








Formula-Supervised Visual-Geometric Pre-training

Ryosuke Yamada^{*1,2} , Kensho Hara^{*1} ,
Hirokatsu Kataoka¹ , Koshi Makihara¹ , Nakamasa Inoue^{1,3} ,
Rio Yokota^{1,3} , and Yutaka Satoh^{1,2} 

¹ AIST, ² University of Tsukuba, ³ Tokyo Institute of Technology

* equal contribution

Project Page: <https://ryosuke-yamada.github.io/fdsl-fsvgp/>

Abstract. Throughout the history of computer vision, while research has explored the integration of images (visual) and point clouds (geometric), many advancements in image and 3D object recognition have tended to process these modalities separately. We aim to bridge this divide by integrating images and point clouds on a unified transformer model. This approach integrates the modality-specific properties of images and point clouds and achieves fundamental downstream tasks in image and 3D object recognition on a unified transformer model by learning visual-geometric representations. In this work, we introduce **Formula-Supervised Visual-Geometric Pre-training (FSVGP)**, a novel synthetic pre-training method that automatically generates aligned synthetic images and point clouds from mathematical formulas. Through cross-modality supervision, we enable supervised pre-training between visual and geometric modalities. FSVGP also reduces reliance on real data collection, cross-modality alignment, and human annotation. Our experimental results show that FSVGP pre-trains more effectively than VisualAtom and PC-FractalDB across six tasks: image and 3D object classification, detection, and segmentation. These achievements demonstrate FSVGP’s superior generalization in image and 3D object recognition and underscore the potential of synthetic pre-training in visual-geometric representation learning. Our project website is available at <https://ryosuke-yamada.github.io/fdsl-fsvgp/>.

Keywords: Visual-Geometric representation · Synthetic pre-training

1 Introduction

Fusing images (visual) and point clouds (geometric) is crucial for developing vision models that enhance understanding of the real world. This is because the visual and geometric modalities are complementary. For example, a vision model that relies only on point clouds cannot distinguish a picture and a poster attached to the wall. However, the difference in texture between these two objects can easily be identified by fusing images in the same 3D scene. Therefore, extracting visual-geometric representations by integrating images and point clouds enhances the recognition capabilities of vision models.

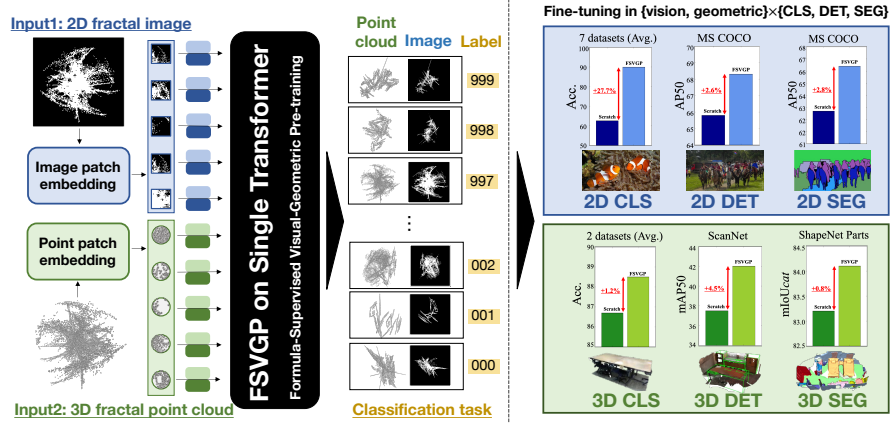


Fig. 1: FSVGP enables pre-training visual and geometric modalities on a unified transformer model by constructing VG-FractalDB from a mathematical formula. VG-FractalDB consists of fractal images, fractal point clouds, and cross-modal supervision called formula-supervised consistency labels. FSVGP simultaneously inputs a fractal image and a fractal point cloud and pre-trains in classification (CLS) tasks based on a formula-supervised consistency label. We show that FSVGP improves six tasks of image and 3D object CLS, detection (DET), and segmentation (SEG).

Despite the ongoing research in visual-geometric representation learning that utilizes images and point clouds, a significant gap exists in developing a unified vision model that effectively trains on both modalities, enhancing both image and 3D object recognition capabilities. Within the realm of visual-geometric representation learning, various studies [11, 20, 23, 36] have pursued improvements in image recognition by integrating visual and geometric data, while others [1, 22, 30, 33, 40] have aimed to enhance 3D object recognition. In 2024, recognition models have emerged that are limited to segmentation tasks yet can address both image and 3D data [13]. The challenge is partly due to the scarcity of large-scale datasets pairing images with point clouds, suggesting extensive paired data is necessary to bridge visual and geometric modalities effectively.

