Reliable-View 2D-3D Key-Part Aligned Transformer with Reinforced Masking for 3D Point Cloud Understanding

Xianglong Jin, Zheng Wang*, Rong Wang, Feiping Nie

School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University
127 West Youyi Road, Beilin District
Xi'an Shaanxi, 710072, P.R.China
xianglong_j@mail.nwpu.edu.cn, zhengwangml@gmail.com, wangrong07@tsinghua.org.cn, feipingnie@gmail.com

Abstract

Self-supervised 3D point cloud understanding is crucial for scene understanding, where Masked Autoencoders (MAE) have achieved excellent performance in point cloud representation learning. However, existing MAE-style methods fail to consider spatial-semantic variations in masking strategies, and joint learning with multi-view images often overlooks view redundancy. To address these challenges, we propose an MAE framework enhanced with reliable multi-view 2D-3D **K**ey-part alignment and **R**einforced masking, named as **KR-MAE**. Our approach comprises three key innovations: Reinforced Masking (RM) strategically samples visible tokens based on semantic saliency to enhance reconstruction fidelity; Reliable Multi-View Selector (RVS) dynamically refines the most informative image subset by filtering occluded or low-texture views, mitigating detrimental redundancy; Reliable-view 2D-3D Key-part Aligned Transformer (KAT) establishes semantic-aligned correspondence between salient 3D point cloud parts and reliable multi-view 2D image patches, leveraging rich texture cues from 2D images to compensate for sparse geometry in point cloud. Extensive experiments on 3D classification and segmentation benchmarks demonstrate that KR-MAE achieves state-of-the-art performance, surpassing prior multi-modal methods.

Code — https://github.com/jinxianglong10/KR-MAE

Introduction

With the rapid advancement of 3D sensing technology, point clouds have emerged as a fundamental modality for capturing geometric and spatial object data, demonstrating significant potential in critical applications such as autonomous driving (Qian, Lai, and Li 2022), robotic navigation (Ren and Jebelli 2025), and embodied AI (Lin et al. 2025). However, the inherent characteristics of point cloud data—including their unordered nature, transformation sensitivity, and structural sparsity—impose significant barriers to acquiring effective representations, often requiring labor-intensive manual annotation for supervised learning. To circumvent the need for costly annotations, self-supervised learning (SSL) paradigms have gained prominence. These approaches broadly fall into two categories: Single-modal SSL (e.g., Point-BERT (Yu et al. 2022), Point-MAE (Pang et al. 2022)) processes only point clouds via pretext tasks like masked reconstruction; Multi-modal SSL (e.g., CrossPoint (Afham et al. 2022), GreenPLM (Tang et al. 2025)) leverages complementary 2D image or text data to enrich 3D representations.

Initial research primarily focused on single-modal learning due to the practical appeal of methods using solely point clouds, valued for simplicity and efficiency. These methods typically leverage pretext tasks such as masked reconstruction (Pang et al. 2022) or contrastive learning (Xie et al. 2020) to extract geometric features exclusively from point clouds. However, the representational richness achievable by such approaches is fundamentally constrained by an intrinsic information bottleneck: they can only process a single modality independently without utilizing implicit correlations across modalities. Compared to the *irregular and sparse* 3D point clouds, 2D images inherently provide *dense and fine-grained* visual signals encompassing both geometric details and semantic context, which offer complementary information to enhance 3D representation learning (Guo et al. 2023). Consequently, relying solely on point clouds often proves insufficient for acquiring highly discriminative and robust features required in complex scenarios.

For these reasons, 2D-3D joint learning methods have emerged as a powerful paradigm to overcome the limitations of single-modal approaches. By integrating complementary information from 2D images with 3D point clouds, these frameworks achieve holistic 3D representations that significantly enhance discriminative power and robustness. Among these, the MAE-style approach has become dominant, with representative techniques advancing cross-modal interaction through innovative strategies such as Joint-MAE (Guo et al. 2023) leverages masked cross-modal reconstruction to jointly optimize image and point cloud features in a unified latent space and PiMAE (Chen et al. 2023) employs complementary masking for point clouds and images, explicitly modeling cross-modal feature complementarity by masking non-overlapping regions across modalities.

Despite these advances, existing MAE-based multi-modal frameworks for 2D-3D joint learning are constrained by two critical issues. First, spatially invariant masking strategies such as random masking in (Guo et al. 2023; Chen et al. 2023) neglect spatial-semantic variations, often retaining

*Corresponding author.

low-information regions while discarding salient structures, thereby reducing the quality of point cloud reconstruction. Second, indiscriminate fusion of multi-view images such as all-view integration in (Zhang et al. 2023; Yu and Song 2024) overlooks view redundancy, where some perspective are redundant or even counterproductive.

To address these challenges, we propose an MAE framework enhanced with reliable multi-view 2D-3D **K**ey-part alignment and **R**einforced masking named as **KR-MAE**. Specifically, our approach introduces a Reinforced Masking (RM) module instead of prior random masking strategies. RM enables higher-quality point cloud reconstruction by strategically preserving geometrically salient regions (e.g., edges, complex structures) while discarding low-information planar surfaces. For multi-modal 2D inputs, we project point clouds into multi-view images for joint learning, but critically address view redundancy through a Reliable Multi-View Selector (RVS) module that dynamically refines the most informative perspective subset, thereby mitigating negative impacts from uninformative or detrimental viewpoints. Building upon these adaptively refined masks and views, we further propose a Reliable-view 2D-3D Key-part Aligned Transformer (KAT) to establish structural correspondence between visible point cloud tokens and multi-view images, leveraging 2D part alignment to guide learning of semantically critical 3D structures and enhancing cross-modal fusion via geometric-texture co-optimization.

