Multimodal Semi-Supervised Learning for 3D Objects

Zhimin Chen[1]
zhiminc@clemson.edu
Longlong Jing[2]
ljing@gradcenter.cuny.edu
Yang Liang [2]
lyang1@ccny.cuny.edu
YingLi Tian [2]
ytian@ccny.cuny.edu
Bing Li (Corresponding author) [1]
bli4@clemson.edu

[1] Clemson University
Clemson, USA
[2] The City University of New York,
New York, USA

Abstract

In recent years, semi-supervised learning has been widely explored and shows excellent data efficiency for 2D data. There is an emerging need to improve data efficiency for 3D tasks due to the scarcity of labeled 3D data. This paper explores how the coherence of different modalities of 3D data (e.g. point cloud, image, and mesh) can be used to improve data efficiency for both 3D classification and retrieval tasks. We propose a novel multimodal semi-supervised learning framework by introducing instance-level consistency constraint and a novel multimodal contrastive prototype (M2CP) loss. The instance-level consistency enforces the network to generate consistent representations for multimodal data of the same object regardless of its modality. The M2CP maintains a multimodal prototype for each class and learns features with small intra-class variations by minimizing the feature distance of each object to its prototype while maximizing the distance to the others. Our proposed framework significantly outperforms all the state-of-the-art counterparts for both classification and retrieval tasks by a large margin on the modelNet10 and ModelNet40 datasets.

1 Introduction

Due to the scarcity of large-scale labeled dataset, in recent years, the semi-supervised learning method has been drawing wide attention and showing the great potential of boosting up the performance of networks by jointly training on both limited labeled data and a large number of unlabeled samples [21, 24, 31, 41, 54]. It can significantly improve the data efficiency of the neural network training by leveraging unlabeled data, such as Pseudo-Labeling [21] which uses the confident prediction of the network as labels to further optimize the network, and FixMatch [37] which optimizes the network to predict consistency output for images with different augmentation from the same image.

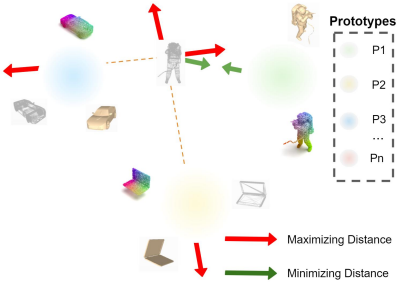

Figure 1: The M2CP maintains a multimodal prototype for each class and learns features with small intra-class variations by minimizing the feature distance of each object to its prototype while maximizing the distance to the others.

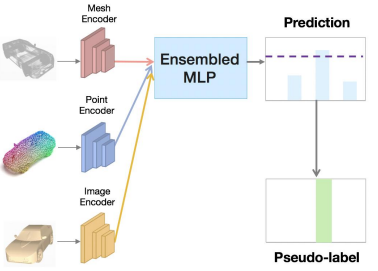

Figure 2: For the unlabeled samples, the features from multiple modalities are aggregated together by an Ensembled-MLP to produce more reliable and consistent pseudo-labels which will be used by the our proposed M2CP loss.

Although many semi-supervised learning methods have been proposed for 2D-related image recognition tasks, semi-supervised learning for 3D-related tasks has not been widely explored. Moreover, we observe that methods that were originally proposed for 2D-related tasks are not able to achieve comparable performance for 3D tasks (e.g., 3D object classification). Different from image and video data, 3D data usually consists of different modalities. The multimodal coherent among these modalities contain rich semantic information of the objects, which can be utilized to advance the 3D semi-supervised learning. Motivated by multimodal learning in other fields [4, 18, 36], we propose a multimodal semi-supervised learning framework based on two novel constraints including instance-level consistency and multimodal contrastive prototype (M2CP) constraints.

