applied sciences

MDPI

Article

Multi-Modal 3D Shape Clustering with Dual Contrastive Learning

Guoting Lin, Zexun Zheng *, Lin Chen, Tianyi Qin and Jiahui Song

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China; gtlin@tju.edu.cn (G.L.); lin_chen@tju.edu.cn (L.C.); tyqin@tju.edu.cn (T.Q.); songjh@tju.edu.cn (J.S.)
* Correspondence: zhengzexun@tju.edu.cn

Abstract: 3D shape clustering is developing into an important research subject with the wide applications of 3D shapes in computer vision and multimedia fields. Since 3D shapes generally take on various modalities, how to comprehensively exploit the multi-modal properties to boost clustering performance has become a key issue for the 3D shape clustering task. Taking into account the advantages of multiple views and point clouds, this paper proposes the first multi-modal 3D shape clustering method, named the dual contrastive learning network (DCL-Net), to discover the clustering partitions of unlabeled 3D shapes. First, by simultaneously performing cross-view contrastive learning within multi-view modality and cross-modal contrastive learning between the point cloud and multi-view modalities in the representation space, a representation-level dual contrastive learning module is developed, which aims to capture discriminative 3D shape features for clustering. Meanwhile, an assignment-level dual contrastive learning module is designed by further ensuring the consistency of clustering assignments within the multi-view modality, as well as between the point cloud and multi-view modalities, thus obtaining more compact clustering partitions. Experiments on two commonly used 3D shape benchmarks demonstrate the effectiveness of the proposed DCL-Net.

Keywords: multi-modal clustering; unsupervised learning; 3D shapes; contrastive learning

check for updates

Citation: Lin, G.; Zheng, Z.; Chen, L.; Qin, T.; Song, J. Multi-Modal 3D Shape Clustering with Dual Contrastive Learning. *Appl. Sci.* **2022**, *12*, 7384. https://doi.org/10.3390/app12157384

Academic Editors: Zhaoqing Pan, Bo Peng and Jinwei Wang

Received: 15 June 2022
Accepted: 11 July 2022
Published: 22 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the development of 3D scanning and modeling technology, 3D shapes have been widely employed in various applications of computer vision and multimedia fields, such as 3D printing, model retrieval, augmented reality, etc. [1–3]. How to effectively analyze large numbers of 3D shapes has become a research hot spot. In recent years, owing to the advanced development of deep learning, a series of deep 3D shape classification methods [4–6] have obtained satisfactory results. However, the success of deep neural networks critically relies on large-scale human-annotated data [7–9], which requires a laborious data annotation procedure. Under these circumstances, clustering has received increasing attention due to its powerful ability to divide massive amounts of unlabeled data [10,11]. Exploring effective 3D shape clustering methods has become a promising approach to overcome the above obstacle.

In practical application scenarios, 3D shapes are generally represented by different modalities due to the diversity of acquisition devices. As two popular 3D modalities, point clouds and multiple views are produced by 3D scanners and RGB cameras respectively, which have the advantages of flexible acquisition and low costs [12]. Specifically, point clouds describe 3D shapes with a series of disordered points, and the positional arrangement of those points preserves the spatial geometry of the 3D shapes [13]. Different from point clouds, multiple views are formed by a series of 2D images corresponding to different camera angles [14,15]. They contain rich visual information of 3D shapes, such as texture and color [16]. Since point clouds and multiple views describe 3D shapes from different perspectives, effectively exploiting the multi-modal properties is conducive to capturing more discriminative descriptions of 3D shapes and better revealing compact 3D shape clustering partitions.

Recently, contrastive learning has shown great success in unsupervised representation learning [17]. The core idea of contrastive learning is to maximize the representation similarities of positives while minimizing those of negatives, thus capturing more effective representations of data. Driven by this, some unsupervised 3D shape representation learning methods [18,19] successfully extract better cross-modal 3D shape representations by performing contrastive learning among different 3D modalities. However, since the above methods lack clustering-oriented learning objectives, the performance is usually limited when directly applying traditional clustering algorithms to the learned representations. In order to learn cross-modal representations that are suitable for clustering, several previous works [20,21] have integrated contrastive learning into multi-modal clustering for text and image data. By maximizing the similarities among the representations or the clustering assignments of different modalities in a contrastive learning manner, these methods have achieved encouraging results. Nonetheless, no existing work has focused on the multi-modal 3D shape clustering task. For the point clouds and multiple views of 3D shapes, in addition to the inter-modal correlations, different views within the multi-view modality also describe different local appearances of 3D shapes from particular angles. Therefore, how to jointly explore the inter-view correlations within the multi-view modality and the inter-modal correlations between the point cloud and multi-view modalities during the learning procedure of contrastive clustering remains a challenging issue for the multi-modal 3D shape clustering task.

To address the above issue, this paper proposes a dual contrastive learning network (DCL-Net) for multi-modal 3D shape clustering. The key motivation behind our design involved two aspects. Firstly, as for a 3D shape, different views within multi-view modality contain diverse appearances from different perspectives. Meanwhile, point cloud and multi-view modalities mainly focus on the geometric and visual information about 3D shapes, respectively. Simultaneously exploring the cross-view consistent representations of different views, as well as the cross-modal consistent representations of point cloud and multi-view modalities, contributes to obtaining more discriminative 3D shape descriptions. Secondly, different views within the multi-view modality and the corresponding point cloud of the same 3D shape all share consistent semantics. In addition to learning consistent representations, exploring the cross-view and cross-modal consistent clustering assignments is beneficial to boosting the robustness of the 3D shape features for clustering, thus further enhancing the compactness of clustering partitions. Therefore, by simultaneously performing cross-view contrastive learning within multi-view modality and cross-modal contrastive learning between point cloud and multi-view modalities at both the representation and clustering assignment levels, a representation-level dual contrastive learning module and an assignment-level dual contrastive learning module were developed in the proposed method. The key contributions of this paper are as follows:

(1) A dual contrastive learning network for multi-modal 3D shape clustering is proposed to discover the underlying clustering partitions of unlabeled 3D shapes. To the best of our knowledge, this is the first deep multi-modal 3D shape clustering method;
(2) By simultaneously ensuring the representation consistency within multi-view modality and between point cloud and multi-view modalities, a representation-level dual contrastive learning module is proposed to capture discriminative 3D shape features for clustering;
(3) To further boost the compactness of clustering partitions, an assignment-level dual contrastive learning module is proposed to simultaneously capture consistent clustering assignments within multi-view modality and between point cloud and multi-view modalities;
(4) Experimental results on two widely used 3D shape benchmark datasets are presented to demonstrate the superior clustering performance of the proposed DCL-Net.

