Received: May 6, 2025. Revised: May 24, 2025.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

BeyondRPC: A Contrastive and Augmentation-Driven Framework for Robust Point Cloud Understanding

Oddy Virgantara Putra[1]* Hanugra Aulia Sidharta[2] Diah Risqiwati[3]
Moch. Iskandar Riansyah[4] Yuni Yamasari[5]

[1]*Department of Informatics, Universitas Darussalam Gontor, Ponorogo, Indonesia*
[2]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia*
[3]*Department of Informatics, Universitas Muhammadiyah Malang, Malang, Indonesia*
[4]*Department of Electrical Engineering, Universitas Telkom, Bandung, Indonesia*
[5]*Department of Informatics, Universitas Negeri Surabaya, Surabaya, Indonesia*
* Corresponding author's Email: oddy@unida.gontor.ac.id

Abstract: Robust perception of 3D point clouds remains a significant challenge in real-world environments where sensor data is often corrupted. While recent models and augmentation strategies have improved robustness individually, their isolated use still limits performance under severe distortions. In this work, we introduce BeyondRPC, a contrastive and augmentation-driven framework for robust point cloud classification. Our approach combines AdaCrossNet for adaptive cross-modal contrastive pretraining with WOLFMix-based fine-tuning to improve generalization under corruption. Specifically, AdaCrossNet employs a dynamic weighting mechanism to balance intra- and cross-modal learning, while WOLFMix integrates both deformation-based and rigid-mix augmentations. Experiments on the ModelNet-C benchmark demonstrate that BeyondRPC achieves a mean Corruption Error of 0.455, outperforming state-of-the-art methods, including RPC, GDANet, and CurveNet, while maintaining high clean overall accuracy at 0.930. These results underscore the importance of joint contrastive representation learning and corruption-aware augmentation for robust 3D point cloud understanding.

Keywords: Contrastive learning, Cross-modal learning, ModelNet-C, Point cloud classification, Point cloud corruption, Robustness, Self-supervised learning, 3D deep learning.

1. Introduction

The ability to robustly perceive and interpret 3D environments is central to a wide range of applications, especially autonomous driving [1, 2], augmented reality [3, 4], and robotics [5, 6]. 3D data can be represented in several formats, such as point clouds, 2.5D images, and volumetric structures. Point clouds are widely used among these because they preserve the original geometric information in Euclidean space without quantization [7]. However, this advantage also brings challenges, especially when visualizing and understanding such data, that are even more complex in applications like autonomous vehicles and human-like robots. Point clouds are challenging to work with due to their unstructured form [8], unordered[9], and high dimensionality [10]. Despite these difficulties, deep learning researchers have made significant progress, primarily supported by public datasets like KITTI [11], ModelNet10, ModelNet40, and ShapeNet [12]. These benchmarks have driven the development of many advanced techniques for classification, detection, tracking, segmentation, registration, and reconstruction of point clouds.

Motivated by the success of robust point cloud classifier (RPC) [13] and the augmentation strength of WOLFMix, we propose a novel framework that integrates adaptive contrastive pretraining with corruption-aware fine-tuning. In particular, we reuse AdaCrossNet for representation learning and apply WOLFMix-based augmentations during supervised

training. Our approach aims to improve generalization and robustness against corruption.

Our contributions are summarized as follows:

- We introduce a resilient point cloud learning framework that combines adaptive contrastive pretraining with corruption-aware fine-tuning
- We integrate our previous AdaCrossNet with strong augmentations inspired by WOLFMix to enhance resilience against real-world corruptions.
- We conduct extensive evaluations on ModelNet-C showing that our approach attains leading robustness performance while sustaining competitive accuracy on clean data.

The remainder of the paper is organized as follows: Section II reviews related work; Section III describes the BeyondRPC framework; Section IV outlines the experimental setup; Section V presents results and ablation analysis; and Section VI concludes with future directions.

2. Related work

2.1 Point cloud classification

A variety of models have been proposed to address point cloud classification. These include MLP-based models such as PointNet [14, 15], convolution-based models [16, 17], graph-based models [8, 18], and the more recent transformer-based models [19-21]. There is a growing interest in enhancing robustness through data augmentation. Researchers have explored mix-based augmentations [22, 23] and auto-augmentations [24]. Self-supervised pre-training has also gained attention as an effective alternative to random initialization. Techniques such as point cloud reconstruction [25] and inpainting [26] have been shown to improve performance on downstream tasks. However, point cloud classifiers still struggle under extreme corruption despite these efforts. Some approaches include subsampled voting [27] and using self-supervision [28].

2.2 Self-supervised learning for point clouds

Self-Supervised Learning (SSL) offers a compelling alternative to supervised learning by leveraging data augmentations to create proxy tasks. Approaches like JigSaw3D [29] and Rotation3D [30] adapt 2D self-supervised methods to the 3D domain. PointContrast [31] uses point-wise contrastive losses to align augmented views of point clouds.

CrossPoint [32] and CrossNet [33] were among the first to explore cross-modal SSL between 3D point clouds and 2D-rendered images. They demonstrated that aligning features across modalities enhances representation quality. However, these methods rely on fixed weighting between intra-modal (IM) and cross modal (CM) losses, making them sensitive to the convergence rate of each branch. AdaCrossNet [34] addressed this limitation by adaptively adjusting the loss contributions using an exponentially weight moving average (EWMA)-based smoothing mechanism, improving both stability and downstream performance.

2.3 Augmentation strategies

Augmentation plays a critical role in enhancing model robustness. Mix-based augmentations such as PointMixUp [22] and RSMix [23] interpolate between two samples to regularize learning. Deformation-based methods like PointWOLF [35] introduce local perturbations to simulate sensor-level noise and spatial inaccuracies.