In summary, the contributions of our paper are as follows:

- We propose KR-MAE, a novel MAE framework that employs Reinforced Masking (RM), a masking strategy that prioritizes semantically rich point cloud regions via gradient-based reward optimization.
- We integrate multi-view 2D images for joint learning via projection and propose a Reliable Multi-View Selector (RVS) module to dynamically select the most informative subset of views, mitigating the impact of redundant or counterproductive perspectives.
- We propose a Reliable-view 2D-3D Key-part Aligned Transformer (KAT) to establish structural correspondence between salient regions of point clouds and aligned multi-view 2D images, leveraging rich texture cues from images to compensate for sparse point cloud.

Related Work

Self-Supervised 3D Representation Learning

Self-supervised learning has emerged as a prominent paradigm for 3D point cloud representation learning, primarily categorized into single-modal and multi-modal approaches. Single-modal methods learn exclusively from point clouds, which can be further divided into contrastive and generative frameworks. Contrastive approaches enforce representation invariance through geometric transformations—exemplified by PointContrast (Xie et al. 2020), which establishes point-level correspondences across rotated views to learn transformation-invariant features. Generative approaches, particularly masked autoencoders (MAE), have gained significant traction: PointBERT (Yu et al. 2022) employs a discrete Variational Autoencoder (dVAE) to tokenize point clouds and predicts masked tokens via a Transformer decoder, while PointMAE (Pang et al. 2022) simplifies this pipeline by directly reconstructing masked point patches. Further enhancing this paradigm, PCP-MAE (Zhang, Zhang, and Yan 2024) guides the model to learn to predict the masked centers and use the predicted centers to replace the directly provided centers and RI-MAE (Su et al. 2025) facilitates self-supervised reconstruction in a rotation-invariant manner. However, existing MAE methods universally rely on indiscriminate random masking strategies, which overlook the varying information density across regions. This limitation often results in suboptimal reconstruction quality and limited generalization capacity in complex scenes. To address this, our KR-MAE introduces Reinforced Masking, a masking strategy that prioritizes semantically rich regions via gradient-based reward optimization, dynamically adjusting mask patterns to maximize information gain during reconstruction.

2D-3D Joint 3D Representation Learning

The inherent information bottleneck in single-modal approaches has prompted extensive research into fusing 2D images with 3D point clouds to leverage cross-modal context. Joint-MAE (Guo et al. 2023) and PiMAE (Chen et al. 2023) align local features between single-view images and point clouds through masked cross-modal reconstruction. I2P-MAE (Zhang et al. 2023) extends this to three orthographic views but suffers from geometric distortion due to coordinate-axis projection, compromising spatial fidelity. MM-Point (Yu and Song 2024) and OpenView (Zhou et al. 2025) incorporate multi-view images but extract only global features through pooling operations, neglecting fine-grained point-pixel correspondences. Despite these advances, existing methods exhibit two critical limitations: viewpoint deficiency persists as most frameworks rely on sparse or fixed viewpoints, where some perspectives are redundant or even counterproductive; coarse feature alignment arises when global feature fusion fails to model localized structural relationships between point clouds and images. To address these challenges, our KR-MAE employs dense multi-view imagery to capture holistic 3D context, integrates a Reliable Multi-View Selector (RVS) module that dynamically prioritizes semantically informative viewpoints, and proposes a Reliable-view 2D-3D Key-part Aligned Transformer (KAT) to establish structural correspondence between key regions of the visible point cloud tokens and localized image regions across diverse perspectives.

Method

The overall pre-training pipeline of KR-MAE is illustrated in Figure 1. Given an input 3D point cloud, we first project it to generate dense multi-view 2D images. These multi-modal data are then unified through Multi-view 2D and 3D Embedding for initial tokenization, converting both point clouds and images into a joint token space. Subsequently, a Reliable Multi-View Selector (RVS) module dynamically filters semantically informative views, discarding redundant or