The remaining of this paper is organized as follows. Section 2 briefly describes the three aspects of current research that are the most relevant to the proposed method. Section 3

introduces the proposed DCL-Net in detail. Section 4 presents a series of experimental results and analyses. In Section 5, the conclusion is summarized.

2. Related Works

2.1. Unsupervised 3D Shape Feature Learning

Due to the rapid growth of 3D shapes, significant progress has been made in unsupervised 3D shape feature learning. Many research works [22–28] have been proposed to learn 3D shape features from various 3D modalities, such as multiple views, point clouds, meshes, and voxels. For instance, Zhao et al. [23] proposed an autoencoder-based 3D point capsule network to extract 3D shape features via point cloud reconstruction. The method in [26] extracted structure-preserving 3D shape features by effectively encoding the local geometry structures of 3D meshes. Han et al. [28] proposed a recurrent neural network (RNN) architecture to learn global 3D features via the multiple view inter-prediction task. Furthermore, considering the multi-modal characteristics of 3D shapes, several cross-modal learning methods [18,19,29,30] have been proposed to boost the quality of 3D shape features by adequately exploiting information from different modalities. For example, Wu et al. [29] proposed a 3D generative adversarial network to capture 3D shape features by reconstructing 3D voxels from 2D images. Girdhar et al. [30] introduced a view encoder into the voxel autoencoder network to learn robust 3D shape features based on image-to-voxel generation. Nonetheless, the above methods are not oriented toward clustering tasks, thus it is difficult to ensure that the learned features are suitable for clustering.

2.2. Deep Multi-Modal Clustering

Multi-modal clustering aims to capture consistent underlying category partitions from multi-modal inputs, such as text [31], images [32], videos [33–35], etc. Due to the powerful feature extraction capability of deep neural networks [36,37], a number of deep multi-modal clustering methods [38–42] have been proposed over recent years. Ngiam et al. [38] introduced a deep autoencoder network to extract consistent representations across different modalities and obtained promising results in speech and vision tasks. Andrew et al. [39] adopted deep canonical correlation analysis (DCCA) to learn cross-modal consistent representations by maximizing the correlations between multi-modal features. Abavisani et al. [40] utilized multiple parallel autoencoders and a shared self-expression layer [43] to capture a joint cross-modal affinity matrix for clustering. The method in [42] adopted deep autoencoders to explore multi-modal shared representation while introducing adversarial training to disentangle the latent space. Zhou et al. [43] designed an adversarial network with an attention mechanism to learn cross-modal consistent representations for clustering. In summary, the current deep multi-modal clustering methods have made remarkable progress. However, the existing works have not focused on the multi-modal 3D shape clustering task. How to sufficiently exploit the advantages of deep learning to design an effective multi-modal 3D shape clustering method still needs to be further investigated.

2.3. Contrastive Learning

As a powerful approach to unsupervised representation learning, contrastive learning has attracted increasing amounts of research attention and several contrastive learning-based works [44–48] have recently emerged. He et al. [45] proposed a momentum contrastive method to facilitate unsupervised representation learning by regarding contrastive learning as a dictionary lookup and building a dynamic dictionary. Chen et al. [46] effectively simplified the framework in [47] by adopting a Siamese network with a prediction head while introducing powerful data augmentations to boost the quality of the learned features. Tian et al. [48] employed two asymmetric networks with an interactive prediction mechanism to learn image representations and avoided model collapse without negative samples. Motivated by the success of contrastive learning in unsupervised representation learning, several methods [20,21,49] have applied contrastive learning to multi-modal learn-

ing tasks. For example, Xu et al. [20] explored the common semantics of different modalities using feature contrastive learning and label contrastive learning. Trosten et al. [21] introduced contrastive learning to align multi-modal representations and achieved effective improvements in clustering performance. Although the above methods have achieved promising results, the existing explorations of multi-modal contrastive clustering have mainly focused on text- and image-related tasks. In contrast, this paper effectively utilizes the characteristics of multiple views and point clouds, and delivers a novel contrastive learning-based multi-modal 3D shape clustering method.

3. The Proposed Method

3.1. Architecture of DCL-Net

The overview architecture of the proposed DCL-Net is illustrated in Figure 1. Let $D = \{I_i^v, P_i\}_{i=1...N}^{v=1...V}$ denote a 3D shape dataset with N shapes, where I_i^v denotes the v-th view in the multiple views of the i-th 3D shape and P_i denotes the corresponding point cloud of the i-th 3D shape. As shown in the figure, the proposed DCL-Net includes a multi-modal feature extractor, a representation-level dual contrastive learning (RDCL) module, and an assignment-level dual contrastive learning (ADCL) module. Taking two views from different angles and the corresponding point cloud as inputs, the multi-modal feature extractor was adopted to extract the view and point cloud features. Afterward, the RDCL module was designed to capture more discriminative 3D shape features by simultaneously applying cross-view contrastive learning within the multi-view modality and cross-modal contrastive learning between the point cloud and multi-view modalities in the representation space. Moreover, by effectively applying cross-view and cross-modal contrastive learning to the clustering assignments, the ADCL module was designed to ensure clustering assignment consistency among different views and the corresponding point cloud, thus further boosting the 3D shape clustering performance. Finally, the clustering results are obtained from the soft labels predicted by the ADCL module.