WOLFMix integrates these two complementary strategies to simulate a broader spectrum of corruptions during training. This augmentation is only applied during fine-tuning to prevent data leakage from corruption-aware augmentations into the pre-training stage. Empirical results show that WOLFMix consistently improves performance on ModelNet-C when paired with both standard and robust backbones.

While RPC has demonstrated robustness through its geometry-aware architecture and corruption evaluation protocol, it does not utilize any form of pre-training which limits its generalization under unseen corruptions. Additionally, RPC relies solely on standard augmentations and lacks corruption-specific adaptation during fine-tuning.

To address these limitations, we propose BeyondRPC, a framework that upgrades RPC by integrating AdaCrossNet, a contrastive self-supervised pre-training, with WOLFMix. This design preserves the effective backbone of RPC while substantially enhancing its feature representation and robustness. As shown in our ablation studies (Table 3 and Table 4), BeyondRPC achieves notable gains in performance compared to RPC and the standalone usage of AdaCrossNet or WOLFMix.

3. Proposed work

Despite the recent progress introduced by RPC, its robustness towards point cloud corruption remains constrained by fixed architecture design and standard augmentation strategies. In this work, we introduce

BeyondRPC, a robust 3D point cloud framework that combines two previously proposed but independently used techniques, AdaCrossNet for contrastive pre-training and WOLFMix for corruption-aware augmentation.

While AdaCrossNet and WOLFMix are not newly proposed in this work, their joint integration within the RPC baseline is novel and non-trivial. AdaCrossNet enhances feature representation via adaptive IM and CM contrastive learning. On the other hand, WOLFMix is applied during fine-tuning to simulate diverse real-world corruptions. The result is a synergistic framework that improves both clean and corruption performance beyond what either module achieves alone. For clarity, the notations used in this section are summarized in Table 1.

3.1 Contrastive pre-training with AdaCrossNet

Contrastive learning is a part of SSL that offers a solution to the challenge of learning from unlabelled point cloud data. AdaCrossNet introduces dynamic weights λ_{IM} and λ_{CM} that are updated during training using an exponentially weighted moving average (EWMA). The final pretraining loss is defined as:

$$\mathcal{L}_{ACM} = \lambda_{IM}.\mathcal{L}_{IM} + \lambda_{CM}.\mathcal{L}_{CM} \quad (1)$$

where each loss term is calculated using cosine similarity over positive and negative pairs. λ_{IM} and λ_{CM} are two dynamic weights which adaptively adjust the relative importance both IM and CM.

Let $\mathcal{L}_{IM}^{(t)}$ and $\mathcal{L}_{CM}^{(t)}$ be the intra-modal and cross-modal contrastive loss at iteration t, respectively. We aim to learn dynamic weights $\lambda_{IM}^{(t)}$ and $\lambda_{CM}^{(t)}$ that adaptively balance these losses. The update rule is based on EWMA principle. We define the relative changes in loss as:

$$\Delta\mathcal{L}^{(t)} = \frac{\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}}{\mathcal{L}^{(t-1)} + \epsilon} \quad (2)$$

where ϵ is a small constant to prevent division by zero. Then, we update the dynamic weight $\lambda^{(t)}$ for each branch by:

$$\lambda(x) = \beta\lambda(x-1) + (1-\beta)\frac{1}{1+e^{-\alpha\Delta L}} \quad (3)$$

where $\lambda(x)$ denotes the dynamic weight at iteration x, $\lambda(x-1)$ is the previous weight, and ΔL represents the change in loss, i.e., between consecutive epochs. The parameter $\beta \in [0,1]$ controls the influence of historical weights. The full procedure of the dynamic weight update mechanism is summarized in Algorithm 1.

3.2 Backbone architecture

In our framework, we adopt the architecture design from the RPC baseline, which integrates two main modules, Point Cloud Transformer (PCT) and Geometry-Disentagle Module (GDM) from GDANet [36]. The RPC model architecture can be seen in Fig. 1.

3.2.1. Point cloud transformer

PCT encoder consists of a coordinate-based input embedding, four stacked attention layers self-attention (S-Attn) and a linear transformation to map the features to the desired representation space. The output feature supplies the input for classification task.

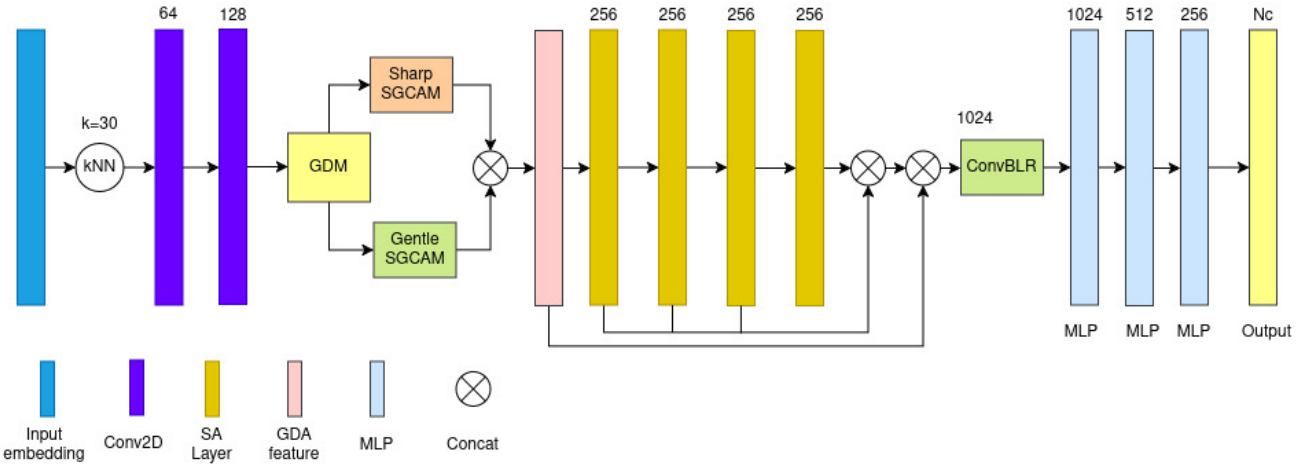

Figure. 1 The overall architecture of the RPC model.The RPC model consists of kNN grouping, 2D convolutions, GDM (sharp/gentle split), Sharp and Gentle SGCAMs, and SA layers. Final features are fused via ConvBLR and classified using MLP layers into Nc classes