As shown in Fig. 1, the M2CP maintains a multimodal prototype for each class and learns to minimize the feature distance of each object to its prototype while maximizing the distance to the others. By minimizing the M2CP loss, the features from different classes are more separable while the features from the same class are closer, and it can potentially benefit the classification and retrieval tasks. Extensive experiments are conducted on two public benchmark datasets (i.e. ModelNet10 and ModelNet40) for both classification and retrieval tasks with three different modalities including point cloud, mesh, and image. Our key contributions are summarized as follows: 1) A novel multimodal contrastive prototype (M2CP) loss is proposed to learn the coherent embedding across multi-modalities. It can simultaneously minimize the intra-class distances and maximize the inter-class distances of embedding features by utilizing prototypes in semi-supervised learning. 2) We propose a novel multimodal semi-supervised framework that further encompasses instance-level consistency loss which enforces the network to generate consistent predictions for multimodal data of the same object regardless of its modality. 3) Our comprehensive experiments and ablation studies demonstrated that our proposed method significantly outperforms the state-of-the-art semi-supervised learning methods for both 3D object classification and retrieval tasks across point cloud, mesh, and image modalities on the ModelNet10 and ModelNet40 datasets.

2 Related Work

3D Object Classification: 3D object classification is a fundamental task for 3D understanding [28, 29, 42, 46, 51, 51, 52, 55, 56]. Since 3D objects usually can be represented in different modalities while each one has its advantages, many methods have been proposed for different modalities including image [40], mesh [9, 11, 33, 45], point cloud [22, 28, 29, 44], etc. However, these methods normally only focus on one modality. Our method is specifically designed to explore the coherence of different modalities of 3D data for semi-supervised learning.

Semi-Supervised Learning: Many semi-supervised learning methods have been proposed, but most of them focused on image-related tasks and some of them are very difficult to be directly transferred to other tasks or data [2, 5, 10, 21, 24, 25, 41]. These methods usually learn by enforcing networks to produce consistent predictions for different views of the same data [1, 20, 37], by minimizing the entropy of the predictions on unlabeled data [21], or by other types of regularization [54]. With the rapid development of self-supervised learning, various self-supervised tasks are used as auxiliary loss for semi-supervised learning [3, 35, 54]. Recently, the 3D semi-supervised learning for tasks including 3D object classification [32, 39], 3D semantic segmentation [6, 23, 26] and 3D object detection [43, 57] start to draw attention from the community. Most of these methods mainly use the data from one modality and the coherence of multimodal data is normally ignored. By utilizing the multimodal data with our proposed constraints, we set a comprehensive benchmark for both 3D classification and retrieval tasks and our proposed model significantly outperforms the most recent state-of-the-art methods [32, 39].

Self-Supervised 3D Feature Learning: Due to the advantage of utilizing unlabeled data, more and more 3D self-supervised learning methods have been proposed [15, 16, 32, 34, 49, 53, 58]. The existing self-supervised learning methods normally learn features by accomplishing pre-defined tasks such as contrasting [19], context prediction [12], orientation prediction [27], etc. The self-supervised learning methods do not require any labels, therefore, they can be used as an auxiliary task for other tasks to help the network to learn more representative features.

Multimodal Feature Learning: Multi-modality has been widely studied in many tasks. Different modalities capture data from different perspectives while the features from multiple modalities are usually complementary with each other. A typical example is the two-stream network [8, 36] for video action classification task which fuses feature extracted from RGB video clips and features from optical flow stacks [7, 17]. The similar idea has been explored in many other tasks including RGBD semantic segmentation [30], video object detection [59], 3D object detection, and sentiment analysis [13] etc. In this paper, we propose a novel multimodal semi-supervised learning framework with two constraints to jointly learn from both labeled and unlabeled data for 3D classification and retrieval tasks.

3 Multimodal Semi-Supervised Learning

Fig. 3 is the overview of our framework, with labeled and unlabeled three modalities of point cloud, mesh, and image as input. The framework consists of three components including supervised learning on labeled data, instance-level consistency learning on unlabeled data, and multimodal contrastive prototype (M2CP) learning on both labeled and unlabeled data. The generalized formulation of our proposed model is described in the following sections.

