Fine-grained Prototypical Voting with Heterogeneous Mixup for Semi-supervised 2D-3D Cross-modal Retrieval

Fan Zhang[1], Xian-Sheng Hua[2], Chong Chen[2], Xiao Luo[3†]
[1]Georgia Tech Shenzhen Institute, Tianjin University (GTSI),
[2]Terminus Group, [3]University of California, Los Angeles
fanzhang@gatech.edu, huaxiansheng@gmail.com, chenchong.cz@gmail.com, xiaoluo@cs.ucla.edu

Abstract

This paper studies the problem of semi-supervised 2D-3D retrieval, which aims to align both labeled and unlabeled 2D and 3D data into the same embedding space. The problem is challenging due to the complicated heterogeneous relationships between 2D and 3D data. Moreover, label scarcity in real-world applications hinders from generating discriminative representations. In this paper, we propose a semi-supervised approach named Fine-grained Prototypcical Voting with Heterogeneous Mixup (FIVE), which maps both 2D and 3D data into a common embedding space for cross-modal retrieval. Specifically, we generate fine-grained prototypes to model intra-class variation for both 2D and 3D data. Then, considering each unlabeled sample as a query, we retrieve relevant prototypes to vote for reliable and robust pseudo-labels, which serve as guidance for discriminative learning under label scarcity. Furthermore, to bridge the semantic gap between two modalities, we mix cross-modal pairs with similar semantics in the embedding space and then perform similarity learning for cross-modal discrepancy reduction in a soft manner. The whole FIVE is optimized with the consideration of sharpness to mitigate the impact of potential label noise. Extensive experiments on benchmark datasets validate the superiority of FIVE compared with a range of baselines in different settings. On average, FIVE outperforms the second-best approach by 4.74% *on 3D MNIST,* 12.94% *on ModelNet10, and* 22.10% *on ModelNet40.*

1. Introduction

3D visual understanding has received growing interest in computer vision and graphics. Among various 3D visual problems, 3D retrieval [12, 52] aims to return similar 3D shapes given a query. Early approaches usually concentrate on single-modal 3D retrieval, which transfers 3D data into

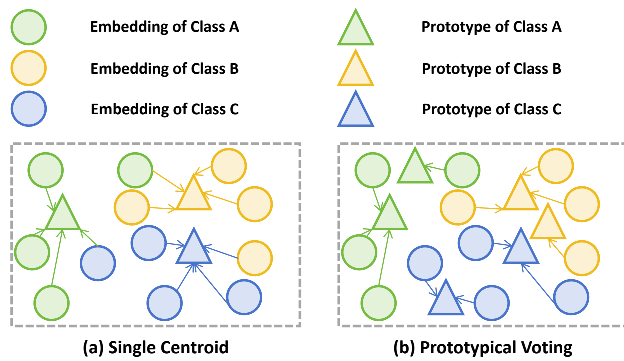

Figure 1. Compared with a single centroid for each class (a), our prototypical voting (b) can provide more reliable and robust pseudo-labels at the distribution boundaries.

an embedding space while maintaining the similarity structure [63]. Recently, due to the massive 2D and 3D data from artificial intelligence generated content, a more challenging problem of 2D-3D cross-modal retrieval has gained popularity, which aims to return data from one modality given a query from the other modality [35, 44].

In literature, researchers have proposed a variety of 2D-3D cross-modal retrieval approaches, which generally map both 2D and 3D samples into a single low-dimensional space with heterogeneous gaps reduced [12, 25, 36, 46, 60, 62]. A portion of these approaches generate ground-truth similarity structures which are then used to match the similarity relationships of deep representations using pairwise or triplet objectives [46, 62]. Another line of the research is to generate anchors for different classes, which can directly guide the optimization process using a point-wise objective [12, 25]. These approaches have also been expanded to address more circumstances including 3D object retrieval [17, 37] and CAD model retrieval [2, 15].

Despite their great success, current cross-modal retrieval approaches [12, 25, 34] usually require large amounts of labeled 2D and 3D data for end-to-end training. However, in practice, annotating a huge number of samples is prohibitively costly or even impossible. For instance, we can-

†Corresponding author.

not appropriately annotate shape data when the color and textual information are missing [22, 23]. A practice solution to the problem is to make use of extensive 2D and 3D unlabeled data at hand. Towards this end, in this paper, we investigate an underexplored but practical problem of semi-supervised 2D-3D cross-modal retrieval, which utilizes both labeled and unlabeled 2D and 3D data to learn joint representations with similarity embedded.

In reality, designing an effective semi-supervised framework for 2D-3D retrieval remains non-trivial since it entails resolving two fundamental challenges as follows. (1) *How to effectively make use of rich 2D and 3D unlabeled data?* Although semi-supervised learning has been studied extensively in classification problems [4, 20], semi-supervised retrieval models are still underexplored with inferior performance in reality. Moreover, existing approaches [56] could generate biased and noisy pseudo-labels when semantic distributions are complicated [3, 55, 69]. (2) *How to align multimodal data in a common embedding space?* Previous approaches [12, 25, 34] usually map 2D and 3D data into the embedding space using different encoders. However, due to the heterogeneous gap, there may be a significant distribution disparity between the two modalities, which would hinder from effective cross-modal retrieval.

To tackle the aforementioned issues, in this paper, we propose a novel approach named Fine-grained Prototypcical Voting with Heterogeneous Mixup (FIVE) for semi-supervised 2D-3D cross-modal retrieval. In particular, our FIVE projects 2D and 3D into a common embedding space using separate encoders and proposes two mechanisms to solve the aforementioned problems. Firstly, we model fine-grained prototypes for each class to illustrate intra-class variance. Then, to overcome label scarcity, we consider each unlabeled sample as a query and then retrieve relevant prototypes, which serve as multiple experts voting for the pseudo-label to guide discriminative learning. Compared with using a single centroid for each class, our voting-based pseudo-labeling strategy characterizes the semantic distribution using a large number of prototypes, which is more robust to perturbation, especially for samples at the distribution boundary (see Figure 1). Secondly, to enhance the modality alignment, we introduce heterogeneous Mixup, which generates composite virtual samples by linearly fusing deep representations with comparable semantics from different modalities. Due to the uncertainty in the quality of prototypes, the generated pseudo-labels often contain noise. To mitigate potential label noise, we propose a soft learning strategy, which maps these fused representations into a new shared embedding space where similarity learning is conducted for sufficient cross-modal discrepancy reduction. We integrate both mechanisms into a bi-level optimization framework [13, 28, 45], which considers the sharpness of the loss objective for robustness and generalization. Extensive experiments on a range of benchmark datasets validate the superiority of the proposed FIVE in comparison to various competing baselines. In brief, the contribution of this paper can be summarized as:

- This paper investigates a less-explored yet practical problem named semi-supervised 2D-3D cross-modal retrieval and develops a novel approach FIVE for the problem.
- On the one hand, FIVE introduces fine-grained prototypes to depict intra-class variance and then votes using retrieved prototypes to generate robust and reliable pseudo-labels. On the other hand, FIVE fuses deep representations from different modalities and conducts similarity learning softly for modality discrepancy reduction.
- Comprehensively experiments on a range of datasets demonstrate the superiority of our proposed FIVE in comparison to various competing baselines.

