Pic@Point: Cross-Modal Learning by Local and Global Point-Picture Correspondence

Vencia Herzog VENCIA.HERZOG@RENUMICS.COM
Karlsruhe Institute of Technology
Renumics GmbH, Karlsruhe, Germany

Stefan Suwelack STEFAN.SUWELACK@RENUMICS.COM
Renumics GmbH, Karlsruhe, Germany

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Self-supervised pre-training has achieved remarkable success in NLP and 2D vision. However, these advances have yet to translate to 3D data. Techniques like masked reconstruction face inherent challenges on unstructured point clouds, while many contrastive learning tasks lack in complexity and informative value. In this paper, we present *Pic@Point*, an effective contrastive learning method based on structural 2D-3D correspondences. We leverage image cues rich in semantic and contextual knowledge to provide a guiding signal for point cloud representations at various abstraction levels. Our lightweight approach outperforms state-of-the-art pre-training methods on several 3D benchmarks.

Keywords: self-supervised pre-training; cross-modal; 2D-3D correspondence; point clouds

1. Introduction

Point clouds are the preferred 3D representation for many applications, including autonomous driving, robotics, AR/VR, and various sensor technologies (LIDAR, SFM, Kinect Structured Light, etc). However, the annotation of point clouds is associated with high costs, and labeled 3D scans are scarce. This poses an obstacle to the development of robust, scalable deep learning models for point cloud analysis.

Recent progress in self-supervised learning has led to significant advances in the areas of image recognition and natural language processing (NLP). By using parts of the input data itself as a guiding signal, self-supervised learning bypasses the annotation bottleneck associated with training large neural networks. This strategy reflects a shift from traditional task-specific training towards pre-training general-purpose representations that are applicable across a wide range of tasks. Extending these advances to 3D data remains challenging due to factors such as irregular information density and the unordered nature of point clouds.

Generative modeling has seen a number of adaptations to 3D data using transformer-style architectures (Yu et al., 2022; Pang et al., 2022; Zhang et al., 2022a). However, these approaches have yet to replace traditional architecture-based methods (Ma et al., 2022; Qian et al., 2022a). Two factors that contribute to the underperformance on point clouds are the lack of inductive biases in Standard Transformers and the use of point set similarity metrics (e.g., Chamfer Distance) in reconstruction, which are imprecise and hard to optimize (Wu et al., 2021; Huang et al., 2023).

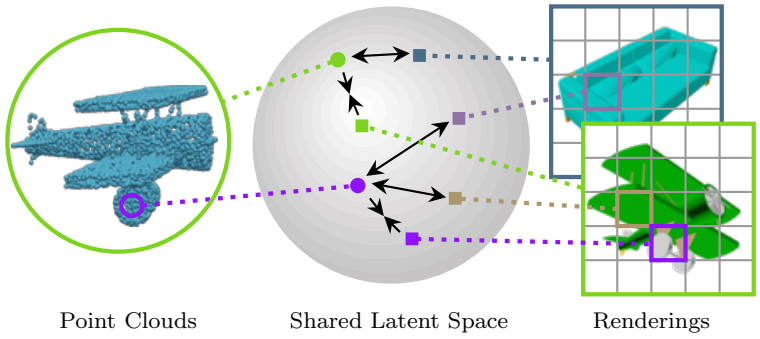

Figure 1: **Illustration of the proposed structural 2D-3D correspondence.** Point cloud and image features of various scales are projected into a shared, representation-invariant latent space, where cross-modal correspondence is enforced.

Contrastive learning approaches, on the other hand, typically aim at learning invariances to transformations or augmentations. They have been shown to be highly competitive to generative modeling (Oord et al., 2018; Chen et al., 2020). However, the quality of learned representations is highly dependent on the complexity and informative value of the contrastive task (Goyal et al., 2019), and many contrastive models only consider global views, neglecting local relationships (Qi et al., 2023).

