Received: October 2, 2024. Revised: October 27, 2024. 134

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

AdaCrossNet: Adaptive Dynamic Loss Weighting for Cross-Modal Contrastive Point Cloud Learning

Oddy Virgantara Putra[1,2] Kohichi Ogata[3] Eko Mulyanto Yuniarno[1,4] Mauridhi Hery Purnomo[1,4]*

[1]*Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia*
[2]*Department of Informatics, Universitas Darussalam Gontor, Ponorogo, 63472, Indonesia*
[3]*Faculty of Advanced Science and Technology, Kumamoto University, Kumamoto, 860-0862, Japan*
[4]*Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia*
* Corresponding author's Email: hery@ee.its.ac.id

Abstract: Manual annotation of large-scale point cloud datasets is laborious due to their irregular structure. While cross-modal contrastive learning methods such as CrossPoint and CrossNet have progressed in utilizing multimodal data for self-supervised learning, they still suffer from instability during training caused by the static weighting of intra-modal (IM) and cross-modal (CM) losses. These static weights fail to account for the varying convergence rates of different modalities. We propose AdaCrossNet, a novel self-supervised learning framework for point cloud understanding that utilizes a dynamic weight adjustment mechanism for IM and CM contrastive learning. AdaCrossNet learns representations by simultaneously enhancing the alignment between 3-D point clouds and their associated 2D-rendered images within a common latent space. Our dynamic weight adjustment mechanism adaptively balances the contributions of IM and CM losses during training, guided by the convergence behavior of each modality. To ensure stability in the training process, we employ an exponentially weighted moving average (EWMA) to smooth the weight updates. We experimented with benchmark datasets, ModelNet40, ShapeNetPart, and ScanObjectNN. The results demonstrate that AdaCrossNet achieves superiority over other methods, with 91.4% accuracy on the ModelNet40 classification task. While on the segmentation task, AdaCrossNet achieved the mIoU score of 85.1% on the ShapeNetPart segmentation task. Additionally, AdaCrossNet, when combined with the DGCNN backbone, showed significant improvements in the ScanObjectNN dataset with 82.1% accuracy. Our method boosts training efficiency while increasing the generalizability of the learned representations across downstream tasks.

Keywords: Adaptive weighting, Contrastive learning, Deep learning, Point cloud understanding, Self-Supervised learning.

1. Introduction

3D vision, essential in applications like autonomous driving, mixed reality, and robots, has garnered significant attention for its capacity to comprehend the human environment. Consequently, study has been abundant in 3D vision issues, including object categorization [1-3], detection [4], and segmentation [2, 5], with point clouds emerging as the predominant approach for 3D data representation in recent years. Nonetheless, the efficacy of deep learning fundamentally depends on extensive annotated datasets. Despite the progress in 3D sensing technologies (e.g., LIDAR) enabling significant gathering of 3D point cloud samples, the inconsistent structure of point clouds renders the manual annotation of large-scale 3D point cloud datasets labor-intensive. Self-supervised learning (SSL) is a leading method for tackling this problem and has demonstrated efficacy in the 2D domain [6-8].

As their SSL success in images [7-9] and videos [10, 11], we have seen that multimodal contexts have been presented for a variety of vision tasks, including object identification [12] and few-shot picture

classifications [13]. CrossPoint [14] first illustrates self-supervised on point cloud which combined with images using pretrained contrastive learning (CL). This is the initial step in self-supervised CL for visualization in three dimensions. CrossPoint can establish suitable initial weights for downstream tasks by understanding the correlation between IM point clouds and CM point clouds and images. Because there have been significant advancements in both the utilization of data and the architecture of networks, this motivates us to investigate cross-modality more.

Despite significant advancements in cross-modal contrastive learning for point cloud data, a key challenge remains in balancing IM and CM learning objectives. Existing methods employ static weighting mechanisms that fail to account for varying convergence rates across different modalities, leading to instability during training [15, 16].

Table 1. List of notations

Symbol	Description
$\mathrm{x}_i^{t_1}, x_i^{t_2}$	Input paired point cloud data
y_i^{rgb}, y_i^{gr}	Embeddings from cross-modal images
L_{sim}	Loss similarity
L_{IM}	Loss for intra-modal
L_{CM}	Loss for cross-modal
α, β	Coefficients for controlling the given weights
$\mathbf{P}_i$	3-D point clouds at i
z	Latent space representation
N	The number of point cloud
$\mathbf{P}_i$	Transformed point cloud
f_{θ_p}	Shared weight network
$g_{\Phi_{rgb}}, g_{\Phi_{gr}}$	MLP projection functions for RGB and Grayscale
$\mathbf{I}_i^{rgb}, \mathbf{I}_i^{gr}$	Input related image of point cloud for RGB and Grayscale
$f_{\Phi_{rgb}}, f_{\Phi_{gr}}$	MLP network for each RGB and Grayscale
$f_{\theta_p}, f_{\theta_l}$	Feature extractor for point cloud and image, respectively
τ	Parameter of temperature in L_{sim}
$s(\cdot)$	Function for cosine similarity
y_i^{rgb}, y_i^{gr}	Color feature and grayscale feature
$\lambda_{CM}, \lambda_{IM}$	Dynamic weights for CM and IM, respectively
$\Delta L_{CM}, \Delta L_{IM}$	Change rate of CM and IM
L_{ACM}	Adaptive cross-modal loss function

The instability results from weights that change too rapidly might cause learning oscillations. Additionally, the interaction between the IM and CM losses adds complexity and leads to an overemphasis on one modality [17]. To address this, we propose a smoothing technique using exponentially weighted moving averages (EWMA) and weight constraints for each IM and CM loss adopted from [14, 18].

Our contributions are summarized as follows:

- We present a novel dynamic weight adjustment technique for IM and CM CL, which allows for a weight change according to each modality's convergence.
- We propose a smoothing technique for loss terms based on EWMA, which is expected to control the jump of the loss terms when the weights are changed noticeably.
- We present a model that adjusts the amounts of contributions of the losses to the overall loss as per the degree of convergence in each modality.
- The method presented facilitates better 3D cognition by embedding point cloud and image features into a common latent space, which allows for modeling geometry and appearance.

The remainder of the article is structured as follows: Section II summarizes the related works. In Section III, we provide the foundational approaches and the comprehensive explanation of our proposed work. Section IV, Experiments, presents and compares our method across multiple benchmarks with other techniques. Section V is the discussion. Section VI is the conclusion.

2. Related work

2.1 Supervised learning on point clouds

Point cloud data, recognized for its precision and depth, is essential in several practical applications, particularly computer vision. Nonetheless, the domain has obstacles, notably the lack of connectivity information between points, which hinders procedures such as surface reconstruction, topological analysis, and geometry [19, 20].