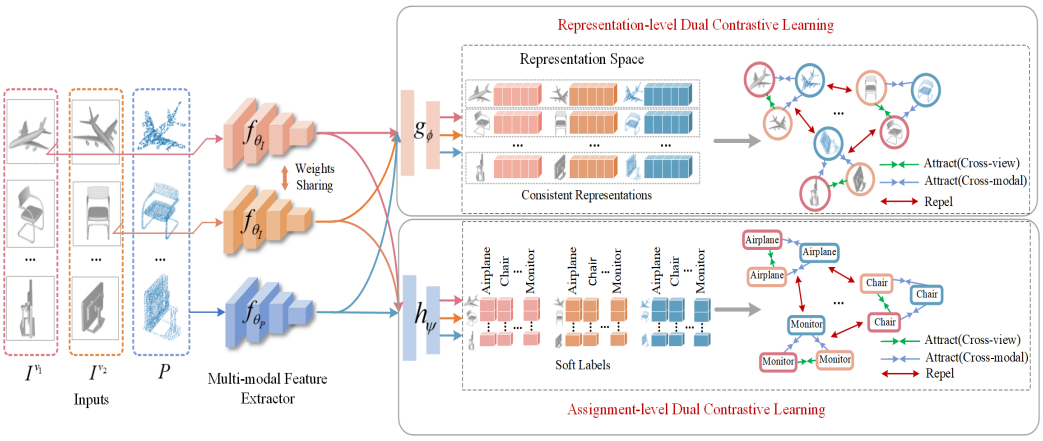


Figure 1. The architecture of the proposed DCL-Net for multi-modal 3D shape clustering.

3.2. Representation-Level Dual Contrastive Learning

For each 3D shape, different views within the multi-view modality and the corresponding point cloud share consistent semantics while containing complementary 3D shape information across both views and modalities. Simultaneously performing cross-view contrastive learning within multi-view modality and cross-modal contrastive learning between point cloud and multi-view modalities for consistent representation is conducive to capturing more discriminative 3D shape information from a comprehensive understanding of 3D shapes. To this end, an RDCL module that adopts both cross-view and cross-modal contrastive representation learning was developed to capture discriminative 3D shape features for clustering.

First, for the i-th 3D shape in a given mini-batch $D = \{I_i^v, P_i\}_{i=1\ldots n}^{v=1\ldots V}$ of size n, two views $\{I_i^{v_1}, I_i^{v_2}\}$ from different angles in the multiple views and the corresponding point cloud P_i are selected as the inputs for the network. Among them, $I_i^{v_1}$ and $I_i^{v_2}$ are arbitrarily chosen from the V views, thereby providing more cross-view and cross-modal combinations for the network learning. After that, to capture latent 3D shape features from the inputs $\{I_i^{v_1}, I_i^{v_2}, P_i\}$, a three-branch multi-modal feature extractor including a shared view encoder $f_{\theta_I}(\cdot)$ and a point cloud encoder $f_{\theta_P}(\cdot)$ is employed. The shared view encoder maps $I_i^{v_1}$ and $I_i^{v_2}$ into the view features $F_i^{v_1}$ and $F_i^{v_2}$ respectively, while the point cloud encoder is responsible for mapping P_i into the point cloud feature F_i^p. The mapping processes are calculated as follows:

$$F_i^{v_1} = f_{\theta_I}(I_i^{v_1}), \tag{1}$$

$$F_i^{v_2} = f_{\theta_I}(I_i^{v_2}), \tag{2}$$

$$F_i^p = f_{\theta_P}(P_i), \tag{3}$$

where θ_I and θ_P denote the parameters of the view encoder and the point cloud encoder, respectively.

Afterward, to effectively ensure the discrimination of the learned 3D shape features, a representation projection head $g_\phi(\cdot)$ is adopted in the RDCL module to further project the learned features into the representation space:

$$Z_i^{v_1} = g_\phi(F_i^{v_1}), \tag{4}$$

$$Z_i^{v_2} = g_\phi(F_i^{v_2}), \tag{5}$$

$$Z_i^p = g_\phi(F_i^p). \tag{6}$$

Considering the semantic consistency among different views and the corresponding point cloud, multiple view representations and the point cloud representation of the same 3D shape should maintain higher similarities than those of different 3D shapes. Therefore, the view representations and point cloud representation of the same 3D shape need to be taken as positives to be pulled together, while those of different 3D shapes need to be regarded as negatives to be pushed apart. In view of this, a representation-level cross-view contrastive loss L_{RCV} and a representation-level cross-modal contrastive loss L_{RCM} are simultaneously applied to the representation space to ensure both the cross-view representation consistency within the multi-view modality and the cross-modal representation consistency between the point cloud and multi-view modalities.

Specifically, the representation-level cross-view contrastive loss for the input view $I_i^{v_1}$ is calculated as follows:

$$L_{RCV_i}^{v_1} = -\log \frac{\exp(s(Z_i^{v_1}, Z_i^{v_2})/\tau_R)}{\sum_{j=1}^{n} [\exp(s(Z_i^{v_1}, Z_j^{v_1})/\tau_R) + \exp(s(Z_i^{v_1}, Z_j^{v_2})/\tau_R)]}, \tag{7}$$

where τ_R is the representation-level temperature parameter and is generally set to 0.5. $s(Z_i^{v_1}, Z_i^{v_2})$ denotes the pair-wise cosine similarity between $Z_i^{v_1}$ and $Z_i^{v_2}$, which is calculated by $s(Z_i^{v_1}, Z_i^{v_2}) = \frac{(Z_i^{v_1}, Z_i^{v_2})^{\mathrm{T}}}{\left\| Z_i^{v_1}, Z_i^{v_2} \right\|}$. By computing the representation-level cross-view contrastive loss for each arbitrarily chosen view in the mini-batch, L_{RCV} can be expressed in the form of:

$$L_{RCV} = \frac{1}{2n} \sum_{i=1}^{n} (L_{RCV_i}^{v_1} + L_{RCV_i}^{v_2}). \tag{8}$$