Table 1. Table of Notations

Symbol	Description
$\mathcal{L}_{IM}, \mathcal{L}_{CM}$	IM and CM contrastive loss
$\lambda_{IM}, \lambda_{CM}$	Dynamic weights for IM, CM loss
β	EWMA smoothing coefficient
$\lambda(x), \Delta\mathcal{L}$	Dynamic weight at x and change in loss
$\mathbf{F}_{in}$	Input feature to attention layer
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	Query, Key, and Value matrices in attention mechanism
$\tilde{A}$	Attention score matrix
$\alpha_{i,j}$	Softmax normalized attention score
$\mathbf{X}_s, \mathbf{X}_g$	Sharp and gentle point cloud features
Θ, Φ	Non-linear transformation functions in GDM
$\mathbf{W}_s, \mathbf{W}_g$	Weight matrices for sharp and gentle components
$\Psi_m(\cdot)$	Non-linear transformation for modality $m \in \{s, g\}$
$\mathbf{Y}_s, \mathbf{Y}_g$	Output features from sharp and gentle components
$\mathbf{Z}$	Concatenated /fused feature from sharp and gentle components

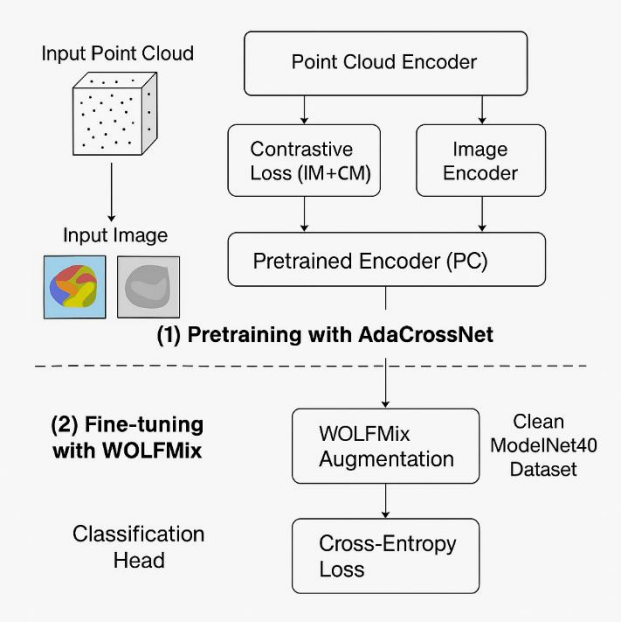

Figure. 2 The end-to-end training pipeline of the proposed Beyond RPC framework

Each attention layer adopts a query-key-value structure. Let $\mathbf{F_{in}}$ be the input feature to the attention layer. The attention operation begins by projecting $\mathbf{F_{in}}$ into query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) matrices using learned linear projections:

$$(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{F}_{in} \cdot (\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v) \quad (4)$$

Next, the attention scores are computed as the dot product between queries and keys:

$$\tilde{A} = \mathbf{Q} \cdot \mathbf{K}^T \quad (5)$$

$$\alpha_{i,j} = \frac{e^{\tilde{a}_{i,j}}}{\sum_k e^{\tilde{a}_{i,k}}} \quad (6)$$

Finally, the attention output is passed through a Linear-BatchNorm-ReLU (LBR) block and added with a residual connection from the input. To capture local geometric details, PCT incorporates a neighbor embedding module that aggregates k-nearest neighbor (k-NN) features using farthest point sampling (FPS) and shared MLP.

3.2.2. Geometry disentangle-module

To obtain geometric representations of 3D point clouds, we employ a GDM. Specifically, GDM finds the sharp and gentle variation components' distinctive traits using the sharp-mild complementary attention module (SCAM). To build a graph, GDM uses *k*-NN. The GDM generates two components that function as inputs for the sharp-gentle complementary attention module (SGCAM). The input point cloud is denoted as $\mathbb{X}_o$, the sharp component as $\mathbb{X}_s$, and the gentle component as $\mathbb{X}_g$.

$$\mathbb{W}_s = \Theta_o(\mathbb{X}_o) \cdot \Theta_s(\mathbb{X}_s)^T \quad (7)$$

$$\mathbb{W}_g = \Phi_o(\mathbb{X}_o) \cdot (\Phi_g(\mathbb{X}_g)^T \quad (8)$$

In the present scenario, Θ_o, Θ_s, Φ_o, and Φ_g represent distinct nonlinear functions, each serving a specific purpose. Subsequently, the fusion of $\mathbb{W}_s$ and $\mathbb{W}_g$ is defined through elementwise operations as outlined below:

Algorithm 1. Dynamic Contrastive Weight Update for AdaCrossNet

Require: Previous weights $\lambda_{IM}^{(t-1)}, \lambda_{CM}^{(t-1)}$, previous losses $\mathcal{L}_{IM}^{(t-1)}, \mathcal{L}_{CM}^{(t-1)}$, current losses $\mathcal{L}_{IM}^{t}, \mathcal{L}_{CM}^{t}$, parameters β, α

Ensure: Updated weights $\mathcal{L}_{IM}^{t}, \mathcal{L}_{CM}^{t}$

1: $\Delta_{IM} \leftarrow \frac{\mathcal{L}_{IM}^{t} - \mathcal{L}_{IM}^{(t-1)}}{\mathcal{L}_{IM}^{(t-1)} + \epsilon}$