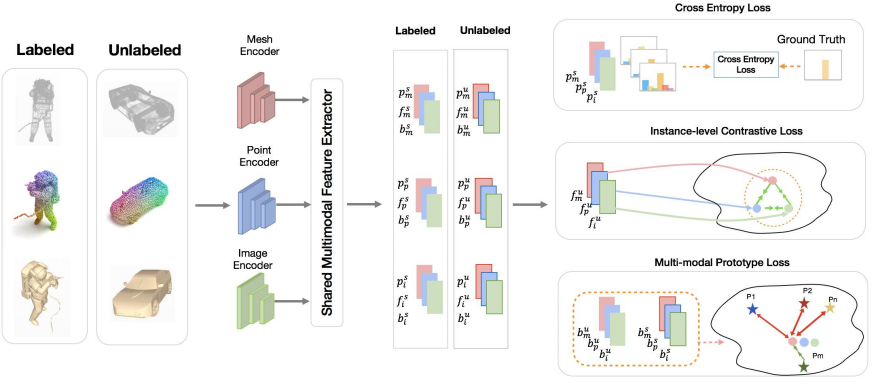

Figure 3: An overview of the proposed framework for multimodal semi-supervised learning. Our model is jointly trained on data with three different modalities from both labeled and unlabeled data. Three loss functions are employed to train the network including (1) a regular cross-entropy loss on the labeled data to learn discriminative features, (2) an instance-level consistency loss to enforce network to predict consistency features for multimodal data, and (3) our proposed novel multimodal contrastive prototype loss which minimize the intra-class distances and maximize the inter-class distances of multimodal features simultaneously.

3.1 Problem Setup

Given a set of limited labeled data X_L and a large amount of unlabeled data X_U, the labeled data X_L is formulated as:

$$X_L = \left\{t_i^l\right\}_{i=1}^{N_l}, t_i^l = \left(s_i^l, y_i\right), s_i^l = \left\{x_i^{lm}\right\}_{m=1}^{M}, \tag{1}$$

where t_i^l is a instance of labeled data containing M modalities $\left\{x_i^{lm}\right\}_{m=1}^{M}$ with label y_i.

The unlabeled data X_U is formulated as:

$$X_U = \{t_i^u\}_{i=1}^{N_u}, t_i^u = (s_i^u), s_i^u = \{x_i^{um}\}_{m=1}^{M}, \tag{2}$$

where each instance in the unlabeled data t_i^u only contains data $\{x_i^{um}\}_{m=1}^{M}$ in M different modalities. Our model is trained on both X_L and X_U for multimodal semi-supervised learning with the proposed constraints.

3.2 Representation Learning

Given each data sample $\mathbf{x}_i^m$ from modality m, a network F_m that designed for modality m produces a general hidden feature vector as:

$$H_i^m = F_m(x_i^m), \tag{3}$$

while H_i^m is the general hidden feature to represent the data x_i^m. There are m different feature encoders in total and one for each modality.

To implicitly enforce the consistency, two shared multimodal feature encoders are employed across all the modalities to map each generally hidden feature vector H_i^m into two outputs. One is a task-specific feature vector and the other is the classification prediction as:

$$E_i^m = W(H_i^m), \quad \hat{y}_i^m = G(H_i^m), \tag{4}$$

where E_i^m is a task-specific feature vector to represent the data x_i^m which will be used for instance-level consistency learning and $\hat{y}_i^m$ is the corresponding classification prediction based on x_i^m. By sharing $W()$ and $G()$ across all the modalities, the network implicitly enforces the feature extractors $F_1, ..., F_m$ to generate consistent features for different modalities of the same object.

Therefore, for each data instance t_i, no matter it is from labeled dataset X_L or unlabeled data X_U, task-specific feature vectors E_i^1, ... E_i^m, and predictions $\hat{y}_i^1$, ..., $\hat{y}_i^m$ are obtained with the shared multimodal encoder from a set of general hidden feature vectors H_i^1, ... H_i^m. Our proposed constraints are optimized over these extracted features and predictions.

3.2.1 Supervised Training

The supervised training over the labeled data X_L helps the network to learn discriminative features. For each sample x_i from the labeled set X_U, the cross-entropy loss is calculated between all the predictions $\hat{y}_i^m$ and the fused prediction result y^f. The corresponding ground truth label:

$$L_e = -\frac{1}{N}(\sum_{i=1}^{N}(\sum_{m=1}^{M} y_i \cdot log(\hat{y}_i^m) + y_i \cdot log(y_i^f)), \tag{5}$$

in which the cross-entropy loss from M and different modalities fused prediction are averaged over N samples to optimize the network.

