REWEIGHTED CONTRASTIVE LEARNING FOR ROBUST POINT CLOUD REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive self-supervised learning has achieved promising performance in 3D point cloud representation learning, typically using the standard NT-Xent loss. However, NT-Xent assigns the strongest repulsive gradients to high-similarity negatives, including false negatives that share similar geometric semantics, which can disrupt intra-class compactness and increase sensitivity to sampling noise. To address this issue, we propose Reweighted Contrastive Loss for Robust Representations (ReCoR^2), derived from a maximum likelihood estimation perspective. ReCoR^2 treats anchor–candidate similarity as evidence of shared latent geometric semantics, applies weaker repulsion to higher-similarity candidates, and retains strong repulsion for low-similarity ones with larger geometric differences. Since similarity estimates can be unreliable early in training, we adopt a two-stage optimization strategy: NT-Xent is first used to establish a semantically structured representation space, followed by optimization with ReCoR^2. Theoretically, we establish an upper bound on the gradient norm of ReCoR^2 with respect to sampling noise, which is strictly lower than the corresponding lower bound for standard NT-Xent, demonstrating improved robustness to sampling perturbations. Integrated into a 3D–2D cross-modal pretraining framework, ReCoR^2 consistently improves performance over representative contrastive baselines on downstream classification and part segmentation tasks.

1 INTRODUCTION

3D point clouds provide rich geometric information for applications such as LiDAR-based perception in autonomous driving (Yang et al., 2024) and augmented reality (Wang et al., 2023b). However, supervised learning relies on costly annotations of large-scale 3D data. Contrastive self-supervised learning alleviates this dependence by constructing supervisory signals from augmented views of unlabeled data.

Contrastive learning commonly adopts the normalized temperature-scaled cross-entropy (NT-Xent) loss (Oord et al., 2018; Chen et al., 2020), which pulls positive pairs closer while pushing negative pairs apart. Consequently, false negatives that share similar geometric semantics are also repelled. Notably, our gradient analysis shows that the repulsive coefficient increases with anchor-negative similarity, causing high-similarity false negatives to receive stronger separation updates than dissimilar true negatives. Such excessive repulsion can weaken intra-class compactness and increase sensitivity to sampling noise. As illustrated in Figure 1(a), representative methods adopting the NT-Xent loss include CrossPoint (Afham et al., 2022) and CrossNet (Wu et al., 2023). In contrast, CCPoint (Xiao et al., 2025) adopts an ob-

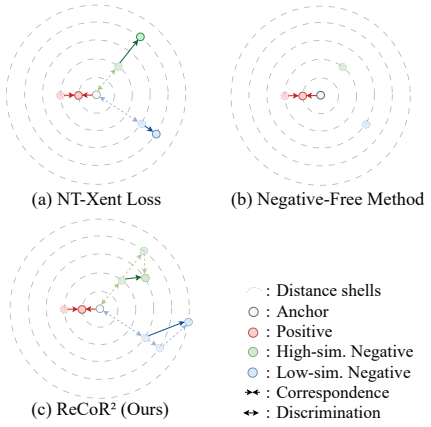


Figure 1: Faded and solid nodes show features before and after one update, respectively. (a) NT-Xent Loss repels similar negatives more strongly. (b) Removing negatives eliminates explicit separation. (c) Our ReCoR^2 reduces potential false-negative repulsion while preserving separation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

jective without explicit negatives, thereby avoiding false-negative repulsion but potentially losing the discriminative information provided by true negatives, as shown in Figure 1(b). These two paradigms reveal a fundamental trade-off: negative samples provide valuable discriminative signals, yet false negatives can introduce harmful repulsion. Our objective addresses this trade-off by suppressing the influence of potential false negatives while preserving strong separation for geometrically dissimilar negatives.

We therefore propose ReCoR^2, a reweighted contrastive loss motivated by a maximum likelihood estimation (MLE) perspective. Our key insight is that, in a semantically structured feature space, higher similarity provides stronger evidence that a candidate shares the latent geometric semantics of the anchor. Accordingly, unlike NT-Xent whose repulsive coefficient increases with similarity, ReCoR^2 assigns a smaller repulsive coefficient to higher-similarity candidates, suppressing potential false negatives while retaining strong repulsion for dissimilar true negatives. Since similarity estimates can be unreliable early in training, we adopt a two-stage optimization strategy that first uses NT-Xent to establish a semantically structured representation space and then applies ReCoR^2 based on the resulting similarities. Theoretically, we establish an upper bound on the sampling-noise gradient norm of ReCoR^2 that is strictly lower than the corresponding lower bound for standard NT-Xent, demonstrating improved robustness to sampling perturbations. Integrated into a 3D–2D cross-modal pretraining framework, ReCoR^2 consistently outperforms representative contrastive baselines on ModelNet40, ScanObjectNN, and ShapeNetPart across linear classification, few-shot classification, and part segmentation, demonstrating its effectiveness across diverse downstream tasks.

The main contributions are as follows.

- We propose ReCoR^2, an MLE-guided reweighted contrastive loss that reduces the repulsion of high-similarity potential false negatives while preserving separation for dissimilar true negatives.
- We theoretically establish an upper bound on the sampling-noise gradient norm of ReCoR^2 that is strictly lower than the corresponding lower bound for standard NT-Xent, providing theoretical support for its improved robustness to sampling perturbations.
- Extensive experiments on ModelNet40, ScanObjectNN, and ShapeNetPart demonstrate consistent improvements across linear classification, few-shot classification, and part segmentation.

2 Related Work

2.1 Self-Supervised Learning for Point Clouds

Point cloud self-supervised learning mainly follows three paradigms: geometric pretext tasks, masked modeling, and contrastive learning. Geometric pretext methods obtain supervision by reconstructing or transforming the input. FoldingNet reconstructs point clouds through a folding-based decoder (Yang et al., 2018), while Latent-GAN models an autoencoder feature space (Achlioptas et al., 2018). Spatial reconstruction (Sauder & Sievers, 2019), occlusion completion (Wang et al., 2021), and rotation prediction (Poursaeed et al., 2020) provide additional geometric supervision.

Masked modeling extends reconstruction-based learning to local point cloud regions (He et al., 2022; Devlin et al., 2019; Joshi et al., 2020; Liang et al., 2024; Chen et al., 2023; Wu et al., 2024). Point-BERT predicts discrete tokens for masked patches (Yu et al., 2022). Point-MAE reconstructs masked coordinates (Pang et al., 2023), and Point-M2AE learns hierarchical representations at multiple scales (Zhang et al., 2022). ReCon combines masked generative and cross-modal contrastive teachers through distillation (Qi et al., 2023). These reconstruction-oriented paradigms often require an additional decoder, tokenizer, or generative teacher, increasing architectural complexity and pretraining cost.