Significant improvements such as PointNet [1] enabled the direct processing of raw point clouds, effectively capturing global properties while encountering difficulties with local structures. PointNet++ [2] tackled this obstacle with hierarchical learning to capture both local and global features, although encountering difficulties with thick point clouds. Alternative graph-based methodologies, such as Dynamic Graph CNN (DGCNN) [3] and

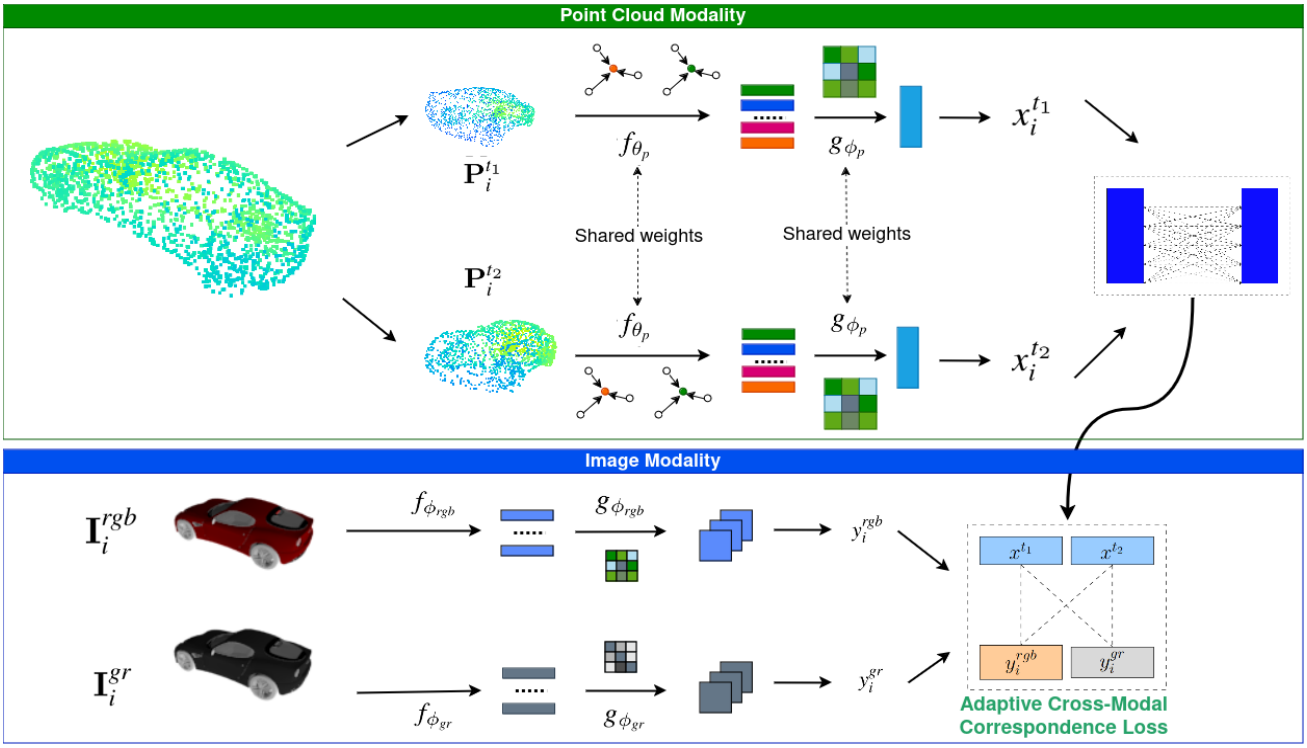

Figure. 1 The diagram illustrates a dual-modality framework combining point cloud and image data. In the Point Cloud Modality (top section), two transformed versions of a point cloud ($P_i^{t_1}$ and $P_i^{t_2}$) undergo feature extraction using a shared-weight network (f_{θ_p}), followed by projection into a lower-dimensional space (g_{ϕ_p}). This results in features $x_i^{t_1}$ and $x_i^{t_2}$, which are compared using intra-modal correspondence loss. In the Image Modality (bottom section), both RGB and grayscale images ($\boldsymbol{I}_i^{rgb}$ and $\boldsymbol{I}_i^{gr}$) pass through separate networks ($f_{\phi_{rgb}}$ and $f_{\phi_{gr}}$) and projection functions ($g_{\phi_{rgb}}$ and $g_{\phi_{gr}}$). The resulting embeddings (y_i^{rgb} and y_i^{gr}) are compared cross-modally with point cloud features using an Adaptive Cross-Modal Correspondence Loss to align features loss from different modalities, enhancing the robustness representations across modalities.

superpoint graph approaches [21], utilize the interrelations among adjacent points to encapsulate intricate geometric configurations. CurveNet [22] introduced a novel curve aggregation to enhance point cloud analysis capturing semantic relationship of those points. However, CurveNet is sensitive to the selection of curve length and quantity. Moreover, multiscale learning methodologies [23, 24] have proven their importance in point cloud processing, enhancing the capacity to acquire and utilize information across many scales for superior task efficacy.

2.2 Unsupervised learning on point clouds

The 3D-GAN approach in [25] effectively generates 3D objects using adversarial networks but suffers from artifacts like fragmented or incomplete shapes. SO-GAN [26] introduces self-organizing maps for point cloud analysis, offering fast training. However, it struggles with scalability for large-scale point clouds due to the complexity of the self-organizing map construction. FoldingNet [27] achieves impressive point cloud reconstruction by folding a 2D grid into a 3D shape. However, its reliance on a fixed 2D grid structure limits flexibility in handling irregular point cloud geometries. Achlioptas et al. [28] propose an autoencoder-based for point cloud learning, but the GAN-based approach poses instability during training. Finally, Zhang [29] presents a graph-based network for unsupervised point cloud learning, which performs well on classification tasks but faces challenges in capturing fine-grained local features. Next year, an unsupervised repeated learning was shown in [30] by reconstructing a partial point cloud obtained from occlusions. However, the generalization is limited due to its reliance on indoor-specific pre-training data.

2.3 Self-Supervised learning on point clouds

SSL is more effectively established in the 2D world [9], leading most current 3D SSL techniques to pursue direct migration without adequately leveraging the distinct characteristics of 3D and 2D data. JigSaw3D [31], while adopting the concept of component rearrangement from 2D JigSaw tasks, needs a more complex task design. For Rotation3D

[32], the arbitrary rearrangement of 3-D components fails to capture the complex geometric relationships that exist in point clouds.

Motivated by the efficacy of self-supervised CL in image comprehension, several studies [33-35] have examined this framework for point cloud analysis. STRL [34], an extension of BYOL [8] for 3D point clouds, explores representations in an unsupervised manner by gaining benefit of the interactions among networks. In contrast to current studies utilizing CL, we provide a supplementary CM contrastive aim that captures multi-dimensional correspondence.

