POINT-S2AE: STRUCTURAL AND SEMANTIC MASKED AUTOENCODERS FOR POINT CLOUD PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked autoencoders have been widely explored in point cloud self-supervised learning, whereby the point cloud is generally divided into visible and masked patches for reconstruction. These methods typically reconstruct masked patches independently based on visible representations, driving the encoder to learn local geometric structures. However, they treat the reconstruction of each masked patch as an isolated task, neglecting the inherent structural dependencies between adjacent patches. Furthermore, their strict reliance on local geometric reconstruction often fails to capture the holistic semantic consistency of 3D objects, limiting the robustness of the learned representations. To overcome these limitations, we propose a novel pre-training framework that augments the standard Masked Point Modeling (MPM) paradigm with two simple yet effective components: Latent Autoregressive Alignment (LAA) and Global Self-Distillation (GSD). Specifically, LAA explicitly models the spatial transitions between masked patches by enforcing an autoregressive predictive relationship between shallow and deep latent representations in the decoder, utilizing an InfoNCE contrastive loss to prevent representation collapse. Concurrently, GSD extracts consistent global semantics by applying a bidirectional self-distillation loss on the class tokens from two distinct masked views of the same point cloud in the encoder. Extensive experiments on standard benchmarks demonstrate that Point-S2AE significantly enhances the learned representations while maintaining high pre-training efficiency. Notably, Point-S2AE achieves great improvement over the Point-MAE baseline, particularly surpassing it by 5.16% on OBJ_BG, 5.34% on OBJ_ONLY, and 5.35% on PB_T50_RS for 3D object classification on the ScanObjectNN dataset, and establishing new state-of-the-art results on ModelNet40 few-shot learning.

1 INTRODUCTION

Recently, Masked Autoencoders (MAE) (19), which mask random patches of an input and reconstruct the missing content, have demonstrated remarkable scalability and superior performance in 2D vision. Inspired by this, Masked Point Modeling (MPM) has become the dominant SSL paradigm in the 3D domain. Methods such as Point-BERT (52) and Point-MAE (33) divide a point cloud into local patches, mask a high proportion of them, and train an asymmetric encoder-decoder architecture to reconstruct the coordinates of the masked patches from the visible ones. By doing so, the network is forced to understand local geometric structures and spatial relationships. The simplicity and effectiveness of this mask-and-reconstruct paradigm have led to numerous extensions, including hierarchical architectures (54) and cross-modal distillation frameworks (12; 36).

Despite these impressive advancements, current MPM architectures still suffer from two fundamental limitations that hinder their ability to learn optimal representations.

First, standard decoders typically treat the reconstruction of each masked patch as an isolated task. While the self-attention mechanism allows tokens to interact, there is no explicit objective enforcing the structural and sequential dependencies between patches in the intermediate latent spaces. Point clouds represent continuous 3D surfaces where adjacent patches exhibit smooth geometric transitions. When the network is only penalized based on the final output coordinates, it tends to memorize local high-frequency patterns without truly understanding the underlying spatial flow and

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.

structural dependencies. This lack of constraint leads to suboptimal representations, as the network fails to capture the intrinsic topological constraints of the 3D shape.

Second, the primary objective of MPM is strictly focused on low-level geometric reconstruction. This local focus often neglects the holistic semantic consistency of the 3D object. For instance, a model might perfectly reconstruct the local geometry of a flat surface without knowing whether it belongs to a table top or a car roof. A robust semantic representation should not only capture local geometric details but also maintain a stable global understanding of the object, regardless of which specific patches are masked or visible during the random masking process. Existing methods either ignore this issue or rely on heavy cross-modal teacher networks, such as pre-trained 2D image models or complex multi-modal alignment pipelines, to inject semantic knowledge. While cross-modal distillation can provide rich semantics, it introduces significant computational overhead, requires paired multi-modal datasets, and limits the applicability of the pre-trained models in pure 3D environments. Therefore, exploring a purely single-modal approach to extract consistent global semantics remains a highly desirable yet challenging goal.

To address these critical limitations, we propose a novel self-supervised pre-training framework that augments the standard MPM paradigm with explicit latent structural regularization and global semantic alignment. Our framework introduces two core components: Latent Autoregressive Alignment (LAA) and Global Self-Distillation (GSD).

To explicitly capture inter-patch dependencies, the LAA module operates within the decoder. It enforces an autoregressive predictive relationship where the deep latent representation of a masked patch is tasked with predicting the shallow representation of an adjacent patch. Instead of operating in the raw coordinate space or relying on discrete tokenization, LAA works entirely in the continuous latent space. To prevent the network from falling into a trivial degenerate solution, such as representation collapse, we formulate this alignment as a contrastive learning task using the InfoNCE loss (32). This contrastive mechanism ensures that the learned features remain diverse and discriminative while explicitly modeling the spatial transitions between patches.

To capture holistic semantic consistency, the GSD mechanism operates within the encoder. Inspired by the success of self-distillation in 2D vision (5), we generate two different masked views of the same input point cloud. The encoder processes both views, and we apply a bidirectional contrastive loss on their respective global [CLS] tokens. This forces the encoder to extract consistent global semantics invariant to the specific masking pattern, effectively bridging the gap between local geometric reconstruction and high-level semantic understanding without relying on heavy cross-modal teacher networks.

Our main contributions can be summarized as follows:

1) Identification of MPM Limitations. We identify the limitations of existing Masked Point Modeling methods, specifically their lack of explicit structural dependency modeling in the latent space and their neglect of global semantic consistency. We provide empirical evidence showing that simply reconstructing coordinates is insufficient for learning robust semantic representations.

2) Latent Autoregressive Alignment (LAA). We propose LAA, which utilizes an InfoNCE contrastive loss to explicitly model spatial transitions between patches in the decoder's latent space while preventing representation collapse. This module forces the network to understand the continuous topological structure of 3D shapes.

3) Global Self-Distillation (GSD). We introduce GSD, a bidirectional distillation mechanism applied to global tokens from different masked views, enforcing holistic semantic understanding within the encoder. This ensures that the learned representations are invariant to the specific masking patterns and capture high-level semantics.

4) State-of-the-Art Performance. Extensive experiments on standard 3D point cloud benchmarks demonstrate that our proposed framework significantly outperforms existing single-modal SSL methods on downstream tasks, including 3D object classification, few-shot learning, and part segmentation. Notably, our single-modal method matches or exceeds the performance of recent heavy cross-modal distillation frameworks.