Contrastive learning directly optimizes agreement between samples and can reduce reliance on reconstruction decoders and discrete tokenizers, providing a comparatively simple and efficient pretraining objective (Ding et al., 2025; Cheng et al., 2024; Fei et al., 2025). PointContrast enforces feature consistency between corresponding points (Xie et al., 2020). STRL uses spatio-temporal prediction (Huang et al., 2021), and DepthContrast learns from depth scans (Chhipa et al., 2022). CrossPoint and CrossNet exploit correspondences between point clouds and rendered images (Afham

et al., 2022; Wu et al., 2023). ToThePoint contrasts features retained and discarded by pooling (Li et al., 2023). ECO-3D incorporates equivariant augmentation (Wang et al., 2023a), while GSPCon uses graph spectral perturbations (Han et al., 2023). CCPoint contrasts original, augmented, and structurally corrupted point clouds (Xiao et al., 2025). Our work focuses on contrastive learning from the perspective of gradient repulsion on false negatives.

2.2 The False Negative Problem in Contrastive Learning

False negatives arise when semantically related samples are treated as negatives. Chuang et al. (Chuang et al., 2020) address the resulting sampling bias through a class-prior correction. Robinson et al. (Robinson et al., 2020) emphasize hard negatives and provide an optional debiasing correction. Chen et al. (Chen et al., 2022) use clustering to detect and remove false negatives as the feature space becomes more informative during training. Du et al. (Du et al., 2021) mine self-similar patches within a point cloud as positive samples, reducing dependence on inter-instance negatives.

Our method uses similarity-dependent gradient reweighting to reduce the influence of potential false negatives. The two-stage strategy shares the motivation of incremental detection that correction should rely on an informative feature space (Chen et al., 2022). The proposed objective retains all candidates and adjusts their gradient coefficients continuously. The analysis relates this adjustment to a geometric margin between false and true negatives.

3 Method

We first analyze the negative-sample gradients of NT-Xent. A reweighted objective is then introduced and integrated into a two-stage cross-modal framework.

3.1 Motivation: The False Negative Problem

Let a batch contain N point cloud samples. Two augmented versions of each sample yield $2N$ samples, each serving once as an anchor. For an anchor z_i, $z_{j(i)}$ denotes its positive, while the other $2N-2$ samples are treated as negatives, resulting in $2N-1$ non-anchor candidates in total. We define $s_{i,k} = z_i^\top z_k$ as cosine similarity and use temperature $\tau > 0$. Averaging over all $2N$ anchors, NT-Xent is

$$\mathcal{L}_{\text{NT-Xent}} = -(1/(2N)) \sum_{i=1}^{2N} \log \left[\exp(s_{i,j(i)}/\tau) \Big/ \sum_{k \neq i} \exp(s_{i,k}/\tau) \right]. \tag{1}$$

For a negative candidate satisfying $k \neq i$ and $k \neq j(i)$, the per-anchor derivative is

$$\partial \mathcal{L}_{\text{NT-Xent},i} / \partial s_{i,k} = q_{i,k}^{\text{NT-Xent}} / \tau, \quad \text{where} \quad q_{i,k}^{\text{NT-Xent}} = \exp(s_{i,k}/\tau) \Big/ \sum_{m \neq i} \exp(s_{i,m}/\tau). \tag{2}$$

The coefficient $q_{i,k}^{\text{NT-Xent}}$ increases strictly with $s_{i,k}$ when other similarities are fixed. Gradient descent therefore applies a stronger separation update to a more similar negative. Sampling-noise-corrupted observations of the same surface can remain highly similar in a trained feature space. Treating such observations as negatives can weaken intra-class compactness and make the encoder more sensitive to sampling-noise-induced variations.

3.2 ReCoR^2: Reweighted Contrastive Loss

From a maximum likelihood estimation (MLE) perspective, similarity in an informative feature space provides evidence that two candidates share a latent geometric class. A high-similarity candidate should therefore receive less separation pressure than a geometrically distinct candidate. To expose the weighting implicit in standard NT-Xent, its denominator can be written as a uniform probability $1/(2N-1)$ scaled by the number of non-anchor candidates:

$$\mathcal{L}_{\text{NT-Xent}} = -(1/(2N)) \sum_{i=1}^{2N} \log \left[\exp(s_{i,j(i)}/\tau) \Big/ \left((2N-1) \cdot (1/(2N-1)) \sum_{k \neq i} \exp(s_{i,k}/\tau) \right) \right]. \tag{3}$$

Thus, every explicit denominator weight equals one:

$$w_{i,k}^{\text{NT-Xent}} = (2N-1)\cdot(1/(2N-1)) = 1, \qquad Z_i^{\text{NT-Xent}} = \sum_{k\neq i} w_{i,k}^{\text{NT-Xent}} \exp(s_{i,k}/\tau). \tag{4}$$

Although the exponential terms depend on similarity, the uniform NT-Xent weights cannot reflect geometric affinity. ReCoR^2 replaces the uniform candidate probability $1/(2N-1)$ with the similarity-dependent probability

$$p_{i,k}^{\text{ReCoR}^2} = \exp(-s_{i,k}/\tau) \Big/ \textstyle\sum_{m\neq i} \exp(-s_{i,m}/\tau), \qquad Z_i^{\text{ReCoR}^2} = \textstyle\sum_{k\neq i} w_{i,k}^{\text{ReCoR}^2} \exp(s_{i,k}/\tau). \tag{5}$$

Under the MLE interpretation, a high-similarity candidate is more likely to share the anchor's latent geometric class and should therefore receive less separation pressure. We accordingly assign lower weights to high-similarity candidates, which requires the weighting function to decrease with similarity. The score $f(s) = -s$ satisfies this requirement on $[-1, 1]$, and its temperature-scaled softmax makes $p_{i,k}^{\text{ReCoR}^2}$ decrease strictly with $s_{i,k}$ when the other similarities are fixed.

To preserve the total explicit weight, ReCoR^2 uses the same scale factor as NT-Xent and sets $w_{i,k}^{\text{ReCoR}^2} = (2N-1)p_{i,k}^{\text{ReCoR}^2}$. The weights therefore sum to $2N-1$. The normalization includes every non-anchor candidate, including the positive sample, and the weights remain differentiable functions of similarity.