3.2.2 Instance-Level Consistency Learning

Normally there are two ways to perform the instance-level consistency learning either from prediction level or from feature level. The existing methods mainly optimize networks by enforcing consistency in the prediction level such as enforcing networks to make the same classification prediction over two views of the same data sample. However, the predictions of networks are often very noisy especially when the labeled data is very limited, and the consistency learning from the prediction level inevitably involves severe noises during the training. Relying on the multimodal attributes of 3D data, we propose to regularize the network with instance-level consistency learning from the feature level in which the network is trained to predict consistent features E_m for data from different modalities of the same object.

We propose to directly maximize the similarity of multimodal features from the same object while minimizing the similarity of multimodal features from different objects. Given the recent progress of contrastive learning, the contrastive loss is employed for the instance-level consistency learning over the task-specific features E_m extracted by our shared multimodal feature encoder. Given any two modalities x_i^{m1} and x_i^{m2} from a total of M different modalities of the data x_i, the multimodal features are firstly extracted and represented as E_i^{m1}, ... E_i^{m2} from the shared multimodal feature encoder. Then the instance-level cross-model consistency loss is optimized through:

$$\mathcal{L}_i(E_i^{m1}, E_i^{m2}) = -\log \frac{h(E_i^{m1}, E_i^{m2})}{\sum_{k=1}^{B} h(E_i^{m1}, E_k^{m2})}, \tag{6}$$

where $h(\mathbf{u}, \mathbf{v}) = \exp\left(\frac{\mathbf{u}^\top \mathbf{v}}{||\mathbf{u}||_2 ||\mathbf{v}||_2} / \tau\right)$ is the exponential of cosine similarity measure, τ is the temperature, and B is the batch size. Each mini-batch consists of data from multiple modalities, and the instance-level consistency loss is calculated on combinations of any of the two modalities.

3.3 Multimodal Contrastive Prototype Learning

The instance-level consistency focuses on the consistency of the instance-level feature representation, but the relations among instance and classes are not utilized. To thoroughly utilize the information hidden on categories, we propose the multimodal contrastive prototype loss to jointly constraint the intra-class and the inter-class variations by using both labeled and unlabeled data.

A prototype P_i is defined for each class i and is learned and updated through the training. Each prototype represents the semantic center for one class in the feature space. For each data sample x_i^m, the distance of features of x_i^m to its corresponding prototype P_i is minimized while the distance with the rest of the prototypes is simultaneously maximized. Formally, for data sample data x_i^m with its general hidden features H_i^m and prototype P_i, the multimodal contrastive prototype loss is formulated as:

$$\mathcal{L}_p = -\log \frac{g(H_i^m, P_i)}{\sum_{k=1}^{C} g(H_i^m, P_k)}, \tag{7}$$

while $g(H_i^m, P_i) = exp(-\frac{\|H_i^m - P_i\|_2^2}{\tau})$, C is the number of prototypes ,and τ is the temperature.

For each data sample from the labeled data, its category is required for the M2CP loss to know the distance with which prototypes should be minimized. To fully utilize a large amount of unlabeled data, we extend this loss to both labeled and unlabeled data by generating pseudo-labels for the unlabeled data during the training. A simple idea would be to select the confident predictions (i.e. if the prediction $\hat{y_i^m}$ is larger than a threshold) and then use these confident predictions as labels for these selected unlabeled data as data to jointly train the multimodal contrastive prototype loss. However, this may lead to inconsistent pseudo-labels for the different modalities of the same object.

To generate consistent pseudo-labels for different modalities of the same object, we propose to fuse the general features F_m from different modalities to obtain an object-level prediction which will then be used as pseudo-labels for all the modalities of this object, as shown in Fig. 2. Formally, the features H_i^1, ..., H_i^m from m modalities are aggregated concatenated together to represent the object t_i to get the final prediction:

$$y_i^f = K(H_i^1, ..., H_i^m), \tag{8}$$

while K has two MLP layers network to predict y_i based on the concatenated features of multiple modalities. Having access to multiple modalities, the prediction y_i is more reliable compared to the predictions from a single modality. The pseudo-labels are further created as:

$$\hat{y}_i^c = \begin{cases} 1, & \text{if } max(y_i^f) \geq \delta \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

while δ is the pre-defined threshold. By generating pseudo-labels for unlabeled data, our proposed multimodal contrastive prototype loss is able to be trained by both labeled data X_L and large-scale unlabeled data X_U.