2. Related Work

Cross-modal Retrieval. With a query from one modality, cross-modal retrieval [5, 19, 25, 27, 48, 70, 71], seeks to return comparable samples from another modality with the goal of generating accurate similarity scores. Existing approaches can be roughly divided into global [10, 32, 38, 39, 50, 65] and fine-grained approaches [11, 19, 29, 31, 40]. Global approaches produce similarity scores by individually encoding input from each modality into a common embedding space. These approaches enjoy high computational efficiency with linear complexity. For example, CLF [25] introduces a cross-modal center loss for modality-invariant representations. RONO [12] further investigates the solution under noisy labels. In contrast, fine-grained approaches concentrate on semantic relationships at the token and patch levels, which can provide a fine-grained understanding of paired semantics [8]. Nevertheless, these approaches often demand a large amount of labeled data, which can be problematic for real-world applications [12]. To address this, this paper introduces a global approach for semi-supervised 2D-3D cross-modal retrieval.

Semi-supervised Learning. To reduce the expense of label annotation, semi-supervised learning has received ever-increasing interest [1, 57] by training a machine learning model using both labeled and unlabeled data, which has been applied to a range of computer vision applications such as image segmentation [6, 49], object detection [21, 68] and image-to-image translation [24, 43]. Current semi-supervised learning approaches can be roughly categorized into pseudo-labeling [3, 4, 20], consistency regularization [41, 59] and hybrid approaches [14, 30, 51, 66]. Pseudo-labeling approaches annotate unlabeled using the model itself to expand the labeled datasets. Consistency regularization aims to promote consistency in predictions under various perturbations including network, input, and feature perturbations. Hybrid approaches combine both

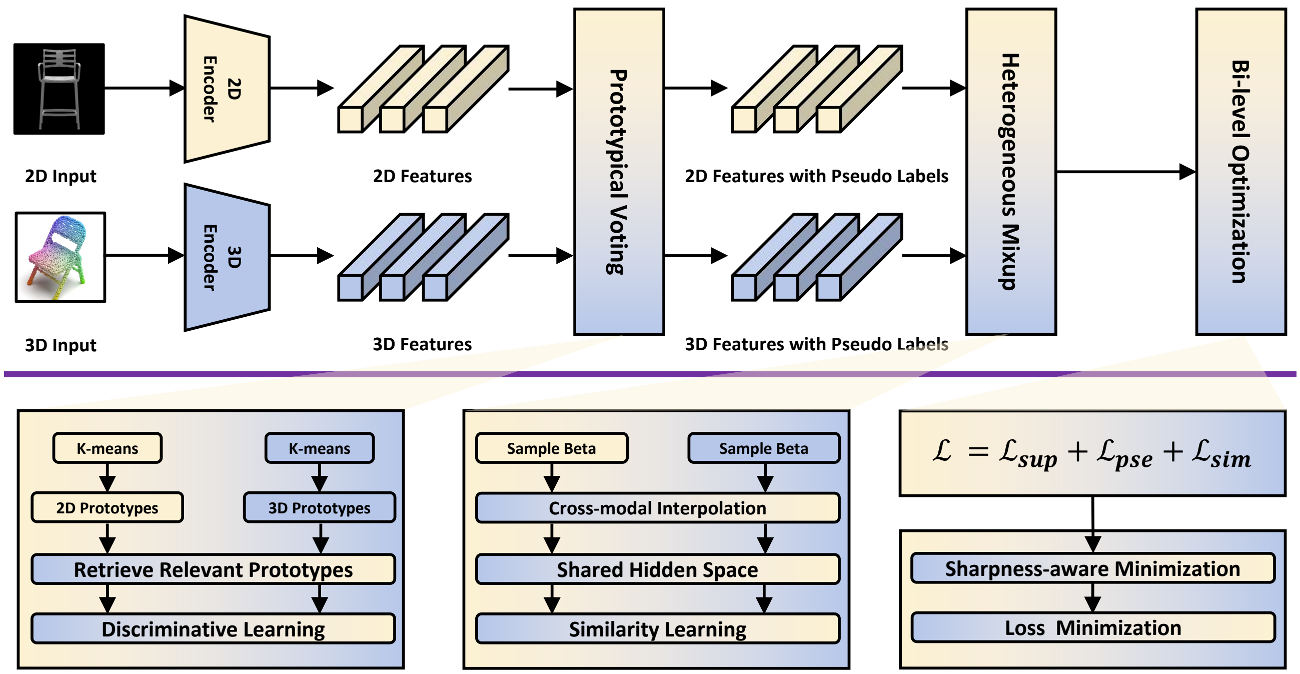

Figure 2. An overview of our FIVE. FIVE utilizes separate encoders to map multimodal input into a common embedding space. For unlabeled samples, FIVE retrieves relevant prototypes to vote for reliable pseudo-labels, which guide robust discriminative learning. In addition, FIVE conducts heterogeneous Mixup for virtual samples and then performs soft similarity learning for modality alignment. The whole FIVE is optimized in a bi-level paradigm for robustness and generalization.

worlds to improve optimization by incorporating data augmentations into consistency learning frameworks. Here, our study focuses on a more complex cross-modal retrieval problem and introduces fine-grained prototypes that serve as experts in generating robust and trustworthy pseudo-labels as guidance.