Similarly, the representation-level cross-modal contrastive loss for the view $I_i^{v_1}$ is calculated as follows:

$$L_{RCM1_i}^{v_1} = -\log \frac{\exp(s(Z_i^{v_1}, Z_i^p)/\tau_R)}{\sum_{j=1}^{n} [\exp(s(Z_i^{v_1}, Z_j^{v_1})/\tau_R) + \exp(s(Z_i^{v_1}, Z_j^p)/\tau_R)]}. \quad (9)$$

Note that, to avoid the bias of the point cloud representation toward a particular view, the cross-modal contrastive loss for the i-th shape is calculated between Z_i^p and $Z_i^{v_1}$ and between Z_i^p and $Z_i^{v_2}$, which are denoted as L_{RCM1_i} and L_{RCM2_i}, respectively. Therefore, the representation-level cross-modal contrastive loss L_{RCM} is calculated as follows:

$$L_{RCM} = \frac{1}{2n} \sum_{i=1}^{n} (L_{RCM1_i}^{v_1} + L_{RCM1_i}^{p} + L_{RCM2_i}^{v_2} + L_{RCM2_i}^{p}). \quad (10)$$

By combining L_{RCV} and L_{RCM}, the overall loss of the RDCL module is expressed as:

$$L_{RDCL} = L_{RCV} + L_{RCM}. \quad (11)$$

Under the constraint of the representation-level dual contrastive loss L_{RDCL}, the network is encouraged to distinguish different 3D shapes according to the cross-modal and cross-view consistent representations, thus effectively promoting the discrimination of the extracted 3D shape features for clustering.

3.3. Assignment-Level Dual Contrastive Learning

The ADCL module was designed to simultaneously ensure cross-view clustering assignment consistency within the multi-view modality and cross-modal clustering assignment consistency between the point cloud and multi-view modalities, thus further boosting the compactness of the learned 3D shape features for clustering. Specifically, for the extracted view and point cloud features $F_i^{v_1}$, $F_i^{v_2}$, and F_i^p of the i-th 3D shape, the ADCL module further maps them into soft labels using an assignment projection head $h_\psi(\cdot)$ with the following process:

$$Y_i^{v_1} = h_\psi(F_i^{v_1}), \quad (12)$$

$$Y_i^{v_2} = h_\psi(F_i^{v_2}), \quad (13)$$

$$Y_i^p = h_\psi(F_i^p). \quad (14)$$

Let $Y^u = [Y_1^u, Y_2^u, \ldots, Y_n^u]$ denote the outputs of the assignment projection head for the n 3D shapes in the mini-batch and G^u denote the transposed matrix of Y^u, where $u \in \{v_1, v_2, p\}$. In G^u, the i-th column vector denotes the soft label of the i-th 3D shape and the k-th row vector denotes the clustering assignment distribution of the cluster k.

Considering that both different views within the multi-view modality and the corresponding point cloud of the same 3D shape contain consistent semantics, the obtained clustering assignment distributions should be similar within the multi-view modality and between the point cloud and multi-view modalities. Namely, the index of different views and the corresponding point cloud that are assigned to a particular cluster should be consistent. To this end, an assignment-level cross-view contrastive loss L_{ACV} and an assignment-level cross-modal contrastive loss L_{ACM} are simultaneously applied to the clustering assignments. The assignment-level cross-view contrastive loss for the $G_k^{v_1}$ is calculated as follows:

$$L_{ACV_k}^{v_1} = -\log \frac{\exp(s(G_k^{v_1}, G_k^{v_2})/\tau_A)}{\sum_{l=1}^{c} [\exp(s(G_k^{v_1}, G_l^{v_1})/\tau_A) + \exp(s(G_k^{v_1}, G_l^{v_2})/\tau_A)]}, \quad (15)$$

where τ_A denotes the assignment-level temperature parameter and is generally set to 1.0. $G_k^{v_1}$ and $G_l^{v_1}$ are the k-th and l-th row vectors of G^{v_1}, respectively. Then, the L_{ACV} is calculated by computing the assignment-level cross-view contrastive loss for each cluster:

$$L_{ACV} = \frac{1}{2c}\sum_{k=1}^{c}(L_{ACV_k}^{v_1} + L_{ACV_k}^{v_2}), \tag{16}$$

where c denotes the number of clusters.

Similar to the representation-level cross-modal contrastive loss, the L_{ACM_k} is also obtained by computing the L_{ACM1_k} between G_k^p and $G_k^{v_1}$ and the L_{ACM2_k} between G_k^p and $G_k^{v_2}$. In this way, the bias of the consistent clustering assignments toward a particular view is effectively removed, thus further boosting the robustness of the clustering assignments. The assignment-level cross-modal contrastive loss for the $G_k^{v_1}$ is calculated as follows:

$$L_{ACM1_k}^{v_1} = -\log \frac{\exp(s(G_k^{v_1}, G_k^p)/\tau_A)}{\sum_{l=1}^{c}[\exp(s(G_k^{v_1}, G_l^{v_1})/\tau_A) + \exp(s(G_k^{v_1}, G_l^p)/\tau_A)]}. \tag{17}$$

Then, the L_{ACM} is naturally expressed as:

$$L_{ACM} = \frac{1}{2c}\sum_{k=1}^{c}(L_{ACM1_k}^{v_1} + L_{ACM1_k}^{p} + L_{ACM2_k}^{v_2} + L_{ACM2_k}^{p}). \tag{18}$$

Afterward, the overall assignment-level dual contrastive loss L_{ADCL} of the ADCL module is obtained by summing L_{ACV} and L_{ACM}:

$$L_{ADCL} = L_{ACV} + L_{ACM}. \tag{19}$$

As can be seen from Equation (19), the L_{ADCL} simultaneously maximizes the cross-view and cross-modal assignment consistency of the same cluster, while minimizing that of different clusters. This effectively enhances the intra-cluster compactness and inter-cluster separation, thus further boosting the multi-modal 3D shape clustering performance.

Finally, the clustering results are easily obtained from the predicted soft labels using the following formula:

$$q_i = \arg\max_k(Y_{ik}^u), i = 1, \ldots, N, k = 1, \ldots, c, \tag{20}$$

where q_i is the final predicted label for the i-th 3D shape.