2.4 Intra-Modal learning on point clouds

CL has arisen as a formidable method for SSL, especially within the two-dimensional domain, as evidenced by numerous studies in computer vision and text-related tasks [9, 36]. An essential element of CL is utilizing an ideal contrastive loss function that proficiently allows the model to differentiate between positive and negative data. PointContrast [30] implements CL at the point level by creating two distinct augmented representations of the same point cloud and ensuring consistency among comparable points.

2.5 Cross-Modal learning on point clouds

Cross-modal retrieval (CMR) pertains to extract relevant information across many modalities, including text and visuals and has attracted considerable attention in recent years [37, 38]. CMR has shown a significant amount of interest in cross-modal learning (CML). At the early stages, CML for point cloud processing leverages the integration of point clouds with different data representations such as voxels [35], RGB images [15], and geometric image features [18]. These methods aimed to enhance point cloud understanding by utilizing additional data from other modalities. Our approach is closely related to the works of CrossPoint [14] and CrossNet [18], which implemented CL to exploit the relationships between 2D and 3D modalities.

3. Proposed work

In traditional contrastive learning frameworks [14, 18] the weights for IM and CM losses are manually set and fixed throughout the training process. However, this static weighting may not be optimal as different modalities (3D point clouds and images) exhibit different convergence behaviors. We propose a dynamic weight adjustment mechanism, called AdaCrossNet which can be seen in Fig. 1, where the contribution of each loss term is updated during training based on the convergence state of each modality, allowing the model to allocate more emphasis on the harder-to-learn modality at different stages of the training process. All notations here are presented in Table 1.

3.1 Preliminaries

ssessing is required. To tackle this issue, SSL can achieve incredible results with no required labeled dataset [39], [40]. Furthermore, since human visual perception can perceive 3D from 2D images, multi-modality is better than singular modality.

Suppose we are given a dataset, $D = \{(\mathbf{P}_i, \mathbf{I}_{rgb}, \mathbf{I}_{gr})\}_{i=1}^{|D|}$, where $\mathbf{P}_i \in \mathbb{R}^{N\times 3}$ represents the 3D point cloud with N points, each with 3D coordinates, $I_{rgb} \in \mathbb{R}^{H\times W\times 3}$ is the rendered RGB image of the corresponding point cloud $\mathbf{P}_i$, and $\mathbf{I}_{gr} \in \mathbb{R}^{H\times W}$ is one-channeled image. The image $\mathbf{I}_{rgb}$ is obtained by rendering the 3D point cloud $\mathbf{P}_i$ from a random camera view-point. The objective is to train a point cloud feature extractor $f_{\theta_p}(\cdot)$ and an image feature extractor $f_{\theta_l}(\cdot)$ in a self-supervised manner. This feature training is transferred to downstream tasks such as 3D classification and segmentation. For both modalities, we apply multi-layer perceptron (MLP) projection $g_{\phi_p}(\cdot)$ and $g_{\phi_l}(\cdot)$ to map the extracted features to a shared latent space for contrastive learning.

Directly evaluate the colored point cloud representation is hard because of its unstructured nature and high dimensionality. Thus, to tackle this issue, the 3D point cloud is represented in RGB and grayscale. We called this feature representation extractor with f_{θ_p} for 3D point cloud, $f_{\theta_{rgb}}$ for RGB images, and $f_{\theta_{gr}}$ for the grayscale images. We employed two famous point cloud extraction methods, PointNet and DGCNN and ResNet as 2D image extractor.

3.2 Intra-Modal learning

The successor works of contrastive learning (CL) for point cloud and images [7, 14, 41] gained attention to others [18, 42]. CL is an approach in deep learning specially designed for tasks where labeled data are scarce. The purpose of CL is to learn an embedding space between corresponding data points by contrasting pairs of positive and negative samples. The term intra modal is come from the pairing point clouds obtained from two random augmentation using translation, rotation, scaling, and jittering.

Given a point cloud input $\mathbf{P}i$, we have two augmented versions, $\mathbf{P}_i^{t_1}$ and $\mathbf{P}_i^{t_2}$

Given two aligned views of a point cloud or 3D object, labeled v^1 and v^2. The views are linked by a point index mapping M, which designates corresponding points in each view. The procedure initiates with the application of a feature extractor f_{θ_p} to the altered point clouds $\mathbf{P}_i^{t_1}$ and $\mathbf{P}_i^{t_2}$, therefore mapping them into an embedding space. The MLP $g_{\phi_p}(\cdot)$ maps these features onto $\mathbb{R}^d$, $d = 256$, yielding $x_i^{t_1}$ and $x_i^{t_2}$, respectively. $x_i^{t_j}$ is derived from the equation $x_i^{t_j} = g_{\phi_p}\left(f_{\theta_p}\left(P_i^{t_j}\right)\right)$. The objective is to enhance the similarity between the corresponding projection vectors $x_i^{t_j}$ while reducing their resemblance to projections from disparate point clouds.

We adapt NT-Xent loss function from [7] to evaluate and separate the positive and negative feature vectors. The intra-modal loss function L_{IM} is defined as:

$$L_{IM} = \frac{1}{2N}\sum_{i=1}^{N}[I_{sim}(x_i^{t_1}, x_i^{t_2}) + I_{sim}(x_i^{t_2}, x_i^{t_1})] \quad (1)$$

where N is the batch size, $x_i^{t_1}$ and $x_i^{t_2}$ are the feature vectors from MLP projection $P_i^{t_1}$ and $P_i^{t_2}$, and I_{sim} is defined as:

$$L_{sim}(x_i^{t_1}, x_i^{t_2}) = -\log\frac{e^{s(x_i^{t_1}, x_i^{t_2})/\tau}}{\sum_{i=1}^{N} 1_{[k\neq 1]} e^{s(x_i^{t_1}, [x_k^{t_1}, x_k^{t_2}])/\tau}} \quad (2)$$

where τ is the parameter of temperature, $s(\cdot)$ is the function to calculate similarity using cosine.

3.3 Cross-Modal learning

Learning spatial-awareness from image modality can improve the learning capability. CrossPoint [14] utilized the joint operation between point cloud and image correspondence. But, it lacks of other image properties such as contour and edge. CrossNet tackled this issue by exploring the properties using RGB and grayscale image features. The interaction of properties between point cloud and both RGB and grayscale image is called Cross Modal (CM). To capture the average score of projected features, we calculate x_i with:

$$x_i = \frac{x^{t_1} + x^{t_2}}{2} \quad (3)$$

By using ResNet as image projection head, we can embed the RGB image $\boldsymbol{I}_{rgb}$ and grayscale image $\boldsymbol{I}_{gr}$ into the feature domain. Given $g_{\phi_{rgb}}$ and $g_{\phi_{gr}}$ as image projection head for RGB and grayscale images, respectively, we have the color feature y_i^{rgb} with:

$$y_i^{rgb} = g_{\phi_{rgb}}\left(f_{\theta_{rgb}}(\mathbf{I}_i^{rgb})\right) \quad (4)$$

and the grayscale features y_i^{gr} with:

$$y_i^{gr} = g_{\phi_{gr}}\left(f_{\theta_{gr}}(\mathbf{I}_i^{gr})\right) \quad (5)$$

where $g_{\phi_{rgb}}$ and $g_{\phi_{gr}}$ are average pooling.