3. Methodology

3.1. Problem Definition

We first provide formal definitions of the problem. $\mathcal{D}^{2d} = \mathcal{D}^{2d,l} \cup \mathcal{D}^{2d,u}$ represents a 2D dataset with labeled samples $\mathcal{D}^{2d,l} = \{(\boldsymbol{x}_i^{2d,l}, y_i^{2d,l})\}_{i=1}^{N_{2d,l}}$ and unlabeled samples $\mathcal{D}^{2d,u} = \{(\boldsymbol{x}_i^{2d,u})\}_{i=1}^{N_{2d,u}}$. $y_i^{2d,l} \in \{1, 2, \cdots, C\}$ denotes the corresponding label of $\boldsymbol{x}_i^{2d,l}$. $\mathcal{D}^{3d} = \mathcal{D}^{3d,l} \cup \mathcal{D}^{3d,u}$ represents a 3D dataset with labeled samples $\mathcal{D}^{3d,l} = \{(\boldsymbol{x}_j^{3d,l}, y_j^{3d,l})\}_{i=1}^{N_{3d,l}}$ and unlabeled samples $\mathcal{D}^{3d,u} = \{(\boldsymbol{x}_j^{3d,u})\}_{j=1}^{N_{3d,u}}\}$. To facilitate effective cross-modal retrieval, we aim to project 2D and 3D samples into a common embedding space with semantics preserved. During evaluation, relevant 2D (3D) examples from a database should be returned given a 3D (2D) query.

3.2. Framework Overview

This work explores the problem of semi-supervised 2D-3D retrieval, which is challenging due to label scarcity and the semantics gap of heterogeneous data. We propose a new approach named FIVE for this problem, which includes two separate encoders $\phi^{2d}(\cdot)$ and $\phi^{3d}(\cdot)$ to transform multimodal data into a common embedding space, i.e.,

$$\boldsymbol{h}_i^{2d} = \phi^{2d}(\boldsymbol{x}_i^{2d}), \boldsymbol{h}_i^{3d} = \phi^{3d}(\boldsymbol{x}_i^{3d}), \tag{1}$$

where $\boldsymbol{x}_i^{2d}$ and $\boldsymbol{x}_i^{3d}$ denote 2D and 3D samples, respectively. As shown in Figure 2, our proposed FIVE is mainly comprised of two essential components: (1) *Fine-grained Prototypical Voting*, which models fine-grained prototypes to depict intra-class variance and then votes by retrieved prototypes to generate pseudo-labels for unlabeled samples, achieving robust discriminative learning under label scarcity. (2) *Heterogeneous Mixup*, which fuses deep representations of cross-modal pairs and conducts similarity learning in a new space for cross-modal discrepancy reduction. The whole framework is optimized using a bi-level paradigm, which includes the sharpness of loss for robust updating. Then, we elaborate on the details of FIVE.

3.3. Fine-grained Prototypical Voting for Robust Discriminative Learning

The pivot goal of semi-supervised 2D-3D retrieval is to make full use of abundant unlabeled data. Previous semi-supervised learning approaches [3, 4, 20] usually leverage the model itself to produce the predictions for unlabeled data and then compare the confidence scores to the threshold. However, the similarity learning framework does not allow us to directly generate pseudo-labels as in previous approaches. Furthermore, these pseudo-labels could be noisy and overconfident due to the class competition nature of the softmax layer [7, 55], which would prevent them from accurately capturing fine-grained semantics needed for effective cross-modal retrieval [33, 42, 69]. To tackle these issues, we provide fine-grained prototypical voting which generates a variety of fine-grained prototypes to characterize intra-class variance and votes by retrieving relevant prototypes for reliable pseudo-labels for successful discriminative learning.

In particular, we first learn from labeled samples for prototype generation. Here, the centroid of every class is summarized using the average of corresponding labeled samples as follows:

$$\boldsymbol{z}_c = \frac{\sum_{i=1}^{N^{2d,l}} \mathbf{1}_{\{y_i^{2d,l}=c\}} \boldsymbol{h}_i^{2d,l} + \sum_{j=1}^{N^{3d,l}} \mathbf{1}_{\{y_j^{3d,l}=c\}} \boldsymbol{h}_j^{3d,l}}{\sum_{i=1}^{N^{2d,l}} \mathbf{1}_{\{y_i^{2d,l}=c\}} + \sum_{j=1}^{N^{3d,l}} \mathbf{1}_{\{y_j^{3d,l}=c\}}}, \tag{2}$$

where $\boldsymbol{h}_i^{2d,l}$ and $\boldsymbol{h}_i^{3d,l}$ denote the representations of $\boldsymbol{x}_i^{2d,l}$ and $\boldsymbol{x}_i^{3d,l}$, respectively. $1_{\{\cdot\}}$ denotes an indicator returning 1 when the condition is satisfied. Then, we enforce each labeled sample to approach its corresponding centroid [25]. In formulation, we have:

$$\mathcal{L}_{sup} = \sum_{i=1}^{N^{2d,l}} \|\boldsymbol{h}_i^{2d,l} - \boldsymbol{z}_{y_i^{2d,l}}\|_2^2 + \sum_{j=1}^{N^{3d,l}} \|\boldsymbol{h}_j^{3d,l} - \boldsymbol{z}_{y_j^{3d,l}}\|_2^2. \tag{3}$$

By utilizing the centroid of each class to align labeled data, we are capable of generating discriminative representations for multimodal data. To make use of abundant unlabeled samples, an intuitive strategy is to find their nearest centroid. Nevertheless, the unlabeled data could not be around the centroid since each class in the hidden space always has a large distribution variance under label scarcity. In particular, the unlabeled samples at the boundary of the distribution could be wrongly annotated. To address these issues, we introduce a number of prototypes initialized using clustering to model fine-grained semantics in each class. Here, K-means is adopted to generate K clusters, i.e., $\mathcal{I}_c^1, \cdots, \mathcal{I}_c^K$ for 2D samples from each class c with the superscript omitted and then utilize the average as the fine-grained prototypes:

$$\boldsymbol{z'}_c^k = \frac{1}{|\mathcal{I}_c^k|} \sum_{\boldsymbol{x}_i \in \mathcal{I}_c^k} \boldsymbol{h}_i, \tag{4}$$

where $\boldsymbol{x}_i$ is a 2D sample with deep features $\boldsymbol{h}_i$. The total fine-grained prototypes can be collected as $\mathcal{P} = \cup_{c=1}^C \{\boldsymbol{z'}_c^1, \cdots, \boldsymbol{z'}_c^K\}$. Then, we view each unlabeled sample as a query $\boldsymbol{x}_i$ and retrieve M prototypes as $\boldsymbol{z'}_{i_1}, \cdots, \boldsymbol{z'}_{i_M}$ from $\mathcal{P}$ by ranking the Euclidean distance in the embedding space. These retrieved prototypes would vote for the pseudo-label of $\boldsymbol{x}_i$ as:

$$\hat{y}_i = argmax_{c=1}^C \sum_{m=1}^M 1\{y'_{i_m} = c\}, \tag{5}$$

where y'_{i_m} denotes the class of the prototype $\boldsymbol{z'}_{i_m}$. Similarly, we can also generate pseudo-labels for 3D samples. With pseudo-labels, we can enforce each unlabeled sample to approach its pseudo-centroid as:

$$\mathcal{L}_{pse} = \sum_{\boldsymbol{x}_i \in \mathcal{D}^{2d,u} \cup \mathcal{D}^{3d,u}} ||\boldsymbol{h}_i - \boldsymbol{z}_{\hat{y}_i}||_2^2, \tag{6}$$

where $\boldsymbol{z}_{\hat{y}_i}$ denotes the centroid corresponding to $\hat{y}_i$. After training the network for a while, we would include all these samples to update the centroids using the momentum strategy as:

$$\boldsymbol{z}_c^{update} = \mu \boldsymbol{z}_c + (1-\mu) \frac{\sum_{\boldsymbol{x}_i \in \mathcal{D}} 1_{\{\bar{y}_i = c\}} \boldsymbol{h}_i}{\sum_{\boldsymbol{x}_i \in \mathcal{D}} 1_{\{\bar{y}_i = c\}}}, \tag{7}$$

where μ is a momentum coefficient set to 0.99 empirically [16] and $\bar{y}_i$ can be labels or pseudo-labels for $\boldsymbol{x}_i$. Compared with previous pseudo-labeling approaches [3, 4, 20], our FIVE does not rely on the classifier, which is more suitable for our similarity learning framework. Moreover, our FIVE can explore the semantic distribution in each class to benefit fine-grained cross-modal retrieval. Incorporating multiple prototypes rather than one centroid for each class can produce smoother pseudo-labels at the distribution boundaries for robust discriminative learning.

3.4. Heterogeneous Mixup for Cross-modal Discrepancy Reduction

Although 2D and 3D samples share identical centroids in the embedding space, their distributions could still vary due to an intrinsic heterogeneous gap, which hinders from effective cross-modal retrieval [12, 25, 34]. To tackle this, we propose heterogeneous Mixup which fuses deep features from different modalities and maximizes the similarity between virtual samples with the same semantics in the projected embedding space for discrepancy reduction [16, 26, 55].

In particular, we fuse samples with the same semantics for simplification. Given two cross-modal samples $\boldsymbol{x}_i^{2d}$ and $\boldsymbol{x}_j^{3d}$ with $\bar{y}_i^{2d} = \bar{y}_j^{3d}$, we first generate their deep representations, $\boldsymbol{h}_i^{2d}$ and $\boldsymbol{h}_j^{3d}$, respectively. Then, we sample a coefficient from the Beta distribution for Mixup:

$$\lambda \sim Beta(\alpha, \beta), \tag{8}$$

where α and β are two coefficients set to 2 as in previous works [54, 64]. Then, the fused virtual samples have representations as:

$$\boldsymbol{h}_t^+ = \lambda \boldsymbol{h}_i^{2d} + (1-\lambda) \boldsymbol{h}_j^{3d}. \tag{9}$$

The label remains the same with $y_t^+ = \bar{y}_i^{2d} = \bar{y}_j^{3d}$. Note that there could be noisy pseudo-labels to influence mixed samples. Therefore, we utilize a soft manner for similarity learning. In particular, we project the fused samples into a new hidden space, and encourage the consistency of hidden embeddings from the same class. In formulation, the index set of positives for $\boldsymbol{h}_t^+$ in the mini-batch $\mathcal{B}$ can be written as:

$$\Pi(t) = \left\{t \mid \boldsymbol{h}_{t'} \in \mathcal{B}, y_{t'}^+ = y_t^+\right\}. \tag{10}$$

Then, the similarity learning objective for mixed samples can be formulated as:

$$\mathcal{L}_{sim} = -\sum_{\boldsymbol{h}_t \in \mathcal{B}} \frac{1}{|\Pi(t)|} \sum_{t' \in \Pi(t)} \log \frac{\exp\left(\boldsymbol{r}_t^+ \cdot \boldsymbol{r}_{t'}^+ / \tau\right)}{\sum_{\boldsymbol{r}_{t''} \in \mathcal{B}} \exp\left(\boldsymbol{r}_t^+ \cdot \boldsymbol{r}_{t''}^+ / \tau\right)}, \tag{11}$$

where $\boldsymbol{r}_t^+ = H(\boldsymbol{h}_t^+)$ denotes the representations in the hidden space and τ is a temperature parameter set to 0.5 empirically [55]. Compared with previous methods [9, 54, 67], our FIVE makes full use of both labeled and unlabeled samples to generate mixed samples and conduct similarity learning of fused multimodal data for cross-modal discrepancy reduction, which provides a Mixup strategy under label scarcity. Even if pseudo-labels of a sample pair are both not correct, they still could be a positive pair, which proves our pairwise similarity learning can provide more accurate supervision. Moreover, we project the representations of different modalities into a new shared embedding space and conduct similarity learning, which can mitigate the influence of wrong pseudo-labels in a soft manner.

3.5. Framework Optimization

In summary, the final objective can be summarized by combining three loss terms as follows:

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{pse} + \mathcal{L}_{sim}. \tag{12}$$

To mitigate the potential label noise resulting from pseudo-labeling with more generalization capacity, we adopt a bi-level optimization paradigm [13, 28, 45] for robustness, which models the sharpness of the loss using the maximum difference after perturbation:

$$\mathcal{L}^S = \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \mathcal{L}(\Theta + \boldsymbol{\epsilon}) - \mathcal{L}(\Theta), \tag{13}$$

where ϵ denotes a small perturbation to model parameters and ρ define the radius. Θ denotes the whole network parameters. Intuitively, a small loss sharpness can provide more robust optimization against potential label noise. Therefore, we simultaneously minimize the total loss and its sharpness as:

$$\mathcal{L}^F(\Theta) \triangleq \mathcal{L}^S + \mathcal{L} = \max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(\Theta + \boldsymbol{\epsilon}). \tag{14}$$

Algorithm 1 Training Algorithm of FIVE

Require: Datasets $\mathcal{D}^{2d}$ and $\mathcal{D}^{3d}$; Number of prototypes for every class K; Number of retrieved prototypes M.
Ensure: Parameters Θ of $\phi^{2d}(\cdot)$ and $\phi^{3d}(\cdot)$.

```
Warm up the 2D and 3D encoders;
Initialize the fine-grained prototypes P;
repeat
    for e=1,···,E do
        Sample labeled and unlabeled data to construct a mini-batch;
        Retrieve relevant prototypes for every unlabeled sample;
        Generate mixed representations using Eqn. 9;
        Generate pseudo-labels using Eqn. 5
        Compute the whole loss by Eqn. 12;
        Conduct inner optimization using Eqn. 17;
        Conduct outer optimization using Eqn. 19;
    end for
    Update centroids using Eqn. 7;
    Update prototypes with clustering;
until convergence
```