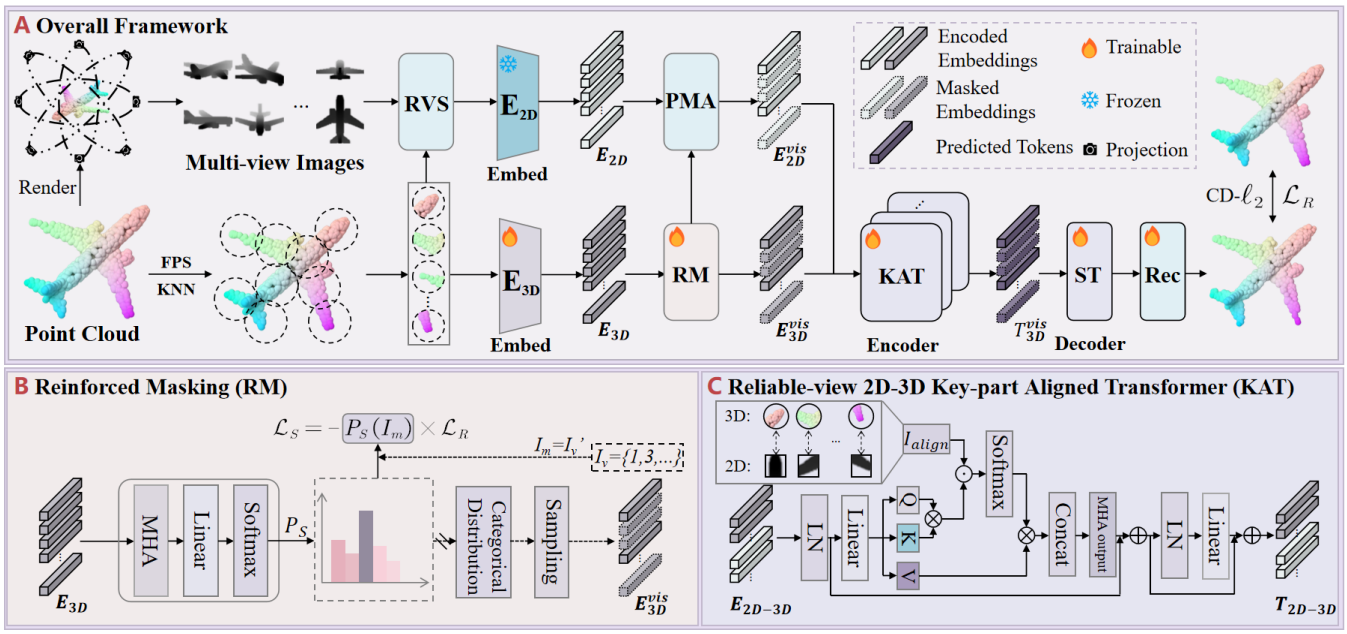

Figure 1: **The pipeline of our KR-MAE**. A is the overall framework, while B and C are sub-modules. Input 3D point clouds are first rendered into multi-view 2D images. Subsequently, a Reliable Multi-View Selector module dynamically selects informative views. Following this, RM and PMA align the selected views with point tokens to generate visible tokens, which are then fused via Reliable-view 2D-3D Key-part Aligned Transformer. Consequently, enhanced 3D tokens are outputted for reconstruction.

low-information perspectives to enhance computational efficiency and representation quality. Then Reinforced Masking (RM) masks the point cloud based on information significance and propagates this masking to corresponding image regions through Part and Masking Alignment (PMA) to generate aligned visible tokens. Finally, these aligned visible tokens are fed into Reliable-view 2D-3D Key-part Aligned Transformer (KAT) to integrate salient 3D regions with multi-view 2D part-aligned features for the following point cloud reconstruction.

Multi-view 2D Rendering from 3D

We adopt the multi-stage 3D-to-2D rendering pipeline from PointCLIP V2 (Zhu et al. 2023) as the continuous projection function $\mathrm{Project}(\cdot)$. This pipeline firstly converts point clouds to volumetric grids through Voxelize, subsequently fills sparse voxels via interpolation in the Densify step, then applies Gaussian filtering for spatial continuity during Smooth, and finally projects volumetric data to 2D planes via Squeeze. Compared to direct point cloud projection, this approach offers two key advantages: generates depth maps from arbitrary viewpoints $\{\theta_i\}_{i=1}^{v}$, mitigating single-view information bias; produces dense and continuous images whereas direct projection yields sparse mappings that hinder 2D feature extraction. These properties enable rendered multi-view images $M = \{m_1, m_2, \ldots, m_v\}$ to retain richer structural information, allowing downstream 2D model to extract more discriminative features. Formally, given an input point cloud $P \in \mathbb{R}^{N\times 3}$ with N points, we select v viewing angles and render:

$$m_i = \mathrm{Project}(P; \theta_i), \quad m_i \in \mathbb{R}^{H\times W\times 3} \tag{1}$$

where m_i denotes the i-th view image, with H and W being the height and width, and 3 corresponding to RGB color channels[1]. This generation method does not require any human intervention and external image resources, which has advanced our self-supervised model.

Multi-view 2D and 3D Embedding

Our multi-view 2D and 3D embedding module aligns multi-view 2D images and 3D point cloud in a unified d^E-dimensional latent space, enabling cross-modal feature interaction through vector space symmetry. For each input image $m_i \in \mathbb{R}^{H\times W\times 3}$, we first partition it into m non-overlapping 16×16 patches, where $m = HW/256$. Each patch is then projected to a d^E-dimensional vector, augmented with learnable positional encodings to preserve spatial relationships. Formally, the i-th view embedding is generated by a Vision Transformer (ViT) (Dosovitskiy et al. 2020) backbone pre-trained with CLIP (Radford et al. 2021):

$$E^i_{\mathrm{2D}} = \mathrm{ViT}(m_i), \quad E^i_{\mathrm{2D}} \in \mathbb{R}^{m\times d^E} \tag{2}$$

This patch-based embedding captures global contextual relationships through self-attention, leveraging ViT's ability to model long-range dependencies in visual data.

For the input point cloud $P \in \mathbb{R}^{N\times 3}$, we adopt Point-BERT's tokenization strategy (Yu et al. 2022). Initially, Farthest Point Sampling (FPS) selects n center points to ensure

[1]All channels share identical depth values in this representation

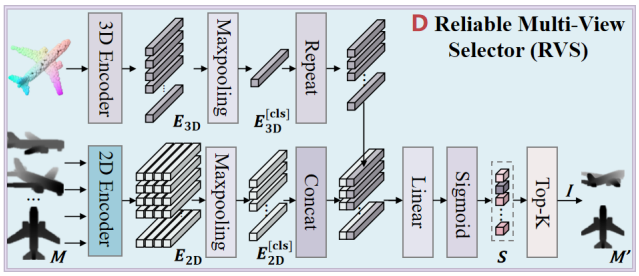

Figure 2: **Reliable Multi-View Selector**. Global [cls] token embeddings are obtained via max-pooling over patch features. These tokens are concatenated and fed through an MLP layer followed by a sigmoid activation, generating cross-modal correlation scores. The Top-K views with the highest scores are selected for downstream fusion.