Since both point clouds and images share mutual embedding, joint learning objective, L_{CM}, is implemented. The purpose of L_{CM} is to find the similarity between y_i^{rgb}, y_i^{gr}, and x_i in the invariant domain. Due to the correspondence of y_i^{rgb} and y_i^{gr} to the same point cloud, we can enhance the model's ability to learn meaningful feature representations to effectively distinguish similar and dissimilar for samples that difficult to distinguish. This condition makes the model learn better at differentiating subtle differences between data points. To find the loss of colored RGB $L_{CM_{rgb}}$, we can calculate with:

$$L_{CM_{rgb}} = \frac{1}{2N}\sum_{i=1}^{N}[C_{sim}(x_i, y_i^{rgb}) + C_{sim}(y_i^{rgb}, x_i)] \quad (6)$$

where

$$C_{sim}(x_i, y_i) = -\log\frac{e^{s(x_i, y_i)/\tau}}{\sum_{k=1}^{N} 1_{[k\neq 1]} e^{s(x_i, [x_k, y_k])/\tau}} \quad (7)$$

where τ is the parameter of temperature, $s(\cdot)$ is the function to calculate similarity using cosine as in Eq. (2). The function of $L_{CM_{gr}}$ can be obtained as in Eq. (6) with:

$$L_{CM_{gr}} = \frac{1}{2N}\sum_{i=1}^{N}[C_{sim}(x_i, y_i^{gr}) + C_{sim}(y_i^{gr}, x_i)] \quad (8)$$

Therefore, the joint loss function is calculated with:

$$L_{CM} = L_{CM_{rgb}} + L_{CM_{gr}} \quad (9)$$

Table 2. SVM Classification Results of AdaCrossNet compared to other methods using ModelNet40 dataset

Model	Accuracy (%)
3D-GAN [25]	83.3
SO-GAN [26]	87.3
LatentGAN [28]	85.7
FoldingNet [27]	88.4
ClusterNet [29]	86.8
PointNet+STRL [34]	88.3
PointNet+OcCo [30]	88.7
PointNet+Rotation [32]	88.6
PointNet+CrossPoint [14]	89.1
PointNet+CrossNet [18]	81.4
DGCNN+STRL [34]	90.4
DGCNN+OcCo [30]	89.2
DGCNN+CrossPoint [14]	88.9
DGCNN+CrossNet [18]	90
DGCNN+AdaCrossNet	**91.4**

3.4 Adaptive cross-modal learning

Dynamic weight adjustment for IM and CM losses comes with its own set of challenges in terms of convergence criterion definition, training stability, and balancing modalities. The rate of convergence for modalities can be different, so it is hard not to bias one modality over the other. Smoothing the weight updates, using relative loss reduction measures, and incremental updating of weights, facilitating stable training and successful employment of both modalities.

Inspired by [43, 44], we introduce two dynamic weights λ_{CM} and λ_{IM} which adjust the relative importance of L_{CM} in Eq. (9) and L_{IM} in Eq. (1). These weights evolve depending on how well each modality is converging. To track convergence, we need to calculate the rage of change from each loss during training. Given ΔL_{IM} and ΔL_{CM} are the rate of change of IM and CM losses, respectively, which represent how quickly the losses decrease. Thus, we have:

$$\Delta L_{IM}(t) = L_{IM}(t) - L_{IM}(t-1) \tag{10}$$

and

$$\Delta L_{CM}(t) = L_{CM}(t) - L_{CM}(t-1) \tag{11}$$

where t is the current epoch value. To avoid noisy updates, we introduce the smoothing weight by using exponentially weighted moving average. Thus, we have:

$$\lambda_{IM}(t) = \beta\lambda_{IM}(t-1) + (1-\beta)\frac{1}{1+e^{-\alpha\Delta L_{IM}}}\lambda_{IM}(t) = \beta\lambda_{IM}(t-1) + (1-\beta)\frac{1}{1+e^{-\alpha\Delta L_{IM}}} \tag{12}$$

and

$$\lambda_{CM}(t) = \beta\lambda_{CM}(t-1) + (1-\beta)\frac{1}{1+e^{-\alpha\Delta L_{CM}}} \tag{13}$$

where α and β are the coefficient that control how much weight is given to past values of λ. Finally, our final total loss function is:

$$L_{ACM} = \lambda_{IM}L_{IM} + \lambda_{CM}L_{CM} \tag{14}$$

4. Experiments and results

4.1 Dataset

Here, we employ the ShapeNet [45] dataset, comprising roughly 43k point clouds across 13 item categories and its corresponding generated images, which is utilized for pretraining, as in [1]. The ima ge data are configured as RGB and grayscale images and then transformed to tensor. We arbitrarily pick a single image from the available images for each point cloud in a random viewpoint. Following CrossPoint and CrossNet, we represent each point cloud with 2048 points containing XYZ coordinates and resize the image to 224 × 224.

In addition, we apply various augmentations to the images, such as random jittering, flipping, cropping, and normalization. Our trials were carried out on a computer that featured a graphics processing unit (GPU) NVIDIA GeForce RTX 4090 24 GB and a central processing unit (CPU) Intel Core i7-12700K 3.6 GHz. Python version 3.9, PyTorch version 1.9.0, CUDA 12 as the GPU driver, and Ubuntu 22.04.

We compare with [1] and [3], two of the most established point cloud approaches, and use ResNet50 [46] as our image feature extractor to enable this fair comparison. The projection head is formed by using a two-layer MLP of (layer one is 512, and the other is 256) for generating the final 256-dimensional feature vectors that are projected to $\mathbb{R}^d$. We pretrain our model for 100 epochs and use a weight decay of 1×10^{-4}, which is initialized at learning rate 1×10^{-3}. We use the cosine annealing to adjust the learning rate to reduce it gradually. After pre-training, we discard the spare module such as $f_{\theta_{rgb}}$ and $f_{\theta_{gr}}$ and two projection heads like g_{ϕ_p}, $g_{\phi_{rgb}}$, and $g_{\phi_{gr}}$. The point cloud feature extractor

f_{θ_p} is aggregated into functions of downstream tasks and could be used as a fine-tuned one.