To minimize Eqn. 14, we adopt a bi-level paradigm with inner and outer optimization steps. In particular, we first conduct the inner optimization step as:

$$\boldsymbol{\epsilon}^*(\Theta) \triangleq \arg\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \mathcal{L}(\Theta + \boldsymbol{\epsilon}). \tag{15}$$

With first-order Taylor expansion, we have:

$$\begin{aligned} \arg\max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(\Theta + \boldsymbol{\epsilon}) &\approx \arg\max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(\Theta) + \boldsymbol{\epsilon}^T \nabla_\Theta \mathcal{L}(\Theta) \\ &= \arg\max_{\|\epsilon\|_2 \leq \rho} \boldsymbol{\epsilon}^T \nabla_\Theta \mathcal{L}(\Theta) \end{aligned} \tag{16}$$

As a classical dual norm problem, Eqn. 16 has the closed solution:

$$\hat{\boldsymbol{\epsilon}}(\Theta) = \rho \frac{\operatorname{sign}\left(\nabla_\Theta \mathcal{L}(\Theta)\right) |\nabla_\Theta \mathcal{L}(\Theta)|}{\|\nabla_\Theta \mathcal{L}(\Theta)\|_2}. \tag{17}$$

Then, for the outer optimization step, we need to calculate the derivative as:

$$\begin{aligned} \nabla_\Theta \mathcal{L}(\Theta + \hat{\boldsymbol{\epsilon}}(\Theta)) &= \frac{d(\Theta + \hat{\boldsymbol{\epsilon}}(\Theta))}{d\Theta} \nabla_\Theta \mathcal{L}(\Theta)|_{\Theta + \hat{\boldsymbol{\epsilon}}(\Theta)} \\ &= \nabla_\Theta \mathcal{L}(\Theta)|_{\Theta + \hat{\boldsymbol{\epsilon}}(\Theta)} + \frac{d\hat{\boldsymbol{\epsilon}}(\Theta)}{d\Theta} \nabla_\Theta \mathcal{L}(\Theta)|_{\Theta + \hat{\boldsymbol{\epsilon}}(\Theta)} \\ &\approx \nabla_\Theta \mathcal{L}(\Theta)|_{\Theta + \hat{\boldsymbol{\epsilon}}(\Theta)}, \end{aligned} \tag{18}$$

where the last approximation aims to drop the high-order term when the perturbation is small. Finally, we can get the gradient update rule for the outer optimization step as:

$$\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}(\Theta)|_{\Theta + \hat{\boldsymbol{\epsilon}}(\Theta)}, \tag{19}$$

Table 1. Performance comparison on different datasets and various amounts of labels. The best results are shown in **boldface** and the second best results are underlined.

Task	Dataset	3D MNIST					ModelNet10					ModelNet40				
	Label	200	400	600	800	Avg	200	400	600	800	Avg	800	1600	2400	3200	Avg
2D → 3D	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
	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
	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-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-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-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
	Ours	**91.68**	**93.51**	**94.03**	**94.78**	**93.50**	**83.97**	**84.22**	**84.37**	**85.18**	**84.44**	**71.32**	**72.18**	**72.86**	**75.40**	**72.94**
3D → 2D	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
	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
	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-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-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-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
	Ours	**89.92**	**91.82**	**92.31**	**93.13**	**91.80**	**83.22**	**83.62**	**84.01**	**84.18**	**83.76**	**70.24**	**71.22**	**71.49**	**72.97**	**71.48**

where η denotes the learning rate. We first warm up the neural network using labeled data and then gradually conduct fine-grained prototypical voting and heterogeneous Mixup. After sampling the mini-batch for E times, we would update the centroids and prototypes. The whole algorithm for learning our FIVE is summarized in Algorithm 1.

4. Experiments

4.1. Experimental Settings

Datasets. To verify the effectiveness of FIVE, we conduct comprehensive experiments on several public datasets. The brief introduction of the datasets is as follows: **3D MNIST** [61] is collected from Kaggle. The entire dataset consists of $6,000$ image-point cloud pairs from 10 different categories. The dataset is split into a training set with $5,000$ pairs and a testing set with $1,000$ pairs. **ModelNet10** [58] contains approximately $5,000$ 3D CAD objects across 10 categories. We split the dataset into two subsets: $3,991$ samples for training and 908 samples for testing. **ModelNet40** [58] and ModelNet10 are similar in nature, consisting of approximately $12,000$ 3D CAD objects from 40 categories. Similarly, the dataset is divided into two subsets: $9,840$ samples for training and $2,468$ samples for testing.

Baselines. We compare our FIVE with nine state-of-the-art (SOTA) cross-modal retrieval methods. These methods consist of seven text-image retrieval approaches (MRL [18], DSCMR [70], ALGCN [47], DA-I-GCN [48], DA-P-GCN [48], DA-I-GAT [48], DA-P-GAT [48]) and two 2D-3D retrieval methods (CLF [25], RONO [12]). As limited options are available for 2D-3D retrieval, we reproduce the text-image retrieval methods according to the corresponding papers and adapt them for the 2D-3D retrieval task.

Evaluation Protocols. We employ the mean average precision (MAP) as the evaluation metric, which is a commonly used criterion for assessing retrieval performance. A higher MAP score indicates enhanced retrieval accuracy.

Implementation Details. All experiments are conducted using the PyTorch framework. To ensure a fair comparison, we employ pre-trained ResNet-18 and DGCNN as the backbone for the image and point cloud networks, respectively, in all approaches. The output feature dimension is consistently set to 256. The networks are optimized using the Adam optimizer with a weight decay of $1e-5$. The learning rate is set to $5e-5$ for the image network and $1e-4$ for the point cloud network. The batch size is set to 50, and the training process is terminated after 50 epochs.