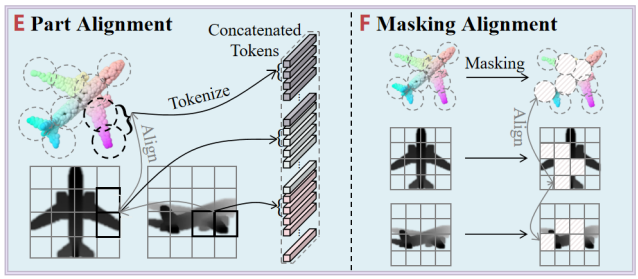

Figure 3: **Part and Masking Alignment**. Part Alignment achieves cross-modal token alignment through mapping relationship between point clouds and multi-view images. Masking Alignment propagates the 3D masking to 2D multi-view images via geometric projection onto spatially correlated positions.

uniform spatial coverage. Subsequently, K-Nearest Neighbors (KNN) gathers p neighboring points around each center, forming n local geometric patches that encapsulate local structural features. Finally, each patch is processed by a lightweight PointNet (Qi et al. 2017a) to generate per-patch embeddings. We simply formulate it as:

$$E_{3D} = \text{PointNet}(P), \quad E_{3D} \in \mathbb{R}^{n \times d^E} \tag{3}$$

Positional encodings of center points are added to embed spatial locality, analogous to the 2D positional encoding mechanism. The output embeddings E^i_{2D} and E_{3D} share identical dimensionality d^E, creating a compatible vector space for subsequent transformer-based part-aligned cross-attention. This symmetry is critical because it enables geometric and semantic alignment between 2D visual features and 3D geometric structures, forming the foundation for joint feature learning.

Reliable Multi-View Selector

While multi-view images provide complementary structural information, certain perspectives are redundant or even counterproductive. Such detrimental redundancy can propagate erroneous spatial cues during cross-modal fusion, ultimately degrading the fidelity of point cloud feature learning. To mitigate this, we propose a Reliable Multi-View Selector module that dynamically identifies and retains the most geometrically consistent views, as illustrated in Figure 2.

The core of RVS is a relation consistency scoring mechanism. We compute the relevance score between the point cloud and the i-th view as:

$$[E^{[cls]}_{3D}; E^{i}_{2D}{}^{[cls]}] = \text{Maxpooling}([\text{E}_{3D}; \text{E}^{i}_{2D}]) \tag{4}$$

$$s_i = \text{Sigmoid}(\text{MLP}([E^{[cls]}_{3D}; E^{i}_{2D}{}^{[cls]}])) \tag{5}$$

where $E^{[cls]}_{3D}, E^{i}_{2D}{}^{[cls]} \in \mathbb{R}^{d^E}$ are global classification token embeddings derived via max pooling over multi-view 2D and 3D embedding. $\text{MLP}(\cdot)$ denotes a multilayer perceptron that projects the concatenated embedding to a scalar. $\text{Sigmoid}(\cdot)$ normalizes the output to $s_i \in [0, 1]$, quantifying geometric alignment confidence. The resultant scores $S = [s_1, s_2, \ldots, s_v]$ are used to rank views and select the Top-K highest-scoring ones. This operation is defined as:

$$I = \{i \in \{1, 2, \ldots, v\} \mid |\{j : s_j > s_i\}| < k\} \tag{6}$$

where I denotes the index set of the Top-K views. The filtered set is then constructed as $M' = \{m_j \mid j \in I\}$ with $|M'| = k$. This selective fusion ensures that only views exhibiting high structural correlation with the point cloud are propagated downstream. The refined view set M' and point cloud features are then forwarded to the RM and KAT for cross-modal alignment.

Reinforced Masking

Conventional random masking uniformly samples tokens to obscure, which risks concealing critical spatial structures while introducing insignificant noise to the decoder. This forces the model to overfit decoder reconstructions—a conflict with MAE's core design, where the lightweight decoder is discarded during fine-tuning. To resolve this misalignment, our Reinforced Masking (RM) employs reinforcement learning strategy to selectively sample important point cloud regions, ensuring reconstruction focuses on structurally informative parts for downstream task transfer.

Given point cloud tokens E_{3D}, we model the token significance distribution via a lightweight Multi-Head Attention (MHA) (Vaswani 2017), which collectively derive adaptive sampling probabilities P_S for visible token selection:

$$P_S = \text{Softmax}\left(\text{Linear}(\text{MHA}(E_{3D}))\right), \quad P_S \in \mathbb{R}^n \tag{7}$$

Here, $P_S = [p_1, p_2, \ldots, p_n]$ satisfies $\sum_{i=1}^{n} p_i = 1$, with p_i quantifying the significance of tokens E_{3D} for downstream sampling tasks. To dynamically sample the visible token set, we define an n-dimensional categorical distribution (Marriott 1990) over P_S ($p \sim \text{Categorical}(n, P_S)$) and perform without replacement sampling to obtain indices I_v and its complement I_m. The sample size $n_v = n \times (1 - \rho)$ is controlled by a predefined masking ratio $\rho \in (0, 1)$. Crucially, probability sampling (vs. deterministic Top-K) injects controlled stochasticity during training, preventing overfitting to fixed token subsets and enhancing model robustness.