To address these limitations, we propose leveraging 2D-3D correspondences for contrastive representation learning on point clouds. We present the **Pic@Point** model, which aims to learn point cloud representations by exploiting structural features of ***pic***tures ***at*** various ***point*** cloud abstraction levels. Specifically, we extract 3D and 2D features at both global and local scales using a generic 3D backbone and a pre-trained 2D backbone, respectively. We then employ global and local, pose-conditioned projection heads to project these features into a common, representation-invariant latent space, as illustrated in Figure 1. This structural 2D-3D contrastive learning approach offers several advantages over existing methods:

- We effectively leverage features from pre-trained vision foundation models. In contrast, generative cross-modal methods (Zhang and Hou, 2023; Wang et al., 2023) generate images from input point clouds using a custom 2D generator atop a 3D backbone, as opposed to integrating a powerful vision model.
- In addition, our method is very lightweight while being highly effective, as it uses no decoder and employs a frozen 2D backbone. Figure 2 shows the size of different pre-training models in relation to linear accuracy on ModelNet40, showcasing the efficiency of our approach.
- Unlike existing contrastive methods (Afham et al., 2022; Wu et al., 2023) which only learn global shape correspondences, Pic@Point provides guidance on a structural level. By exploiting local correspondences, it provides pose-aware, positional guidance while benefiting from a larger number of contrastive samples.

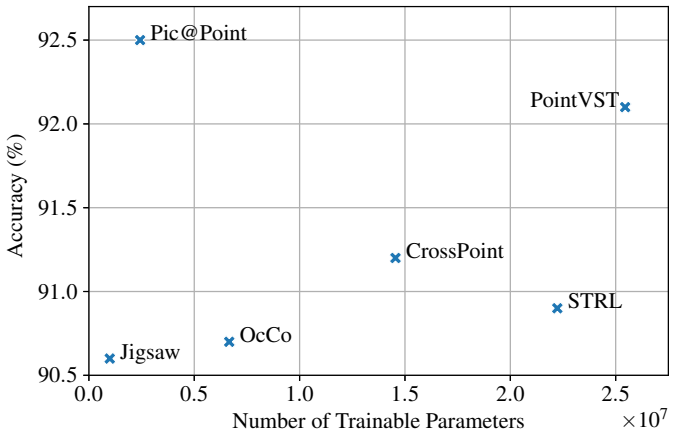

Figure 2: **Size of Pre-training Model vs. Linear SVM Accuracy on ModelNet40.** Reported is the number of trainable parameters of self-supervised pre-training models with DGCNN backbone.

2. Related Work

Point Cloud Analysis Early efforts to apply neural networks to points clouds involved converting the input into structured forms such as voxel grids (Wu et al., 2015; Maturana and Scherer, 2015) or multi-view images (Su et al., 2015). Sparse voxel-based methods (Riegler et al., 2017; Choy et al., 2019) remain a prevalent approach for large-scale scene analysis.

With the introduction of PointNet (Qi et al., 2017a), point-based methods began to apply neural architectures directly to raw point sets without any conversion or preprocessing steps. A key concern is ensuring permutation invariance, often addressed by aggregating information using symmetric functions. Deep set architectures can be divided into three categories: MLP-methods (Qi et al., 2017a,b; Wang et al., 2019; Ma et al., 2022), convolution-based methods (Li et al., 2018; Thomas et al., 2019), and attention-based methods (Yu et al., 2022; Pang et al., 2022; Zhang et al., 2022a).

Self-supervised Point Cloud Pre-training Pre-training is a widely used transfer learning approach in which a model is initially trained on a *pretext* task and subsequently fine-tuned on a *downstream* task. Self-supervised pre-training does not use any labels during the pre-training phase, which enables a wide scope of available pre-training data.

Pretext tasks are typically either generative or contrastive. A common generative approach is autoencoding, where the input is recovered under some form of corruption (Chen et al., 2021) or masking (Wang et al., 2021). Transformer architectures (Devlin et al., 2018; He et al., 2022) combine masked reconstruction with multi-head self-attention mechanisms. Point-BERT (Yu et al., 2022) and Point-MAE (Pang et al., 2022) adapt language and vision Transformers to point clouds. Generative models typically reconstruct directly in point cloud space, which is computationally expensive and hard to optimize (Wu et al., 2021; Huang et al., 2023).

Contrastive learning uses available pairs of similar and dissimilar data points to learn an embedding space where the distance between data points reflects a measure of their similarity. This is done by contrasting a sample with augmented versions (He et al., 2020; Chen et al., 2020) or by capturing the relationship between local features and their global context (Oord et al., 2018; Hjelm et al., 2018).

Contrastive methods benefit from diverse, informative data to mitigate overfitting issues and counterbalance representation deficits. To this end, we propose a cross-modal method leveraging image cues rich in structural and semantic context.