4.2. Experimental Results

Quantitative Comparisons. Table 1 illustrates the comparison between our FIVE and various SOTA cross-modal retrieval approaches on three datasets, with different amounts of labeled data. Based on these results, the following conclusions can be drawn: **First**, previous research on cross-modal retrieval has focused on effectively utilizing a large amount of labeled data while overlooking the potential of unlabeled data. This has resulted in performance degradation when only a portion of labeled data is available. 2D-3D retrieval methods such as RONO [12] and CLF [25] often outperform the modified text-image retrieval methods, except for ALGCN [47]. The superior performance

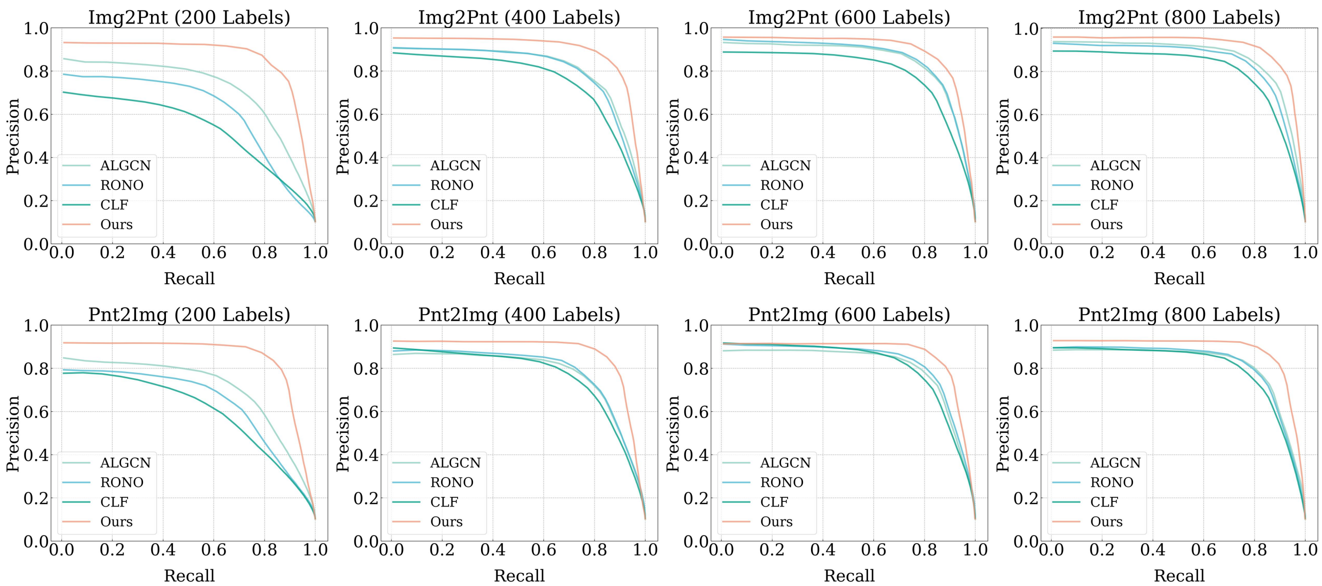

Figure 3. The Precision-Recall curve with various amounts of labels on the 3D MNIST dataset. 2D-to-3D results are plotted in the first row, and 3D-to-2D results are plotted in the second row.

of ALGCN [47] mainly stems from the use of an additional graph neural network (GNN) as guidance. **Second**, FIVE opens the door to utilizing a small amount of labeled data and a large amount of unlabeled data together to enhance 2D-3D retrieval performance. We achieve performance improvement across all three datasets and various scenarios with different amounts of labeled data, surpassing all current SOTA approaches. **Furthermore**, unlike methods like ALGCN [47] that incorporate additional GNN structures, we do not introduce extra parameters to the network but rather make improvements in algorithm design and training strategies. This indicates the effectiveness of our approach in addressing the challenge of semi-supervised 2D-3D cross-modal retrieval. Particularly when the amount of labeled data is extremely scarce (e.g., 200 labels), our FIVE outperforms other methods with a greater improvement.

Qualitative Comparisons. In addition, we make qualitative comparisons between our FIVE and the other three top-performing methods (i.e., ALGCN [47], RONO [12], CLF [25]) by plotting the Precision-Recall curves in scenarios with the number of labeled data varying from 200 to 800 in Figure 3. More comprehensive comparisons can be found in the supplementary material. Precision and recall are a pair of contradictory metrics, where an increase in one often results in a decrease in the other. In the precision-recall curve plot, methods represented by curves located higher on the graph are generally considered to have better performance. From the results, it can be observed that our FIVE consistently outperforms the other approaches in both 2D-to-3D and 3D-to-2D retrieval tasks across the four scenarios with varying amounts of labeled data. The fewer labeled data there are, the more pronounced the advantage of our FIVE compared to other approaches. This further confirms the effectiveness and robustness of our approach.

Ablation Study. In this section, we examine the contribution of each proposed component in Table 2. **FIVE w/o SL** refers to the removal of supervised learning for both 2D and 3D modalities, resulting in no semantic information injected into the two modalities. From the results, we can observe that without this module, the performance experiences a significant decline. **FIVE w/o PV** indicates the absence of prototypical voting, instead using the single centroid to generate pseudo-labels. We can observe that using a single centroid can lead to erroneous proximity between features and low-confident centroids, resulting in incorrect predictions. Therefore, the retrieval performance is significantly affected. **FIVE w/o BO** signifies the exclusion of bi-level optimization, opting for the conventional Adam optimizer. The results confirm that by simultaneously optimizing the loss value and loss sharpness, the model exhibits better generalization performance while providing robustness to certain low-confidence samples. **FIVE w/o HM** denotes the removal of the heterogeneous Mixup module. Heterogeneous Mixup can be seen as a cross-modal data augmentation strategy in the feature space, which introduces more diverse samples. As a result, it brings about some performance improvement. Lastly, by incorporating all the components, **Full Model** demonstrates the best performance. The results of ablation experiments confirm that each proposed component contributes to addressing the 2D-3D cross-modal retrieval problem under label scarcity conditions. Additionally, the individual contributions of these

Table 2. Ablation study on each proposed component with 200 labels, 200 labels, and 800 labels on three datasets.

Dataset	**3D MNIST**		**ModelNet10**		**ModelNet40**	
Task	**I2P**	**P2I**	**I2P**	**P2I**	**I2P**	**P2I**
FIVE w/o SL	19.47	17.42	19.71	18.19	5.39	4.76
FIVE w/o PV	84.62	82.53	74.50	73.14	65.66	64.80
FIVE w/o BO	89.91	88.47	76.95	76.48	68.81	67.13
FIVE w/o HM	90.91	89.10	82.18	81.81	70.34	70.02
Full Model	**91.68**	**89.92**	**83.97**	**83.22**	**71.32**	**70.24**

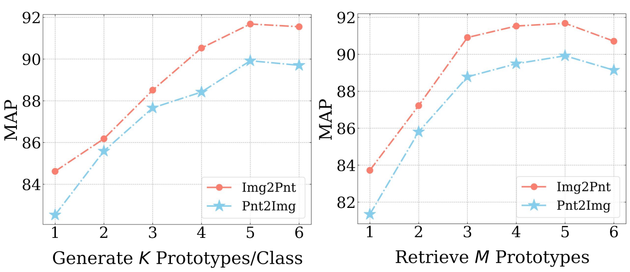

Figure 4. Sensitivity analysis of hyper-parameters M and K with 200 labels on 3D MNIST.