The RM network implements a reinforcement learning framework (Williams 1992), formulating token selection as a sequential decision process. To be more specific, under the MAE environment, we treat the selection of mask regions as an action and utilize the reconstruction error as the reward. The objective function of RM is formulated as:

$$\mathcal{L}_{\mathrm{S}} = -\mathbb{E}_{I_m \sim P}\left[\mathcal{L}_{\mathrm{R}}\right] = -\sum_{i \in I_m} \log p_i \cdot \mathcal{L}_{\mathrm{R}}^i \tag{8}$$

Notably, logarithmic probabilities $\log p_i$ enhance backpropagation stability. Through the training loss $\mathcal{L}_{\mathrm{S}}$, we can transform non-differentiable sampling into a trainable process that progressively prioritizes informative critical tokens.

Through Reinforced Masking, we refine the 3D mask. To propagate 3D structural priors to multi-view images, we apply Part and Masking Alignment (PMA) illustrated in Figure 3. The PMA comprises two stages: Part Alignment establishes cross-modal token correspondence by aligning point cloud parts with multi-view image regions; Masking Alignment transfers the 3D mask to perspective images by projecting it onto corresponding 2D spatial positions.

Reliable-view 2D-3D Key-part Aligned Transformer

For the adaptively selected masks and views, we further propose a Reliable-view 2D-3D Key-part Aligned Transformer (KAT) to establish structural correspondence between key regions selected by RM of point cloud and multi-view images. KAT leverages rich texture cues from multi-view 2D images to compensate for sparse geometry in point cloud.

To model intra-modality relationships within visible point cloud features, we first apply self-attention on the visible 3D tokens $E_{\text{3D}}^{\text{vis}} \in \mathbb{R}^{n_v \times d^E}$. The query Q_{3D}, key K_{3D}, and value V_{3D} projections are computed as:

$$Q_{\text{3D}} = E_{\text{3D}}^{\text{vis}} W_Q\,,\ \ K_{\text{3D}} = E_{\text{3D}}^{\text{vis}} W_K\,,\ \ V_{\text{3D}} = E_{\text{3D}}^{\text{vis}} W_V \tag{9}$$

where $W_Q, W_K \in \mathbb{R}^{d^E \times d'}$ and $W_V \in \mathbb{R}^{d^E \times d^E}$ are learnable projection matrices. The updated 3D feature is obtained via scaled dot-product attention:

$$T_{\text{3D-Self}}^{\text{vis}} = \mathrm{softmax}(\frac{Q_{\text{3D}}(K_{\text{3D}})^\top}{\sqrt{d'}})V_{\text{3D}} \tag{10}$$

Similarly, for multi-view image features $E_{\text{2D}}^{\text{vis}} \in \mathbb{R}^{n_v \times d^E}$, the query, key and value is computed as:

$$Q_{\text{2D}} = E_{\text{2D}}^{\text{vis}} W_Q\,,\ \ K_{\text{2D}} = E_{\text{2D}}^{\text{vis}} W_K\,,\ \ V_{\text{2D}} = E_{\text{2D}}^{\text{vis}} W_V \tag{11}$$

However, unlike standard cross-attention that only change self-attention for 3D features as Q, 2D features as K and V , we propose a key-part aligned fusion mechanism. After acquiring point cloud keypoints via RM, KAT selectively attends to geometrically correlated regions in multi-view images for feature fusion. This is implemented by a binary alignment mask $I_{\text{align}} \in \{0,1\}^{n_v \times n_v}$, where:

$$I_{\text{align}}(i,j) = \begin{cases} 1 & \text{if point } i \text{ projects to patch } j \text{ in any view} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Then the multi-view 2D-3D key-part aligned cross-modal attention is formulated as:

$$[T_{\text{3D-Cross}}^{\text{vis}}]^i = [\mathrm{softmax}(\frac{Q_{\text{3D}}(K_{\text{2D}})^\top}{\sqrt{d'}} \odot I_{\text{align}})V_{\text{2D}}]^i \tag{13}$$

where $\odot$ represents the Hadamard product and i is the image of the i-th perspective. By virtue of the proposed Part and Masking Alignment strategy, the binary mask I_{align} becomes equivalent to the identity matrix I_{n_v}. Consequently, the final KAT representation integrates intra-modal features with geometrically aligned cross-modal components:

$$\begin{aligned} T_{\text{3D}}^{\text{vis}} &= T_{\text{3D-Self}}^{\text{vis}} + \sum_{i=1}^{k} [T_{\text{3D-Cross}}^{\text{vis}}]^i \\ &= \mathrm{softmax}(\frac{Q_{\text{3D}}(K_{\text{3D}})^\top}{\sqrt{d'}})V_{\text{3D}} \\ &+ \sum_{i=1}^{k} [\mathrm{softmax}(\frac{Q_{\text{3D}}(K_{\text{2D}})^\top}{\sqrt{d'}} \odot I_{\text{n}_v})V_{\text{2D}}]^i \end{aligned} \tag{14}$$

where k is the number of refined views from RVS. Through Reliable-view 2D-3D Key-part Aligned Transformer, we enforce geometry-aware cross-modal interaction by confining attention to locally aligned regions. This spatial sparsification, governed by the binary mask I_{align}, yields two fundamental advantages: fine-grained feature aggregation where each 3D token selectively attends to semantically congruent 2D patches via projective correspondence, intensively integrating high-frequency texture details, while simultaneously achieving distraction suppression by pruning non-relevant cross-modal interactions through masking geometrically unrelated positions, thereby eliminating noise propagation from heterogeneous structures.