4.2 Point cloud object classification

We assess the transferability of AdaCrossNet on two prevalent downstream tasks in point cloud representation learning: 3D object shape recognition and partial segmentation. We choose a dataset for validation for each downstream task: We conduct an experiment for 3D object categorization using dataset ModelNet40 [47] and ScanObjectNN [48]. ModelNet40 is a 3D CAD-based point cloud model containing 40 classes. For the such task, the dataset is split into train and test, containing roughly 10k/2.5k items. ScanObjectNN is a practical dataset comprising point clouds characterized by occlusion and noise, encompassing 15 categories and having 2,3k training objects and 0.5k test objects. The second classification task employed the ShapeNetPart [49] dataset for part segmentation. This dataset comprises 14k training items and 2.8k test objects over 16 categories, each divided into 2 to 6 pieces, resulting in 50 parts.

4.2.1. Linear evaluation for object classification

First, we set up a straightforward linear Support Vector Machine (SVM) designed to classify with the pretrained feature extractor SSL point cloud. We then train the classifier on the training splits of both the ModelNet40 and ScanObjectNN datasets. We sample 1024 points for each object in the dataset, which will be used during the training and testing of the classification task. To make the learning and computations as efficient as possible, we will train with a batch size of 128. Finally, the robustness and generalization of our proposed AdaCrossNet architecture are evaluated on different backbone networks. All these are set up together, allowing us to evaluate performance when generalizing to unseen data and benefit from the improved feature extraction we get with pre-trained point cloud models. The comparison of the SVM classification results of AdaCrossNet with other models on the ModelNet40 dataset is shown in Table 2. Our model is the highest-performing, which achieved an accuracy of 91.4%, while RotNet and DGCNN+STRL yielded 90.4%. Even when used with the DGCNN, our method dramatically improved accuracy compared to the existing methods. Conversely, 3D-GAN and SO-GAN exhibit suboptimal performance, achieving 83.3% and 87.3%, respectively. In current designs, experimental results indicate that CrossNet and AdaCrossNet generally outperform simpler models, such as DGCNN. Furthermore, models such as PointNet+CrossNet and DGCNN+CrossPoint demonstrate robust performance; nonetheless, DGCNNbased methodologies generally surpass their PointNet counterparts. DGCNN+CrossNet achieves 90%, while PointNet+CrossNet attains 89.1%, demonstrating that DGCNN is a more effective architecture when combined with crossmodal techniques. These results highlight the importance of selecting the suitable backbone architecture and learning methodology combination to improve point cloud classification effectiveness.

4.2.2. Fine-Tuning on object classification

We evaluated AdaCrossNet using other supervised methods to support the classification from SVM. After we pre-trained DGCNN with AdaCrossNet, we get the represented selfsupervised model. As initialization, the weights from the model are used for the backbone in DGCNN. We tested the fine-tuned model with the dataset ModelNet40 and ScanObjectNN. For the ModelNet40 dataset, our model, AdaCrossNet, became a top-notch with the enhanced score by +0.6% compared to CrossNet, as seen in Table 3. While in the ScanObjectNN, AdaCrossNet stands on top with 82.1% as in Table 4. This improvement more substantial than prior methods and demonstrates competitive performance against state-of-the-art approaches.

Table 3. Fine-Tuning Classification Results of AdaCrossNet compared to other methods using ModelNet40 dataset

Task	Method	Acc (%)
Supervised Learning	PointNet [1]	84.1
	PointNet++ [2]	90.7
	CurveNet [22]	91.9
	DGCNN [3]	85.8
Self-Supervised Learning	DGCNN+JigSaw3D [31]	92.4
	DGCNN+STRL [34]	90.9
	DGCNN+CrossNet [18]	92.5
	DGCNN+AdaCrossNet	**93.1**

Table 4. Fine-Tuning Classification Results of AdaCrossNet compared to other methods using ScanObjectNN dataset

Task	Method	Accuracy (%)
Supervised Learning	PointNet[1]	69.9
	PointNet++ [2]	77.9
	CurveNet [22]	79.8
	DGCNN [3]	78.6
Self-Supervised Learning	DGCNN+STRL [34]	77.9
	DGCNN+CrossNet[18]	79
	DGCNN+AdaCrossNet	**82.1**

Table 5. Fine-Tuning Part Segmentation Results of AdaCrossNet compared to other methods using ShapeNetPart dataset

Models	mIoU	airplane	bag	cap	car	chair	Earphone	guitar	knife	lamp	laptop	motorbike	mug	pistol	rocket	skateboard	table
PointNet [1]	83.6	**83.4**	72.4	62.1	76.2	89.8	69.1	90.8	85.6	81.2	94.9	**62.3**	92.4	**82.3**	48.4	70.8	82.0
PointNet++ [2]	83.8	81.5	78.2	80.2	73.3	89.8	72.9	89.2	85.3	82.8	95.2	55.4	**94.1**	77.4	**56.0**	70.8	82.2
DGCNN [3]	83	80.2	66.8	**82.1**	75.1	89.8	69.4	89.9	86.5	80.9	95.1	44.4	92.4	75.8	48.7	71.4	81.8
CrossNet [18]	83.1	80.6	**78.5**	78.6	74.6	90.0	71.5	89.7	87.1	82.4	95.3	52.1	90.4	77.3	49.5	72.7	81.0
AdaCrossNet	**85.1**	82.7	75.3	80.8	**78.0**	**90.8**	**77.4**	**91.6**	**87.8**	**85.0**	**95.8**	60.9	93.5	81.5	54.5	**75.3**	**82.6**

4.3 Point cloud object part segmentation

Here, we evaluate our model with other methods for object part segmentation using the ShapeNetPart dataset [49]. The models compared are PointNet, PointNet++, DGCNN, and CrossNet.

Table 5 illustrates the average Intersection over Union (mIoU) of various models on point cloud segmentation using the ShapeNet dataset. Here, each model is evaluated using a distinct approach to segmentation. AdaCrossNet distinguishes itself among the models with the greatest mIoU of 85.1 and outperforms others, including CrossNet, with a mIoU of 83.1. Additional significant achievements encompass PointNet++ and DGCNN, achieving mIoUs of 83.8 and 83, respectively. AdaCrossNet enforces IM CM correspondence to collect detailed part-level attributes effectively, which is crucial for part segmentation.

We further performed a qualitative analysis of PointNet++, CrossNet, and our proposed model AdaCrossNet. Fig. 2 depicts overlayed segmentation errors in heatmap form with regions of red color depicting the difference between the predicted and actual labels, calculated from the absolute differences. The grey dots represent regions with low segmentation errors, while the red dots indicate high-error regions. Across all objects, AdaCrossNet achieves superior segmentation accuracy with fewer red dots compared to the other models. PointNet++ struggles with errors in the wings and body. For earphone objects, both PointNet++ and CrossNet scattered errors in part segmentation. PointNet++ exhibits dense errors on the front and back of car objects, while CrossNet has fewer dense errors. AdaCrossNet shows the fewest errors across the car's body in this object. These qualitative results demonstrate that AdaCrossNet outperforms existing methods.