Now, the complete per-anchor ReCoR^2 objective is

$$\mathcal{L}_{\text{ReCoR}^2,i} = -s_{i,j(i)}/\tau + \log \sum_{k\neq i} (2N-1) p_{i,k}^{\text{ReCoR}^2} \exp(s_{i,k}/\tau). \tag{6}$$

Using the identity $\exp(-s/\tau)\exp(s/\tau) = 1$, the weighted denominator simplifies, and the per-anchor objective can be rewritten as

$$\mathcal{L}_{\text{ReCoR}^2,i} = -s_{i,j(i)}/\tau - \log \sum_{m\neq i} \exp(-s_{i,m}/\tau) + 2\log(2N-1). \tag{7}$$

Since the final term is independent of the encoder parameters, removing this constant and averaging over anchors yields the optimization-equivalent objective

$$\widetilde{\mathcal{L}}_{\text{ReCoR}^2} = -(1/(2N)) \sum_{i=1}^{2N} \left[s_{i,j(i)}/\tau + \log \sum_{m\neq i} \exp(-s_{i,m}/\tau) \right]. \tag{8}$$

The positive term increases positive-pair similarity. For the negative term, when the feature space captures geometric affinity, low-similarity candidates are more likely to be true negatives and are therefore assigned larger coefficients, whereas high-similarity potential false negatives receive smaller coefficients. Appendix A provides the intermediate algebra and positive-pair derivative.

Gradient allocation. Direct comparison of the two derivatives for a negative candidate gives

$$\partial \mathcal{L}_{\text{NT-Xent},i}/\partial s_{i,k} = q_{i,k}^{\text{NT-Xent}}/\tau; \qquad \partial \mathcal{L}_{\text{ReCoR}^2,i}/\partial s_{i,k} = p_{i,k}^{\text{ReCoR}^2}/\tau. \tag{9}$$

Here $q_{i,k}^{\text{NT-Xent}} \propto \exp(s_{i,k}/\tau)$ increases with similarity, whereas $p_{i,k}^{\text{ReCoR}^2} \propto \exp(-s_{i,k}/\tau)$ decreases with similarity. Let $\mathbf{d}_{i,k}^{(t)} = -\nabla_\theta s_{i,k}$ denote the similarity-descent direction at parameters $\theta^{(t)}$, and let $\eta_t > 0$ be the learning rate. The per-anchor contributions of a fixed negative pair are

$$\Delta\theta_{\text{NT-Xent},i,k}^{(t)} = (\eta_t/\tau) q_{i,k}^{\text{NT-Xent}} \mathbf{d}_{i,k}^{(t)}, \qquad \Delta\theta_{\text{ReCoR}^2,i,k}^{(t)} = (\eta_t/\tau) p_{i,k}^{\text{ReCoR}^2} \mathbf{d}_{i,k}^{(t)}. \tag{10}$$

Both coefficients are positive, so both objectives preserve the repulsive direction for each negative pair. Reweighting changes only the repulsion strength, i.e., NT-Xent increases it with similarity, whereas ReCoR^2 assigns weaker repulsion to high-similarity candidates and stronger repulsion to low-similarity candidates.

Conditional sampling-noise gradient bounds. We next compare the sensitivity of the two objectives to sampling noise. The analysis introduces three geometric assumptions, derives a positive lower bound for NT-Xent and an upper bound for ReCoR^2, and then gives a deterministic condition under which the ReCoR^2 bound is smaller.

Let a fixed anchor have the candidate partition $\mathcal{F} \dot\cup \mathcal{T} \dot\cup \{j(i)\}$, where $\mathcal{F}$ and $\mathcal{T}$ denote the false- and true-negative sets. Define $\mathcal{H} = \mathcal{F} \cup \{j(i)\}$, $h = |\mathcal{H}|$, and $t = |\mathcal{T}| > 0$. We adopt the local perturbation and semantic-margin conditions from the supplementary analysis, stated as Assumptions 1–3 below.

Assumption 1 (Bounded Local Sampling Noise)**.** *For every $k \in \mathcal{F}$, $s_{i,k} \geq 1 - \delta$, where $\delta \in (0, 1)$ bounds the feature variation induced by local sampling noise.*

Assumption 2 (Global Semantic Margin)**.** *For every $m \in \mathcal{T}$, $s_{i,m} \leq \gamma$; true negatives therefore remain outside the high-similarity neighborhood of the anchor.*

Assumption 3 (Manifold Separability)**.** *The false- and true-negative regions have a positive similarity margin $\Delta = (1 - \delta) - \gamma > 0$.*

Assumption 1 states that samples sharing the anchor's underlying geometry remain in a high-similarity neighborhood despite local sampling noise. Assumption 2 requires geometrically distinct samples to remain below a global similarity threshold. Assumption 3 combines the two conditions into a positive margin Δ, ensuring that the false- and true-negative regions do not overlap in the feature space.

Lemma (Robustness Bound against Sampling Noise)**.** *Under Assumptions 1–3, let z_k denote the feature of a false-negative sample $k \in \mathcal{F}$, and let $z_{\mathcal{M}}$ denote its sampling-invariant representation on the underlying geometric manifold. We model sampling noise as $z_k = z_{\mathcal{M}} + \varepsilon_k$, where ε_k is the sampling-noise component. Holding the other features fixed and letting $\tau > 0$, we obtain*

$$\|\nabla_{\varepsilon_k} \mathcal{L}_{\text{NT-Xent},i}\|_2 = q_{i,k}^{\text{NT-Xent}}/\tau \geq L_{\text{NT-Xent}} = 1/\{\tau[h \exp(\delta/\tau) + t \exp(-\Delta/\tau)]\}, \tag{11}$$

$$\|\nabla_{\varepsilon_k} \mathcal{L}_{\text{ReCoR}^2,i}\|_2 = p_{i,k}^{\text{ReCoR}^2}/\tau \leq U_{\text{ReCoR}^2} = \exp(-\Delta/\tau)/\{\tau[h \exp((\gamma - 1)/\tau) + t]\}, \tag{12}$$

Where $L_{\text{NT-Xent}}$ denotes the lower bound on the NT-Xent sampling-noise gradient magnitude, whereas U_{ReCoR^2} denotes the upper bound on the corresponding ReCoR^2 magnitude.