Training Objective

The decoder of MAE process will be discarded during the fine-tuning stage, so we do not specifically design the decoder. Following Point-MAE (Pang et al. 2022), we use the Standard Transformer (ST) and a reconstruction head (Rec) to obtain the reconstructed point cloud.

$$P_{\text{pred}} = \mathrm{Rec}(\mathrm{ST}(T_{\text{3D}}^{\text{vis}})) \tag{15}$$

The primary point cloud MAE objective reconstructs spatial coordinates by minimizing the symmetric Chamfer Distance (Fan, Su, and Guibas 2017) between predicted patches P_{pred} and ground truth P_m:

$$\mathcal{L}_{\mathrm{R}} = \frac{1}{|P_{\text{pred}}|} \sum_{a \in P_{\text{pred}}} \min_{b \in P_m} \|a-b\|_2^2 + \frac{1}{|P_m|} \sum_{b \in P_m} \min_{a \in P_{\text{pred}}} \|a-b\|_2^2 \tag{16}$$

The joint optimization objective integrates $\mathcal{L}_{\mathrm{R}}$ and $\mathcal{L}_{\mathrm{S}}$ through a balancing coefficient λ, formulated as: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathrm{R}} + \lambda \mathcal{L}_{\mathrm{S}}$. This co-adaptation mechanism enables robust feature learning from adaptively selected tokens while dynamically refining token significance estimation. Consequently, it suppresses low-information pathways and meaningless feature propagation, achieving efficient learning of geometrically consistent representations.

Method	OBJ-BG	OBJ-ONLY	PB-T50-RS
Supervised Learning Only			
PointNet (Qi et al. 2017a)	73.3	79.2	68.0
PointNet++ (Qi et al. 2017b)	82.3	84.3	77.9
DGCNN (Wang et al. 2019)	82.8	86.2	78.1
SimpleView (Goyal et al. 2021)	-	-	80.5
PointMLP (Ma et al. 2022)	-	-	85.2
Single-Modal Self-Supervised Learning			
Point-BERT (Yu et al. 2022)	87.43	88.12	83.07
Point-MAE (Pang et al. 2022)	90.02	88.29	85.18
Point-M2AE (Zhang et al. 2022)	91.22	88.81	86.43
PointGPT (Chen et al. 2024)	91.60	90.00	86.90
PointDif (Zheng et al. 2024)	93.29	91.91	87.61
Cross-Modal Self-Supervised Learning			
ACT (Dong et al. 2022)	93.29	91.91	88.21
TAP (Wang et al. 2023)	90.36	89.50	85.67
Joint-MAE (Guo et al. 2023)	90.94	88.86	86.07
MM-Point (Yu and Song 2024)	-	-	87.80
KR-MAE	**94.45**	**91.91**	**89.28**
Improvement	+4.43	+3.62	+4.10

Table 1: **Classification accuracy on ScanObjectNN**. Three variants are evaluated on the ScanObjectNN dataset and we record the improvement over Point-MAE. All values represent accuracy (%).

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
Supervised Learning Only				
PointNet	52.0±3.8	57.8±4.9	46.6±4.3	35.2±4.8
DGCNN	31.6±2.8	40.8±4.6	19.9±2.1	16.9±1.5
OcCo	90.6±2.8	92.5±1.9	82.9±1.3	86.5±2.2
Single-Modal Self-Supervised Learning				
Point-BERT	94.6±3.1	96.3±2.7	91.0±5.4	92.7±5.1
MaskPoint	95.0±3.7	97.2±1.7	91.4±4.0	93.4±3.5
Point-MAE	96.3±2.5	97.8±1.8	92.6±4.1	95.0±3.0
Point-M2AE	96.8±1.8	98.3±1.4	92.3±4.5	95.0±3.0
Cross-Modal Self-Supervised Learning				
ACT	96.8±2.3	98.0±1.4	93.3±4.0	95.6±2.8
Joint-MAE	96.7±2.2	97.9±1.8	92.6±3.7	95.1±2.6
I2P-MAE	97.0±1.8	98.3±1.3	92.6±5.0	95.5±3.0
MM-Point	96.5±2.8	97.2±1.4	90.3±2.1	94.1±1.9
KR-MAE	**96.6±1.7**	**98.3±1.6**	**93.0±4.3**	**95.2±2.8**
Improvement	+0.3	+0.5	+0.4	+0.2

Table 2: **Few-shot classification results on ModelNet40.** We conduct 10 independent experiments and report the average classification accuracy (%) with the standard deviation.

Experiments

Pre-training Setup

Following existing works (Yu et al. 2022; Pang et al. 2022), we pre-train our KR-MAE on the ShapeNet dataset (Chang et al. 2015). We randomly sample 1,024 points per object as our 3D point cloud input. To generate multi-view images, we adopt the rendering technique used in PointCLIP V2 (Zhu et al. 2023) and the resolution is set as 112×112.

Fine-tuning Setup

After pre-training, the reconstruction decoder is removed, and only the 3D encoder is utilized for downstream task fine-tuning. Following common practice in prior works (Pang et al. 2022; Liu, Cai, and Lee 2022; Yu et al. 2022), we initialize the encoder parameters with those obtained during pre-training as the starting point for fine-tuning.