2 Related Works

Self-Supervised Learning for Point Clouds. Inspired by the recent successes of Self-Supervised Learning (SSL) in NLP and 2D vision, an increasing amount of research has explored SSL in the realm of 3D point clouds. Early 3D SSL methods primarily relied on contrastive learning (49; 2; 22; 38; 26; 13), which discriminates between positive and negative pairs of point cloud augmentations or views. For instance, PointContrast (49) pioneered point-level invariant mapping across transformed views, while CrossPoint (2) introduced cross-modal contrastive objectives to align 3D point clouds with 2D images. However, these contrastive methods heavily rely on carefully designed data augmentations, which can sometimes alter the underlying semantics of the 3D shapes.

More recently, generative learning, particularly Masked Point Modeling (MPM), has become the dominant paradigm. Point-BERT (52) introduced a discrete variational autoencoder to generate point tokens for masked modeling. Point-MAE (33) simplified this by directly reconstructing the 3D coordinates of masked patches using a standard asymmetric autoencoder, achieving remarkable efficiency and performance. Following this, various extensions and point-based generative paradigms have been proposed (23; 28; 62; 50; 44; 1; 24). For example, Point-M2AE (54) utilizes a hierarchical architecture to capture multi-scale features. Point-FEMAE (53) extends global random masking to two branches to capture latent semantic features. I2P-MAE (55) introduces a 2D-guided masking strategy, PCP-MAE (63) introduces center predicting task, and Point-PQAE (64) utilizes cross-reconstruction. Meanwhile, others explore cross-modal learning frameworks such as ULIP (51), TAP (46), and Cross-BERT (25). Representation learning has also benefited from diverse methods such as score mixtures (47), ensemble networks (57), deep learning foundations (21; 15; 3), and positional query models (56; 59; 60). Furthermore, innovations in data pruning (41; 40), contrastive mixing (61; 58), and sample affinity (42; 39) have advanced self-supervised methods across various domains. While these methods achieve impressive results, they almost exclusively focus on output-space reconstruction. Point-S2AE explicitly regularizes the intermediate latent spaces and enforces global semantic consistency, which is orthogonal and complementary to existing MPM architectures.

Contrastive Learning and Self-Distillation. Contrastive learning and self-distillation have driven significant progress in representation learning. Methods like MoCo (20), BYOL (17), DINO (6), SimCLR (10), SimSiam (11), and iBOT (65) have shown that matching representations from different augmented views of the same image can yield powerful semantic features without requiring negative samples. In 3D point clouds, some recent works have explored cross-modal distillation to enhance representations. ACT (12) acquires knowledge from other modalities by adopting cross-modal auto-encoders as teachers. ReCon (36) harmoniously combines the generative framework and contrastive framework to transfer knowledge from pre-trained 2D image models or 3D teacher networks to a 3D student network. Unlike these methods that rely on complex cross-modal alignment, large 2D datasets, or heavy teacher models, our Global Self-Distillation (GSD) is entirely single-modal and intra-view. We apply a bidirectional DINO-style loss directly on the global tokens of two differently masked views of the same point cloud, achieving strong semantic consistency with minimal computational overhead.

Autoregressive Modeling. Autoregressive modeling, popularized by the GPT family (37; 4), predicts the next token in a sequence given the previous context. This paradigm has been successfully adapted to vision tasks, including ImageGPT (9), and point clouds. For example, PointGPT (8) adopts an autoregressive generative approach for point clouds by sorting patches and predicting the next discrete token. These methods typically apply autoregressive prediction in the raw data space or discrete token space as the primary generative objective, requiring strict causal masking in the encoder and breaking bidirectional context. In contrast, our Latent Autoregressive Alignment (LAA) module applies the autoregressive constraint strictly in the *continuous latent space* of the decoder. By predicting shallow latent features from deep ones, we explicitly model the structural dependencies between patches while preserving the full bidirectional attention capabilities of the encoder.

3 Method

In this section, we present our proposed framework, which builds upon the Masked Autoencoding (MAE) paradigm for point clouds. To overcome the limitations of independent patch reconstruction and the lack of global semantic consistency, we introduce two novel components: a Latent Autore-

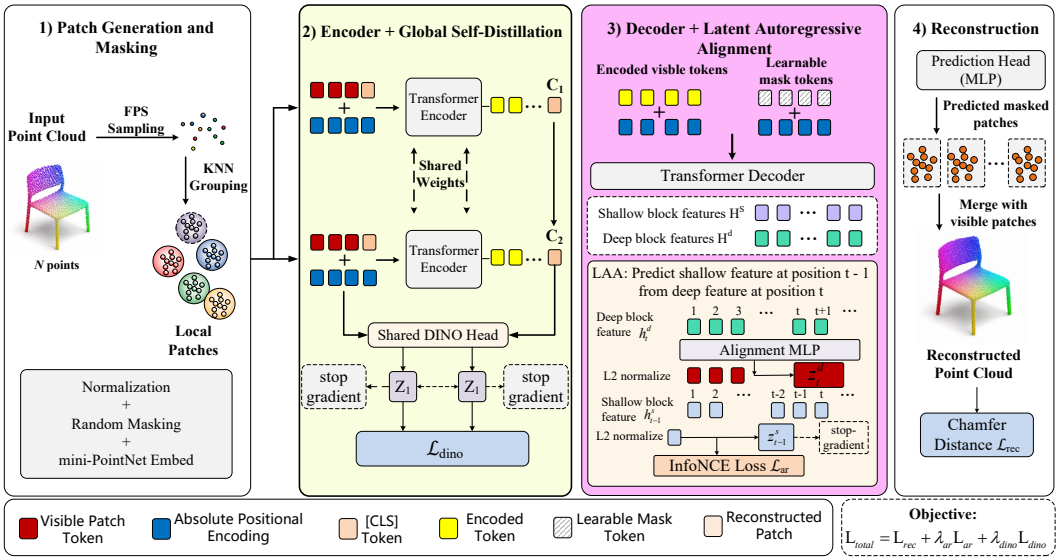

Figure 1: Overview of our proposed framework. The input point cloud is divided into patches and randomly masked. The encoder processes two different masked views and applies Global Self-Distillation (GSD) on the [CLS] tokens to ensure semantic consistency. The decoder reconstructs the masked patches while enforcing Latent Autoregressive Alignment (LAA) between deep and shallow masked token representations to capture structural dependencies.

gressive Alignment (LAA) module in the decoder and a Global Self-Distillation (GSD) mechanism in the encoder. The overall architecture consists of patch generation, an encoder equipped with GSD, and a decoder equipped with LAA.