Proof. Because $s_{i,k} = z_i^\top (z_{\mathcal{M}} + \varepsilon_k)$, we have $\nabla_{\varepsilon_k} s_{i,k} = z_i$. The chain rule and $\|z_i\|_2 = 1$ yield

$$\|\nabla_{\varepsilon_k} \mathcal{L}_i\|_2 = |\partial \mathcal{L}_i / \partial s_{i,k}| \, \|z_i\|_2 = |\partial \mathcal{L}_i / \partial s_{i,k}|. \tag{13}$$

Thus, bounding the similarity derivative is equivalent to bounding the sampling-noise gradient magnitude. For NT-Xent, write the denominator of $q_{i,k}^{\text{NT-Xent}}$ as the sum over $\mathcal{H}$ and $\mathcal{T}$. Assumption 1 lower-bounds its numerator by $\exp((1-\delta)/\tau)$. Since normalized cosine similarities are at most one and Assumption 2 gives $s_{i,m} \leq \gamma$ for $m \in \mathcal{T}$, the two denominator sums are upper-bounded by $h \exp(1/\tau)$ and $t \exp(\gamma/\tau)$, respectively. Hence

$$q_{i,k}^{\text{NT-Xent}}/\tau \geq \exp((1 - \delta)/\tau) \Big/ \{\tau[h \exp(1/\tau) + t \exp(\gamma/\tau)]\} = L_{\text{NT-Xent}}, \tag{14}$$

which proves the positive lower bound for NT-Xent. For ReCoR^2, Assumption 1 upper-bounds the numerator of $p_{i,k}^{\text{ReCoR}^2}$ by $\exp(-(1 - \delta)/\tau)$. The bounds $s_{i,m} \leq 1$ for $m \in \mathcal{H}$ and $s_{i,m} \leq \gamma$ from Assumption 2 for $m \in \mathcal{T}$ lower-bound the two denominator sums by $h \exp(-1/\tau)$ and $t \exp(-\gamma/\tau)$. Therefore,

$$p_{i,k}^{\text{ReCoR}^2}/\tau \leq \exp(-(1 - \delta)/\tau) \Big/ \{\tau[h \exp(-1/\tau) + t \exp(-\gamma/\tau)]\} = U_{\text{ReCoR}^2}, \tag{15}$$

which proves the ReCoR^2 upper bound in Eq. (12). Consequently,

$$U_{\text{ReCoR}^2} \leq \exp(-\Delta/\tau)/(\tau t) \longrightarrow 0 \quad (\tau \to 0), \tag{16}$$

while $L_{\text{NT-Xent}} > 0$ for every fixed $\tau > 0$. □

Theorem (Deterministic Gradient Dominance)**.** *Under Assumptions 1–3, if $1 - \gamma > 2\delta$ and $t \geq h$, then $U_{\text{ReCoR}^2} < L_{\text{NT-Xent}}$ for every $\tau > 0$.*

Proof. The condition $1 - \gamma > 2\delta$ is equivalent to $\Delta > \delta$, meaning that the false–true similarity margin exceeds the high-similarity tolerance. The condition $t \geq h$ means that the batch contains at least as many true negatives as high-similarity candidates, including the positive. Define $u = \exp(-1/\tau)$, $\alpha = \Delta - \delta > 0$, and $\beta = \delta > 0$. The inequality $U_{\text{ReCoR}^2} < L_{\text{NT-Xent}}$ is equivalent to $t/h > \Phi(\tau)$, where

$$\Phi(\tau) = u^\alpha (1 - u^{2\beta})/[1 - u^{2(\alpha+\beta)}] < 1. \tag{17}$$

The final inequality follows because the denominator minus the numerator equals $(1 - u^\alpha)(1 - u^{\alpha+2\beta}) > 0$; $t/h \geq 1$ then proves $U_{\text{ReCoR}^2} < L_{\text{NT-Xent}}$. □

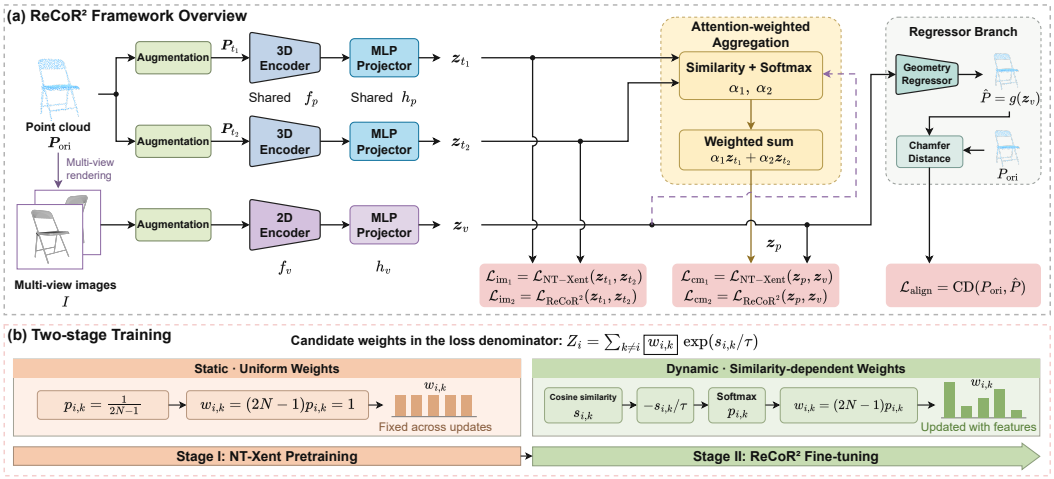


Figure 2: Overview of the proposed ReCoR^2 framework. (a) Multi-view images are rendered from the input point cloud. Parallel point-cloud and image branches optimize intra-modal, cross-modal, and geometric alignment objectives, with attention-weighted aggregation connecting the two modalities. (b) The two-stage schedule uses static uniform weights with NT-Xent during Stage I and adopts dynamic, similarity-dependent weights with ReCoR^2 during Stage II.

3.3 Training Framework and Two-Stage Strategy

Figure 2 summarizes the framework, which jointly processes point clouds and rendered images. Rotation, jitter, and point dropout first produce two point-cloud versions P_{t_1} and P_{t_2}, while a rendered image I of the same object receives photometric augmentation. The point encoder f_p and image encoder f_v then extract features, and MLP heads h_p and h_v project the features into a shared space.

$$z_{t_j} = h_p(f_p(P_{t_j})), \quad j \in \{1, 2\}, \qquad z_v = h_v(f_v(I)). \tag{18}$$

The two point cloud versions can differ in their geometric alignment with the image. Selecting only one view can therefore introduce view-selection bias. We aggregate the point embeddings using attention based on their cosine similarities to the image embedding:

$$[\alpha_1, \alpha_2] = \mathrm{softmax}([z_{t_1}^\top z_v, z_{t_2}^\top z_v]), \qquad z_p = \alpha_1 z_{t_1} + \alpha_2 z_{t_2}. \tag{19}$$

Thus the weights are nonnegative and sum to one, assigning greater weight to the better-aligned point cloud view and thereby enabling adaptive cross-modal representation. In addition, a geometry regressor g maps the image embedding to a point set, whose geometric alignment guides the image features toward the underlying point-cloud geometry.