Classification. To demonstrate the adaptability of our model to real-world scenarios, we evaluate its classification performance on ScanObjectNN (Uy et al. 2019), a challenging dataset comprising approximately 15,000 real-scanned objects across 15 categories. We conduct experiments under its three distinct settings: OBJ-BG, OBJ-ONLY, and PB-T50-RS. As summarized in Table 1, our KR-MAE model achieves exceptional performance. It outperforms Point-MAE significantly, with improvements of 4.43%, 3.62%, and 4.10% on the three variants respectively, and demonstrates competitive results against other sophisticated multi-modal methods.

Few-shot Learning. Following standard n-way m-shot protocols, we evaluate KR-MAE on ModelNet40 (Wu et al. 2015), where n denotes sampled categories and m specifies support set size per category. For each task, models are trained on $n \times m$ samples and evaluated on 20 queries per category, with results averaged over 10 trials. As Table 2 shows, KR-MAE outperforms state-of-the-art methods (including Point-MAE) across four settings, achieving accuracy gains of 0.3%, 0.5%, 0.4% and 0.2%, demonstrating superior transferability under data scarcity.

Part Segmentation. KR-MAE's fine-grained representation capability is evaluated on ShapeNetPart (Yi et al. 2016) (16,881 objects, 50 part labels). Following (Pang et al. 2022), we sample 2,048 points partitioned into 128 patches. As Table 3 indicates, KR-MAE achieves 0.2% higher class-average mIoU_C than Point-MAE, highlighting its ability to capture semantically consistent part structures under geometric variations, proving enhanced geometric-semantic consistency for part segmentation.

Ablation Studies

Main components in the KR-MAE. Compare to Point-MAE which only performs point cloud reconstruction, our KR-MAE has two main differences: 1) Masking strategy: we use RM to preserve geometrically salient point cloud regions. 2) Multi-view 2D input: we incorporates multi-view images to enhance point cloud feature learning, with RVS to eliminate view redundancy and KAT to establish semantic alignment between 3D point cloud key parts and multi-view 2D image patches. To validate the effectiveness of RM, RVS and KAT, four experimental configurations progressively integrate these components as shown in Table 4. A adopts Point-MAE as the baseline, B incorporates multi-view 2D input integrated via KAT for cross-modal feature fusion, C further introduces RVS to refine informative imaging perspectives, and D represents the complete KR-MAE framework integrating all components. Additionally, config-

Method	Reference	mIoU_c	mIoU_I
Supervised Learning Only			
PointNet (Qi et al. 2017a)	CVPR'17	80.4	83.7
PointNet++ (Qi et al. 2017b)	NeurIPS'17	81.9	85.1
PointMLP (Ma et al. 2022)	ICLR'22	84.6	86.1
Single-Modal Self-Supervised Learning			
Transformer (Vaswani 2017)	NeurIPS'17	83.4	84.7
OcCo (Wang et al. 2021)	ICCV'21	83.4	85.1
Point-BERT (Yu et al. 2022)	CVPR'22	84.1	85.6
Point-MAE (Pang et al. 2022)	ECCV'22	84.2	86.1
PointGPT (Chen et al. 2024)	NeurIPS'23	84.1	86.2
Cross-Modal Self-Supervised Learning			
ACT (Dong et al. 2022)	ICLR'23	84.7	86.1
Recon (Qi et al. 2023)	ICML'23	84.8	86.4
MM-Point (Yu and Song 2024)	AAAI'24	-	85.7
KR-MAE		**84.4**	**86.2**
Improvement		+0.2	+0.1

Table 3: **Part segmentation results on the ShapeNetPart.** Mean intersection over union for all classes mIoU_C (%) and all instances mIoU_I (%) are reported.

	KAT	RVS	RM	views	OBJ-BG(%)
A	-	-	-	-	92.94
B	✓	-	-	3	93.46
C	✓	✓	-	3	93.98
D	✓	✓	✓	3	94.45
E	✓	✓	✓	1	93.98
F	✓	✓	✓	3	**94.45**
G	✓	✓	✓	6	94.15
H	✓	✓	✓	9	93.98

Table 4: Effects of the main components and view numbers in the proposed KR-MAE.

urations E–H configure varying selected numbers of RVS to determine the optimal viewpoint quantity for KR-MAE.

Table 4 reveals three core findings: removing multi-view input reduces accuracy by 0.52%, confirming its critical contribution. The incremental addition of RVS and RM each yields approximately 0.5% gains, demonstrating their efficacy in eliminating redundancy and preserving structural integrity. Performance peaks at 3 views, with slight declines observed at fewer or more views, indicating KR-MAE's robustness to view-number variations.

Furthermore, we quantify the pre-training losses of our masking strategy in Table 5. The reconstruction error of our method remains consistently lower than that of random masking across mask ratios ranging from 0.5 to 0.9. This demonstrates the superiority of reconstruction performance and implies enhanced representation learning efficacy.

Effects of the multi-view 2D images and 3D point cloud feature fusion method. The Reliable-view 2D-3D Key-part Aligned Transformer employ a part-aligned cross-attention mechanism to establish precise geometric correspondence between modalities. Specifically, this mechanism constrains each 3D token to attend exclusively to 2D patches sharing identical spatial regions, which simultaneously achieves high-frequency texture integration while minimizing misaligned feature interference through geometric constraints.

Masking	0.5	0.6	0.7	0.8	0.9
Random	0.869	0.872	0.876	0.882	0.889
Reinforced	0.790	0.803	0.812	0.836	0.866
	-0.079	-0.069	-0.064	-0.046	-0.023

Table 5: The pre-training losses of random masking and our reinforced masking under different mask ratios.