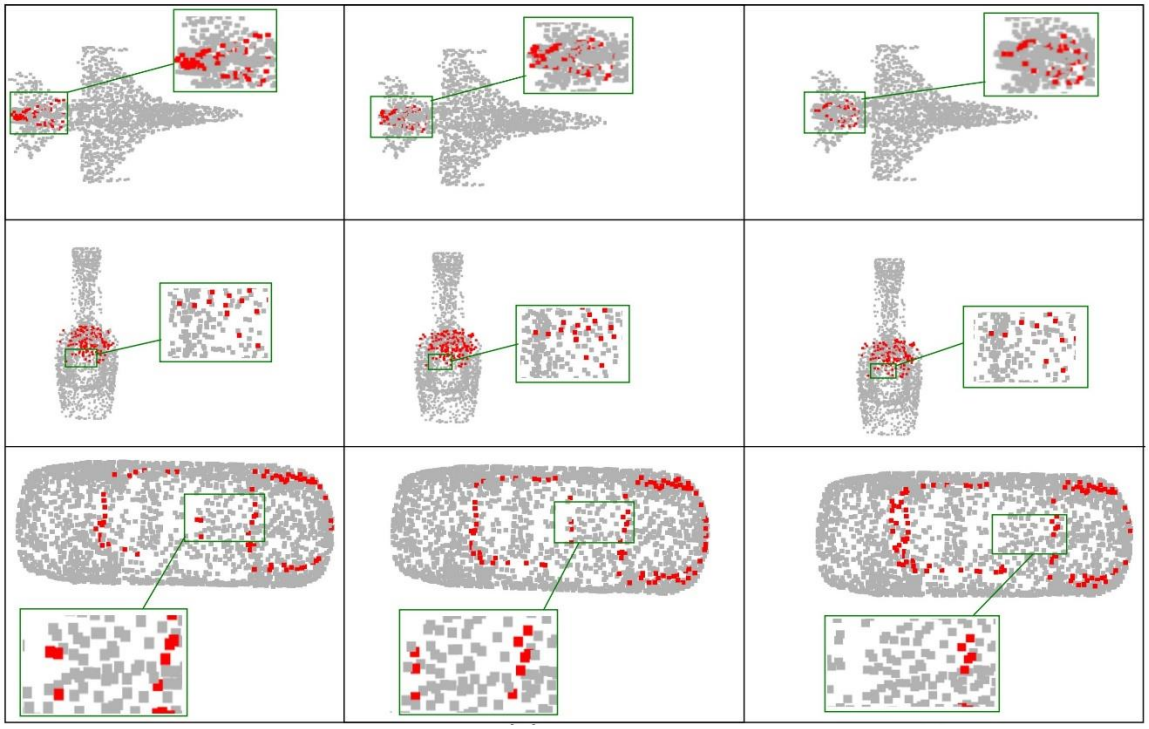

Figure. 2 Comparison of our proposed work AdaCrossNet (right) with PointNet++ (left), and CrossNet (middle) using dataset ShapeNetPart. The first row is the airplane model, the second row is the earphone model, and the third row is the car model viewed from top

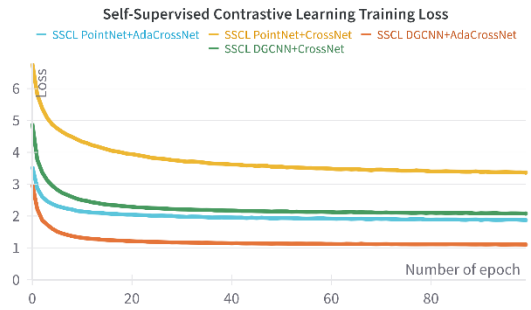

Figure. 3 Self-Supervised Contrastive Learning Training Loss for SSCL PointNet and DGCNN with AdaCrossNet and CrossNet.

4.4 Ablation study

We investigate the impact of the proposed dynamic weight adjustment mechanism, AdaCrossNet, on self-supervised contrastive learning (SSCL) training performance. The graph in Fig. 3 compares the training loss trajectories across different configurations: SSCL PointNet with AdaCrossNet and CrossNet and SSCL DGCNN with AdaCrossNet and CrossNet. The dynamic weight adjustment mechanism balances the learning focus between IM and CM tasks, resulting in improved convergence and more efficient learning. This study highlights the benefit of the AdaCrossNet mechanism in reducing the overall loss, especially when combined with the DGCNN architecture.

As shown in Fig. 3, the models incorporating AdaCrossNet exhibit lower training loss than counterparts using CrossNet. Specifically, SSCL DGCNN+AdaCrossNet achieves the fastest convergence, reaching the lowest loss by the end of training. The results validate the strength of dynamically adjusting the weights of IM and CM objectives during training, leading to improved learning efficiency and better overall performance. This confirms that our dynamic weight adjustment strategy is particularly beneficial in multi-modal contrastive learning scenarios.

5. Conclusion

This section outlines our concerns about AdaCrossNet for CM contrastive learning. One key disadvantage is that it is difficult to state a suitable convergence criterion for the dynamic adjustment of the IM and CM losses. Also, the dynamic weight adjustment system depends on the convergence tracking of every modality, which may cause issues during training. For instance, excessive damping of the response to a change in any of the loss values, if the convergence is not achieved, could lead to unstable learning and increased iterations with even possible failure to converge, especially in sector weight adjustments. This kind of response time is detrimental because it also means one has to adjust the parameters, the smoothing factor, and the weight changing, therefore making the training process more elaborate than it was before.

Another drawback is increasing the complexity of the training phase because of the need to proportionate the contribution of each modality. Typically, one modality suppresses the other during learning, which is undesirable. This leads to representations skewed towards a more dominant modality and limits the model's performance on tasks that require both modalities to be active. This is partially resolved with dynamic weight adjustment in our implementation; however, more work is required to ensure both modalities are balanced in the final representation, especially in noise or unequal class distribution. Dynamic weight adjustments during training can also be computationally expensive, increasing training time and making it inconvenient for large-scale projects.

6. Conclusion

In this work, we have proposed AdaCrossNet, a novel dynamic weight modulation for CM contrastive learning of 3D point clouds and 2D images. The main contribution is implementing an adaptive intra-modal and cross-modal loss weighting strategy, which helps the model control the learning rate in different modalities during training. We have implemented exponentially weighted moving averages (EWMA) to improve the stability of weight modifications during training. Our contribution has been proved by extensive experimentation on various benchmark datasets. AdaCrossNet attained a 91.4% accuracy on the ModelNet40 classification challenge, exceeding the performance of CrossNet and CrossPoint.