Stage I: Feature pretraining. Following CrossPoint (Afham et al., 2022) and CCPoint (Xiao et al., 2025), stage I uses NT-Xent to establish an informative feature space before reweighting:

$$\mathcal{L}_{\mathrm{S1}} = \mathcal{L}_{\mathrm{im}_1} + \mathcal{L}_{\mathrm{cm}_1} + \lambda \mathcal{L}_{\mathrm{align}}. \tag{20}$$

Here $\mathcal{L}_{\mathrm{S1}}$ denotes the total Stage I objective, whose three terms provide intra-modal contrast $\mathcal{L}_{\mathrm{im}_1}$, cross-modal contrast $\mathcal{L}_{\mathrm{cm}_1}$, and geometric alignment $\mathcal{L}_{\mathrm{align}}$, respectively. Specifically, $\mathcal{L}_{\mathrm{im}_1} = \mathcal{L}_{\text{NT-Xent}}(z_{t_1}, z_{t_2})$ encourages view invariance, while $\mathcal{L}_{\mathrm{cm}_1} = \mathcal{L}_{\text{NT-Xent}}(z_p, z_v)$ aligns point and image representations. We use the Chamfer loss $\mathcal{L}_{\mathrm{align}} = \mathrm{CD}(P_{\mathrm{ori}}, g(z_v))$ (Fan et al., 2017) for geometric alignment and set $\lambda = 0.1$.

Stage II: Reweighted fine-tuning. After 300 pretraining epochs, the learned feature space provides informative similarities for candidate-pair reweighting. Stage II therefore resumes from the Stage I checkpoint with ReCoR^2:

$$\mathcal{L}_{\mathrm{S2}} = \mathcal{L}_{\mathrm{im}_2} + \mathcal{L}_{\mathrm{cm}_2} + \lambda \mathcal{L}_{\mathrm{align}}. \tag{21}$$

Here $\mathcal{L}_{\mathrm{S2}}$ denotes the total Stage II objective. It retains the three-part structure of $\mathcal{L}_{\mathrm{S1}}$ while replacing both contrastive terms with ReCoR^2: $\mathcal{L}_{\mathrm{im}_2} = \mathcal{L}_{\text{ReCoR}^2}(z_{t_1}, z_{t_2})$ and $\mathcal{L}_{\mathrm{cm}_2} = \mathcal{L}_{\text{ReCoR}^2}(z_p, z_v)$ reweight candidate pairs by similarity. The alignment term and λ remain unchanged. Both stages use end-to-end optimization with Adam (Kingma & Ba, 2014) and cosine annealing.

4 EXPERIMENTS

We evaluate the learned representations through linear classification, few-shot classification, and part segmentation. The classification experiments cover both synthetic objects and real-world scans.

4.1 EXPERIMENTAL SETUP

Implementation. The 3D branch uses DGCNN (Wang et al., 2019) or PointNet (Qi et al., 2017a), and the image branch uses ResNet-50 (He et al., 2016). MLP heads project the encoder outputs into the shared space, and the geometry regressor g predicts a point set from the image embedding for the alignment objective. Both backbone configurations are pretrained on the ShapeNet rendering dataset, which contains 55 object categories (Chang et al., 2015). Following CrossPoint (Afham et al., 2022), each CAD model provides a surface-sampled point cloud and images rendered from evenly distributed viewpoints. These paired inputs provide the cross-modal correspondence used during pretraining.

Table 1: Linear classification accuracy (%) on ModelNet40 and ScanObjectNN. PN and DG denote PointNet and DGCNN. Bold and underlined values indicate the best and second-best results in each column. Appendix B provides the full comparisons.

	ModelNet40		ScanObjectNN	
Method	PN	DG	PN	DG
Jigsaw (2019)	87.3	90.6	55.2	59.5
STRL (2021)	88.3	90.9	74.2	77.9
OcCo (2021)	88.7	89.2	69.5	78.3
CrossPoint (2022)	89.1	91.2	75.6	81.7
ToThePoint (2023)	85.6	89.2	74.7	81.9
CrossNet (2023)	89.5	91.5	<u>76.8</u>	83.9
GSPCon (2023)	<u>89.8</u>	91.5	75.8	83.9
CCPoint (2025)	**90.2**	<u>92.4</u>	–	<u>86.2</u>
ReCoR^2 **(Ours)**	**90.2**	**92.6**	**77.0**	**87.8**

Stage I uses NT-Xent for 300 epochs with $\tau = 0.1$. Stage II continues for 100 epochs with $\mathcal{L}_{\text{ReCoR}^2}$, keeping the other settings unchanged. Both stages use Adam with a learning rate of 0.001, weight decay of 5×10^{-6}, and cosine annealing. Each batch contains 16 objects, and each point cloud contains 2048 sampled points. Point cloud augmentations include random rotation, jitter, and point dropout. Image augmentations include random resizing, color jitter, and horizontal flipping.

Evaluation and baselines. The pretrained encoder is frozen during downstream evaluation, and a task-specific head is trained on labeled data (Wu et al., 2023; Xiao et al., 2025). Linear classification uses a linear SVM. Few-shot evaluation covers 5-way and 10-way tasks with 10-shot and 20-shot support sets, and results are averaged over 10 runs. We compare with CrossPoint (Afham et al., 2022), CrossNet (Wu et al., 2023), GSPCon (Han et al., 2023), and CCPoint (Xiao et al., 2025). The tables also include additional methods for reference.

4.2 LINEAR CLASSIFICATION

ModelNet40. Table 1 reports linear classification accuracy on ModelNet40, a synthetic benchmark with 40 object categories (Wu et al., 2015). The proposed method achieves 90.2% with PointNet, matching CCPoint and exceeding CrossPoint by 1.1 percentage points. With DGCNN, the method achieves 92.6%, exceeding CCPoint by 0.2 percentage points and CrossPoint by 1.4 percentage points. The improvements across both backbones suggest that our proposed ReCoR^2 can benefit different encoder architectures.

ScanObjectNN. The PB_T50_RS split of ScanObjectNN contains real-world scans with background clutter and occlusion (Uy et al., 2019). The proposed method achieves 77.0% with PointNet and 87.8% with DGCNN (Table 1). The improvements over the strongest reported contrastive baselines are 0.2 and 1.6 percentage points, respectively. The larger improvement with DGCNN than on ModelNet40 is consistent with the motivation to reduce sensitivity to geometric variation.

4.3 FEW-SHOT CLASSIFICATION

Table 2 reports the mean accuracy and standard deviation for each few-shot setting, restricted to self-supervised baselines. Few-shot classification evaluates whether the pretrained representation

supports recognition from a limited number of labeled examples. The evaluation uses the same pretrained features across the different support-set sizes.

On ModelNet40, the proposed method achieves 96.1% in the 5-way 10-shot setting and 97.9% in the 5-way 20-shot setting. The corresponding improvements over CCPoint are 0.5 and 0.1 percentage points. In the 10-way settings, the method achieves 90.1% with 10 shots and 93.6% with 20 shots. Both values exceed CCPoint by 0.2 percentage points.