Our entire framework is jointly optimized on both labeled and unlabeled data with the three weighted loss functions as:

$$Loss = \alpha L_e + \beta L_i + \lambda L_p, \tag{10}$$

where α, β, and λ are the weights for each loss term.

3.4 Architecture

Feature Extractor: There are three feature extractors and one for each modality. Due to the powerful ability to learn image features, ResNet [14] is used as backbone networks for image feature extractors. For the point cloud feature extractor, the DGCNN [44] is employed due to its powerful ability to capture local structures with the KNN graph module from the point cloud. The MeshNet [9] is chosen as a feature encoder for mesh modality, and it takes the n faces and its normal vectors as inputs. The shared multimodal feature encoder consists of two parallel branches while one has three MLP layers with a size of $512, 256, 256$ and the other with a size of $512, 256, C$ while C is the number of classes.

4 Experimental Results

Datasets: The proposed framework is evaluated on two datasets including ModelNet10 [47] and ModelNet40 [47] with different percentages of labeled samples. The ModelNet40 dataset is a 3D object benchmark that consists of $12,311$ samples belong to 40 different categories while $9,843$ for training and $2,468$ for testing. The ModelNet10 dataset consists of $4,900$ samples belong to 10 categories with $3,991$ for training and 909 for testing.

Setup and Training Details On the ModelNet40 dataset, our model is trained with an SGD optimizer with a learning rate of 0.01 for a total of $10,0000$ iterations. The learning rate is reduced by 90% every $4,0000$ iterations. On the ModelNet10 dataset, the model is trained for a total of $6,0000$ iterations and the learning rate starts from 0.01 and is reduced by 90% every $2,0000$ iterations. The weights are 1, 2, and 9 for cross-entropy, instance-level contrastive loss, and multimodal contrastive prototype loss respectively. For all the experiments, a batch size of 48 is used while half of them are labeled data and the other half are unlabeled data.

4.1 Performance on Semi-Supervised 3D Object Classification

Since there are only few existing semi-supervised learning methods specifically designed for the 3D object classification task and without a comprehensive semi-supervised benchmark, we first compare with methods that were originally designed for 2D semi-supervised learning but apply them to the 3D object classification task. We compare with the supervised baseline and three different semi-supervised methods include Pseudo-Labeling (PL) [21], FixMatch [37], and S4L [54] under the same settings. Worth to note that the FixMatch-based methods achieve state-of-the-art performance on many semi-supervised tasks [57], [38], [50], [48]. For a fair comparison, all the methods are using the same backbones and data augmentations as our methods, and only one modality is available during the inference phase for all the methods.

The performance comparison with the above-mentioned four methods on the ModelNet40 and ModelNet10 datasets for 3D object classification are shown in Table 1. To extensively evaluate the performance, we compare with the other methods under different percentages (2%, 5%, and 10%) of labeled data for both datasets. The following conclusions can be drawn from the comparison: 1) Since the state-of-the-art semi-supervised learning methods were mainly designed for image-based tasks, they perform well on the image modality while having a negligible impact on the point cloud and mesh modalities. Directly adapting these methods to the 3D classification task obtains unsatisfied results due to the lack of specific constraints. 2) By jointly training our framework with multiple modalities, the performance with all three modalities is significantly improved regardless of the amount of

labeled data. 3) With the different amounts of labeled data, our method consistently significantly improves the classification performance and outperforms all three state-of-the-art semi-supervised learning methods. These results confirm the advantage of our proposed multimodal semi-supervised method.

ModelNet40						ModelNet10					
Test Modality	Baseline	PL [21]	FixMatch [37]	S4L [54]	Ours	Test Modality	Baseline	PL [21]	FixMatch [37]	S4L [54]	Ours
2% of Labeled data											
Image	69.61	72.20	75.69	80.96	**82.78**	Image	73.57	74.04	70.44	80.51	**85.46**
Point	63.21	65.68	65.56	68.64	**79.86**	Point	77.31	78.41	78.74	81.61	**84.36**
Mesh	48.18	49.33	52.59	51.72	**78.81**	Mesh	66.41	63.00	61.44	65.40	**86.13**
5% of Labeled data											
Image	82.29	83.18	85.49	84.12	**88.61**	Image	83.70	86.56	84.36	88.33	**92.14**
Point	76.62	79.74	79.38	78.69	**85.29**	Point	83.59	86.34	85.13	84.91	**89.87**
Mesh	71.19	73.49	72.76	77.02	**86.51**	Mesh	80.51	80.73	82.60	79.74	**90.75**
10% of Labeled data											
Image	85.90	86.95	89.02	87.16	**91.61**	Image	91.85	91.74	90.86	92.07	**93.95**
Point	82.86	84.04	84.81	83.75	**88.49**	Point	87.44	88.00	88.33	88.16	**91.63**
Mesh	80.39	82.47	81.36	82.42	**88.29**	Mesh	84.25	87.11	83.29	81.83	**92.84**