3.1 Patch Generation and Masking

Following the standard patch-based learning paradigm established by Point-MAE (33), we first divide the input point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$ into local point patches. Specifically, we use the Farthest Point Sampling (FPS) algorithm to sample M center points $\mathcal{C} \in \mathbb{R}^{M\times 3}$, ensuring a relatively uniform coverage of the entire 3D shape. Then, the K-Nearest Neighbors (KNN) algorithm groups the k nearest points around each center to form M patches $\mathcal{X} \in \mathbb{R}^{M\times k\times 3}$. To ensure translation invariance, the coordinates of points within each patch are normalized relative to their center, allowing the network to focus on local geometric structures rather than absolute spatial locations.

During pre-training, we apply a random masking strategy with a high masking ratio m, typically between 60% and 80%. The patches are divided into visible patches $\mathcal{X}_v$ and masked patches $\mathcal{X}_m$. A lightweight mini-PointNet, which consists of a few MLP layers and a max-pooling operation, embeds the visible patches into D-dimensional tokens $\mathbf{E}_v \in \mathbb{R}^{(1-m)M\times D}$. The coordinates of the patch centers are embedded using a separate MLP to serve as positional encodings, which are added to the tokens to retain global spatial information.

3.2 Encoder and Global Self-Distillation

The encoder is a standard asymmetric Transformer that processes only the visible tokens $\mathbf{E}_v$. To capture global shape information and aggregate features across all visible patches, we prepend a learnable [CLS] token to the sequence. Absolute positional encodings are added to all tokens. The encoder applies multiple self-attention blocks to extract high-level latent representations. Formally, let $\mathbf{Z}_0 = [\text{CLS}, \mathbf{E}_v] + \mathbf{PE}$ be the input to the encoder. The l-th Transformer block performs the following operations:

$$\mathbf{Z}'_l = \mathbf{Z}_{l-1} + \text{MSA}(\text{LN}(\mathbf{Z}_{l-1})) \quad (1)$$

$$\mathbf{Z}_l = \mathbf{Z}'_l + \text{MLP}(\text{LN}(\mathbf{Z}'_l)) \quad (2)$$

where MSA denotes Multi-Head Self-Attention, LN denotes Layer Normalization, and MLP is a two-layer feed-forward network. The final output corresponding to the [CLS] token serves as the global representation of the visible point cloud.

Global Self-Distillation (GSD). While the MAE reconstruction task focuses on local geometric details, it often struggles to capture the holistic semantic understanding of the 3D object. To address this, we introduce a GSD mechanism inspired by DINO (5). During the forward pass, we generate two different masked views of the same input point cloud by applying the random masking strategy twice independently. Let $\mathcal{X}_{v1}$ and $\mathcal{X}_{v2}$ denote the visible patches of the two views. The encoder processes both views, yielding two different [CLS] tokens, denoted as c_1 and c_2, which represent the global semantics of the two views.

We then project these tokens using a shared projection head f_{dino}, which is a multi-layer perceptron with GELU activations. The GSD loss is formulated as a symmetric cross-entropy loss between the student and a momentum-updated teacher. For simplicity and efficiency in our symmetric design, we do not maintain a separate teacher network; instead, we apply the DINO loss bidirectionally using stop-gradients:

$$\mathcal{L}_{\text{dino}} = \frac{1}{2}\left(\mathcal{H}(f_{\text{dino}}(c_1), \text{sg}(f_{\text{dino}}(c_2))) + \mathcal{H}(f_{\text{dino}}(c_2), \text{sg}(f_{\text{dino}}(c_1)))\right) \tag{3}$$

where $\mathcal{H}(s,t) = -\sum t^{(i)} \log s^{(i)}$ is the cross-entropy loss with temperature scaling and centering applied to the target t, and sg denotes the stop-gradient operation. This mechanism forces the encoder to extract consistent global semantics regardless of which specific patches are masked, effectively bridging the gap between local geometric reconstruction and high-level semantic understanding. By doing so, the encoder learns a representation that is robust to severe occlusions and missing parts.

3.3 Decoder and Latent Autoregressive Alignment

The decoder aims to reconstruct the coordinates of the masked patches. It takes the encoded visible tokens and a set of learnable mask tokens corresponding to the masked positions as input, along with their positional encodings. The decoder consists of several standard Transformer blocks with self-attention and cross-attention mechanisms.

Latent Autoregressive Alignment (LAA). Standard decoders treat the reconstruction of each masked patch independently, ignoring the structural dependencies between them. To explicitly model these spatial transitions and enforce a sequential understanding of the 3D geometry, we introduce the LAA module within the decoder.

Let $H^s = [h_1^s, h_2^s, \ldots, h_{M_m}^s]$ and $H^d = [h_1^d, h_2^d, \ldots, h_{M_m}^d]$ denote the intermediate representations of the masked tokens extracted from a shallow block and a deep block of the decoder, respectively, where M_m is the number of masked tokens. These tokens are ordered based on a predefined spatial sequence, such as a Morton curve or the order generated by FPS. We enforce an autoregressive predictive relationship: the deep representation of the t-th masked token, h_t^d, should predict the shallow representation of the $(t-1)$-th masked token, h_{t-1}^s. This forces the network to understand the geometric transition from one patch to its neighbor.