2: $\Delta_{CM} \leftarrow \frac{\mathcal{L}_{CM}^{t} - \mathcal{L}_{CM}^{(t-1)}}{\mathcal{L}_{CM}^{(t-1)} + \epsilon}$

3: $\lambda_{IM}^{t} \leftarrow \beta \cdot \lambda_{IM}^{(t-1)} + (1-\beta) \cdot \frac{1}{1+\exp(\alpha \cdot \Delta_{IM})}$

4: $\lambda_{CM}^{t} \leftarrow \beta \cdot \lambda_{CM}^{(t-1)} + (1-\beta) \cdot \frac{1}{1+\exp(\alpha \cdot \Delta_{CM})}$

5: **return** $\lambda_{IM}^{t}, \lambda_{CM}^{t}$

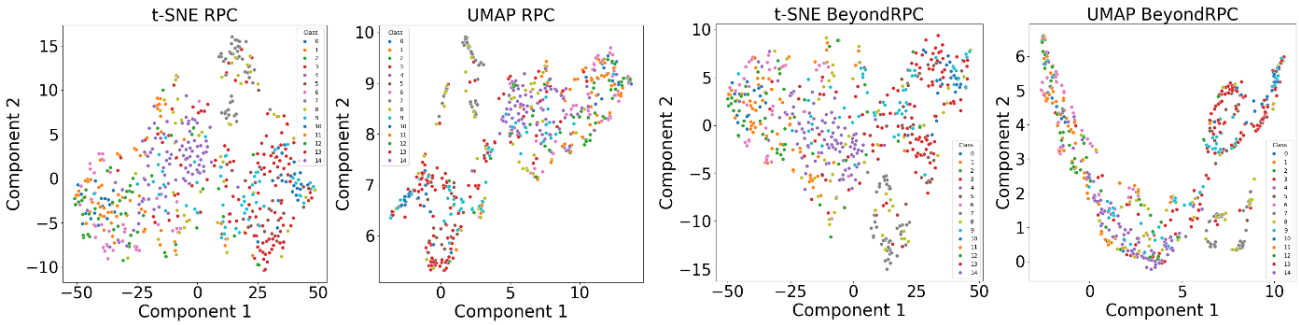

Figure. 3 Comparison of 2D embeddings on ScanObjectNN test set using t-SNE and UMAP. Left: Baseline RPC. Right: BeyondRPC

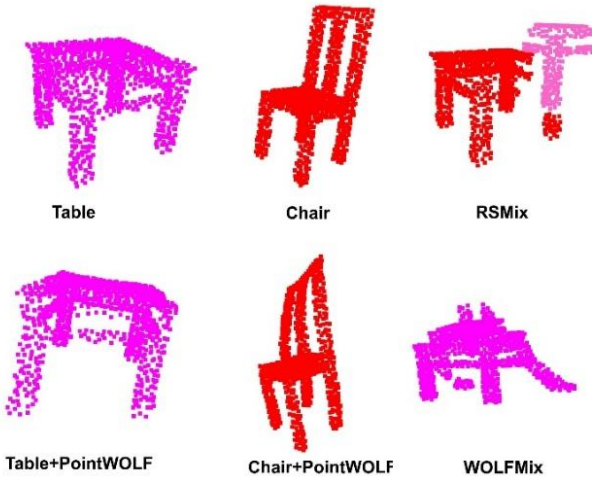

Figure. 4 Visualization of sample augmentations on ModelNet40 point clouds using "table" and "chair" classes. The top row is the original samples and RSMix. The bottom row is individual PointWOLF augmentations using WOLFMix

$$(\mathbb{Y}_m)_i = (\mathbb{X}_o)_i + \sum_{j=1}^{M} (\mathbb{W}_m)_{i,j} \cdot \Psi_m\big((\mathbb{X}_m)_j\big) \quad (9)$$

where $m \in \{s, g\}$ indicates the sharp and gentle geometric components, respectively. $\mathbb{X}_o$ is the shared input feature extracted from the point cloud, $\mathbb{X}_m$ denotes the input corresponding to each component, $\mathbb{W}_m$ is the weight matrix specific to component m, and $\Psi_m(\cdot)$ is a nonlinear transformation function for modality m.

The output features obtained from the sharp ($\mathbb{Y}_s$) and gentle ($\mathbb{Y}_g$) components are integrated using the operation specified in Eq. (10):

$$\mathbb{Z} = \mathbb{Y}_s \oplus \mathbb{Y}_g \quad (10)$$

The concatenation process ($\oplus$) joins important geometric features with their corresponding complementary counterparts.

3.3 RPC baseline overview

RPC serves as our baseline model and integrates two key modules: the PCT and the GDM. The PCT captures global contextual information using offset-attention and neighbor embedding, while the GDM focuses on learning disentangled geometric features by separating sharp and gentle components. We adopt this architecture as the foundation for our framework and evaluate its robustness with and without AdaCrossNet pretraining and WOLFMix augmentations. While retaining the core principles of point cloud encoding from PCT and GDM, BeyondRPC introduces architectural refinements aimed at improving feature representation under corruption. The whole beyond RPC pipeline can be seen in Fig. 3.

3.4 Evaluation metrics

To evaluate the corruption error (CE), we need to calculate the mean CE. Before that, we first find solution for CE:

$$CE_i = \frac{\sum_{l=1}^{m} 1 - OA_i}{\sum_{l=1}^{m} 1 - OA_{i,l}^{bl}} \quad (11)$$

where OA is the overall accuracy for the corruption set at i with given level of l. OA^{bl} is the OA of baseline model. The baseline model can be any point cloud model such as PointNet, DGCNN, and others.