3.4. Implementation Details

In the proposed DCL-Net, except for the designed representation-level dual contrastive loss L_{RDCL} and the assignment-level dual contrastive loss L_{ADCL}, an additional regularization loss L_{RL} [20] is imposed on the predicted soft labels from the ADCL, so as to avoid trivial solutions in deep clustering. Therefore, the total loss of the DCL-Net is calculated by summing the representation-level dual contrastive loss, the assignment-level dual contrastive loss, and the regularization loss:

$$L_{total} = L_{RDCL} + \lambda_1 L_{ADCL} + \lambda_2 L_{RL}, \tag{21}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are trade-off parameters to balance the roles of the different loss terms.

For the multi-modal feature extractor in the DCL-Net, ResNet18 [50] and PointNet [51] are adopted as the encoder networks for the selected views and point clouds, respectively. For the RDCL module, a multi-layer perceptron (MLP) with the dimensions of 512–128–128 is utilized as the representation projection head. For the ADCL module, another MLP

followed by a Softmax operation are utilized as the assignment projection head, in which the dimensions are set to 512–128–c.

4. Experimental Results

4.1. Experimental Setup

The proposed DCL-Net was implemented on the PyTorch platform using a GeForce GTX 1080 Ti GPU and an Intel i7-8700K processor at 3.70 GHz. During the network training phase, Adam [52] was adopted as the optimizer and the learning rate was set to 1.0×10^{-4}. The batch size was set to 128 for all of the experiments and the trade-off parameters λ_1 and λ_2 were fixed to 1 and 5, respectively. After the training phase, Equation (20) was utilized to calculate the final clustering results.

To evaluate the clustering performance of the proposed DCL-Net, experiments were conducted on two widely used 3D shape benchmark datasets: ModelNet10 [53] and ModelNet40 [53]. The ModelNet10 dataset consists of 4,899 3D CAD models from 10 classes, while the ModelNet40 dataset includes 12,311 3D CAD models from 40 classes. Following the experimental settings of [12], 1,024 points from the surface of each CAD model were sampled to form the point clouds and twelve 2D views of each CAD model were rendered to obtain multiple views. Note that this paper focused on unsupervised multi-modal 3D shape clustering, thus the class labels were not provided in all of the experiments.

Following the previous clustering work [54], four commonly used evaluation metrics, i.e., accuracy (ACC), normalized mutual information (NMI), adjusted rand index (ARI), and F-score, were employed to evaluate the clustering performance of the proposed DCL-Net and comparison methods. Different metrics were used to measure the consistency between the predicted labels and the ground truth labels from different perspectives. Specifically, the ACC represented the proportion of correctly predicted samples in the total samples. The NMI was used as a normalized measure of the correlations between the distributions of the predicted labels and the ground truth labels. The ARI was a modified version of RI [55] and indicated the distribution correlations between the predicted labels and the ground truth labels. Finally, the F-score was the harmonic mean of the precision and recall, where precision and recall represented the fraction of correctly predicted samples in the total positive predictions and the actual positives, respectively. For all of these metrics, higher values indicated a better clustering performance.

4.2. Comparison Results

To demonstrate the performance of the proposed DCL-Net, several existing multi-modal clustering methods were adopted for comparison, including DMSC [40], EAMC [43], and CoMVC [21]. Note that the selected methods were not designed for the multi-modal 3D shape clustering task, thus it was incapable of directly comparing them with the proposed method. To this end, they were extended as DMSC*, EAMC*, and CoMVC* to adapt to the multi-modal 3D shape clustering task. Specifically, the feature extractor corresponding to the input point cloud was replaced with PointNet [51], which was consistent with the proposed DCL-Net. Then, the point cloud and an arbitrarily selected view of each 3D shape were fed into the corresponding feature extractors of different modalities to obtain the final clustering results. Additionally, to ensure the reliability of the experimental results, all of the experiments were repeated ten times with random initializations to reduce the effects of randomness and the mean results of the repeated experiments are reported in this paper.

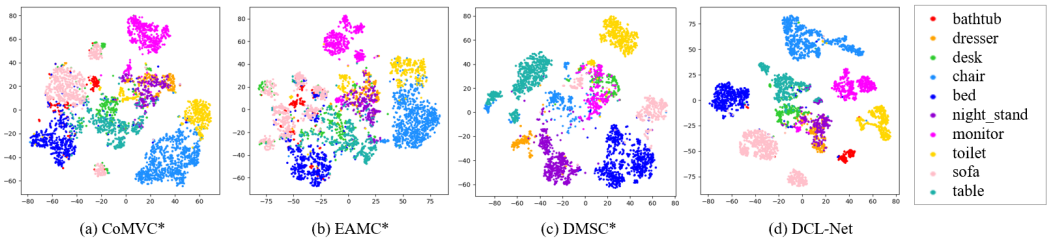
The quantitative comparison results on the ModelNet10 and ModelNet40 datasets are shown in Table 1. As shown in the table, the proposed DCL-Net achieved a better clustering performance than the comparison methods on both the ModelNet10 and ModelNet40 datasets, which effectively proved the superiority of the proposed method. In particular, the DCL-Net significantly outperformed CoMVC* and EAMC* by large margins. Even compared to the advanced method DMSC*, the proposed method also achieved the performance improvements of 2.94%, 9.27%, 2.20%, and 1.68% for the ACC, NMI, ARI, and F-score metrics on the ModelNet10 and the performance improvements of 6.04%,

3.99%, 5.25%, and 4.66% for the ACC, NMI, ARI, and F-score metrics on the ModelNet40. This was mainly because the comparison methods were proposed for general multi-modal clustering and directly transferring them into the 3D shape clustering task failed to leverage the characteristics of 3D shapes, thus resulting in unsatisfactory clustering performances. In contrast, the proposed method took full advantage of the inter-view correlations within the multi-view modality, as well as the inter-modal correlations between the point cloud and the multi-view modalities, and developed a dual contrastive learning network. By jointly exploring the cross-view and cross-modal consistent representations and clustering assignments, the proposed method was more suitable for the multi-modal 3D shape clustering task. Additionally, it is worth mentioning that the clustering accuracy of the comparison methods dropped by more than 20% when the benchmark dataset was changed from ModelNet10 to ModelNet40. The main reason for this was that ModelNet40 held more classes and more imbalanced data distributions than ModelNet10, making it challenging for the multi-modal clustering task. Nevertheless, the proposed method obtained the values of 61.20%, 72.46%, 57.61%, and 60.22% for the ACC, NMI, ARI, and F-score metrics on the ModelNet40 dataset respectively, which further proved the robustness of our method.