To prevent representation collapse, which is a common issue where the network outputs a constant vector to minimize the loss trivially, we formulate this alignment as a contrastive learning task using the InfoNCE loss (32). We first project the deep representations using an alignment MLP f_{align} and ℓ_2-normalize both the projected deep features and the target shallow features:

$$z_t^d = \frac{f_{\text{align}}(h_t^d)}{\|f_{\text{align}}(h_t^d)\|_2}, \quad z_{t-1}^s = \frac{h_{t-1}^s}{\|h_{t-1}^s\|_2} \tag{4}$$

The LAA loss is then computed over all valid pairs as:

$$\mathcal{L}_{\text{ar}} = -\frac{1}{M_m - 1}\sum_{t=2}^{M_m} \log \frac{\exp(z_t^d \cdot z_{t-1}^s/\tau)}{\sum_{j=1}^{M_m-1}\exp(z_t^d \cdot z_j^s/\tau)} \tag{5}$$

where τ is a temperature hyperparameter controlling the sharpness of the distribution. A stop-gradient operation is applied to the shallow targets H^s to provide a stable learning objective. This contrastive mechanism ensures that the learned features remain diverse and discriminative while explicitly

modeling the spatial transitions between patches. Unlike discrete autoregressive models that require complex tokenization, LAA operates smoothly in the continuous latent space.

Finally, a prediction head reconstructs the coordinates of the masked patches from the decoder's final output. The reconstruction loss $\mathcal{L}_{\text{rec}}$ is computed using the ℓ_2 Chamfer Distance (CD) (14) between the predicted patches and the ground-truth masked patches.

3.4 Objective Function

The overall pre-training objective of our framework combines the masked point reconstruction loss, the latent autoregressive alignment loss, and the global self-distillation loss. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{ar}}\mathcal{L}_{\text{ar}} + \lambda_{\text{dino}}\mathcal{L}_{\text{dino}} \tag{6}$$

where λ_{ar} and λ_{dino} are balancing weights that control the contribution of each module. During downstream fine-tuning, the decoder, the alignment MLP, and the DINO projection heads are discarded, and only the pre-trained encoder is utilized to extract representations for various tasks.

4 Experiment

To evaluate the effectiveness of our proposed framework, we conduct extensive experiments on standard point cloud understanding benchmarks. We compare Point-S2AE against state-of-the-art self-supervised learning approaches on various downstream tasks, including 3D object classification, few-shot learning, and part segmentation.

4.1 Implementation Details

Pre-training. Our model architecture is built upon the Transformer-based PointMAE baseline. We sample 1024 points from each input point cloud using FPS and partition them into 64 patches with 32 points each. The model is pre-trained on the ShapeNet dataset, which contains over 50,000 unique 3D models across 55 categories, for 300 epochs using the AdamW optimizer (30) with a batch size of 128. We employ a cosine learning rate decay strategy (29). More detailed pre-training hyper-parameters and network architecture setups are provided in the Appendix.

Fine-tuning. During the fine-tuning stage, we discard the decoder, the alignment MLP, and the DINO head. We append task-specific heads to the pre-trained encoder and fine-tune the model on downstream datasets, including ScanObjectNN, ModelNet40, and ShapeNetPart. Detailed fine-tuning configurations are also provided in the Appendix.

4.2 Downstream Tasks

3D Object Classification on Real-World and Synthetic Datasets. We evaluate the learned representations on the real-world ScanObjectNN dataset (43) and the synthetic ModelNet40 dataset (48). ScanObjectNN is a highly challenging dataset containing approximately 15,000 real-world objects across 15 categories. Unlike synthetic datasets, ScanObjectNN features background noise, occlusions, and missing parts, making it a robust testbed for evaluating the transferability and robustness of pre-trained models. For ScanObjectNN, we report the accuracy on three challenging variants: OBJ_BG which includes background objects, OBJ_ONLY without background, and PB_T50_RS featuring perturbed bounding boxes with 50% translation and random rotation.

As shown in Table 1, Point-S2AE achieves state-of-the-art performance among single-modal methods. Specifically, on the ScanObjectNN dataset, Point-S2AE achieves 95.18% on OBJ_BG, 93.63% on OBJ_ONLY, and 90.53% on PB_T50_RS. These results match or exceed the performance of recent heavy cross-modal distillation methods like I2P-MAE, while our approach remains entirely single-modal and highly efficient. Specifically, Point-S2AE outperforms the baseline Point-MAE by large margins of 5.16%, 5.34%, and 5.35% across the three ScanObjectNN variants, respectively. This highlights that our LAA and GSD modules effectively empower the network to learn noise-resistant and semantically rich representations.

On the synthetic ModelNet40 dataset, which contains 12,311 meticulously crafted CAD models across 40 categories, our method achieves 94.2% without voting and 94.5% with voting. Improvements in

Table 1: 3D Object Classification performance on ScanObjectNN and ModelNet40. We report the overall accuracy (%).

Method	ScanObjectNN			ModelNet40	
	OBJ_BG	OBJ_ONLY	PB_T50_RS	w/o Vote	w/ Vote
Supervised Learning Only					
PointNet (34)	73.3	79.2	68.0	89.2	-
PointNet++ (35)	82.3	84.3	77.9	90.7	-
DGCNN (45)	82.8	86.2	78.1	92.9	-
PointMLP (31)	-	-	85.4	94.5	-
SimpleView (16)	-	-	80.5	93.9	-
Single-Modal Self-Supervised Learning					
Point-BERT (52)	87.43	88.12	83.07	92.7	93.2
MaskPoint (27)	89.30	88.10	84.30	-	93.8
Point-MAE (33)	90.02	88.29	85.18	93.2	93.8
Point-M2AE (54)	91.22	88.81	86.43	93.4	94.0
PointGPT (8)	91.60	90.00	86.90	-	94.0
Point-FEMAE (53)	**95.18**	93.29	90.22	94.0	**94.5**
ReCon (Single-Modal) (36)	94.49	93.29	90.35	93.9	94.2
Point-S2AE (Ours)	**95.18**	**93.63**	**90.53**	**94.2**	**94.5**
Improve (over Point-MAE)	+5.16	+5.34	+5.35	+1.0	+0.7
Cross-Modal Self-Supervised Learning					
ACT (12)	93.29	91.91	88.21	93.2	93.7
Joint-MAE (18)	90.94	88.86	86.07	-	94.0
I2P-MAE (55)	94.14	91.57	90.11	93.7	94.1

Point-S2AE over Point-MAE can be observed, yielding 1.0% and 0.7% absolute gains for without and with voting respectively. This demonstrates the effectiveness of combining local autoregressive alignment and global self-distillation for learning robust geometric features even on clean, synthetic shapes.