Cross-Modal 3D Representation Learning The potential of joint 2D-3D learning has been recognized by previous works. One line of work adapts pre-trained vision and language models for 3D point cloud analysis using strategies such as prompt tuning (Zhang et al., 2022b; Wang et al., 2022), 2D-to-3D architectural modification (Xu et al., 2022; Qian et al., 2022b) and knowledge distillation (Dong et al., 2022; Qi et al., 2023) techniques. These methods cannot be directly applied to generic 3D backbones because they require specialized architectures or adaptation processes, often having large memory requirements and difficulties with 3D extensibility.

A different approach is to design cross-modal pretext tasks directly on 3D models, allowing the use of generic 3D backbones for downstream tasks. PointVST (Zhang and Hou, 2023) and TAP (Wang et al., 2023) employ a generative cross-modal approach that generates images from point clouds at specified camera views, but they lack knowledge integration from pre-trained vision models. CrossPoint (Afham et al., 2022) performs cross-modal and inter-modal contrastive learning to align point clouds with 2D renderings and with augmented versions. They leverage only global correspondences, resulting in a coarse guiding signal. Conversely, point-pixel level correspondence methods such as Tran et al. (2022) require costly upsampling layers and loss computation, while being unnecessarily fine-grained for learning meaningful contextual relationships.

3. Proposed Method

We propose structural 2D-3D correspondence learning for self-supervised pre-training of point cloud representations. The importance of incorporating structural knowledge has been demonstrated in prior uni-modal research (Hjelm et al., 2018; Oord et al., 2018; Rao et al., 2020). Our novel cross-modal approach enriches point cloud information with structured, semantic image cues to provide a comprehensive guiding signal for 3D understanding.

3.1. Overview

Let $\mathcal{D} = \{(P_i, I_i)\}_{i=1}^{|\mathcal{D}|}$ denote an unlabeled dataset of point clouds $P_i \in \mathbb{R}^{N\times 3}$, where N is the number of points, and shape renderings $I_i \in \mathbb{R}^{H\times W\times 3}$ of size $H \times W$. Rendering I_i is captured from a random camera view point with view matrix $M_{view} \in \mathbb{R}^{3\times 3}$ and projection matrix $M_{proj} \in \mathbb{R}^{3\times 3}$. Figure 3 depicts the overall architecture of our proposed Pic@Point model. It consists of the following modules: 1) a *3D Backbone* extracts local and global point cloud features, obtained as top-level positional features f^{3d} and final shape embeddings $\mathcal{A}(f^{3d})$, after a global pooling function $\mathcal{A}$, 2) a frozen *2D Backbone* returns top-level local and global image features as extracted by a pre-trained vision model (e.g.,

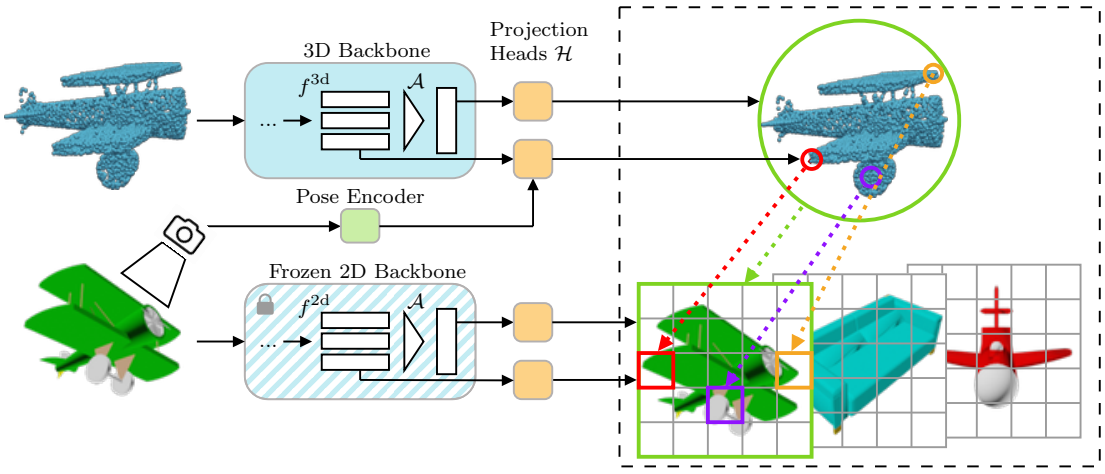

Figure 3: **Overview of the proposed Pic@Point workflow.** We extract top-level local and aggregated global point cloud and image features using a generic 3D backbone and a frozen 2D backbone. Subsequently, the features are processed through global and local, pose-conditioned projection heads $\mathcal{H}$.