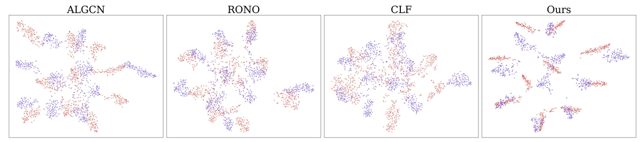

Figure 5. The t-SNE visualization with 200 labels on 3D MNIST. 2D modality is colored red, and 3D modality is colored blue.

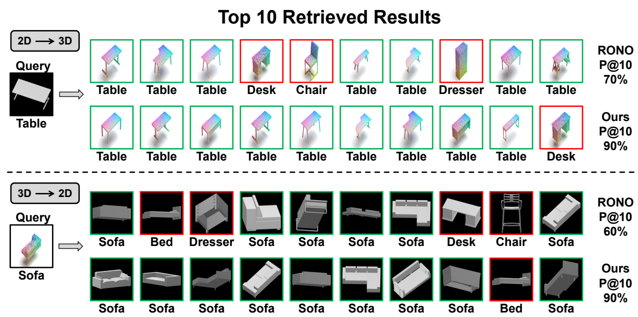

Figure 6. Top 10 retrieved results with 800 labels on ModelNet10. Green boxes signify accurate retrieval, whereas red boxes signify erroneous retrieval.

modules complement and promote each other, and their combination leads to maximum performance improvement.

Sensitivity Analysis. In Figure 4, we investigate the impact of two hyperparameters, K and M, on the 3D MNIST dataset. K represents the number of confident prototypes generated per class, while M represents the number of prototypes with the highest occurrence in a single retrieval process. Both hyperparameters directly affect feature classification and alignment, thus influencing the final retrieval performance. First, we gradually increase K from 1 to 6. When $K = 1$, the prototype approximation is similar to the intra-class centroid feature. Increasing K gradually adds more tolerance to the model. We can observe that the model exhibits the best performance when $K = 5$. Next, with K fixed at 5, we increase M from 1 to 6. M represents the range of retrieval within a single process. Values of M that are too large or too small can harm performance. For a class with K prototypes, when $M = K$, the ideal scenario within the sampling range is that all prototypes belong to the same class. As M increases further, erroneous prototypes will inevitably be sampled. From the experimental results, the optimal value for M is 5, validating our hypothesis.

4.3. Visualization

t-SNE Visualization. In Figure 5, we present t-SNE [53] visualizations of four methods on the 3D MNIST dataset. The dispersion of individual modalities reflects the discriminability of features for different classes, while the overlap degree between the two modalities indicates the degree of modality-invariant features. Compared to the other three methods, the features of FIVE are significantly dispersed into 10 clusters, representing the 10 classes of 3D MNIST. Additionally, our FIVE exhibits the highest overlap degree between the two modalities, indicating successful alignment of the 2D and 3D modalities in the feature space.

Case Study. We visualize the top 10 retrieved results of our FIVE and the compared baseline RONO [12] in Figure 6. It is evident that FIVE retrieves more relevant objects for both the 2D-to-3D and 3D-to-2D tasks. For instance, when given a 2D query ‘Table’, RONO [12] retrieves unrelated results such as ‘Desk’, ‘Chair’, and ‘Dresser’. Similarly, when given a 3D query ‘Sofa’, RONO [12] retrieves unrelated objects like ‘Bed’, ‘Dresser’, and ‘Desk’. In contrast, FIVE exhibits significantly lower error rates and more balanced results, whether using 2D or 3D data as the query.

5. Conclusion

This paper investigates the problem of semi-supervised 2D-3D retrieval and proposes a novel approach FIVE, which maps both 2D and 3D data into a common embedding space. FIVE introduces fine-grained prototypes and then retrieves relevant prototypes to vote for reliable pseudo-labels, which accomplish discriminative learning under label scarcity. Additionally, FIVE combines cross-modal pairs with comparable semantics in the embedding space and then performs soft similarity learning for effective cross-modal discrepancy reduction. Extensive experiments on benchmark datasets validate the advantage of FIVE over a variety of baselines. In future works, we would extend our approach to more complicated scenarios such as zero-shot cross-modal retrieval and 3D object understanding.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8443–8452, 2021. 2

[2] Armen Avetisyan, Angela Dai, and Matthias Nießner. End-to-end cad model retrieval and 9dof alignment in 3d scans. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pages 2551–2560, 2019. 1

[3] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 2, 3, 4

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 2, 3, 4

[5] Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, and Mihai Datcu. A simplified framework for zero-shot cross-modal sketch data retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 182–183, 2020. 2

[6] Duowen Chen, Yunhao Bai, Wei Shen, Qingli Li, Lequan Yu, and Yan Wang. Magicnet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23869–23878, 2023. 2

[7] Liang Chen, Yihang Lou, Jianzhong He, Tao Bai, and Minghua Deng. Evidential neighborhood contrastive learning for universal domain adaptation. 2022. 4

[8] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. Vista: vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5184–5193, 2022. 2

[9] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 95–110. Springer, 2020. 5

[10] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1218–1226, 2021. 2

[11] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9346–9355, 2019. 2

[12] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, and Peng Hu. Rono: Robust discriminative learning with noisy labels for 2d-3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11610–11619, 2023. 1, 2, 4, 6, 7, 8

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 2, 5

[14] Ashima Garg, Shaurya Bagga, Yashvardhan Singh, and Saket Anand. Hiermatch: Leveraging label hierarchies for improving semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1015–1024, 2022. 2

[15] Can Gümeli, Angela Dai, and Matthias Nießner. Roca: Robust cad model retrieval and alignment from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4022–4031, 2022. 1

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 4

[17] Xinwei He, Tengteng Huang, Song Bai, and Xiang Bai. View n-gram network for 3d object retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7515–7524, 2019. 1

[18] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5403–5413, 2021. 6

[19] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[20] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: Similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108, 2021. 2, 3, 4