Here, we also utilized the relative mCE (RmCE) to determine the drop performance compared to the clean one. We can find RmCE with:

$$RCE_i = \frac{\sum_{l=1}^{m} OA_{clean} - OA_i}{\sum_{l=1}^{m} OA_{clean}^{bl} - OA_{i,l}^{bl}} \quad (12)$$

RCE for a given corruption type i is calculated by taking the average performance drop of the model

Table 2. Comparison of classification accuracy (OA) and robustness (mCE) on ModelNet-C with individual corruption results

Model	OA	mCE	Scale	Jitter	Drop-G	Drop-L	Add-G	Add-L	Rotate
DGCNN [8]	0.926	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
PointNet++ [15]	0.930	1.072	0.872	1.177	0.641	1.802	0.614	0.993	1.405
RSCNN [16]	0.923	1.130	1.074	1.171	0.806	1.517	0.712	1.153	1.479
GDANet [36]	0.928	0.892	1.043	0.744	1.012	0.623	0.478	0.480	**0.493**
CurveNet [18]	0.933	0.927	0.894	0.674	0.706	1.135	1.525	1.015	0.809
PAConv [17]	**0.936**	1.104	0.904	1.465	1.000	1.005	1.085	1.298	0.967
PCT [19]	0.919	0.779	1.170	**0.570**	0.496	1.005	0.929	0.938	0.526
GTNet [21]	0.897	0.944	**0.844**	0.863	0.864	0.774	0.618	0.807	0.812
RPC [13]	0.926	0.863	0.840	0.892	0.492	0.797	0.929	1.011	1.079
BeyondRPC	0.930	**0.455**	1.032	0.810	**0.476**	**0.531**	**0.325**	**0.447**	0.507

across five severity levels of that corruption type, relative to the performance drop experienced by a baseline model under the same conditions.

A $RCE_i = 1$ indicates the model degrades similarly to DGCNN; a value less than 1 indicates better robustness (i.e., less degradation). RmCE provides a single scalar summary of the model's robustness relative to the baseline across various corruption types.

4. Experimental setup

AdaCrossNet is evaluated on standard 3D benchmarks under clean and corrupted conditions. Pretraining uses the ShapeNet [37] dataset containing roughly 50,000 3D CAD models. We also collected the RGB images from [38] which contains around 43K point clouds with RGB and grayscale renderings, with each sample represented by 2048 XYZ points and one 224×224 image.

For each 3D point cloud, a single 2D image is randomly chosen from the available rendered views, each captured from different arbitrary viewpoints. We adopt PointNet [14] and DGCNN [8] as the backbone networks for extracting point cloud features to ensure a fair comparison with prior approaches. At the same time, ResNet-50 [39] is used to extract visual features from the selected 2D image. Both modalities are passed through dedicated 2-layer MLP projection heads that map features into a shared 256-dimensional latent space. The training is conducted using the Adam optimizer [40] with a weight decay of 1×10^{-4} and an initial learning rate of 1×10^{-3}, regulated by a cosine annealing schedule [41] over 100 epochs. The batch size is set to 32. After pre-training, the image encoder f_{θ_I} and both projection heads g_{ϕ_P} and g_{ϕ_I} are discarded. The downstream evaluation relies solely on the pre-trained point cloud encoder f_{θ_P}.

Data augmentations include jittering, flipping, cropping, and normalization. Training runs for 200 epochs with cosine annealing (initial LR 1e-3, weight decay 1e-4) on an RTX 4090. After pre-training, image branches are removed, and the point cloud encoder is fine-tuned or evaluated via linear SVM on ModelNet40 and ScanObjectNN [42]. Robustness is assessed on ModelNet-C using RCE and RmCE, compared to a DGCNN baseline.

5. Results and discussion

Here, we discuss the experimental results of our proposed method, BeyondRPC, compared against several state-of-the-art models, including DGCNN, PointNet++, RSCNN, GDANet, CurveNet, PAConv, PCT, GTNet, and RPC. We first analyze robustness under corruption using the ModelNet-C benchmark.

5.1 Performance on modelnet-c

We evaluate BeyondRPC using the ModelNet-C benchmark, which simulates real-world corruption across seven types and five severity levels. The results are measured using overall accuracy (OA), mean Corruption Error (mCE), and individual corruption sensitivity (e.g., jitter, scale, drop/add global/local, rotation).

Table 2 compares classification accuracy (OA) and robustness (mCE) across models on the ModelNet-C benchmark. BeyondRPC achieves the best robustness with the lowest mCE (0.455) and high OA (0.930), outperforming others in five of seven corruption types. Its strength comes from adaptive cross-modal pre-training via AdaCrossNet, which improves generalization under geometric corruptions. Compared to RPC (mCE: 0.863) and GDANet (0.892), BeyondRPC shows notable gains. While CurveNet has the highest OA (0.933), its mCE

(0.927) is less favorable. Overall, BeyondRPC offers the best trade-off between accuracy and robustness.

Scores for DGCNN, RSCNN, PointNet++, and PAConv are taken directly from the official RPC benchmark [13]. GDANet, CurveNet, PCT, and RPC were reproduced using publicly available code and evaluated under the same ModelNet-C corruption setup for consistency. Where original ModelNet-C scores were not reported (e.g., for CurveNet or PCT), we reproduced results using official implementations and corruption settings from RPC [13].

BeyondRPC's superiority is due to its two-stage design: AdaCrossNet enables better generalization via dynamic contrastive learning, while WOLFMix simulates real-world corruptions during fine-tuning. These components complement each other—contrastive pre-training improves global feature alignment, and WOLFMix reinforces local robustness. Unlike static models such as PAConv [17] or GTNet [21], BeyondRPC actively learns to adapt to both clean and corrupted data conditions. This theoretical design rationale explains its strong results across all corruption types in Table 2, especially in Add-G, Add-L, and Drop-L scenarios.