ResNet, ViT), denoted as f^{2d} and $\mathcal{A}(f^{2d})$, and 3) the *Projection Heads* $\mathcal{H}$ project the point cloud and image features into a shared, representation-invariant latent space $\chi \subseteq \mathbb{R}^d$.

3.2. Projection into Shared Latent Space

To unify knowledge and produce rich, transferable representations, we project the modal features into a shared latent space, where a contrastive loss is applied between projected features within the mini-batch. The global projection heads $\mathcal{H}^{3d}_{glb}, \mathcal{H}^{2d}_{glb}$ consist of simple multi-layer perceptrons (MLPs) with two layers. The local projection heads $\mathcal{H}^{3d}_{lcl}$, $\mathcal{H}^{2d}_{lcl}$ use Conv1d and Conv2d layers with kernel size 1 and stride 1, respectively, to apply transformations in the channel dimension while preserving the spatial dimensions. All projected features are L2-normalized.

Pose Encoding We facilitate spatially aware correspondence between local point cloud and image features by integrating a pose encoding into the local point cloud projection head $\mathcal{H}^{3d}_{lcl}$. This is necessary because while it is assumed that a complete shape can be uniquely identified in a representation of any modality, this may not hold true for a local shape region. For instance, symmetric features like the red circled plane wing tip depicted in Figure 1 could potentially belong to any of the image snippets depicting a plane wing tip, if no additional pose information is given. The pose encoding is calculated via a two-layer MLP that transforms M_{view} into a 64-dimensional vector that is concatenated to the output of the first convolutional layer in $\mathcal{H}^{3d}_{lcl}$.

3.3. Cross-Modal Correspondence Task

Point-to-Pixel Mapping For the local cross-modal correspondence task, we begin by establishing the ground truth mapping from points to image positions using projective transformations with $M_{view}, M_{proj} \in \mathbb{R}^{4\times4}$. Given a local point cloud embedding z_l, let $(u, v) \in [0, 1]^2$ denote the image position projected from its center point. With top-level downsampled image regions of size 7×7, such as produced by ResNet (He et al., 2016), the indexed position (i, j) of the corresponding image region embedding q_{ij} is determined as $(i, j) = \lfloor (u \cdot 7, v \cdot 7) \rfloor$.

Subsequently, a contrastive learning objective can be applied to the cross-modal correspondences. For each of the L local point cloud region embeddings $\{z_l\}_{l=1,\dots,L}$, we pull close the corresponding image region embedding q^+ of the same object, while pushing away all other image region embeddings $\left\{q_{ij}^k \neq q^+\right\}_{i,j=1,\dots,7}^{k=1,\dots,m}$ within the mini-batch of size m. To achieve this, we apply the InfoNCE loss function (Oord et al., 2018) with temperature hyper-parameter τ:

$$\mathcal{L}_{\text{lcl}} = \frac{1}{L} \sum_l \mathcal{L}_{\text{lcl}}^l, \qquad \text{with} \tag{1}$$

$$\mathcal{L}_{\text{lcl}}^l = -\log \frac{\exp(z_l^\top q^+/\tau)}{\sum_{i,j,k} \exp(z_l^\top q_{ij}^k/\tau)}. \tag{2}$$

Similarly, for global point cloud embedding z, we pull close the global image embedding q^+ of the same object, while pushing away all other global image embeddings $\left\{q^k \neq q^+\right\}^{k=1,\dots,m}$,

$$\mathcal{L}_{\text{glb}} = -\log \frac{\exp(z^\top q^+/\tau)}{\sum_k \exp(z^\top q^k/\tau)}. \tag{3}$$

The overall loss function is given by $\mathcal{L} = \mathcal{L}_{\text{lcl}} + \mathcal{L}_{\text{glb}}$.

4. Experiments

In the following, we present our experimental setups and results of Pic@Point pre-training using different point cloud backbones. We conduct experiments on four standard benchmarks: *ModelNet40* (Wu et al., 2015) and *ScanObjectNN* (Uy et al., 2019) for object classification, *ShapeNetPart* (Yi et al., 2016) for part segmentation, and *S3DIS* (Armeni et al., 2016) for semantic scene segmentation.

4.1. Pre-Training Setup

Pre-training Dataset Following common practice, we pre-train on the ShapeNet (Chang et al., 2015) dataset, which consists of more than 50 000 CAD models from 55 semantic categories. We obtain point clouds by randomly sampling 1024 points from each object. Furthermore, we obtain renderings of size 224×224 from 20 different view points placed in a regular dodecahedron around the object, saving projection and camera matrices $M_{\text{proj}}, M_{\text{view}}$. We randomly select a single rendering per object at each iteration and apply random rotation as augmentation.

