song, R.; Liang, J.; and Lu, X. 2023b. Boosting Semi-Supervised Learning by Exploiting All Unlabeled Data. In *CVPR*, 7548–7557.

Cole, E.; Mac Aodha, O.; Lorieul, T.; Perona, P.; Morris, D.; and Jojic, N. 2021. Multi-label learning from single positive labels. In *CVPR*, 933–942.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *AAAI*, volume 35, 1218–1226.

Dong, J.; Li, X.; Xu, C.; Ji, S.; He, Y.; Yang, G.; and Wang, X. 2019. Dual encoding for zero-example video retrieval. In *CVPR*, 9346–9355.

Feng, Y.; Zhu, H.; Peng, D.; Peng, X.; and Hu, P. 2023. RONO: Robust Discriminative Learning With Noisy Labels for 2D-3D Cross-Modal Retrieval. In *CVPR*, 11610–11619.

Garg, A.; Bagga, S.; Singh, Y.; and Anand, S. 2022. Hiermatch: Leveraging label hierarchies for improving semi-supervised learning. In *WACV*, 1015–1024.

Hu, P.; Huang, Z.; Peng, D.; Wang, X.; and Peng, X. 2023. Cross-Modal Retrieval with Partially Mismatched Pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hu, P.; Peng, X.; Zhu, H.; Zhen, L.; and Lin, J. 2021a. Learning cross-modal retrieval with noisy labels. In *CVPR*, 5403–5413.

Hu, Q.; Wu, J.; Cheng, J.; Wu, L.; and Lu, H. 2017. Pseudo label based unsupervised deep discriminative hashing for image retrieval. In *ACMMM*, 1584–1590.

Hu, Z.; Yang, Z.; Hu, X.; and Nevatia, R. 2021b. Simple: Similar pseudo label exploitation for semi-supervised classification. In *CVPR*, 15099–15108.

Hua, W.; Liang, D.; Li, J.; Liu, X.; Zou, Z.; Ye, X.; and Bai, X. 2023. SOOD: Towards Semi-Supervised Oriented Object Detection. In *CVPR*, 15558–15567.

Hwang, U.; Lee, J.; Shin, J.; and Yoon, S. 2024. Source-free Domain Adaptation Through the Lens of Data Augmentation. *arXiv preprint arXiv:2403.10834*.

Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2019. Label propagation for deep semi-supervised learning. In *CVPR*, 5070–5079.

Jain, A.; Mildenhall, B.; Barron, J. T.; Abbeel, P.; and Poole, B. 2022. Zero-shot text-guided object generation with dream fields. In *CVPR*, 867–876.

Jaritz, M.; Vu, T.-H.; De Charette, R.; Wirbel, É.; and Pérez, P. 2022. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1533–1544.

Jing, L.; Vahdani, E.; Tan, J.; and Tian, Y. 2021. Cross-modal center loss for 3d cross-modal retrieval. In *CVPR*, 3142–3151.

Kim, J. M.; Koepke, A.; Schmid, C.; and Akata, Z. 2023. Exposing and Mitigating Spurious Correlations for Cross-Modal Retrieval. In *CVPR*, 2584–2594.

Kim, Y.; Yim, J.; Yun, J.; and Kim, J. 2019. Nlnl: Negative learning for noisy labels. In *ICCV*, 101–110.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *ECCV*, 201–216.

Li, H.; Wang, N.; Yang, X.; Wang, X.; and Gao, X. 2022. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *CVPR*, 4166–4175.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019. Visual semantic reasoning for image-text matching. In *ICCV*, 4654–4662.

Li, P.; Xie, C.-W.; Zhao, L.; Xie, H.; Ge, J.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023a. Progressive spatio-temporal prototype matching for text-video retrieval. In *ICCV*, 4100–4110.

Li, S.; Jin, W.; Wang, Z.; Wu, F.; Liu, Z.; Tan, C.; and Li, S. Z. 2023b. Semireward: A general reward model for semi-supervised learning. *arXiv preprint arXiv:2310.03013*.

Li, Y.; Su, H.; Qi, C. R.; Fish, N.; Cohen-Or, D.; and Guibas, L. J. 2015. Joint embeddings of shapes and images via cnn image purification. *ACM transactions on graphics (TOG)*, 34(6): 1–12.

Lin, M.-X.; Yang, J.; Wang, H.; Lai, Y.-K.; Jia, R.; Zhao, B.; and Gao, L. 2021. Single image 3d shape retrieval via cross-modal instance and category contrastive learning. In *ICCV*, 11405–11415.

Liu, A.-A.; Zhang, C.; Li, W.; Gao, X.; Sun, Z.; and Li, X. 2022. Self-supervised auxiliary domain alignment for unsupervised 2d image-based 3d shape retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8809–8821.

Liu, C.; Mao, Z.; Zhang, T.; Xie, H.; Wang, B.; and Zhang, Y. 2020. Graph structured network for image-text matching. In *CVPR*, 10921–10930.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2837–2845.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993.

Nie, W.; Jiao, C.; Chang, R.; Qu, L.; and Liu, A.-A. 2023. CPG3D: Cross-modal Priors Guided 3D Object Reconstruction. *IEEE Transactions on Multimedia*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*, 4195–4205.

Pham, Q.-H.; Uy, M. A.; Hua, B.-S.; Nguyen, D. T.; Roig, G.; and Yeung, S.-K. 2020. Lcd: Learned cross-domain descriptors for 2d-3d matching. In *AAAI*.