Results for DGCNN and PointBERT are directly taken from the official RPC benchmark [13]. Other results were reproduced under the same ModelNet-C corruption setup using publicly released code and checkpoints.

5.2 Performance on real-world scanobjectnn dataset

To assess real-world robustness, we evaluate BeyondRPC on the ScanObjectNN dataset, which contains naturally corrupted objects with background clutter, occlusion, and viewpoint variation. As shown in Table 5, BeyondRPC achieves the highest overall accuracy of 84.7%, outperforming state-of-the-art baselines such as PointBERT (83.1%), APPNet (84.1%), and RPC (83.6%). BeyondRPC maintains consistent superiority over augmentation-based (SageMix) and contrastive learning-based methods (CrossPoint), highlighting the complementary effect of AdaCrossNet pretraining and WOLFMix fine-tuning. This demonstrates the model's strong generalization to real-world scenarios beyond synthetic benchmarks like ModelNet40.

5.3 Ablation study

To analyze the effectiveness of the AdaCrossNet-based pretraining in BeyondRPC, we compare it against other well-established pre-training methods, including OcCo and PointBERT.

Table 3. Comparison of pre-training strategies under consistent augmentation (WOLFMix) on ModelNet-C

Model	Pre-train	mCE	ΔmCE
RPC	Baseline	0.637	0.000
RPC	CrossPoint	0.605	-0.032
RPC	CrossNet	0.599	-0.038
DGCNN	OcCo	1.047	+0.41
PointBERT [26]	PointBERT	1.248	+0.611
BeyondRPC	AdaCrossNet	0.455	-0.182

Table 4. Paired t-test results on ModelNet-C (mCE averaged over 5 seeds)

Model	Mean (mCE)	Std
RPC	0.8758	± 0.0371
BeyondRPC	0.6502	± 0.0111
t-stat	11.3532	
p-value	0.0003	

5.3.1. Effect on different pre-training strategies

Table 3 presents the effect of different pre-training strategies on model robustness under the WOLFMix augmentation pipeline. BeyondRPC achieves the lowest mCE score of 0.455 among all evaluated methods, outperforming all baselines and recent contrastive approaches. BeyondRPC also achieves the most significant robustness gain with a ΔmCE of -0.182 over the RPC baseline. In contrast, CrossNet and CrossPoint show minor gains (ΔmCE of -0.038 and -0.032), while OcCo and PointBERT degrade performance (+0.41 and +0.611). These results demonstrate that AdaCrossNet performs well with WOLFMix and the fusion-aware architecture.

To validate the robustness gain is statistically significant, we conducted a paired t-test comparing mCE scores of BeyondRPC and RPC across five different random sees. The test yields a t-statistic of 11.35 and a p-value of 0.0003. This result confirms the performance improvement is statistically significant ($p < 0.01$). The detailed results are summarized in Table 4.

5.3.2. Effect of different augmentation strategies

We perform an ablation study to assess the impact of fine-tuning augmentations by comparing PointWOLF, RSMix, and their combination in WOLFMix using the same AdaCrossNet pre-training. As shown in Table 6, WOLFMix achieves the best results with the lowest mCE (0.590) and highest mOA (0.870), indicating its effectiveness in capturing both local and global corruption patterns. Individually, RSMix provides better robustness

(mCE = 0.755), while PointWOLF yields higher clean accuracy (mOA = 0.837). These findings confirm their complementary roles, with WOLFMix offering the best trade-off between accuracy and robustness.

5.4 Feature representation visualization (t-sne and umap)

Here, we visualize the embeddings of the ScanObjectNN test set using t-SNE and UMAP for our model compared to the baseline (RPC). Fig. 4 visualizes ScanObjectNN test embeddings using t-SNE and UMAP for RPC (top row) and BeyondRPC (bottom row). BeyondRPC forms tighter, more distinct clusters in t-SNE and better-separated, denser groupings in UMAP, indicating improved feature discrimination.

To support these visual findings quantitatively, we compute the Silhouette Score and Davies-Bouldin Index on the PCA-reduced embeddings [45]. As in Table 7, BeyondRPC achieves a slightly higher Silhouette Score (-0.1791 vs. -0.1763) and a lower Davies-Bouldin Index (4.8504 vs. 4.8752) compared to RPC.

To demonstrate the synergistic effect of AdaCrossNet and WOLFMix, we compare BeyondRPC with its components. Table 3 shows that BeyondRPC outperforms AdaCrossNet-based pre-training alone, even when using the same WOLFMix augmentation.

Table 5. Results of the proposed BeyondRPC compared to other methods for ScanObjectNN dataset

Model	OA %
PointNet [14]	68.2
PointNet++ [15]	77.9
PointNet++ + SageMix [43]	83.7
DGCNN+CrossPoint [32]	81.7
DGCNN+SageMix [43]	83.6
DGCNN+AdaCrossNet [34]	82.1
CurveNet [18]	79.8
GDANet [36]	79.9
PointBERT [26]	83.1
APPNet [44]	84.1
RPC [13]	83.6
BeyondRPC	84.7

Table 6. Ablation of Augmentation Strategies during Fine-Tuning on ModelNet-C using BeyondRPC

Model	Augmentation	mCE	mOA
BeyondRPC	PointWOLF	0.888	0.789
BeyondRPC	RSMix	0.755	0.837
BeyondRPC	WOLFMix	0.590	0.870

Table 7. Clustering Quality Metrics on ScanObjectNN Embeddings

Model	Silhouette Score	Davies-Bouldin Index
RPC	-0.1763	48.752
BeyondRPC	**-0.1791**	**48.504**

Likewise, Table 6 confirms that WOLFMix achieves better robustness when used within BeyondRPC than when applied on top of other augmentations. These results suggest combining adaptive contrastive learning and corruption-specific fine-tuning leads to consistent and non-trivial performance gains.