Few-shot Learning. We further assess the generalization ability and data efficiency of our pre-trained model on the ModelNet40 few-shot classification benchmark following standard protocols. We evaluate under four settings: 5-way 10-shot, 5-way 20-shot, 10-way 10-shot, and 10-way 20-shot. For each setting, we perform 10 independent runs and report the average accuracy along with the standard deviation. The results in Table 2 show that Point-S2AE significantly outperforms the PointMAE baseline across all settings. Notably, in the 5-way 20-shot and 10-way 20-shot settings, our method achieves 98.9% and 95.9% accuracy, respectively, establishing new state-of-the-art results. Specifically, 1.0%, 1.1%, 0.1%, and 0.9% performance gains over Point-MAE on the four respective settings are obtained. This indicates that the explicit modeling of structural dependencies via LAA and global semantics via GSD is particularly beneficial when labeled data is scarce, allowing the model to quickly adapt to new categories with minimal fine-tuning.

Part Segmentation. We also evaluate our method on the dense prediction task of 3D part segmentation using the ShapeNetPart dataset. This dataset contains 16,881 3D models from 16 categories, annotated with 50 part labels. The task requires the model to assign a part label to each individual point, demanding a fine-grained understanding of local geometries and part boundaries. The results in Table 3 demonstrate that our approach achieves state-of-the-art performance with 84.9% class mIoU and 86.4% instance mIoU. Particularly, Point-S2AE improves upon Point-MAE by 0.7% and 0.3% in class mIoU and instance mIoU, respectively. This significant improvement over the PointMAE baseline proves that our latent autoregressive alignment module effectively captures fine-grained geometric details and spatial transitions, which are crucial for dense point-level prediction tasks.

Table 2: Few-shot classification performance on the ModelNet40 dataset. We report the average accuracy (%) and standard deviation over 10 independent runs.

Method	5W10S	5W20S	10W10S	10W20S
Supervised Learning Only				
PointNet (34)	52.0±3.8	57.8±4.9	46.6±4.3	35.2±4.8
DGCNN (45)	31.6±2.8	40.8±4.6	19.9±2.1	16.9±1.5
OcCo (44)	90.6±2.8	92.5±1.9	82.9±1.3	86.5±2.2
Single-Modal Self-Supervised Learning				
Point-BERT (52)	94.6±3.1	96.3±2.7	91.0±5.4	92.7±5.1
MaskPoint (27)	95.0±3.7	97.2±1.7	91.4±4.0	93.4±3.5
Point-MAE (33)	96.3±2.5	97.8±1.8	92.6±4.1	95.0±3.0
Point-M2AE (54)	96.8±1.8	98.3±1.4	92.3±4.5	95.0±3.0
PointGPT (8)	96.8±2.0	98.6±1.1	92.6±4.6	95.2±3.4
Point-FEMAE (53)	97.2±1.9	98.6±1.3	**94.0±3.3**	95.8±2.8
Point-S2AE (Ours)	**97.3±2.1**	**98.9±1.1**	92.7±4.5	**95.9±2.8**
Improve (over Point-MAE)	+1.0	+1.1	+0.1	+0.9
Cross-Modal Self-Supervised Learning				
ACT (12)	96.8±2.3	98.0±1.4	93.3±4.0	95.6±2.8
Joint-MAE (18)	96.8±2.1	98.4±1.2	92.8±4.2	95.4±2.9
I2P-MAE (55)	97.0±1.8	98.5±1.2	93.1±4.1	95.8±2.5

Table 3: Part segmentation performance on the ShapeNetPart dataset. We report the class mIoU and instance mIoU (%).

Method	Cls. mIoU	Inst. mIoU
Supervised Learning Only		
PointNet (34)	80.4	83.7
PointNet++ (35)	81.9	85.1
DGCNN (45)	82.3	85.2
PointMLP (31)	84.6	86.1
Single-Modal Self-Supervised Learning		
Point-BERT (52)	84.1	85.6
MaskPoint (27)	84.4	86.0
Point-MAE (33)	84.2	86.1
PointGPT (8)	84.1	86.2
Point-FEMAE (53)	**84.9**	86.3
Point-S2AE (Ours)	**84.9**	**86.4**
Improve (over Point-MAE)	+0.7	+0.3
Cross-Modal Self-Supervised Learning		
ACT (12)	84.7	86.1
ReCon (36)	84.8	86.4

4.3 Ablation Study

We conduct extensive ablation studies to validate the individual contributions of the Latent Autoregressive Alignment (LAA) module and the Global Self-Distillation (GSD) mechanism. All ablation models are pre-trained on ShapeNet and fine-tuned on the respective downstream datasets under identical settings.

Table 4: Ablation study on different alignment loss functions for the LAA module, evaluated on ScanObjectNN.

Loss Function	OBJ_BG	OBJ_ONLY	PB_T50_RS
Cosine Similarity	71.94	75.04	70.74
Triplet Loss	90.60	88.98	86.07
InfoNCE (Point-S2AE)	**93.98**	**91.73**	**89.76**

Table 5: Ablation study on the core components of our proposed method. We evaluate the impact of the Latent Autoregressive Alignment (LAA) and Global Self-Distillation (GSD) modules.

Components		ScanObjectNN			ModelNet40		Few-shot (ModelNet40)				PartSeg	
LAA	GSD	OBJ_BG	OBJ_ONLY	PB_T50_RS	w/o Vote	w/ Vote	5W10S	5W20S	10W10S	10W20S	Cls.	Inst.
		90.02	88.29	85.18	93.2	93.8	96.3±2.5	97.8±1.8	92.6±4.1	95.0±3.0	84.2	86.1
✓		93.98	91.73	89.76	93.6	93.9	96.5±2.9	98.8±1.1	91.2±4.5	94.8±3.2	84.6	86.2
	✓	94.49	91.57	90.08	93.1	93.4	95.2±4.4	98.1±1.5	91.2±4.4	94.8±3.2	84.8	86.3
✓	✓	**95.18**	**93.63**	**90.53**	**94.2**	**94.5**	**97.3±2.1**	**98.9±1.1**	**92.7±4.5**	**95.9±2.8**	**84.9**	**86.4**

Component Analysis. As shown in Table 5, the baseline model PointMAE achieves reasonable performance relying solely on the reconstruction loss. Adding the LAA module improves the results significantly, providing gains of 3.96% on ScanObjectNN OBJ_BG and 4.58% on PB_T50_RS, validating that explicitly modeling inter-patch dependencies in the latent space helps the network understand spatial transitions and local structural continuity. Alternatively, integrating the GSD mechanism also enhances the performance by 4.47% on OBJ_BG and 4.90% on PB_T50_RS by enforcing global semantic consistency across different masked views, making the representation more robust to noise and occlusions. Finally, the combination of both modules yields the best results across almost all metrics, demonstrating their complementary nature: LAA focuses on local structural continuity and fine-grained transitions, while GSD ensures holistic semantic stability. This synergy allows the network to learn a comprehensive representation that excels in both global classification and dense segmentation tasks.