	2D-Input	Fusion Method	OBJ-BG(%)
A	✓	MLP	92.60
B	✓	Self-Attention	93.12
C	✓	Cross-Attention	94.15
D	✓	Aligned-Attention	**94.45**

Table 6: Effects of the multi-view 2D images and 3D point cloud feature fusion method.

To evaluate the significance of this design, we conducted comparative experiments replacing it with three alternative fusion strategies as shown in Table 6. A: MLP fusion directly concatenates features but suffers severe performance degradation due to inherent modality incompatibility. B: while self-attention captures intra-modal relationships, it fails to establish meaningful cross-modal interactions. C: standard cross-attention accommodates inter-modal fusion but permits attention to geometrically unrelated regions, resulting in suboptimal performance.

Quantitatively, our part-aligned attention outperforms these alternatives by 0.30%-1.85%, with the critical advantage stemming from its dual mechanisms: geometric constraints via binary mask I_{align} prune non-relevant cross-modal interactions, while spatially focused attention prevents propagation of heterogeneous structural noise. Consequently, this approach uniquely balances texture detail preservation with structural consistency maintenance.

Conclusions

This study proposes KR-MAE, a multi-view masked autoencoder overcoming single-modal limitations via point cloud and multi-view images fusion. Three innovations establish geometric cross-modal consistency, with Reliable Multi-View Selector (RVS) optimizing perspectives through relation scores between point cloud and multi-view images, Reinforced Masking (RM) focusing computation on salient regions, and Reliable-view 2D-3D Key-part Aligned Transformer (KAT) leveraging 3D-to-2D geometric correspondence for part-aware feature fusion. Extensive experiments validate KR-MAE's superior 3D point cloud representation capability. We expect this work to inspire unified frameworks combining multi-modality learning with 3D point cloud MAE in the future.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62406250, 62236001, 62576277 and the Fundamental Research Funds for the Central Universities.

References

Afham, M.; Dissanayake, I.; Dissanayake, D.; Dharmasiri, A.; Thilakarathna, K.; and Rodrigo, R. 2022. 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*, 9902–9912.

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

Chen, A.; Zhang, K.; Zhang, R.; Wang, Z.; Lu, Y.; Guo, Y.; and Zhang, S. 2023. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5291–5301.

Chen, G.; Wang, M.; Yang, Y.; Yu, K.; Yuan, L.; and Yue, Y. 2024. Pointgpt: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems*, 36.

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

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fan, H.; Su, H.; and Guibas, L. J. 2017. 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*, 605–613.

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

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

Lin, X.; Lin, T.; Huang, L.; Xie, H.; and Su, Z. 2025. Bip3d: Bridging 2d images and 3d perception for embodied intelligence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9007–9016.

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

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

Marriott, F. H. 1990. *A Dictionary of Statistical Terms*. Longman Scientific & Technical, 5 edition. Accessed via secondary sources.

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

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

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

Qi, Z.; Dong, R.; Fan, G.; Ge, Z.; Zhang, X.; Ma, K.; and Yi, L. 2023. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, 28223–28243. PMLR.

Qian, R.; Lai, X.; and Li, X. 2022. 3D object detection for autonomous driving: A survey. *Pattern Recognition*, 130: 108796.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ren, T.; and Jebelli, H. 2025. Efficient 3D robotic mapping and navigation method in complex construction environments. *Computer-Aided Civil and Infrastructure Engineering*, 40(12): 1580–1605.

Su, K.; Wu, Q.; Cai, P.; Zhu, X.; Lu, X.; Wang, Z.; and Hu, K. 2025. RI-MAE: Rotation-Invariant Masked AutoEncoders for Self-Supervised Point Cloud Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7015–7023.

Tang, Y.; Han, X.; Li, X.; Yu, Q.; Xu, J.; Hao, Y.; Hu, L.; and Chen, M. 2025. More text, less point: Towards 3d data-efficient point-language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7284–7292.

Uy, M. A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, T.; and Yeung, S.-K. 2019. 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*, 1588–1597.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.

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

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3): 229–256.

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

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 574–591. Springer.

Yi, L.; Kim, V. G.; Ceylan, D.; Shen, I.-C.; Yan, M.; Su, H.; Lu, C.; Huang, Q.; Sheffer, A.; and Guibas, L. 2016. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6): 1–12.

Yu, H.-T.; and Song, M. 2024. Mm-point: Multi-view information-enhanced multi-modal self-supervised 3d point cloud understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6773–6781.

Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2022. 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*, 19313–19322.

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

Zhang, R.; Wang, L.; Qiao, Y.; Gao, P.; and Li, H. 2023. 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*, 21769–21780.

Zhang, X.; Zhang, S.; and Yan, J. 2024. PCP-MAE: Learning to Predict Centers for Point Masked Autoencoders. *arXiv preprint arXiv:2408.08753*.

Zheng, X.; Huang, X.; Mei, G.; Hou, Y.; Lyu, Z.; Dai, B.; Ouyang, W.; and Gong, Y. 2024. Point cloud pre-training with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22935–22945.

Zhou, Z.; Wang, P.; Liang, Z.; Bai, H.; and Zhang, R. 2025. Cross-Modal 3D Representation with Multi-View Images and Point Clouds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3728–3739.

Zhu, X.; Zhang, R.; He, B.; Guo, Z.; Zeng, Z.; Qin, Z.; Zhang, S.; and Gao, P. 2023. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2639–2650.