6. Conclusion

This paper proposes BeyondRPC, designed to improve 3D point cloud classification robustness by combining adaptive contrastive pretraining with strong corruption-aware augmentations. Built upon the AdaCrossNet backbone—which dynamically balances intra- and cross-modal learning signals using EWMA—BeyondRPC enhances feature alignment and training stability. During fine-tuning, the WOLFMix augmentation strategy effectively simulates diverse real-world corruptions, enabling the model to generalize better to out-of-distribution scenarios.

Extensive evaluations on the ModelNet-C benchmark demonstrate that BeyondRPC achieves state-of-the-art robustness, with the lowest mCE (0.455) among all compared models, while maintaining competitive clean accuracy (0.930). Additionally, BeyondRPC attains the highest classification accuracy on the real-world ScanObjectNN dataset (84.7%), highlighting its effectiveness under both synthetic corruption and realistic sensor noise.

Ablation studies confirm that AdaCrossNet and WOLFMix contribute significantly to these performance gains. These findings underscore the potential of contrastive and augmentation-based approaches to bridge the gap between clean accuracy and robustness in 3D vision.

In future work, we plan to extend this framework to semantic segmentation tasks and explore multi-modal fusion beyond RGB, including depth and language-based supervision, to enhance applicability in more diverse and information-rich 3D environments. While ScanObjectNN provides a challenging real-world testbed, we acknowledge the need for broader evaluation. As such, we also intend to expand BeyondRPC to include domain-shift

experiments and real-world scene-level datasets such as SemanticKITTI and S3DIS.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Oddy Virgantara Putra and Hanugra Aulia Sidharta; methodology, Oddy Virgantara Putra and Diah Risqiwati; validation, Oddy Virgantara Putra and Moch. Iskandar Riansyah; software and resources, Hanugra Aulia Sidharta and Yuni Yamasari; investigation, Hanugra Aulia Sidharta, Diah Risqiwati, and Yuni Yamasari; writing-original draft preparation, Oddy Virgantara Putra; writing-review and editing, Oddy Virgantara Putra, Diah Risqiwati, and Moch. Iskandar Riansyah; visualization, Diah Risqiwati and Yuni Yamasari. All authors read and approved the final manuscript.

References

[1] X. Wang, K. Li, and A. Chehri, "Multi-Sensor Fusion Technology for 3D Object Detection in Autonomous Driving: A Review", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 25, No. 2, pp. 1148-1165, 2024, doi: 10.1109/TITS.2023.3317372.

[2] H. Chen, H. Yan, X. Yang, H. Su, S. Zhao, and F. Qian, "Efficient Adversarial Attack Strategy Against 3D Object Detection in Autonomous Driving Systems", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 25, No. 11, pp. 16118-16132, 2024, doi: 10.1109/TITS.2024.3410038.

[3] K. A. Szczurek, R. M. Prades, E. Matheson, J. Rodriguez-Nogueira, and M. Di Castro, "Multimodal Multi-User Mixed Reality Human-Robot Interface for Remote Operations in Hazardous Environments", *IEEE Access*, Vol. 11, pp. 17305-17333, 2023, doi: 10.1109/ACCESS.2023.3245833.

[4] J. Liang, H. He, and Y. Wu, "Bare-Hand Depth Perception Used in Augmented Reality Assembly Supporting", *IEEE Access*, Vol. 8, pp. 1534-1541, 2020, doi: 10.1109/ACCESS.2019.2962112.

[5] Y. Cheng, J. Su, M. Jiang, and Y. Liu, "A Novel Radar Point Cloud Generation Method for Robot Environment Perception", *IEEE Transactions on Robotics*, Vol. 38, No. 6, pp. 3754-3773, 2022, doi: 10.1109/TRO.2022.3185831.

[6] J. H. Park, Y. E. Lim, J. H. Choi, and M. J. Hwang, "Trajectory-Based 3D Point Cloud ROI Determination Methods for Autonomous Mobile Robot", *IEEE Access*, Vol. 11, pp. 8504-8522, 2023, doi: 10.1109/ACCESS.2023.3238824.

[7] G. Jiang, W. Y. Yan, and D. D. Lichti, "A Maximum Entropy-Based Optimal Neighbor Selection for Multispectral Airborne LiDAR Point Cloud Classification", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61, pp. 1-18, 2023, doi: 10.1109/TGRS.2023.3323963.

[8] 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.

[9] X. Zheng *et al.*, " Point Cloud Pre-Training with Diffusion Models ", in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22935-22945, 2024, doi: 10.1109/CVPR52733.2024.02164.

[10] O. V. Putra *et al.*, "Enhancing LiDAR-Based Object Recognition Through a Novel Denoising and Modified GDANet Framework", *IEEE Access*, Vol. 12, pp. 7285-7297, 2024, doi: 10.1109/ACCESS.2023.3347033.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite", In: *Proc. of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354-3361, 2012, doi: 10.1109/CVPR.2012.6248074.

[12] 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: 10.1109/CVPR.2015.7298801.

[13] J. Ren, L. Pan, and Z. Liu, "Benchmarking and Analyzing Point Cloud Classification under Corruptions", In: *Proc. of the 39th International Conference on Machine Learning (ICML)*, PMLR, pp. 18559-18575, 2022, [Online]. Available: https://proceedings.mlr.press/v162/ren22c.html

[14] 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: 10.1109/CVPR.2017.16.

[15] 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, pp. 5105-5114, 2017, doi: 10.5555/3295222.3295263.

[16] Y. Liu, B. Fan, S. Xiang, and C. Pan, " Relation-Shape Convolutional Neural Network for Point Cloud Analysis ", *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8887-8896, 2019, doi: 10.1109/CVPR.2019.00910.