Table 1. The comparison of results from different clustering methods on the ModelNet10 and ModelNet40 datasets.

Method	ModelNet10				ModelNet40			
	ACC	**NMI**	**ARI**	**F-score**	**ACC**	**NMI**	**ARI**	**F-score**
CoMVC*	0.6703	0.6520	0.5920	0.6408	0.4191	0.5769	0.3608	0.3841
EAMC*	0.7040	0.6583	0.5890	0.6455	0.4563	0.5403	0.3259	0.3685
DMSC*	0.7638	0.7291	0.7426	0.7766	0.5516	0.6847	0.5236	0.5556
DCL-Net	**0.7932**	**0.8218**	**0.7646**	**0.7934**	**0.6120**	**0.7246**	**0.5761**	**0.6022**

To further evaluate the superiority of the proposed DCL-Net over the comparison methods, t-SNE [56] visualizations of the 3D shape features utilized for clustering in the different methods were provided on the ModelNet10 dataset. The visualization results are shown in Figure 2, in which the different colors indicate different classes. As shown in Figure 2a–c, the features extracted by the comparison methods were quite dispersed and the boundaries between the different classes were inconspicuous. By contrast, the proposed DCL-Net provided more clear and compact clustering partitions, which further demonstrated the effectiveness of the proposed method.



(a) CoMVC* (b) EAMC* (c) DMSC* (d) DCL-Net

Figure 2. The t-SNE visualization results of the latent 3D shape features on the ModelNet10 dataset.

4.3. Evaluation of Key Components

4.3.1. Evaluation of the Proposed RDCL Module and ADCL Module

In this paper, a representation-level dual contrastive learning module and an assignment-level dual contrastive learning module were developed to discover the clustering partitions of unlabeled 3D shapes by jointly learning consistent 3D shape representations and clustering assignments. To validate the effectiveness of the two modules, evaluations were conducted on the ModelNet10 and ModelNet40 datasets. The results are shown in Table 2, in which "*w/o* ADCL" indicates the proposed method without the ADCL module and "*w/o*

RDCL" indicates the proposed method without the designed RDCL module. When the ADCL module was removed, the clustering results could not be directly predicted by the model, hence the k-means algorithm [57] was introduced to perform clustering on the consistent 3D shape representations obtained by the RDCL. As shown in the table, the performance of the "*w/o* ADCL" method dropped significantly on both two datasets compared to the DCL-Net and the ACC value dropped sharply by more than 10% on ModelNet10. This was mainly because the removal of the ADCL module disconnected the procedure of the representation learning and clustering, thus the obtained 3D shape features were irrelevant to the subsequent clustering. Similarly, the "*w/o* RDCL" method also obtained unsatisfactory clustering performances on both two datasets. The main reason was that removing the RDCL module made it difficult to ensure the feature consistency between different views and the corresponding point cloud, thus damaging the intra-cluster compactness and inter-cluster separation of the latent 3D shape features for clustering. Comparatively, the clustering results of the DCL-Net were consistently improved on both two datasets when using the RDCL and ADCL modules simultaneously. This sufficiently reflected the significance of both the representation-level dual contrastive learning and the assignment-level dual contrastive learning for multi-modal 3D shape clustering.

Table 2. The evaluation of the proposed RDCL module and ADCL module.

Method	ModelNet10				ModelNet40			
	ACC	**NMI**	**ARI**	**F-score**	**ACC**	**NMI**	**ARI**	**F-score**
w/o ADCL	0.6901	0.7437	0.6538	0.6958	0.5572	0.7241	0.4573	0.4779
w/o RDCL	0.7921	0.7872	0.7476	0.7782	0.4966	0.6581	0.4321	0.4641
DCL-Net	**0.7932**	**0.8218**	**0.7646**	**0.7934**	**0.6120**	**0.7246**	**0.5761**	**0.6022**

4.3.2. Evaluation of the Cross-View and Cross-Modal Contrastive Learning

By adequately exploiting the characteristics of 3D shapes, the proposed method simultaneously performed cross-view contrastive learning within the multi-view modality and cross-modal contrastive learning between the point cloud and the multi-view modalities, so as to better explore the consistent semantic information of 3D shapes. To evaluate the effectiveness of the cross-view and cross-modal contrastive learning, validation experiments were conducted on the ModelNet10 and ModelNet40 datasets. The experimental results are reported in Table 3, in which "*w/o* cross-modal contrastive learning" denotes removing the point cloud branch and only combining the cross-view contrastive losses for the network constraints, and "*w/o* cross-view contrastive learning" denotes removing one of the view branches and constraining the network via the cross-modal contrastive losses of the remaining view branch with the point cloud branch. As shown in the table, the "*w/o* cross-modal contrastive learning" and "*w/o* cross-view contrastive learning" methods only achieved limited clustering performances on the two datasets compared to the DCL-Net. The main reason was that removing either the cross-view contrastive learning or the cross-modal contrastive learning prevented the network from exploring the consistent semantics from more comprehensive 3D shape information. Specifically, when the cross-modal contrastive learning was removed, the network was incapable of perceiving the spatial geometry of 3D shapes, which made it challenging to explore the discriminative 3D shape descriptions from harder contrastive positives. Similarly, when the cross-view contrastive learning was removed, the network could not observe the richer visual information about the 3D shape from different angles, thus failing to ensure the compactness of view features of the same 3D shape and misleading the extraction of consistent information. Therefore, both the cross-view and cross-modal contrastive learning adopted in the proposed DCL-Net were crucial for the 3D shape clustering.