On ScanObjectNN, the proposed method achieves 86.1% and 89.6% in the two 5-way settings. The 10-way results are 72.1% and 80.0%. Compared with CCPoint, the improvements range from 0.8 to 0.9 percentage points across the four settings. The mean accuracies indicate a consistent improvement on the real-world benchmark.

Table 2: Few-shot classification among self-supervised methods on ModelNet40 and ScanObjectNN. Values are mean accuracy (%) and standard deviation over 10 runs. Bold means indicate the best results.

Dataset	Setting	Jigsaw	OcCo	CrossPoint	GSPCon	CCPoint	ReCoR^2 **(Ours)**
ModelNet40	5-way 10-shot	34.3±1.3	90.6±2.8	92.5±3.0	93.6±2.9	95.6±2.6	**96.1**±2.1
	5-way 20-shot	42.2±3.5	92.5±1.9	94.9±2.1	95.3±2.5	97.8±1.3	**97.9**±1.6
	10-way 10-shot	26.0±2.4	82.9±1.3	83.6±5.3	86.1±4.7	89.9±1.8	**90.1**±3.4
	10-way 20-shot	29.9±2.6	86.5±2.2	87.9±4.2	89.0±3.7	93.4±1.8	**93.6**±3.5
ScanObjectNN	5-way 10-shot	65.2±3.8	72.4±1.4	74.8±1.5	78.1±3.9	85.2±5.4	**86.1**±4.1
	5-way 20-shot	72.2±2.7	77.2±1.4	79.0±1.2	81.5±3.5	88.8±3.3	**89.6**±3.5
	10-way 10-shot	45.6±3.1	57.0±1.3	62.9±1.7	64.8±3.8	71.2±4.4	**72.1**±4.0
	10-way 20-shot	48.2±2.8	61.6±1.2	73.9±2.2	75.3±2.1	79.1±5.2	**80.0**±3.0

4.4 Part Segmentation

Table 3 reports the part segmentation results on ShapeNetPart (Yi et al., 2016). The proposed method achieves an overall accuracy of 94.5% and a mean intersection over union of 85.7%. The overall accuracy matches the highest value in the comparison, and the mean IoU matches GSPCon (Han et al., 2023).

Part segmentation requires discrimination among local geometric regions, whereas the contrastive objective primarily targets global representations. The competitive segmentation results indicate that the learned features preserve useful local geometric information, complementing the classification experiments from the perspective of local-level discrimination. Overall, our method improves classification performance while maintaining part segmentation performance comparable to the strongest methods in the table.

Table 3: ShapeNetPart segmentation (OA and mIoU, %). DG denotes DGCNN. Bold and underlined values indicate the best and second-best results in each column.

Method	OA	mIoU
Supervised		
PointNet (2017a)	–	83.7
PointNet++ (2017b)	–	85.1
DG (2019)	–	85.1
Self-supervised pretraining		
DG + OcCo (2021)	<u>94.4</u>	85.0
DG + CrossPoint (2022)	<u>94.4</u>	85.3
DG + ToThePoint (2023)	<u>94.4</u>	85.5
DG + CrossNet (2023)	–	85.5
DG + GSPCon (2023)	–	**85.7**
DG + CCPoint (2025)	**94.5**	<u>85.6</u>
DG + ReCoR^2 **(Ours)**	**94.5**	**85.7**

4.5 Empirical Validation of the Similarity Prior

The Stage II weighting strategy assumes that higher feature similarity provides stronger evidence that two samples share a semantic class. We evaluate this prior using the checkpoint obtained after Stage I, before reweighted optimization begins. Figure 3 presents two representative views selected from a broader set of diagnostic plots.

Figure 3 (a) compares the similarity densities of same- and different-class pairs. The same-class distribution is shifted toward higher similarity and has a higher peak. Figure 3 (b) estimates the conditional probability of class agreement within each similarity interval, which increases mono-

tonically with similarity. Together, the two observations support similarity as a practical signal for assigning candidate weights in Stage II.

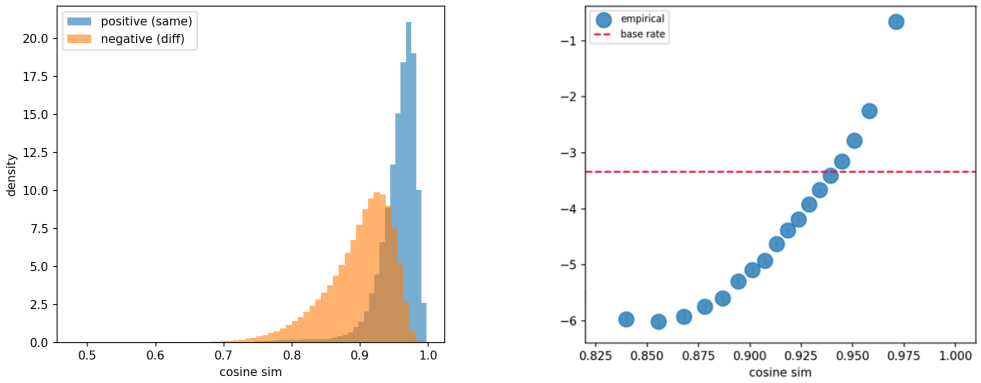


Figure 3: Validation of the similarity prior using Stage I representations. (a) Similarity densities for same- and different-class pairs. (b) Empirical conditional probability of class agreement, shown on a logit scale.

4.6 Ablation Study

The ablation uses DGCNN for ModelNet40 linear classification with matched architectures and training budgets. For compactness, $\mathcal{L}_{\mathrm{cm}_{1,2}}$ denotes the cross-modal terms $\{\mathcal{L}_{\mathrm{cm}_1}, \mathcal{L}_{\mathrm{cm}_2}\}$ across the two stages, and $\mathcal{L}_{\mathrm{im}_{1,2}}$ analogously denotes the intra-modal terms. Removing $\mathcal{L}_{\mathrm{align}}$, $\mathcal{L}_{\mathrm{cm}_{1,2}}$, or $\mathcal{L}_{\mathrm{im}_{1,2}}$ reduces accuracy from 92.6% to 92.1%, 91.8%, or 91.4%, respectively, showing complementary supervision (Table 4). Applying reweighting from epoch zero yields 91.1%, supporting its delayed use after the feature space becomes informative.

Table 4: Ablation on ModelNet40 with DGCNN (%). Each row removes one component from both stages or removes the two-stage schedule. Bold and underlined values indicate the best and second-best accuracies.