[17] M. Xu, R. Ding, H. Zhao, and X. Qi, "PAConv: Position Adaptive Convolution with Dynamic Kernel Assembling on Point Clouds", In: *Proc. of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3172-3181, 2021, doi: 10.1109/CVPR46437.2021.00319.

[18] T. Xiang, C. Zhang, Y. Song, J. Yu, and W. Cai, "Walk in the Cloud: Learning Curves for Point Clouds Shape Analysis", 2021, [Online]. Available: https://curvenet.github.io/.

[19] J.-X. and L. Z.-N. and M. T.-J. and M. R. R. and H. S.-M. Guo Meng-Hao and Cai, "PCT: Point cloud transformer", *Comput Vis Media (Beijing)*, Vol. 7, No. 2, pp. 187-199, 2021, doi: 10.1007/s41095-021-0229-5.

[20] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point Transformer", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16239-16248, 2021, doi: 10.1109/ICCV48922.2021.01595.

[21] W. Zhou, Q. Wang, W. Jin, X. Shi, and Y. He, "Graph Transformer for 3D point clouds classification and semantic segmentation", *Comput Graph*, Vol. 124, p. 104050, 2024, doi: 10.1016/j.cag.2024.104050.

[22] V. T. and G. E. and M. T. and M. P. and Y. P. and S. C. G. M. Chen Yunlu and Hu, "PointMixup: Augmentation for Point Clouds", In: *Proc. of Computer Vision - ECCV 2020*, pp. 330-345, 2020.

[23] D. Lee *et al.*, "Regularization Strategy for Point Cloud via Rigidly Mixed Sample", In: *Proc. of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15895-15904, 2021, doi: 10.1109/CVPR46437.2021.01564.

[24] R. Li, X. Li, P.-A. Heng, and C.-W. Fu, "PointAugment: An Auto-Augmentation Framework for Point Cloud Classification", In: *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6377-6386, 2020, doi: 10.1109/CVPR42600.2020.00641.

[25] H. Wang, Q. Liu, X. Yue, J. Lasenby, and M. J. Kusner, "Unsupervised Point Cloud Pre-training via Occlusion Completion ", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9762-9772, 2021, doi: 10.1109/ICCV48922.2021.00964.

[26] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling", In: *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19291-19300, 2022, doi: 10.1109/CVPR52688.2022.01871.

[27] H. Liu, J. Jia, and N. Z. Gong, "PointGuard: Provably Robust 3D Point Cloud Classification", In: *Proc. of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6182-6191, 2021, doi: 10.1109/CVPR46437.2021.00612.

[28] J. Sun *et al.*, "Adversarially robust 3D point cloud recognition using self-supervisions", In: *Proc. of the 35th International Conference on Neural Information Processing Systems*, 2021.

[29] J. Sauder and B. Sievers, "Self-Supervised Deep Learning on Point Clouds by Reconstructing Space", *Advances in Neural Information Processing Systems*, 2019.

[30] J. Yu, C. Zhang, and W. Cai, "Rethinking Rotation Invariance with Point Cloud Registration", In: *Proc. of the AAAI Conference on Artificial Intelligence*, pp. 3313-3321, 2023, doi: 10.1609/aaai.v37i3.25438.

[31] 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, Ed., Cham: Springer International Publishing, pp. 574-591, 2020, doi: 10.1007/978-3-030-58580-8_34.

[32] 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.

[33] 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.

[34] O. V. Putra, K. Ogata, E. M. Yuniarno, and M. H. Purnomo, "AdaCrossNet: Adaptive Dynamic Loss Weighting for Cross-Modal Contrastive Point Cloud Learning", *International Journal of*

Intelligent Engineering and Systems, Vol. 18, No. 1, pp. 134-146, 2025, doi: 10.22266/ijies2025.0229.11.

[35] S. Kim, S. Lee, D. Hwang, J. Lee, S. J. Hwang, and H. J. Kim, "Point Cloud Augmentation with Weighted Local Transformations", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 528-537, 2021, doi: 10.1109/ICCV48922.2021.00059.

[36] M. Xu, J. Zhang, Z. Zhou, M. Xu, X. Qi, and Y. Qiao, "Learning Geometry-Disentangled Representation for Complementary Understanding of 3D Object Point Cloud", *AAAI*, 2021.

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

[38] W. Wang, Q. Xu, D. Ceylan, R. Mech, and U. Neumann, "DISN: deep implicit surface network for high-quality single-view 3D reconstruction", In: *Proc. of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", In: *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.

[40] D. Kinga, J. B. Adam, and others, "Adam: A method for stochastic optimization", In: *Proc. of International Conference on Learning Representations (ICLR)*, 2015.

[41] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts", In: *Proc. of International Conference on Learning Representations*, 2017.

[42] 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.

[43] S. Lee, M. Jeon, I. Kim, Y. Xiong, and H. J. Kim, "SageMix: Saliency-Guided Mixup for Point Clouds", *Advances in Neural Information Processing Systems*, 2022.

[44] T. Lu, C. Liu, Y. Chen, G. Wu, and L. Wang, "APP-Net: Auxiliary-Point-Based Push and Pull Operations for Efficient Point Cloud Recognition", *Trans. Img. Proc.*, Vol. 32, pp. 6500-6513, 2023, doi: 10.1109/TIP.2023.3333191.

[45] A. Rizwan, N. Iqbal, A. N. Khan, R. Ahmad, and D. H. Kim, "Toward Effective Pattern Recognition Based on Enhanced Weighted K-Mean Clustering Algorithm for Groundwater Resource Planning in Point Cloud", *IEEE Access*, Vol. 9, pp. 130154-130169, 2021, doi: 10.1109/ACCESS.2021.3111112.