Furthermore, in point cloud segmentation on the ShapeNetPart dataset, AdaCrossNet surpassed other models, achieving a mean Intersection over Union (mIoU) score of 85.1%, in contrast to CrossNet's 83.1% and PointNet++'s 83.8%. AdaCrossNet exhibited a notable enhancement in the ScanObjectNN dataset, achieving an accuracy of 82.1%, illustrating its resilience across diverse workloads and datasets. The results validate that AdaCrossNet provides significant enhancements over alternative approaches, enhancing cross-modal contrastive learning and overall performance. With implications for broader use in 3D vision tasks, this

research improves the use of dynamic weighting in self-supervised learning for 3D point clouds.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Oddy Virgantara Putra, Eko Mulyanto Yuniarno and Mauridhi Hery Purnomo; methodology, Oddy Virgantara Putra, Kohichi Ogata, Eko Mulyanto Yuniarno; validation, Oddy Virgantara Putra, Eko Mulyanto Yuniarno; software and resources, Oddy Virgantara Putra, Eko Mulyanto Yuniarno; investigation, Oddy Virgantara Putra, Kohichi Ogata, Eko Mulyanto Yuniarno, and Mauridhi Hery Purnomo; writing-original draft preparation, Oddy Virgantara Putra;writing-review and editing, Oddy Virgantara Putra, Kohichi Ogata, Eko Mulyanto Yuniarno, and Mauridhi Hery Purnomo; visualization, Oddy Virgantara Putra, Kohichi Ogata, and Eko Mulyanto Yuniarno; supervision, Eko Mulyanto Yuniarno and Mauridhi Hery Purnomo; All authors read and approved the final manuscript.

Acknowledgments

This work was supported and fully funded by Lembaga Pengelola Dana Pendidikan (LPDP) Indonesia.

References

[1] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”, In: *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77-85, 2017, doi: https://doi.org/10.1109/CVPR.2017.16.

[2] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”, In: *Proc. of the 31st International Conference on Neural Information Processing Systems*, in NIPS’17. Red Hook, NY, USA, pp. 5105-5114, 2017, [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295263

[3] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic Graph CNN for Learning on Point Clouds”, *ACM Trans. Graph.*, Vol. 38, No. 5, 2019, doi: 10.1145/3326362.

[4] I. Misra, R. Girdhar, and A. Joulin, “An End-to-End Transformer Model for 3D Object Detection”, In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2886-2897, 2021, doi: 10.1109/ICCV48922.2021.00290.

[5] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, “KPConv: Flexible and Deformable Convolution for Point Clouds”, In: *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6410-6419, 2019, doi: 10.1109/ICCV.2019.00651.

[6] M. Caron *et al.*, “Emerging Properties in Self-Supervised Vision Transformers”, In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630-9640, 2021, doi: 10.1109/ICCV48922.2021.00951.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations”, In: *Proc. of the 37th International Conference on Machine Learning*, 2020.

[8] J.-B. Grill *et al.*, “Bootstrap Your Own Latent A New Approach to Self-Supervised Learning”, In: *Proc. of the 34th International Conference on Neural Information Processing Systems*, in NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum Contrast for Unsupervised Visual Representation Learning”, In: *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726-9735, 2020, doi: 10.1109/CVPR42600.2020.00975.

[10] S. Chang, Y. Li, S. Shen, J. Feng, and Z. Zhou, “Contrastive Attention for Video Anomaly Detection”, *IEEE Trans Multimedia*, Vol. 24, pp. 4067-4076, 2022, doi: 10.1109/TMM.2021.3112814.

[11] M. Li, P.-Y. Huang, X. Chang, J. Hu, Y. Yang, and A. Hauptmann, “Video Pivoting Unsupervised Multi-Modal Machine Translation”, *IEEE Trans Pattern Anal Mach Intell*, Vol. 45, No. 3, pp. 3918-3932, 2023, doi: 10.1109/TPAMI.2022.3181116.

[12] B. A. Plummer *et al.*, "Revisiting Image-Language Networks for Open-Ended Phrase Detection", *IEEE Trans Pattern Anal Mach Intell*, Vol. 44, No. 4, pp. 2155-2167, 2022, doi: 10.1109/TPAMI.2020.3029008.

[13] X. Zhong, C. Gu, M. Ye, W. Huang, and C.-W. Lin, "Graph Complemented Latent Representation for Few-Shot Image Classification", *IEEE Trans Multimedia*, Vol. 25, pp. 1979-1990, 2023, doi: 10.1109/TMM.2022.3141886.

[14] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo, "CrossPoint: Self-Supervised Cross-Modal Contrastive Learning for 3D Point Cloud Understanding", In: *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9892-9902, 2022, doi: 10.1109/CVPR52688.2022.00967.

[15] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis", In: *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 12, pp. 10790-10797, 2021, doi: 10.1609/aaai.v35i12.17289.

[16] Z. Ma and T. Pan, "Adaptive Weight Tuning of EWMA Controller via Model-Free Deep Reinforcement Learning", *IEEE Transactions on Semiconductor Manufacturing*, Vol. 36, No. 1, pp. 91-99, 2023, doi: 10.1109/TSM.2022.3225480.

[17] P. Chattopadhyay, Y. Balaji, and J. Hoffman, "Learning to Balance Specificity and Invariance for In and Out of Domain Generalization", In: *Proc. of Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, Berlin, pp. 301-318, 2020, doi: 10.1007/978-3-030-58545-7_18.

[18] Y. Wu *et al.*, "Self-Supervised Intra-Modal and Cross-Modal Contrastive Learning for Point Cloud Understanding", *IEEE Trans Multimedia*, Vol. 26, pp. 1626-1638, 2024, doi: 10.1109/TMM.2023.3284591.

[19] Q. Meng, W. Wang, T. Zhou, J. Shen, Y. Jia, and L. Van Gool, "Towards a Weakly Supervised Framework for 3D Point Cloud Object Detection and Annotation", *IEEE Trans Pattern Anal Mach Intell*, Vol. 44, No. 8, pp. 4454-4468, 2022, doi: https://doi.org/10.1109/TPAMI.2021.3063611.

[20] Y. Wu, X. Chen, X. Huang, K. Song, and D. Zhang, "Unsupervised Distribution-aware Keypoints Generation from 3D Point Clouds", *Neural Networks*, Vol. 173, p. 106158, 2024, doi: https://doi.org/10.1016/j.neunet.2024.106158.

[21] L. Landrieu and M. Simonovsky, "Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs", In: *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4558-4567, 2018, doi: https://doi.org/10.1109/CVPR.2018.00479.

[22] T. Xiang, C. Zhang, Y. Song, J. Yu, and W. Cai, "Walk in the Cloud: Learning Curves for Point Clouds Shape Analysis", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV*, pp. 895-904*)*, 2021, doi: 10.1109/ICCV48922.2021.00095.

[23] L. Hui, H. Yang, M. Cheng, J. Xie, and J. Yang, "Pyramid Point Cloud Transformer for Large-Scale Place Recognition", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6078-6087, 2021, doi: 10.1109/ICCV48922.2021.00604.