Architectures We conduct experiments using two prominent point cloud backbones: DGCNN (Wang et al., 2019) and PointNeXt (Qian et al., 2022a). DGCNN is a standard architecture used for benchmarking many existing self-supervised methods. PointNeXt is a more recent hierarchical model achieving state-of-the-art results across many 3D tasks. Besides drawing comparison to related contrastive and generative cross-modal methods using the same backbones, we also compare against recent transformer-style methods. For the image backbone we use a light-weight ResNet-18 (He et al., 2016).

During the pre-training phase, the global and local point cloud features are extracted directly from the 3D backbone. Subsequently, the features are projected to shared latent space with dimension $d = 512$. After pre-training, the projection heads are dropped, and the 3D backbone is used exclusively.

Implementation Details The experiments are implemented in PyTorch using the OpenPoints framework (Qian et al., 2022a). We utilize the Adam optimizer with CosineAnnealing and a weight decay of $1e^{-6}$ for pre-training. For the point cloud branch we use an initial learning rate of $1e^{-3}$, for the image branch we set a lower learning rate of $5e^{-5}$. We train with a batch size of 32. For the downstream tasks, we follow the training and evaluation settings of Qian et al. (2022a).

4.2. Downstream Tasks

In the following, we present experimental results on object classification, part segmentation and scene segmentation.

Method	ModelNet40	ScanObjectNN
FoldingNet (Yang et al., 2018)	88.4	-
VIP-GAN (Han et al., 2019)	90.2	-
Point-BERT (Yu et al., 2022)	87.4	-
Point-MAE (Pang et al., 2022)	91.0	77.7
Point-M2AE (Zhang et al., 2022a)	92.9	84.1
DGCNN+Jigsaw (Sauder and Sievers, 2019)	90.6	59.5
DGCNN+STRL (Huang et al., 2021)	90.9	77.9
DGCNN+OcCo (Wang et al., 2021)	90.7	78.3
DGCNN+CrossPoint (Afham et al., 2022)	91.2	81.7
DGCNN+PointVST (Zhang and Hou, 2023)	92.1	-
DGCNN+Pic@Point (ours)	**92.5**	**85.7**
DGCNN+Pic@Point (ours, w/ normals)	**92.9**	-

Table 1: **Linear SVM Classification on ModelNet40 and ScanObjectNN (OBJ_BG).** We compare against methods using a specialized point cloud architecture (*top*), and methods using a DGCNN backbone (*bottom*). We report the overall accuracy (%) with 1024 points.

4.2.1. Object Classification

To evaluate the effectiveness of our proposed Pic@Point method for object classification, we conduct experiments on the synthetic dataset ModelNet40 (Wu et al., 2015) and the real-world scanned dataset ScanObjectNN (Uy et al., 2019). ModelNet40 consists of 12 331 3D CAD models from 40 object categories. ScanObjectNN contains 15 categories of real-world indoor scans with 2903 unique object instances. We evaluate on all three common variants of the the ScanObjectNN dataset: OBJ_BG contains complete object scans, OBJ_ONLY uses background cropping, and PB_T50_RS contains perturbed versions of the scans. We sample 1024 points on each object for both training and testing.

We use two transfer learning protocols to evaluate the effectiveness of our proposed Pic@Point model: linear probing with an SVM and fine-tuning on the downstream dataset.

Linear SVM Results We test the representation capability of Pic@Point by fitting a linear SVM on the features extracted from the pre-trained point cloud backbone. In Table 1 we report our linear classification results on ModelNet40 and ScanObjectNN (OBJ_BG). The upper part of the table shows existing self-supervised methods employing specialized point cloud architectures, including transformer-style methods. These methods focus primarily on point cloud encoding architectures rather than on the design of pretext tasks. The lower part of the table shows architecture-agnostic pre-training methods, which are compared using a DGCNN backbone.

Pic@Point significantly outperforms competing methods on a DGCNN backbone with 92.5% and 85.7% linear classification accuracy on ModelNet40 and ScanObjectNN (OBJ_BG), respectively. By incorporating normals as additional input, we further improve to 92.9% accuracy on ModelNet40. We achieve margins of +0.4% and +4.0% over the second best methods PointVST (Zhang and Hou, 2023) and CrossPoint (Afham et al., 2022), respectively. This underscores the effectiveness of our proposed structural 2D-3D correspondence learning over existing cross-modal approaches such as PointVST, a generative method, and CrossPoint, a contrastive method that relies solely on global correspondences.