Alignment Loss Function. We further explore different loss functions for the Latent Autoregressive Alignment (LAA) module. The objective is to explicitly enforce structural dependencies between the deep representation of a target masked token (h_t^d) and the shallow representation of an adjacent source token (h_{t-1}^s). We compare three alternatives: Cosine Similarity, Triplet Loss, and the InfoNCE loss. As shown in Table 4, optimizing the alignment directly via mean Cosine Similarity, given by $\mathcal{L}_{\text{align}} = 1 - \frac{1}{M_m - 1} \sum_{t=2}^{M_m} \cos(h_t^d, h_{t-1}^s)$, suffers from severe representation collapse, dropping to 71.94% on OBJ_BG. While the Triplet Loss alleviates this issue by employing hard negatives within a specified margin, its performance (90.60%) still trails our proposed method. The InfoNCE loss effectively treats all other tokens in the sequence as negatives, dynamically pushing them apart in the continuous latent space. This prevents trivial solutions and forces the network to learn diverse and discriminative representations, ultimately achieving the highest accuracy (93.98%) across all ScanObjectNN variants.

5 Conclusion

In this paper, we identify the limitations of independent patch reconstruction and the lack of holistic semantic consistency in existing Masked Point Modeling (MPM) methods. To address these issues, we propose a novel framework that introduces Latent Autoregressive Alignment (LAA) and Global Self-Distillation (GSD). LAA explicitly models the structural dependencies between masked patches in the continuous latent space, avoiding representation collapse through contrastive learning. Simultaneously, GSD extracts consistent global semantics invariant to specific masking patterns by applying bidirectional self-distillation. Extensive experiments showcase the superior performance of Point-S2AE, setting new state-of-the-art results across various 3D representation learning benchmarks and demonstrating high-capacity data efficiency. We hope our framework inspires further exploration of latent structural constraints and single-modal semantic alignment in point cloud representation learning.

REFERENCES

[1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018.

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9902–9912, 2022.

[3] Y. Bengio and Y. LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[6] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[7] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[8] G. Chen, M. Wang, Y. Yang, K. Yu, L. Yuan, and Y. Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems*, 36:29667–29679, 2023.

[9] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[12] R. Dong, Z. Qi, L. Zhang, J. Zhang, J. Sun, Z. Ge, L. Yi, and K. Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022.

[13] B. Du, X. Gao, W. Hu, and X. Li. Self-contrastive learning with hard negative sampling for self-supervised point cloud learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3133–3142, 2021.

[14] H. Fan, H. Su, and L. 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*, pages 605–613, 2017.

[15] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*, volume 1. MIT Press, 2016.

[16] A. Goyal, H. Law, B. Liu, A. Newell, and J. Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *International Conference on Machine Learning*, pages 3809–3820. PMLR, 2021.

[17] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.

[18] Z. Guo, R. Zhang, L. Qiu, X. Li, and P.-A. Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *arXiv preprint arXiv:2302.14007*, 2023.

[19] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[21] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[22] S. Huang, Y. Xie, S.-C. Zhu, and Y. Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021.

[23] G. Li, H. Zheng, D. Liu, C. Wang, B. Su, and C. Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *NeurIPS*, 2022.

[24] J. Li, B. M. Chen, and G. H. Lee. So-net: Self-organizing network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9397–9406, 2018.

[25] X. Li, P. Li, Z. Wei, Z. Zhu, M. Wei, J. Hou, L. Nan, J. Qin, H. Xie, and F. L. Wang. Cross-bert for point cloud pretraining. *arXiv preprint arXiv:2312.04891*, 2023.

[26] F. Liu, G. Lin, C.-S. Foo, C. K. Joshi, and J. Lin. Point discriminative learning for data-efficient 3d point cloud analysis. In *2022 International Conference on 3D Vision (3DV)*, pages 42–51. IEEE, 2022.

[27] H. Liu, M. Cai, and Y. J. Lee. Masked discrimination for self-supervised learning on point clouds. In *European Conference on Computer Vision*, pages 657–675. Springer, 2022.

[28] Y. Liu, C. Chen, C. Wang, X. King, and M. Liu. Regress before construct: Regress autoencoder for point cloud self-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1738–1749, 2023.

[29] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[30] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[31] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022.

[32] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[33] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, and L. Yuan. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pages 604–621. Springer, 2022.

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[36] Z. Qi, R. Dong, G. Fan, Z. Ge, X. Zhang, K. Ma, and L. Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. *arXiv preprint arXiv:2302.02318*, 2023.

[37] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[38] A. Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 626–642. Springer, 2020.

[39] H. Tan, C. Wang, S. Wu, X.-Y. Zhang, F. Yin, and C.-L. Liu. Ensemble quadratic assignment network for graph matching. *International Journal of Computer Vision (IJCV)*, 2024.

[40] H. Tan, S. Wu, F. Du, Y. Chen, Z. Wang, F. Wang, and X. Qi. Data pruning via moving-one-sample-out. In *Neural Information Processing Systems (NeurIPS)*, 2023.

[41] H. Tan, S. Wu, W. Huang, S. Zhao, and X. Qi. Data pruning by information maximization. In *International Conference on Learning Representations (ICLR)*, 2025.

[42] H. Tan, S. Wu, Z. Tian, Y. Chen, X. Qi, and J. Jia. Saco loss: Sample-wise affinity consistency for vision-language pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[43] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597, 2019.

[44] H. Wang, Q. Liu, X. Yue, J. Lasenby, and M. J. Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9782–9792, 2021.

[45] 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 Transactions on Graphics (tog)*, 38(5):1–12, 2019.

[46] Z. Wang, X. Yu, Y. Rao, J. Zhou, and J. Lu. Take-a-photo: 3d-to-2d generative pre-training of point cloud models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5640–5650, 2023.