[24] D. Lu, Q. Xie, K. Gao, L. Xu, and J. Li, "3DCTN: 3D Convolution-Transformer Network for Point Cloud Classification", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 12, pp. 24854-24865, 2022, doi: 10.1109/TITS.2022.3198836.

[25] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling", In: *Proc. of the 30th International Conference on Neural Information Processing Systems*, pp. 82-90, 2016.

[26] J. Li, B. M. Chen, and G. Lee, "SO-Net: Self-Organizing Network for Point Cloud Analysis", In: *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, pp. 9397-9406, 2018, doi: 10.1109/CVPR.2018.00979.

[27] Y. Yang, C. Feng, Y. Shen, and D. Tian, "FoldingNet: Point Cloud Auto-Encoder via Deep Grid Deformation", In: *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 206-215, 2018, doi: 10.1109/CVPR.2018.00029.

[28] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, "Learning Representations and Generative Models for 3D Point Clouds", In: *Proc. of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., in Proceedings of Machine Learning Research, Vol. 80. PMLR, pp. 40-49, 2018, [Online]. Available: https://proceedings.mlr.press/v80/achlioptas18a.html

[29] L. Zhang and Z. Zhu, "Unsupervised Feature Learning for Point Cloud Understanding by Contrasting and Clustering Using Graph Convolutional Neural Networks", In: *Proc. of 2019 International Conference on 3D Vision (3DV)*, pp. 395-404, 2019, doi: 10.1109/3DV.2019.00051.

[30] J. and G. D. and Q. C. R. and G. L. and L. O. Xie Saining and Gu, "PointContrast: Unsupervised Pre-training for 3D Point Cloud Understanding", In: *Proc. of Computer Vision - ECCV 2020*, H. and B. T. and F. J.-M. Vedaldi Andrea and Bischof, pp. 574-591, 2020, doi: 10.1007/978-3-030-58580-8_34.

[31] J. Sauder and B. Sievers, "Self-Supervised Deep Learning on Point Clouds by Reconstructing Space", *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/993edc98ca87f7e08494eec37fa836f7-Paper.pdf

[32] O. Poursaeed, T. Jiang, H. Qiao, N. Xu, and V. G. Kim, "Self-Supervised Learning of Point Clouds via Orientation Estimation", In: *Proc. of 2020 International Conference on 3D Vision (3DV)*, pp. 1018-1028, 2020, doi: 10.1109/3DV50981.2020.00112.

[33] B. Du, X. Gao, W. Hu, and X. Li, "Self-Contrastive Learning with Hard Negative Sampling for Self-supervised Point Cloud Learning", In: *Proc. of the 29th ACM International Conference on Multimedia*, in MM '21. New York, NY, USA: Association for Computing Machinery, pp. 3133-3142, 2021, doi: 10.1145/3474085.3475458.

[34] S. Huang, Y. Xie, S.-C. Zhu, and Y. Zhu, "Spatio-Temporal Self-Supervised Representation Learning for 3D Point Clouds", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6515-6525, 2021, doi: 10.1109/ICCV48922.2021.00647.

[35] Z. Zhang, R. Girdhar, A. Joulin, and I. Misra, "Self-Supervised Pretraining of 3D Features on any Point-Cloud", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10232-10243, 2021, doi: 10.1109/ICCV48922.2021.01009.

[36] X. Hu, Y. Wu, X. Liu, Z. Li, Z. Yang, and M. Li, "Intra- and Inter-Modal Graph Attention Network and Contrastive Learning for SAR and Optical Image Registration", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61, pp. 1-16, 2023, doi: 10.1109/TGRS.2023.3328368.

[37] Q. Chen, G. Huang, and Y. Wang, "The Weighted Cross-Modal Attention Mechanism With Sentiment Prediction Auxiliary Task for Multimodal Sentiment Analysis", *IEEE/ACM Trans Audio Speech Lang Process*, Vol. 30, pp. 2689-2695, 2022, doi: 10.1109/TASLP.2022.3192728.

[38] Z. Han, A. Bin Azman, M. Rina Binti Mustaffa, and F. Binti Khalid, "Cross-Modal Retrieval: A Review of Methodologies, Datasets, and Future Perspectives", *IEEE Access*, Vol. 12, pp. 115716-115741, 2024, doi: 10.1109/ACCESS.2024.3444817.

[39] S. T. and K. M. and M. A. and K. K. Rani Veenu and Nabi, "Self-supervised Learning: A Succinct Review", *Archives of Computational Methods in Engineering*, Vol. 30, No. 4, pp. 2761-2775, 2023, doi: 10.1007/s11831-023-09884-2.

[40] J. Gui *et al.*, "A Survey on Self-supervised Learning: Algorithms, Applications, and Future Trends", *IEEE Trans Pattern Anal Mach Intell*,

pp. 1-20, 2024, doi: 10.1109/TPAMI.2024.3415112.

[41] I. Misra and L. van der Maaten, "Self-Supervised Learning of Pretext-Invariant Representations", In: *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6706-6716, 2020, doi: 10.1109/CVPR42600.2020.00674.

[42] L. Shi, G. Zhang, Q. Cao, L. Zhang, Y. Cen, and Y. Cen, "DCPoint: Global-Local Dual Contrast for Self-Supervised Representation Learning of 3-D Point Clouds", *IEEE Sens J*, Vol. 24, No. 14, pp. 23224-23238, 2024, doi: 10.1109/JSEN.2024.3405079.

[43] X. Wang, M. Zhang, B. Chen, D. Wei, and Y. Shao, "Dynamic Weighted Multitask Learning and Contrastive Learning for Multimodal Sentiment Analysis", *Electronics (Basel)*, Vol. 12, No. 13, 2023, doi: 10.3390/electronics12132986.

[44] Z. Ma, T. Pan, and J. Tian, "Deep reinforcement learning optimized double exponentially weighted moving average controller for chemical mechanical polishing processes", *Chemical Engineering Research and Design*, Vol. 197, pp. 419-433, 2023, doi: 10.1016/j.cherd.2023.07.049.

[45] A. X. Chang *et al.*, "ShapeNet: An Information-Rich 3D Model Repository", *arXiv.*, 2015, doi: 10.48550/ARXIV.1512.03012.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", In: *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, doi: 10.1109/CVPR.2016.90.

[47] Z. Wu *et al.*, "3D ShapeNets: A Deep Representation for Volumetric Shapes", In: *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1912-1920, 2015, doi: https://doi.org/10.1109/CVPR.2015.7298801.

[48] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, "Revisiting Point Cloud Classification: A New Benchmark Dataset and Classification Model on Real-World Data", In: *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1588-1597, 2019, doi: 10.1109/ICCV.2019.00167.

[49] L. Yi *et al.*, "A Scalable Active Framework for Region Annotation in 3D Shape Collections", *ACM Trans. Graph.*, Vol. 35, No. 6, 2016, doi: 10.1145/2980179.2980238.