Notably, the improvements over existing methods are more pronounced on ScanObjectNN than on ModelNet40. This may be attributed to Pic@Point's enhanced generalization ability through learning modality-invariant features, making it more robust on real-world data with irregular sampling and noise. We outperform methods employing larger transformer-style architectures on ScanObjectNN, achieving a margin of +1.6% over the second best method Point-M2AE (Zhang et al., 2022a), which uses a large multi-scale Transformer architecture.

Fine-tuning Results Next, we perform extensive fine-tuning experiments on all three variants of the ScanObjectNN dataset, which has established itself as a favored classification benchmark (Qian et al., 2022a; Ma et al., 2022). The results are shown in Table 2. The top part of the table shows supervised 3D models trained from scratch. The middle part shows 2D-to-3D methods that use large-scale vision foundation models with specialized architectures and 3D downstream adaptations such as visual prompt tuning or multi-stage knowledge distillation. We note that our model does not directly compete with these methods due to their use of specialized architectures with large memory requirements and 2D-to-3D adaptations. Nevertheless, we include them to provide a comprehensive overview of related

Method	#Params (M)	OBJ_BG	OBJ_ONLY	PB_T50_RS
PointNet (Qi et al., 2017a)	3.5	73.3	79.2	68.2
DGCNN (Wang et al., 2019)	1.8	82.8	86.2	78.1
PointNeXt-S (Qian et al., 2022a)	1.4	-	-	87.7
PointMLP (Ma et al., 2022)	12.6	-	-	85.4
Pix4Point (Qian et al., 2022b)	22.1	-	-	87.9
ACT (Dong et al., 2022)	22.1	93.3	91.9	88.2
P2P (Wang et al., 2022)	195.8	-	-	89.3
I2P-MAE (Zhang et al., 2023)	12.9	94.2	91.6	90.1
ReCon (Qi et al., 2023)	43.6	95.2	93.6	90.6
[T]ransformer (Vaswani et al., 2017)	22.1	79.9	80.6	77.2
[T]+OcCo (Yu et al., 2022)	22.1	84.9	85.5	78.8
Point-BERT (Yu et al., 2022)	22.1	87.4	88.1	83.1
Point-MAE (Pang et al., 2022)	22.1	90.0	88.3	85.2
Point-M2AE (Zhang et al., 2022a)	15.3	91.2	88.8	86.4
[T]+PointVST (Zhang and Hou, 2023)	22.1	-	-	86.6
[T]+TAP (Wang et al., 2023)	22.1	90.4	89.5	85.7
PointMLP+TAP (Wang et al., 2023)	12.6	-	-	88.5
PointNeXt-S (reproduce)	1.4	91.2	89.9	87.2
+ Pic@Point	1.4	**94.0** (+2.8)	**92.6** (+2.7)	**88.1** (+0.9)

Table 2: **Real-world 3D Classification on ScanObjectNN Variants.** We report the number of inference model parameters (M) and the overall accuracy (%) with 1024 points.

research. The lower part of the table lists self-supervised methods employing state-of-the-art 3D point cloud models, many of which are based on Transformer architectures.

We perform experiments using the PointNeXt-S (Qian et al., 2022a) backbone. We observe significant improvements compared to training from scratch: +2.8% on OBJ_BG, +2.7% on OBJ_ONLY and +0.9% on PB_T50_RS. Pic@Point surpasses all self-supervised methods with transformer-style point cloud encoders on all three variants of ScanObjectNN, while having a factor $\times\frac{1}{10}$ smaller model size. On PB_T50_RS, Pic@Point is outperformed only by TAP (Wang et al., 2023) using a PointMLP backbone. On OBJ_BG and OBJ_ONLY, Pic@Point outperforms all competing 3D methods.

Visualization In Figure 4 we show visualizations of local 2D-3D correspondences learned during Pic@Point pre-training. For each object, the correspondences between the object's image patches and two exemplar points are depicted. Two key observations can be made: Firstly, Pic@Point successfully learns to correlate structures of point cloud regions with image patches depicting semantically similar features. Secondly, it accurately matches corresponding image features by incorporating pose information.