[47] S. Wu, H. Tan, Y. Chen, S. Zhang, J. Li, B. Yu, X. Qi, and J. Jia. Mixture-of-scores: Robust image-text data quality score via three lines of code. In *International Conference on Computer Vision (ICCV)*, 2025.

[48] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[49] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020*, pages 574–591. Springer, 2020.

[50] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.

[51] L. Xue, M. Gao, C. Xing, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1189, 2023.

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322, 2022.

[53] Y. Zha, H. Ji, J. Li, R. Li, T. Dai, B. Chen, Z. Wang, and S.-T. Xia. Towards compact 3d representations via point feature enhancement masked autoencoders. *arXiv preprint arXiv:2312.10726*, 2023.

[54] R. Zhang, Z. Guo, P. Gao, R. Fang, B. Zhao, D. Wang, Y. Qiao, and H. Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35:27061–27074, 2022.

[55] R. Zhang, L. Wang, Y. Qiao, P. Gao, and H. Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21769–21780, 2023.

[56] S. Zhang, J. Huang, Q. Zhou, F. Wang, J. Luo, J. Yan, et al. Continuous-multiple image outpainting in one-step via positional query and a diffusion-based approach. In *ICLR*, 2024.

[57] S. Zhang, M. Liu, and J. Yan. The diversified ensemble neural network. *Advances in Neural Information Processing Systems*, 33:16001–16011, 2020.

[58] S. Zhang, M. Liu, J. Yan, H. Zhang, L. Huang, X. Yang, and P. Lu. M-mix: Generating hard negatives via multi-sample mixing for contrastive learning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2461–2470, 2022.

[59] S. Zhang, Q. Zhou, Z. Wang, F. Wang, and J. Yan. Patch-level contrastive learning via positional query for visual pre-training. In *ICML*, 2023.

[60] S. Zhang, Q. Zhou, S. Wu, H. Tan, Z. Wang, J. Huang, and J. Yan. Cr2pq: Continuous relative rotary positional query for dense visual representation learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[61] S. Zhang, F. Zhu, J. Yan, R. Zhao, and X. Yang. Zero-cl: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *International Conference on Learning Representations*, 2021.

[62] S. Zhang, F. Zhu, R. Zhao, and J. Yan. Contextual image masking modeling via synergized contrasting without view augmentation for faster and better visual pretraining. In *ICLR*, 2023.

[63] X. Zhang, S. Zhang, and J. Yan. Pcp-mae: Learning to predict centers for point masked autoencoders. *Advances in Neural Information Processing Systems*, 37:80303–80327, 2024.

[64] X. Zhang, S. Zhang, and J. Yan. Towards more diverse and challenging pre-training for point cloud learning: Self-supervised cross reconstruction with decoupled views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28696–28706, 2025.

[65] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

A Additional Experimental Details

Table 6: Training details for pre-training and downstream fine-tuning.

Config	**ShapeNet**	**ScanObjectNN**	**ModelNet40**	**Few-shot**	**ShapeNetPart**
optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
learning rate	5e-4	2e-5	1e-5	5e-4	2e-4
weight decay	5e-2	5e-2	5e-2	5e-2	5e-2
learning rate scheduler	cosine	cosine	cosine	cosine	cosine
training epochs	300	300	300	300	300
warmup epochs	10	10	10	10	10
batch size	128	32	32	32	16
drop path rate	0.1	0.2	0.2	0.1	0.1
number of points	1024	2048	1024	1024	2048
number of point patches	64	128	64	64	128
point patch size	32	32	32	32	32
augmentation	Scale&Trans	Rotation	Scale&Trans	-	-
GPU device	A100	A100	A100	A100	A100

Pre-training details. We use ShapeNet (7) as the pre-training dataset. ShapeNet is a clean set of 3D CAD object models with rich annotations, including 51K unique 3D models from 55 common object categories. The overall pre-training includes 300 epochs, and we use a cosine learning rate (29) of 5e-4 warming up for 10 epochs. The AdamW optimizer (30) is used, and the batch size is set to 128. We apply a random masking ratio of 60%. For our proposed modules, we set the Latent Autoregressive Alignment (LAA) loss weight $\lambda_{\text{ar}} = 0.001$, the InfoNCE temperature $\tau = 0.07$, and the Global Self-Distillation (GSD) loss weight $\lambda_{\text{dino}} = 1.0$, where student and teacher temperatures are set to 0.1 and 0.04 respectively. We run all experiments with a single A100 GPU (80GB). More hyperparameter details are shown in Table 6.

Model details. The encoder comprises 12 Transformer blocks with a hidden dimension of 384 and 6 attention heads. The decoder consists of 4 Transformer blocks with 6 attention heads each. The projector is composed of MLP layers, LayerNorm, ReLU, and another MLP layer, sequentially. During downstream fine-tuning, we add specific MLPs with dropout and batch normalization as task heads.

B Generality of Our Method

To demonstrate that the effectiveness of our proposed Latent Autoregressive Alignment (LAA) and Global Self-Distillation (GSD) loss functions is not limited to a specific network design, we conduct experiments to validate their universality. We integrate our LAA and GSD objectives into another representative Masked Point Modeling architecture, PointGPT (8), in addition to our default Point-MAE baseline. PointGPT utilizes an autoregressive generative framework based on discrete point tokens. It predicts the next discrete token based on a sorted sequence of patches, mirroring the standard GPT-style autoregressive pre-training. However, this prediction is performed in the discrete sequence space and strictly relies on causal masking, ignoring the structural transitions in continuous latent spaces and global bidirectional semantics. By simply introducing our continuous latent autoregressive alignment loss and global distillation loss during its pre-training phase, without fundamentally altering its core architecture, we evaluate the modified model on the ScanObjectNN classification and ModelNet40 few-shot learning tasks.

As shown in Table 7, the LAA and GSD loss functions serve as highly versatile and universal regularization terms that yield consistent performance improvements across entirely different baseline architectures. For the Point-MAE baseline, our loss functions act as substantial enhancers, notably increasing the accuracy by 5.16% on OBJ_BG, 5.34% on OBJ_ONLY, and 5.35% on the challenging PB_T50_RS variant. More importantly, when the identical loss functions are applied to PointGPT, we observe stable and robust enhancements: the classification accuracy improves by 0.8% on OBJ_BG, 2.3% on OBJ_ONLY, and 1.3% on PB_T50_RS. Additionally, the modified PointGPT shows consistent improvements across various few-shot learning settings. These results clearly

indicate that the roles of the LAA and GSD loss functions possess broad generality. They can function as plug-and-play auxiliary objectives to seamlessly integrate with and enhance latent structural dependencies and global semantic consistency across diverse point cloud pre-training frameworks, ranging from continuous coordinate reconstruction (Point-MAE) to discrete token prediction (PointGPT).