Variant	$\mathcal{L}_{\mathrm{im}_{1,2}}$	$\mathcal{L}_{\mathrm{cm}_{1,2}}$	$\mathcal{L}_{\mathrm{align}}$	Acc. (%)
w/o $\mathcal{L}_{\mathrm{align}}$	✓	✓	×	<u>92.1</u>
w/o $\mathcal{L}_{\mathrm{cm}_{1,2}}$	✓	×	✓	91.8
w/o $\mathcal{L}_{\mathrm{im}_{1,2}}$	×	✓	✓	91.4
w/o two-stage	✓	✓	✓	91.1
ReCoR^2 **(Ours)**	✓	✓	✓	**92.6**

Table 5: Weighting-function comparison on ModelNet40 with DGCNN (%). Bold and underlined values indicate the best and second-best accuracies.

Function	Acc. (%)
$-x^2$	<u>92.3</u>
$-x^3$	92.1
$-\sin x$	<u>92.3</u>
$\ln(2-x)$	92.2
e^{-x}	91.9
$-x$ **(Ours)**	**92.6**

Weighting function. Under the same configuration and ℓ_p normalization, alternative scores obtain 91.9–92.3%, whereas the monotone-decreasing linear score $-x$ reaches 92.6% (Table 5). The narrow range indicates robustness to the weighting form, with $-x$ performing best.

5 Conclusion

We present ReCoR^2, a similarity-aware reweighted contrastive objective for robust point cloud representation learning. By assigning weaker repulsion to high-similarity potential false negatives while preserving strong separation for dissimilar true negatives, ReCoR^2 achieves a better balance between mitigating false-negative repulsion and retaining discriminative information. A two-stage optimization strategy further enables reliable similarity-based reweighting after establishing a semantically structured representation space. Theoretically, our gradient analysis shows that ReCoR^2 is less sensitive to sampling perturbations than standard NT-Xent. Extensive experiments on ModelNet40, ScanObjectNN, and ShapeNetPart demonstrate consistent improvements across classification and part segmentation tasks.

AI USE STATEMENT

Generative AI tools were used to improve the readability and linguistic presentation of the manuscript and to assist in searching for and identifying potentially relevant literature. All retrieved references were independently reviewed and verified by the authors using the original sources.

REFERENCES

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pp. 40–49. PMLR, 2018.

Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9902–9912, 2022.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems*, 36:29667–29679, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dDjSKKA5TP1.

Haozhe Cheng, Xu Han, Pengcheng Shi, Jihua Zhu, and Zhongyu Li. Multi-trusted cross-modal information bottleneck for 3d self-supervised representation learning. *Knowledge-Based Systems*, 283:111217, 2024.

Prakash Chandra Chhipa, Richa Upadhyay, Rajkumar Saini, Lars Lindqvist, Richard Nordenskjold, Seiichi Uchida, and Marcus Liwicki. Depth contrast: Self-supervised pretraining on 3dpm images for mining material classification. In *European Conference on Computer Vision*, pp. 212–227. Springer, 2022.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Weichao Ding, Zehao Yang, Fei Luo, Chunhua Gu, and Wenbo Dong. Multi-modal self-supervised contrastive representation learning for three-dimensional point cloud understanding. *Engineering Applications of Artificial Intelligence*, 160:111919, 2025.

Bi'an Du, Xiang Gao, Wei Hu, and Xin Li. Self-contrastive learning with hard negative sampling for self-supervised point cloud learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3133–3142, 2021.

Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 605–613, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Fan_A_Point_Set_CVPR_2017_paper.html.

Ben Fei, Liwen Liu, Tianyue Luo, Weidong Yang, Lipeng Ma, Zhijun Li, and Wen-Ming Chen. Point patches contrastive learning for enhanced point cloud completion. *IEEE Transactions on Multimedia*, 27:581–596, 2025.

Yuehui Han, Jiaxin Chen, Jianjun Qian, and Jin Xie. Graph spectral perturbation for 3d point cloud contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5389–5398, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6535–6545, 2021.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Xinglin Li, Jiajing Chen, Jinhui Ouyang, Hanhui Deng, Senem Velipasalar, and Di Wu. Tothepoint: Efficient contrastive learning of 3d point clouds via recycling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21781–21790, 2023.

Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *Advances in neural information processing systems*, 37:32653–32677, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Yatian Pang, Eng Hock Francis Tay, Li Yuan, and Zhenghua Chen. Masked autoencoders for 3d point cloud self-supervised learning. *World Scientific Annual Review of Artificial Intelligence*, 1: 2440001, 2023.

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

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

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, pp. 28223–28243. PMLR, 2023.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in neural information processing systems*, 32, 2019.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1588–1597, 2019.

Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9782–9792, 2021.

Ruibin Wang, Xianghua Ying, Bowei Xing, and Jinfa Yang. Eco-3d: equivariant contrastive learning for pre-training on perturbed 3d point cloud. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2626–2634, 2023a.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12, 2019.

Zeyu Wang, Cuong Nguyen, Paul Asente, and Julie Dorsey. Pointshopar: Supporting environmental design prototyping using point cloud in augmented reality. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–15, 2023b.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4840–4851, 2024.

Yue Wu, Jiaming Liu, Maoguo Gong, Peiran Gong, Xiaolong Fan, A Kai Qin, Qiguang Miao, and Wenping Ma. Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding. *IEEE Transactions on Multimedia*, 26:1626–1638, 2023.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Xiaoyang Xiao, Shaoyi Du, Zhiqiang Tian, Meiqin Liu, and Xinhu Zheng. Ccpoint: Contrasting corrupted point clouds for self-supervised representation learning. *IEEE Transactions on Multimedia*, 2025.

Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, pp. 574–591. Springer, 2020.

Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 206–215, 2018.

Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14673–14684, 2024.

Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19313–19322, 2022.

Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35:27061–27074, 2022.

A Additional Algebra for the Reweighted Objective

The derivations below provide the intermediate algebra for the reweighted objective in Section 3.2. Section 3.3 describes how the objective is used in the training framework and two-stage schedule.

A.1 Simplification of the Weighted Denominator

Define $M = 2N - 1$ and $S_i^- = \sum_{m \neq i} \exp(-s_{i,m}/\tau)$. Equation (5) gives $p_{i,k}^{\text{ReCoR}^2} = \exp(-s_{i,k}/\tau)/S_i^-$. Each candidate contributes the same value after multiplication by its positive-similarity exponential:

$$M p_{i,k}^{\text{ReCoR}^2} \exp(s_{i,k}/\tau) = M/S_i^-. \tag{22}$$

Summing over the M non-anchor candidates gives

$$Z_i = \sum_{k \neq i} M p_{i,k}^{\text{ReCoR}^2} \exp(s_{i,k}/\tau) = M^2/S_i^-. \tag{23}$$

Taking the logarithm yields $\log Z_i = 2 \log M - \log S_i^-$. Substitution into Eq. (6) gives Eq. (7). The removed constant depends on batch size but has zero derivative with respect to the encoder parameters.