Ablation Studies Table 3 presents ablation studies on key components. First, we evaluate the local and global correspondence losses $\mathcal{L}_{\text{lcl}}$ and $\mathcal{L}_{\text{gbl}}$, both of which have equal impact on the results. Second, we confirm that the pose encoding aids in learning from local cor-

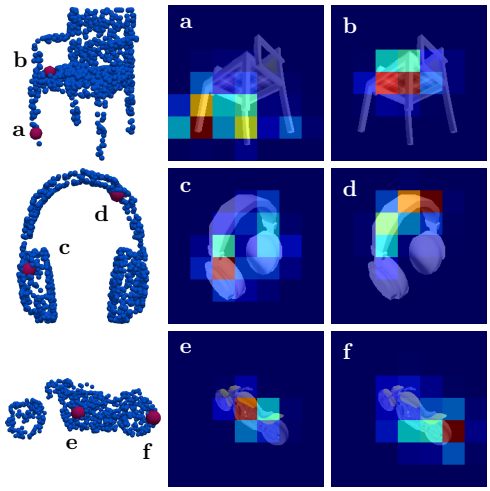

Figure 4: **Visualization of local 2D-3D Correspondences**. The first column displays point clouds with two example points highlighted. The second and third columns display heatmaps of the feature distances between one of these points and the image patches of the corresponding object in latent space.

$\mathcal{L}_{\text{lcl}}$	$\mathcal{L}_{\text{glb}}$	Pose Cond.	Normals	OA (%)
✗	✓	✓	✓	92.3 (−0.2)
✓	✗	✓	✓	92.3 (−0.2)
✓	✓	✗	✓	91.8 (−0.7)
✓	✓	✓	✓	92.8 (+0.3)
✓	✓	✓	✗	92.5

Table 3: **Ablation Studies.** Contributions of local and global correspondence, the impact of pose conditioning, and the use of point normals are evaluated via linear SVM classification on ModelNet40 using a DGCNN backbone.

respondences. A drop of −0.7% accuracy is observed when pose information is omitted. Lastly, while our standard Pic@Point model excludes normals to ensure fair comparison with related methods, we show that incorporating normal information during pre-training and downstream fine-tuning, when available, provides significant benefits.

4.2.2. Part Segmentation

ShapeNetPart (Yi et al., 2016) is a widely used dataset for object part segmentation. It contains 16 881 pre-aligned CAD models from 16 object classes and has a total of 50 part categories. We use the same pre-training setup as for classification, pre-training on 1024 randomly sampled points per object on the ShapeNet dataset. For downstream fine-tuning on ShapeNetPart, we train and test on 2048 points. Table 4 reports the mean intersection

Method	ins. mIoU (%)	cls. mIoU (%)
PointNet (Qi et al., 2017a)	83.7	80.4
DGCNN (Wang et al., 2019)	85.2	82.3
PointNeXt-S (Qian et al., 2022a)	86.7	84.4
PointMLP (Ma et al., 2022)	86.1	84.6
Transformer (Vaswani et al., 2017)	85.1	83.4
Point-BERT (Yu et al., 2022)	85.6	84.1
Point-MAE (Pang et al., 2022)	86.1	84.2
DGCNN+Jigsaw (Sauder and Sievers, 2019)	85.3	82.3
DGCNN+OcCo (Wang et al., 2021)	85.0	-
DGCNN+CrossPoint (Afham et al., 2022)	85.5	-
DGCNN+PointVST (Zhang and Hou, 2023)	**87.4**	-
DGCNN+Pic@Point (ours)	85.8	**83.0**

Table 4: **Part Segmentation on ShapeNetPart.** We report the mean IoU across all instances (ins.) and across all classes (cls.) with 2048 points.

over union (mIoU) averaged over all instances (ins.) and averaged over all object classes (cls.) with DGCNN backbone.

Overall, ShapeNetPart results are less distinguishable compared to other benchmarks. Both Jigsaw (Sauder and Sievers, 2019) and OcCo (Wang et al., 2021) show no significant improvement over training from scratch using DGCNN. Pic@Point demonstrates slightly better performance than previous pre-training methods on a DGCNN backbone, except for PointVST (Zhang and Hou, 2023), which reports an exceptional result of 87.4%. Compared to CrossPoint (Afham et al., 2022), which uses global 2D correspondences, Pic@Point shows an improvement of +0.3%.

4.2.3. Scene Segmentation

Semantic segmentation on large 3D scenes challenges the understanding of contextual relationships and coherent semantic interpretation. S3DIS (Armeni et al., 2016) is a scene segmentation benchmark consisting of 6 types of large scanned indoor areas with 13 semantic categories. Following common practice, we test on the largest area, Area 5, and fine-tune on the remaining areas. For training, the point clouds are downsampled with a voxel size of 0.04m and sub-sampled to 24 000 points. Testing is conducted on the entire scene.