Qian, S.; Xue, D.; Fang, Q.; and Xu, C. 2021. Adaptive label-aware graph convolutional networks for cross-modal retrieval. *IEEE Transactions on Multimedia*, 24: 3520–3532.

Qian, S.; Xue, D.; Fang, Q.; and Xu, C. 2022. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4794–4811.

Qiao, P.; Wei, Z.; Wang, Y.; Wang, Z.; Song, G.; Xu, F.; Ji, X.; Liu, C.; and Chen, J. 2023. Fuzzy Positive Learning for Semi-Supervised Semantic Segmentation. In *CVPR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Saberi, A. H.; Najafi, A.; Heidari, A.; Movasaghinia, M. H.; Motahari, A. S.; and Khalaj, B. H. 2024. Out-Of-Domain Unlabeled Data Improves Generalization. In *ICLR*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 596–608.

Song, D.; Zhang, C.-M.; Zhao, X.-Q.; Wang, T.; Nie, W.-Z.; Li, X.-Y.; and Liu, A.-A. 2023. Self-supervised image-based 3d model retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2): 1–18.

Sun, H.; Xu, L.; Jin, S.; Luo, P.; Qian, C.; and Liu, W. 2024. PROGRAM: PROtotype GRAph Model based Pseudo-Label Learning for Test-Time Adaptation. In *ICLR*.

Vahdat, A.; Williams, F.; Gojcic, Z.; Litany, O.; Fidler, S.; Kreis, K.; et al. 2022. Lion: Latent point diffusion models for 3d shape generation. *NeurIPS*, 35: 10021–10039.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial cross-modal retrieval. In *ACMMM*.

Wang, H.; Sun, J.; Wei, X.; Zhang, S.; Chen, C.; Hua, X.-S.; and Luo, X. 2023. Dance: Learning a domain adaptive framework for deep hashing. In *TheWebConf*.

Wu, J.; Yang, H.; Gan, T.; Ding, N.; Jiang, F.; and Nie, L. 2023. CHMATCH: Contrastive Hierarchical Matching and Robust Adaptive Threshold Boosted Semi-Supervised Learning. In *CVPR*, 15762–15772.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3D ShapeNets: A deep representation for volumetric shapes. In *CVPR*, 1912–1920.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. *NeurIPS*, 33: 6256–6268.

Xu, R.; Han, Z.; Hui, L.; Qian, J.; and Xie, J. 2022. Domain disentangled generative adversarial network for zero-shot sketch-based 3d shape retrieval. In *AAAI*, volume 36, 2902–2910.

Xu, X.; Dehghani, A.; Corrigan, D.; Caulfield, S.; and Moloney, D. 2016. Convolutional Neural Network for 3D object recognition using volumetric representation. In *International Workshop on Sensing, Processing and Learning for Intelligent Machines*, 1–5. IEEE.

Xu, X.; Lin, K.; Lu, H.; Gao, L.; and Shen, H. T. 2020. Correlated features synthesis and alignment for zero-shot cross-modal retrieval. In *SIGIR*, 1419–1428.

Xue, L.; Gao, M.; Xing, C.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J. C.; and Savarese, S. 2023. ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding. In *CVPR*, 1179–1189.

Yang, J.; Xian, K.; Wang, P.; and Zhang, Y. 2019. A performance evaluation of correspondence grouping methods for 3D rigid data matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6): 1859–1874.

Yang, W.; Zhang, R.; Chen, J.; Wang, L.; and Kim, J. 2023. Prototype-guided pseudo labeling for semi-supervised text classification. In *ACL*, 16369–16382.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, 18408–18419.

Zhang, F.; Hua, X.-S.; Chen, C.; and Luo, X. 2024a. Fine-grained Prototypical Voting with Heterogeneous Mixup for Semi-supervised 2D-3D Cross-modal Retrieval. In *CVPR*, 17016–17026.

Zhang, F.; Zhou, H.; Hua, X.-S.; Chen, C.; and Luo, X. 2024b. Hope: A Hierarchical Perspective for Semi-supervised 2D-3D Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; Lin, X.; Zhang, W.; Wang, K.; Tan, X.; Han, J.; Ding, E.; Wang, J.; and Li, G. 2023. Semi-DETR: Semi-Supervised Object Detection With Detection Transformers. In *CVPR*, 23809–23818.

Zhang, R.; Fan, Z.; Yao, J.; Zhang, Y.; and Wang, Y. 2024c. Domain-Inspired Sharpness-Aware Minimization Under Domain Shifts. *arXiv preprint arXiv:2405.18861*.

Zhen, L.; Hu, P.; Wang, X.; and Peng, D. 2019. Deep supervised cross-modal retrieval. In *CVPR*, 10394–10403.

Zheng, S.; Chen, C.; Cai, X.; Ye, T.; and Tan, W. 2022. Dual decoupling training for semi-supervised object detection with noise-bypass head. In *AAAI*, volume 36, 3526–3534.

Zhu, J.-Y.; Zhang, Z.; Zhang, C.; Wu, J.; Torralba, A.; Tenenbaum, J.; and Freeman, B. 2018. Visual object networks: Image generation with disentangled 3d representations. *NeurIPS*, 31.

Zhu, L.; Wu, X.; Li, J.; Zhang, Z.; Guan, W.; and Shen, H. T. 2022. Work together: correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*.

Zhu, X.; and Ghahramani, Z. 2002. Learning from labeled and unlabeled data with label propagation. *ProQuest number: information to all users*.