Table 1: Performance comparison for the 3D object classification task with the state-of-the-art semi-supervised learning methods on the ModelNet40 and ModelNet 10 dataset with different percentages of labeled data.

ModelNet40						ModelNet10					
Test Modality	Baseline	PL [21]	FixMatch [37]	S4L [54]	Ours	Test Modality	Baseline	PL [21]	FixMatch [37]	S4L [54]	Ours
2% of Labeled data											
Image	63.01	62.84	73.27	73.13	**81.50**	Image	68.90	65.87	73.34	78.52	**83.82**
Point	55.43	60.42	61.57	59.41	**78.45**	Point	72.81	76.42	76.96	68.94	**84.85**
Mesh	50.50	52.39	53.37	54.81	**80.31**	Mesh	64.17	70.99	72.62	67.23	**84.09**
5% of Labeled data											
Image	73.68	74.38	78.87	74.90	**85.71**	Image	79.00	77.73	82.49	80.25	**87.67**
Point	57.92	64.02	63.09	61.11	**82.05**	Point	72.98	76.60	75.25	69.39	**87.62**
Mesh	56.98	63.94	63.59	58.81	**84.84**	Mesh	75.81	81.62	79.00	72.95	**88.43**
10% of Labeled data											
Image	78.15	79.11	82.24	79.28	**86.96**	Image	85.33	85.28	87.68	83.39	**90.74**
Point	60.20	64.50	63.60	64.96	**84.16**	Point	72.93	76.33	74.70	71.18	**89.75**
Mesh	60.20	73.22	73.74	72.02	**84.29**	Mesh	81.01	84.70	78.03	70.92	**90.61**

Table 2: Performance comparison for the 3D object retrieval task with other state-of-the-art semi-supervised learning methods on the ModelNet40 and ModelNet10 dataset with different percentages of labeled data.

4.2 Performance on Semi-Supervised Object Retrieval

We further evaluate the performance of the 3D object retrieval task and compare it with the state-of-the-art semi-supervised methods on both ModelNet40 and ModelNet10 datasets. Following the convention, the mean Average Precision (mAP) is used to indicate the performance.

We report the retrieval performance with different amounts of labeled data for all three modalities. As shown in Table 2, all these three state-of-the-art semi-supervised learning methods can only improve the performance on the image modality while the performances for other modalities sometimes are even worse than the baseline, which is probably due to the noises during the training. Benefited from our novel constraints, our method significantly improves the performance consistently for all the modalities by using different percentages

Modality	L_e	L_e, L_i	L_e, L_p	L_e, L_i, L_p
3D Object Retrieval				
Image	78.15	76.74	85.26	**86.96**
Point	60.20	70.40	82.51	**84.16**
Mesh	60.20	74.91	78.04	**84.29**
3D Object Classification				
Image	85.90	89.34	89.79	**91.61**
Point	82.86	85.78	87.03	**88.49**
Mesh	80.39	**88.61**	85.82	88.29

Table 3: Ablation study for the combination of losses to the 3D object classification and retrieval tasks on ModelNet40 dataset with 10% of labeled data.

Modality	Mesh-Image	Image-Point	Point-Mesh	All
3D Object Retrieval				
Image	86.15	86.23	—	**86.96**
Point	—	82.90	82.47	**84.16**
Mesh	81.27	—	79.68	**84.29**
3D Object Classification				
Image	89.95	90.48	—	**91.61**
Point	—	87.48	86.47	**88.49**
Mesh	86.83	—	84.93	**88.29**

Table 4: Ablation study for the number of modalities to the 3D object classification and retrieval tasks on ModelNet40 dataset with 10% of labeled data.

of labeled data. These results demonstrate the effectiveness of our proposed framework and the generalizability in the 3D object retrieval task.