Table 7: Generality of our proposed method. We apply our Latent Autoregressive Alignment (LAA) and Global Self-Distillation (GSD) modules to different baseline architectures (Point-MAE and PointGPT) and evaluate the performance on ScanObjectNN and ModelNet40 Few-shot classification.

Method	ScanObjectNN			Few-shot (ModelNet40)			
	OBJ_BG	OBJ_ONLY	PB_T50_RS	5W10S	5W20S	10W10S	10W20S
Point-MAE (33)	90.02	88.29	85.18	96.3±2.5	97.8±1.8	92.6±4.1	95.0±3.0
+ Ours (Point-S2AE)	**95.18**	**93.63**	**90.53**	**97.3±2.1**	**98.9±1.1**	**92.7±4.5**	**95.9±2.8**
PointGPT (8)	91.60	90.00	86.90	96.8±2.0	**98.6±1.1**	92.6±4.6	95.2±3.4
+ Ours	**92.40**	**92.30**	**88.20**	**97.1±2.1**	**98.6±0.7**	**92.8±4.0**	**95.5±3.0**

C Per-Category Accuracy Results

To provide a more comprehensive understanding of our model's performance, we present the detailed per-category accuracy results of Point-S2AE on the ScanObjectNN (PB_T50_RS) and ModelNet40 datasets.

Table 8 shows the accuracy for the 15 object categories in the challenging ScanObjectNN PB_T50_RS variant. Our model demonstrates robust classification capabilities across various categories, achieving over 95% accuracy on classes like *Table*, *Cabinet*, and *Bag*, highlighting its ability to handle noisy and occluded real-world data effectively.

Table 8: Per-category accuracy (%) on the ScanObjectNN (PB_T50_RS) dataset.

Category	Accuracy	Category	Accuracy	Category	Accuracy
Bag	75.90	Chair	84.67	Shelf	85.45
Bed	91.96	Desk	92.16	Sink	93.33
Bin	75.19	Display	94.29	Sofa	83.33
Box	91.67	Door	93.78	Table	96.19
Cabinet	95.38	Pillow	78.52	Toilet	92.94

Table 9 details the per-category accuracy for all 40 object classes in the synthetic ModelNet40 dataset. The results emphasize the generalizability of our framework, as Point-S2AE achieves perfect (100%) or near-perfect accuracy on a large majority of the categories, such as *Airplane*, *Bed*, *Car*, and *Laptop*.

Table 9: Per-category accuracy (%) on the ModelNet40 dataset.

Category	Accuracy	Category	Accuracy	Category	Accuracy	Category	Accuracy
airplane	100.00	cup	70.00	laptop	100.00	sofa	100.00
bathtub	98.00	curtain	90.00	mantel	98.00	stairs	100.00
bed	100.00	desk	94.19	monitor	100.00	stool	90.00
bench	70.00	door	100.00	night_stand	80.23	table	87.00
bookshelf	99.00	dresser	87.21	person	100.00	tent	95.00
bottle	99.00	flower_pot	5.00	piano	98.00	toilet	100.00
bowl	90.00	glass_box	96.00	plant	94.00	tv_stand	91.00
car	100.00	guitar	99.00	radio	85.00	vase	85.00
chair	98.00	keyboard	100.00	range_hood	95.00	wardrobe	85.00
cone	100.00	lamp	90.00	sink	90.00	xbox	100.00

D LIMITATIONS AND FUTURE WORK

Point-S2AE is a simple and effective framework that improves standard Masked Point Modeling (MPM) by additionally learning to align latent representations and distill global semantics, which significantly boosts performance across diverse tasks. However, there are some limitations in our Point-S2AE approach, which may be two-fold. First, while our method effectively models local structures and global semantics for object-centric point clouds (such as CAD models or scanned single objects), its extension to large-scale, dense 3D scenes remains underexplored. The autoregressive ordering of patches and the global [CLS] token might require further structural adaptation to handle the complex, unconstrained, and expansive nature of scene-level environments. Second, Point-S2AE is a purely single-modal self-supervised method. While this ensures computational efficiency and wider applicability without relying on paired multi-modal data, the current 3D point cloud datasets are relatively constrained in size compared to massive 2D image or text corpora. This data scale limitation might eventually bound the potential capability of purely single-modal representation learning.

To address these issues, future works could focus on developing a hierarchical Point-S2AE framework that can effectively scale to large 3D scenes by modeling local and global structures at multiple resolutions. Additionally, exploring a hybrid model that effectively combines our latent structural constraints and global semantics with cross-modal knowledge distillation could potentially yield even more powerful representations by leveraging the vast knowledge embedded in large-scale multi-modal foundation models.

E BROADER IMPACTS

Our proposed Point-S2AE demonstrates a marked improvement over existing baseline methods like Point-MAE and PointGPT, establishing new state-of-the-art benchmarks across a variety of downstream tasks in 3D point cloud understanding. This enhancement is pivotal in advancing technologies heavily reliant on precise spatial recognition. For instance, in autonomous driving, accurate 3D environmental perception and robust object classification are paramount for safety and efficiency. Moreover, the versatility and robustness of our learned representations extend to other critical real-world applications, including robotics manipulation, automated industrial inspection, and urban planning where detailed and scalable 3D modeling is essential. In augmented and virtual reality, such advancements can enhance interactive experiences by seamlessly integrating highly accurate digital assets with the physical world.

While Point-S2AE marks a substantial advancement in point cloud understanding with broad beneficial applications ranging from autonomous systems to smart cities, it is also susceptible to potential negative societal impacts. The primary concerns stem from its potential misuse in unauthorized or unregulated 3D surveillance systems, where enhanced and highly accurate spatial recognition capabilities could lead to privacy infringements in public or private spaces. It is imperative that the deployment of such advanced 3D perception technologies be accompanied by strict ethical guidelines, robust legal frameworks, and comprehensive privacy-preserving measures to prevent misuse.