Table 3. The evaluation of the cross-view and cross-modal contrastive learning.

Method	ModelNet10				ModelNet40			
	ACC	**NMI**	**ARI**	**F-score**	**ACC**	**NMI**	**ARI**	**F-score**
w/o cross-modal contrastive learning	0.7644	0.7881	0.7214	0.7557	0.5666	0.7080	0.5204	0.5400
w/o cross-view contrastive learning	0.7791	0.8142	0.7526	0.7828	0.5566	0.6926	0.4940	0.5315
DCL-Net	**0.7932**	**0.8218**	**0.7646**	**0.7934**	**0.6120**	**0.7246**	**0.5761**	**0.6022**

5. Conclusions

3D shape clustering has become a promising research topic in computer vision and multimedia fields due to its powerful ability to divide unlabeled 3D shape data. However, little effort has been put into solving the 3D shape clustering task in previous works. To this end, a novel DCL-Net for 3D shape clustering was proposed in this paper. Taking full advantage of the data characteristics of multiple views and point clouds, the proposed DCL-Net is the first deep multi-modal 3D shape clustering method. Specifically, a representation-level dual contrastive learning module was first designed to extract discriminative 3D shape features for clustering by ensuring cross-view representation consistency within multi-view modality, as well as cross-modal representation consistency between point cloud and multi-view modalities. Meanwhile, by simultaneously performing cross-view and cross-modal contrastive learning at the clustering assignment level, an assignment-level dual contrastive learning module was designed to further obtain consistent clustering assignments based on the robust learned 3D shape features. Under the joint effects of the two modules, the proposed DCL-Net is able to sufficiently exploit the consistency and complementarity within multi-view modality as well as between point cloud and multi-view modalities, thus obtaining more compact category partitions. As the first attempt at solving the multi-modal 3D shape clustering task, the proposed DCL-Net achieved remarkable performances on two widely used 3D shape benchmark datasets, which would bring enlightening investigations in future unsupervised 3D shape analysis research.

Author Contributions: Conceptualization, G.L. and Z.Z.; methodology, G.L.; software, G.L. and L.C.; investigation, T.Q. and J.S.; resources, Z.Z.; data curation, L.C.; writing—original draft preparation, G.L.; writing—review and editing, G.L., Z.Z., L.C., T.Q. and J.S.; supervision, Z.Z., T.Q. and J.S.; project administration, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Postdoctoral Science Foundation (grant number: 2021TQ0244), and the Tianjin Research Innovation Project for Postgraduate Students (grant number: 2021YJSB106).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ma, C.; Guo, Y.; Yang, J.; An, W. Learning multi-view representation with LSTM for 3D shape recognition and retrieval. *IEEE Trans. Multimed.* **2018**, *21*, 1169–1182. [CrossRef]
2. Dai, G.; Xie, J.; Fang, Y. Deep correlated holistic metric learning for sketch-based 3D shape retrieval. *IEEE Trans. Image Process.* **2018**, *27*, 3374–3386. [CrossRef] [PubMed]
3. Bu, S.; Wang, L.; Han, P.; Liu, Z.; Li, K. 3D shape recognition and retrieval based on multi-modality deep learning. *Neurocomputing* **2017**, *259*, 183–193. [CrossRef]
4. Qiu, S.; Anwar, S.; Barnes, N. Geometric back-projection network for point cloud classification. *IEEE Trans. Multimed.* **2021**, *24*, 1943–1955 . [CrossRef]
5. Han, Z.; Lu, H.; Liu, Z.; Vong, C.; Liu, Y.; Zwicker, M.; Han, J.; Chen, C. 3D2SeqViews: Aggregating sequential views for 3D global feature learning by CNN with hierarchical attention aggregation. *IEEE Trans. Image Process.* **2019**, *28*, 3986–3999. [CrossRef]