4.3 Ablation Study

Ablation Study for Loss Functions: Our proposed framework is jointly trained with three loss functions including cross-entropy loss L_e, instance-level consistency loss L_i, and a novel multimodal prototype loss L_p. To understand the impact of each loss term, we conduct ablation studies with four combinations of different loss functions including: 1) L_e, 2) $L_e + L_i$, 3) $L_e + L_p$, 4) $L_e + L_i + L_p$. We report the performance using different amounts of labeled data for both 3D classification and 3D retrieval tasks on the ModelNet40 dataset in Table 3. From the result of Table 3, we can draw the conclusion that: 1) When only the cross-entropy loss L_e is used, the performance for both classification and retrieval tasks with all the modalities are very low due to the very limited labeled samples. 2) When the M2CP loss L_p is jointly used with the cross-entropy loss L_e, the performances for all the tasks are significantly improved compared to the baseline, as well as significantly outperforms the performance of $L_e + L_i$. The results are consistent with our hypothesis since L_i does not use the category information and only enforces the instance-level consistency, while the M2CP utilizes the category information to regularize the hidden features. 3) The best performances are achieved for all the tasks when all the three losses are used indicating that they are indeed complementary with each other.

Ablation Study for Number of Modalities: Compared to other semi-supervised methods, our method is designed to explicitly leverage the multimodal coherence of multimodal data with the proposed novel constraints. Our model is jointly trained from three Modalities including point cloud, mesh, and image. When more modality data are available, better performance should be achieved since more multimodal constraints are available to the networks. To verify the impact of the number of Modalities, we conduct experiments by training with three different modality combinations including (1) point cloud and image, (2) point cloud and mesh, and (3) image and mesh. The performance for both classification and retrieval tasks on the ModelNet40 dataset is shown in Table 4. The performances for both tasks are the best when all three Modalities are used in training. This confirms our assumption that more Modalities can provide more constraints which produces better performance.

Method	Percentage	Modality	Accuracy
Info3D [32]	2%	Point cloud	71.06%
Ours	2%	Point cloud	**79.86**%
Info3D [32]	5%	Point cloud	80.48%
Ours	5%	Point cloud	**85.29**%
Deep Co-training [39]	10%	Point cloud	83.50%
FixMatch [37]	10%	Point cloud	84.81%
Ours	10%	Point cloud	**88.49**%
Deep Co-training [39]	10%	Image	89.00%
FixMatch [37]	10%	Image	89.02%
Ours	10%	Image	**91.61**%

Table 5: Comparison with the most recent state-of-the-art 2D and 3D semi-supervised methods [32, 37, 39]. Our model significantly outperforms all of them with different modalities and different settings on ModelNet40.

5 Comparison with the State-of-the-Art 3D Semi-Supervised Methods

To further demonstrate the capability of our proposed methods, we compare with the state-of-the-art methods [32, 39] that are specifically designed for the 3D semi-supervised object classification task. The performance comparison with these methods on the ModelNet40 dataset is shown in Table 5.

Our method significantly outperforms the state-of-the-art methods [32, 39] with different modalities under different settings demonstrating the effectiveness of our proposed method. Our method outperforms the Info3D [32] by almost 9% when only 2% labeled data is available during training. The Deep Co-training [39] is specifically designed for 3D semi-supervised learning which mainly uses the consistency from the **prediction level** of the multimodal data as constraints however, the results are comparable with the 2D based method FixMatch [37]. Relying on our proposed constraints learning directly from **feature level**, our model significantly outperforms state-of-the-art 2D and 3D semi-supervised methods by a large margin with both modalities under the same setting on the ModelNet40 dataset. Moreover, with our proposed novel M2CP loss directly optimizing from the feature level, our model is able to learn features with small intra-class variations which can achieve state-of-the-art results for the 3D retrieval tasks.

6 Conclusion

We have proposed a novel multimodal semi-supervised learning method for 3D objects based on the coherence of multimodal data. The network jointly learns from both labeled and unlabeled data mainly using the proposed instance-level consistency and multimodal contrastive prototype (M2CP) constraints. Our proposed method remarkably outperforms the state-of-the-art semi-supervised learning methods and the baseline on both 3D classification and retrieval tasks. These results demonstrate that it is a promising direction to study how to apply the multimodal for 3D semi-supervised learning tasks.