Pic@Point consistently improves performance over training from scratch, increasing mIoU by +0.7% and mAcc by +0.6%. It outperforms leading transformer-style methods by a margin of +2.8% mIoU. The achieved scene segmentation results indicate that Pic@Point is superior in performing dense prediction tasks through its integration of rich semantic cues and structural, pose-aware guidance.

Method	mIoU (%)	mAcc (%)
PointNet (Qi et al., 2017a)	41.1	49.0
DGCNN (Wang et al., 2019)	47.9	-
PointNeXt-S (Qian et al., 2022a)	63.4	-
PointNeXt-XL (Qian et al., 2022a)	70.5	-
Transformer (Vaswani et al., 2017)	60.0	68.6
Point-BERT (Yu et al., 2022)	60.8	69.9
Point-MAE (Pang et al., 2022)	60.8	69.9
PointNeXt-S (reproduce)	62.9	69.5
+ Pic@Point	**63.6** (+0.7)	**71.1** (+0.6)

Table 5: **Scene-Level Semantic Segmentation on S3DIS Area 5.** We report mean IoU and mean accuracy.

5. Conclusions

This paper presents Pic@Point, a self-supervised pre-training method that leverages 2D-3D correspondences at local and global scales. Our proposed method uses a simple contrastive learning framework to integrate point cloud and image features across various abstraction levels, providing a guiding signal that is rich in semantic and structural knowledge. Pic@Point pre-training significantly outperforms existing self-supervised pre-training methods, including those based on Transformer architectures, in various 3D understanding tasks. Its lightweight, architecture-agnostic design offers distinct practical advantages and benefits from future advancements in point cloud technologies.

Acknowledgments

This work was supported by funding from the Federal Ministry of Education and Research (BMBF) through the research project DAVIS within the "Research, Development and Use of Artificial Intelligence Methods in SMEs" program.

References

Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9902–9912, 2022.

Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016.

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

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

Ye Chen, Jinxian Liu, Bingbing Ni, Hang Wang, Jiancheng Yang, Ning Liu, Teng Li, and Qi Tian. Shape self-correction for unsupervised point cloud understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8382–8391, 2021.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022.

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the ieee/cvf International Conference on computer vision*, pages 6391–6400, 2019.

Zhizhong Han, Mingyang Shang, Yu-Shen Liu, and Matthias Zwicker. View inter-prediction gan: Unsupervised representation learning for 3d shapes by learning global shape memories to support local view predictions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8376–8384, 2019.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross 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.

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

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021.

Tianxin Huang, Zhonggan Ding, Jiangning Zhang, Ying Tai, Zhenyu Zhang, Mingang Chen, Chengjie Wang, and Yong Liu. Learning to measure the point cloud reconstruction loss in a representation space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12208–12217, 2023.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022.

Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015.

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

Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pages 604–621. Springer, 2022.

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

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

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

Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35:23192–23204, 2022a.

Guocheng Qian, Xingdi Zhang, Abdullah Hamdi, and Bernard Ghanem. Pix4point: Image pretrained transformers for 3d point cloud understanding. *arXiv preprint arXiv:2208.12259*, 2022b.

Yongming Rao, Jiwen Lu, and Jie Zhou. Global-local bidirectional reasoning for unsupervised representation learning of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5376–5385, 2020.

Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3577–3586, 2017.

Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32, 2019.

Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.

Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019.

Bach Tran, Binh-Son Hua, Anh Tuan Tran, and Minh Hoai. Self-supervised learning with multi-view rendering for 3d point cloud analysis. In *Proceedings of the Asian Conference on Computer Vision*, pages 3086–3103, 2022.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

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

Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *Advances in neural information processing systems*, 35:14388–14402, 2022.

Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen 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.

Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Density-aware chamfer distance as a comprehensive metric for point cloud completion. *arXiv preprint arXiv:2111.12702*, 2021.

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

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

Chenfeng Xu, Shijia Yang, Tomer Galanti, Bichen Wu, Xiangyu Yue, Bohan Zhai, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2point: 3d point-cloud understanding with 2d image pretrained models. In *European Conference on Computer Vision*, pages 638–656. Springer, 2022.

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

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

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

Qijian Zhang and Junhui Hou. Pointvst: Self-supervised pre-training for 3d point clouds via view-specific point-to-image translation. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

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

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022b.

Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng 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.