6. Chen, S.; Zheng, L.; Zhang, Y.; Sun, Z.; Xu, K. VERAM: View-enhanced recurrent attention model for 3D shape classification. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 3244–3257. [CrossRef] [PubMed]
7. Peng, B.; Lei, J.; Fu, H.; Zhang, C.; Chua, T.; Li, X. Unsupervised video action clustering via motion-scene interaction constraint. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *30*, 131–144. [CrossRef]
8. Kumar, K.; Shrimankar, D. Deep event learning boost-up approach: DELTA. *Multimed. Tools. Appl.* **2018**, *77*, 26635–26655. [CrossRef]
9. Lei, J.; Li, X.; Peng, B.; Fang, L.; Ling, N.; Huang, Q. Deep spatial-spectral subspace clustering for hyperspectral image. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2686–2697. [CrossRef]
10. Kumar, K.; Shrimankar, D.; Sing, N. Equal partition based clustering approach for event summarization in videos. In Proceedings of the International Conference on Signal-Image Technology & Internet-Based Systems, Naples, Italy, 28 November–1 December 2016.
11. Peng, B.; Lei, J.; Fu, H.; Shao, L.; Huang, Q. A recursive constrained framework for unsupervised video action clustering. *IEEE Trans. Industr. Inform.* **2020**, *16*, 555–565. [CrossRef]
12. You, H.; Feng,Y.; Zhao, X.; Zou, C.; Ji, R.; Gao, Y. PVRNet: Point-view relation neural network for 3D shape recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
13. Yu, C.; Lei, J.; Peng, B.; Shen, H.; Huang, Q. SIEV-Net: A structure-information enhanced voxel network for 3D object detection from LiDAR point clouds. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5703711. [CrossRef]
14. Kumar, K.; Shrimankar, D.; Singh, N. Event BAGGING: A novel event summarization approach in multiview surveillance video. In Proceedings of the International Conference on Innovations in Electronics, Signal Processing and Communication, Shillong, India, 6–7 April 2017.
15. Kumar, K.; Shrimankar, D. F-DES: Fast and deep event summarization. *IEEE Trans. Multimed.* **2017**, *20*, 323–334. [CrossRef]
16. Pan, Z.; Yu, W.; Lei, J.; Ling, N.; Kwong, S. TSAN: Synthesized view quality enhancement via two-stream attention network for 3D-HEVC. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 345–358. [CrossRef]
17. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–18283. [CrossRef] [PubMed]
18. Afham, M.; Dissanayake, I.; Dissanayake, D.; Dharmasiri, A.; Thilakarathna, K.; Rodrigo, R. CrossPoint: Self-supervised cross-modal contrastive learning for 3D point cloud understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022.
19. Jing, L.; Zhang, L.; Tian, Y. Self-supervised feature learning by cross-modality and cross-view correspondences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021.
20. Xu, J.; Tang, H.; Ren, Y.; Zhu, X.; He, L. Contrastive multi-modal clustering. *arXiv* **2021**, arXiv:2106.11193.
21. Trosten, D.; Lokse, S.; Jenssen R.; Kampffmeyer, M. Reconsidering representation alignment for multi-view clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021.
22. Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; Guibas, L. Learning representations and generative models for 3D point clouds. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
23. Zhao, Y.; Birdal, T.; Deng, H.; Tombari, F. 3D point capsule networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Los Angeles, CA, USA, 15–21 June 2019.
24. Yang, Y.; Feng, C.; Shen, Y.; Tian, D. FoldingNet: Point cloud auto-encoder via deep grid deformation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–22 June 2018.
25. Sanghi, A. Info3D: Representation learning on 3D objects using mutual information maximization and contrastive learning. In Proceedings of the European Conference on Computer Vision, Virtual, 23–28 August 2020.
26. Han, Z.; Liu, Z.; Han, J.; Vong, C.; Bu, S.; Chen, C. Mesh convolutional restricted Boltzmann machines for unsupervised learning of features with structure preservation on 3-D meshes. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2268–2281. [CrossRef]
27. Park, J.; Florence, P.; Straub, J.; Newcombe, R.; Lovegrove, S. DeepSDF: Learning continuous signed distance functions for shape representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Los Angeles, CA, USA, 15–21 June 2019.
28. Han, Z.; Shang, M.; Liu, Y.; Zwicker, M. View Inter-Prediction GAN: Unsupervised representation learning for 3D shapes by learning global shape memories to support local view predictions. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
29. Wu, J.; Zhang, C.; Xue, T.; Freeman, B.; Tenenbaum, J. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In Proceedings of the Conference and Workshop on Neural Information Processing System, Barcelona, Spain, 5–10 December 2016.
30. Girdhar, R.; Fouhey, D.F.; Rodriguez, M.; Gupta, A. Learning a predictable and generative vector representation for objects. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
31. Kumar, K. Text query based summarized event searching interface system using deep learning over cloud. *Multimed. Tools. Appl.* **2021**, *80*, 11079–11094. [CrossRef]
32. Chang, J.; Wang, L.; Meng, G.; Xiang, S.; Pan, C. Deep adaptive image clustering. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.

33. Kumar, K.; Shrimankar, D.; Singh, N. Eratosthenes sieve based key-frame extraction technique for event summarization in videos. *Multimed. Tools. Appl.* **2018**, *77*, 7383–7404. [CrossRef]
34. Kumar, K. Event video skimming using deep keyframe. *J. Vis. Commun. Image Represent.* **2019**, *58*, 345–352. [CrossRef]
35. Peng, B.; Zhang, X.; Lei, J.; Zhang, Z.; Ling, N.; Huang, Q. LVE-S2D: Low-light video enhancement from static to dynamic. *IEEE Trans. Circuits Syst. Video Technol.* **2022**. [CrossRef]
36. Pan, Z.; Yuan, F.; Lei, J.; Fang, Y.; Shao, X.; Kwong, S. VCRNet: Visual Compensation Restoration Network for No-Reference Image Quality Assessment. *IEEE Trans. Image Process.* **2022**, *31*, 1613–1627. [CrossRef]
37. Pan, Z.; Yuan, F.; Yu, W.; Lei, J.; Ling, N.; Kwong, S. RDEN: Residual distillation enhanced network-guided lightweight synthesized view quality enhancement for 3D-HEVC. *IEEE Trans. Circuits Syst. Video Technol.* 2022, *Early Access*. [CrossRef]
38. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A. Multi-modal deep learning. In Proceedings of the International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011.
39. Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. Deep canonical correlation analysis. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013.
40. Abavisani, M.; Patel, V. Deep multimodal subspace clustering networks. *IEEE J. Sel. Top. Sign. Process.* **2018**, *12*, 1601–1614. [CrossRef]
41. Peng, B.; Lei, J.; Fu, H.; Jia, Y.; Zhang, Z.; Li, Y. Deep video action clustering via spatio-temporal feature learning. *Neurocomputing* **2021**, *456*, 519–527. [CrossRef]
42. Li, Z.; Wang, Q.; Tao, Z.; Gao, Q.; Yang, Z. Deep adversarial multi-view clustering network. In Proceedings of the International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019.
43. Zhou, R.; Shen, Y. End-to-end adversarial-attention network for multi-modal clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020.
44. Zhuang, C.; Zhai, A.; Yamins, D. Local aggregation for unsupervised learning of visual embeddings. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019.
45. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 16–18 June 2020.
46. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual, 12–18 July 2020.
47. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021.
48. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Guo, Z.; Azar, M.G.; et al. Bootstrap your own latent-a new approach to self-supervised learning. In Proceedings of the Conference and Workshop on Neural Information Processing System, Virtual, 6–12 December 2020.
49. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multi-view coding. In Proceedings of the European Conference on Computer Vision, Virtual, 23–28 August 2020.
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
51. Qi, C.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017.
52. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
53. Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3D ShapeNets: A deep representation for volumetric shapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015.
54. Zhang, C.; Liu, Y.; Fu, H. AE^2-Nets: Autoencoder in autoencoder networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Los Angeles, CA, USA, 15–21 June 2019.
55. Rand, W. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **1971**, *66*, 846–850. [CrossRef]
56. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
57. Hartigan, J.A.; Wong, M.A. Algorithm as 136: A k-means clustering algorithm. *J. R. Stat. Soc.* **1979**, *28*, 100–108. [CrossRef]