7 Acknowledgement

This work was supported in part by U.S. DOT UTC grant 69A3551747117.

References

[1] 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] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

[3] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. *arXiv preprint arXiv:2101.08482*, 2021.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[6] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[9] Yutong Feng, Yifan Feng, Haoxuan You, Xibin Zhao, and Yue Gao. Meshnet: mesh neural network for 3d shape representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8279–8286, 2019.

[10] Atin Ghosh and Alexandre H Thiery. On data-augmentation and consistency-based semi-supervised learning. *arXiv preprint arXiv:2101.06967*, 2021.

[11] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[12] Kaveh Hassani and Mike Haley. Unsupervised multi-task feature learning on point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8160–8171, 2019.

[13] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. *arXiv preprint arXiv:2005.03545*, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[15] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021.

[16] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? *arXiv preprint arXiv:2104.11225*, 2021.

[17] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[18] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss. *arXiv preprint arXiv:2008.03561*, 2020.

[19] Longlong Jing, Ling Zhang, and Yingli Tian. Self-supervised feature learning by cross-modality and cross-view correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1591, 2021.

[20] Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. Featmatch: Feature-based augmentation for semi-supervised learning. In *European Conference on Computer Vision*, pages 479–495. Springer, 2020.

[21] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

[22] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9267–9276, 2019.

[23] Jilin Mei, Biao Gao, Donghao Xu, Wen Yao, Xijun Zhao, and Huijing Zhao. Semantic segmentation of 3d lidar data in dynamic scene using semi-supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 21(6):2496–2509, 2019.

[24] 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.

[25] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv preprint arXiv:1804.09170*, 2018.

[26] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[27] Omid Poursaeed, Tianxing Jiang, Han Qiao, Nayun Xu, and Vladimir G Kim. Self-supervised learning of point clouds via orientation estimation. In *2020 International Conference on 3D Vision (3DV)*, pages 1018–1028. IEEE, 2020.

[28] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[29] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.

[30] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5199–5208, 2017.

[31] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018.

[32] Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *European Conference on Computer Vision*, pages 626–642. Springer, 2020.

[33] Jonas Schult, Francis Engelmann, Theodora Kontogianni, and Bastian Leibe. Dualconvmesh-net: Joint geodesic and euclidean convolutions on 3d meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[34] Charu Sharma and Manohar Kaul. Self-supervised few-shot learning on point clouds. *arXiv preprint arXiv:2009.14168*, 2020.

[35] Chenyang Si, Xuecheng Nie, Wei Wang, Liang Wang, Tieniu Tan, and Jiashi Feng. Adversarial self-supervised learning for semi-supervised 3d action recognition. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020.

[36] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

[37] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[38] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.

[39] Mofei Song, Yu Liu, and Xiao Fan Liu. Semi-supervised 3d shape recognition via multimodal deep co-training. In *Computer Graphics Forum*, volume 39, pages 279–289. Wiley Online Library, 2020.

[40] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

[42] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6411–6420, 2019.

[43] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14615–14624, 2021.

[44] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019.

[45] Ruben Wiersma, Elmar Eisemann, and Klaus Hildebrandt. Cnns on surfaces using rotation-equivariant features. *ACM Transactions on Graphics (TOG)*, 39(4):92–1, 2020.

[46] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2019.

[47] 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 conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[48] Rongchang Xie, Chunyu Wang, Wenjun Zeng, and Yizhou Wang. Humble teacher and eager student: Dual network learning for semi-supervised 2d human pose estimation. *arXiv preprint arXiv:2011.12498*, 2020.

[49] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020.

[50] Bo Xiong, Haoqi Fan, Kristen Grauman, and Christoph Feichtenhofer. Multi-view pseudo-labeling for semi-supervised learning from video. *arXiv preprint arXiv:2104.00682*, 2021.

[51] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13706–13715, 2020.

[52] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018.

[53] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–215, 2018.

[54] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019.

[55] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5565–5573, 2019.

[56] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. *arXiv preprint arXiv:2012.09164*, 2020.

[57] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020.

[58] Yongheng Zhao, Tolga Birdal, Haowen Deng, and Federico Tombari. 3d point capsule networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1009–1018, 2019.

[59] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 408–417, 2017.