A.2 Negative- and Positive-Pair Derivatives

For a negative index satisfying $k \neq i$ and $k \neq j(i)$, only S_i^- depends on $s_{i,k}$ in the nonconstant part of Eq. (7). Differentiation gives

$$\partial \mathcal{L}_{\text{ReCoR}^2,i}/\partial s_{i,k} = -(1/S_i^-)[-\exp(-s_{i,k}/\tau)/\tau] = p_{i,k}^{\text{ReCoR}^2}/\tau. \tag{24}$$

For the positive index, the explicit attraction term also contributes:

$$\partial \mathcal{L}_{\text{ReCoR}^2,i}/\partial s_{i,j(i)} = -1/\tau + p_{i,j(i)}^{\text{ReCoR}^2}/\tau = -(1 - p_{i,j(i)}^{\text{ReCoR}^2})/\tau. \tag{25}$$

The positive derivative is nonpositive, so the objective encourages positive-pair attraction. Equations (24) and (25) require differentiation through the similarity-dependent weights. Differentiating through the weights yields the gradient allocation in Eq. (9).

A.3 Aggregation in Parameter Space

The chain rule combines the similarity derivatives with the encoder derivatives. For a fixed anchor,

$$\nabla_\theta \mathcal{L}_i = \sum_{k \neq i} (\partial \mathcal{L}_i/\partial s_{i,k}) \nabla_\theta s_{i,k}. \tag{26}$$

The batch update under the averaging convention of Eq. (1) is

$$\theta^{(t+1)} = \theta^{(t)} - (\eta_t/(2N)) \sum_{i=1}^{2N} \nabla_\theta \mathcal{L}_i. \tag{27}$$

Equation (10) gives the contribution of a negative pair, and Eq. (27) aggregates the pairwise contributions into the batch update.

B Supplementary Experiments

B.1 Classification Comparisons

Table 1 compares methods reported on both classification benchmarks. The tables below provide the complete ModelNet40 and ScanObjectNN comparisons, including the additional reconstruction-based methods and rotation-prediction baseline. The additional comparisons supplement the principal results in Section 4.2. The main text describes the evaluation protocol and discusses the principal findings.

Table 6: Linear classification on ModelNet40. The table reports accuracy (%). Bold and underlined values indicate the best and second-best results within each backbone group.

Method	Accuracy (%)
SO-Net (CVPR 2018)	87.3
FoldingNet (CVPR 2018)	88.4
DepthContrast (ECCV 2022)	85.4
ClusterNet (CVPR 2019)	86.8
PointNet-based methods:	
PointNet + Jigsaw (NeurIPS 2019)	87.3
PointNet + STRL (ICCV 2021)	88.3
PointNet + Rotation (3DV 2020)	88.6
PointNet + OcCo (ICCV 2021)	88.7
PointNet + CrossPoint (CVPR 2022)	89.1
PointNet + ToThePoint (CVPR 2023)	85.6
PointNet + CrossNet (TMM 2023)	89.5
PointNet + GSPCon (ACM MM 2023)	<u>89.8</u>
PointNet + CCPoint (TMM 2025)	**90.2**
PointNet + ReCoR^2**(Ours)**	**90.2**
DGCNN-based methods:	
DGCNN + Jigsaw (NeurIPS 2019)	90.6
DGCNN + STRL (ICCV 2021)	90.9
DGCNN + Rotation (3DV 2020)	90.8
DGCNN + OcCo (ICCV 2021)	89.2
DGCNN + CrossPoint (CVPR 2022)	91.2
DGCNN + ToThePoint (CVPR 2023)	89.2
DGCNN + CrossNet (TMM 2023)	91.5
DGCNN + GSPCon (ACM MM 2023)	91.5
DGCNN + CCPoint (TMM 2025)	<u>92.4</u>
DGCNN + ReCoR^2**(Ours)**	**92.6**

Table 7: Linear classification on the PB_T50_RS split of ScanObjectNN. The table reports overall accuracy (%) for PointNet and DGCNN.

Method	Backbone	
	PointNet	DGCNN
Jigsaw (NeurIPS 2019)	55.2	59.5
OcCo (ICCV 2021)	69.5	78.3
STRL (ICCV 2021)	74.2	77.9
CrossPoint (CVPR 2022)	75.6	81.7
TothePoint (CVPR 2023)	74.7	81.9
CrossNet (TMM 2023)	76.8	83.9
GSPCon (ACM MM 2023)	75.8	83.9
CCPoint (TMM 2025)	-	86.2
ReCoR^2**(Ours)**	**77.0**	**87.8**

B.2 Controlled False-Negative Ratios

We manually inject false negatives at fixed ratios and compare NT-Xent with the reweighted objective. Table 8 shows that the reweighted objective maintains accuracies between 92.46% and 92.50%, while the NT-Xent accuracy decreases as the false-negative ratio increases. The improvement is non-decreasing across the evaluated ratios and reaches 0.44 percentage points at a ratio of 0.5. The trend indicates that reweighting becomes more beneficial as false negatives constitute a larger fraction of the negative set.

Table 8: Classification accuracy (%) under controlled false-negative ratios. Improvement is the accuracy difference between the reweighted objective and NT-Xent in percentage points.

False-negative ratio	NT-Xent	Reweighted	Improvement
0.1	92.18	92.46	0.28
0.2	92.14	92.50	0.36
0.3	92.14	92.50	0.36
0.5	92.02	92.46	0.44

B.3 Analysis of Two-Stage Training Dynamics

Figure 4 tracks training loss and ModelNet40 linear SVM accuracy. During Stage I, NT-Xent loss falls while accuracy rises, indicating increasingly useful features. Switching to reweighting at epoch 300 briefly raises the fine-tuning loss; accuracy then exceeds the baseline, supporting Stage I initialization for reweighted Stage II.

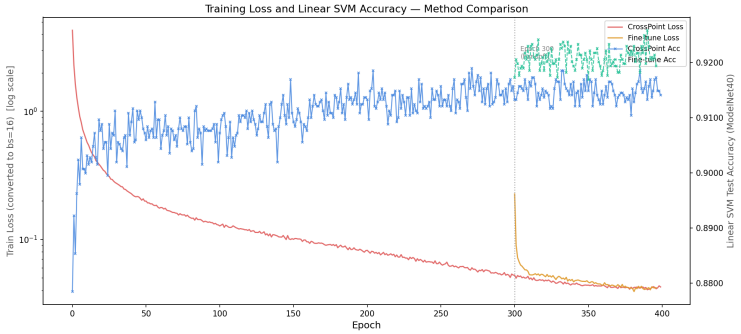


Figure 4: Training loss and ModelNet40 linear SVM accuracy. The dashed line marks the switch to reweighting at epoch 300; loss is shown logarithmically.