[21] Wei Hua, Dingkang Liang, Jingyu Li, Xiaolong Liu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Sood: Towards semi-supervised oriented object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15558–15567, 2023. 2

[22] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 2

[23] Maximilian Jaritz, Tuan-Hung Vu, Raoul De Charette, Émilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1533–1544, 2022. 2

[24] Yuxin Jiang, Liming Jiang, Shuai Yang, and Chen Change Loy. Scenimefy: Learning to craft anime scene via semi-supervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7357–7367, 2023. 2

[25] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3142–3151, 2021. 1, 2, 4, 6, 7

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 4

[27] Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2584–2594, 2023. 2

[28] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021. 2, 5

[29] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision.*, pages 201–216, 2018. 2

[30] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9475–9484, 2021. 2

[31] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4654–4662, 2019. 2

[32] Pandeng Li, Chen-Wei Xie, Liming Zhao, Hongtao Xie, Jiannan Ge, Yun Zheng, Deli Zhao, and Yongdong Zhang. Progressive spatio-temporal prototype matching for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4100–4110, 2023. 2

[33] Xiaoyu Li, Xiaoxue Chen, Zuming Huang, Lele Xie, Jingdong Chen, and Ming Yang. Fine-grained pseudo labels for scene text recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5786–5795, 2023. 4

[34] Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J Guibas. Joint embeddings of shapes and images via cnn image purification. *ACM transactions on graphics (TOG)*, 34(6):1–12, 2015. 1, 2, 4

[35] Ming-Xian Lin, Jie Yang, He Wang, Yu-Kun Lai, Rongfei Jia, Binqiang Zhao, and Lin Gao. Single image 3d shape retrieval via cross-modal instance and category contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11405–11415, 2021. 1

[36] An-An Liu, Chenyu Zhang, Wenhui Li, Xingyu Gao, Zhengya Sun, and Xuanya Li. 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, 2022. 1

[37] An-An Liu, He-Yu Zhou, Xuanya Li, and Lanjun Wang. Vulnerability of feature extractors in 2d image-based 3d object retrieval. *IEEE Transactions on Multimedia*, 2022. 1

[38] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930, 2020. 2

[39] Yang Liu, Qingchao Chen, and Samuel Albanie. Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14954–14964, 2021. 2

[40] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2

[41] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 2

[42] Daniele Mugnai, Federico Pernici, Francesco Turchini, and Alberto Del Bimbo. Fine-grained adversarial semi-supervised learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s): 1–19, 2022. 4

[43] Aamir Mustafa and Rafał K Mantiuk. Transformation consistency regularization–a semi-supervised paradigm for image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 599–615, 2020. 2

[44] Weizhi Nie, Chuanqi Jiao, Rihao Chang, Lei Qu, and An-An Liu. Cpg3d: Cross-modal priors guided 3d object reconstruction. *IEEE Transactions on Multimedia*, 2023. 1

[45] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. 2, 5

[46] Quang-Hieu Pham, Mikaela Angelina Uy, Binh-Son Hua, Duc Thanh Nguyen, Gemma Roig, and Sai-Kit Yeung. Lcd: Learned cross-domain descriptors for 2d-3d matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11856–11864, 2020. 1

[47] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Adaptive label-aware graph convolutional networks for cross-modal retrieval. *IEEE Transactions on Multimedia*, 24: 3520–3532, 2021. 6, 7

[48] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. 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, 2022. 2, 6

[49] Pengchong Qiao, Zhidan Wei, Yu Wang, Zhennan Wang, Guoli Song, Fan Xu, Xiangyang Ji, Chang Liu, and Jie Chen.

Fuzzy positive learning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15465–15474, 2023. 2

[50] 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 *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 2

[51] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proceedings of the Conference on Neural Information Processing Systems*, 2020. 2

[52] Dan Song, Chu-Meng Zhang, Xiao-Qian Zhao, Teng Wang, Wei-Zhi Nie, Xuan-Ya Li, and An-An Liu. Self-supervised image-based 3d model retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19 (2):1–18, 2023. 1

[53] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[54] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019. 4, 5

[55] Haixin Wang, Jinan Sun, Xiang Wei, Shikun Zhang, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Dance: Learning a domain adaptive framework for deep hashing. In *Proceedings of the ACM Web Conference 2023*, pages 3319–3330, 2023. 2, 4, 5

[56] Zhenyu Wang, Ya-Li Li, Ye Guo, and Shengjin Wang. Combating noise: semi-supervised learning by region uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:9534–9545, 2021. 2

[57] Jianlong Wu, Haozhe Yang, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. Chmatch: Contrastive hierarchical matching and robust adaptive threshold boosted semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15762–15772, 2023. 2

[58] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1912–1920. IEEE Computer Society, 2015. 6

[59] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 2

[60] Rui Xu, Zongyan Han, Le Hui, Jianjun Qian, and Jin Xie. Domain disentangled generative adversarial network for zero-shot sketch-based 3d shape retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2902–2910, 2022. 1

[61] Xiaofan Xu, Alireza Dehghani, David Corrigan, Sam Caulfield, and David Moloney. Convolutional neural network for 3d object recognition using volumetric representation. In *International Workshop on Sensing, Processing and Learning for Intelligent Machines*, pages 1–5. IEEE, 2016. 6

[62] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1189, 2023. 1

[63] Jiaqi Yang, Ke Xian, Peng Wang, and Yanning Zhang. A performance evaluation of correspondence grouping methods for 3d rigid data matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1859–1874, 2019. 1

[64] Nan Yin, Li Shen, Mengzhu Wang, Xiao Luo, Zhigang Luo, and Dacheng Tao. Omg: towards effective graph classification against label noise. *IEEE Transactions on Knowledge and Data Engineering*, 2023. 4

[65] Zhixiong Zeng, Shuai Wang, Nan Xu, and Wenji Mao. Pan: Prototype-based adaptive network for robust cross-modal retrieval. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1125–1134, 2021. 2

[66] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 18408–18419, 2021. 2

[67] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations*, 2018. 5

[68] Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Semi-detr: Semi-supervised object detection with detection transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23809–23818, 2023. 2

[69] Ziyi Zhang, Weikai Chen, Chaowei Fang, Zhen Li, Lechao Chen, Liang Lin, and Guanbin Li. Rankmatch: Fostering confidence and consistency in learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1644–1654, 2023. 2, 4

[70] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019. 2, 6

[71] Lei Zhu, Xize Wu, Jingjing Li, Zheng Zhang, Weili Guan, and Heng Tao Shen. Work together: correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2