However, given the scarcity of high-quality 3D data on the web, collecting paired images and point clouds proves significantly more challenging and costly. Furthermore, accurate annotations often necessitate manual labeling by experts who can interpret complex spatial information of 3D data. In addition, aligning images with point clouds requires significant correspondence and pre-processing costs to address distortions in raw point clouds and their associated image projections, which are prone to distortion. Consequently, the construction of large-scale datasets for visual and geometric modalities poses a formidable challenge, demanding substantial human resources and specialized expertise. Moreover, copyright and ethical biases are becoming an increasing concern on real datasets.

We facilitate visual-geometric representation learning by employing formula-driven supervised learning (FDSL) to address these challenges. FDSL [16] automatically generates synthetic data and supervision from a mathematical formula

based on principles such as fractal geometry. Furthermore, FDSL helps circumvent common issues associated with real data, including manual labeling costs, copyright concerns, and ethical biases.

Hence, we introduce a visual-geometric pre-training method called **Formula-Supervised Visual-Geometric Pre-training (FSVGP)**. FSVGP enables synthetic pre-training through a unified transformer model by automatically generating aligned synthetic images and point clouds, as shown in Figure 1. We first developed the visual-geometric fractal database (VG-FractalDB), which employs fractal geometry to generate fractal point clouds and their corresponding fractal images automatically, processed simultaneously by a unified transformer model. VG-FractalDB provides a formula-supervised consistency label as cross-modality supervision between visual and geometric modalities. The formula-supervised consistency labels ensure correspondences between fractal images and fractal point clouds, facilitating supervised pre-training for classification tasks on a unified transformer model. For the transformer model, we made minimal modifications to the input processing—drawing upon the Vision Transformer (ViT) [8] and Point Transformer (PointT) [42]—to maintain its flexibility. Thus, FSVGP achieves synthetic pre-training, effectively learning visual-geometric representations to integrate images and point clouds into a unified transformer model. This facilitates image and 3D object recognition using a unified transformer model.

In summary, FSVGP is a novel supervised synthetic pre-training designed to train the VG-FractalDB on a unified transformer model. Our contributions are as follows: (i) Our experimental results show that FSVGP improves fine-tuning performance across six tasks, including image and 3D object classification, detection, and segmentation. (ii) We demonstrate that FSVGP surpasses the latest FDSL method (VisualAtom) in image classification, detection, and segmentation tasks. (iii) We show that FSVGP is superior to the latest FDSL method (PC-FractalDB) in 3D object classification, detection, and segmentation tasks.

2 Related work

In this section, we limit the discussion of related work to that closely related to the proposed FSVGP.

In image recognition, recent self-supervised learning (SSL) methods [2, 5, 10] using massive datasets such as JFT-300M [31] or ImageNet-21k have begun to surpass the longstanding de facto standard of ImageNet-1k for supervised pre-training. Various SSL methods [21, 28, 41] in 3D object recognition, utilizing ShapeNet [3], have been proposed, showcasing certain pre-training effects in downstream tasks. In addition, large-scale 3D datasets such as Objaverse-XL [6] have recently been introduced in 3D vision tasks. Nevertheless, using large-scale real datasets raises ethical concerns, including manual labeling costs, copyright issues, and biases. It is possible to delete 3D objects from the web by their creators in Objaverse-XL. The importance of open-source datasets becomes evident as models like ViT-22B [5] are often trained with non-public datasets, underscoring the need for transparency and accessibility in computer vision.

Visual-geometric representation learning with images and point clouds aims to improve recognition performance over a single modality vision model. This research is categorized into two main areas: enhancing visual and geometric recognition. For the former, the goal is to integrate geometric information from 3D data into visual representations to enrich the understanding of 3D scenes [11, 20, 23, 36]. For instance, Pri3D [11] utilizes contrastive learning to fuse the relationship between corresponding point clouds and pixels. Conversely, methods to enhance 3D object recognition leverage visual knowledge derived from massive images [1, 22, 30, 33, 40]. CrossPoint [1] uses contrastive learning between point clouds and multi-view images.

Furthermore, FDSL is a notable method that addresses pre-training real dataset limitations [15, 16, 32, 38]. Unlike SSL, which assigns pseudo-labels to unlabeled data, FDSL uses a mathematical formula to generate synthetic data and corresponding labels for pre-training. For instance, VisualAtom-21k [32] utilizes FDSL to generate synthetic images with complex contours to pre-train ViT, demonstrating effectiveness in image classification, detection, and segmentation. Similarly, PC-FractalDB-1k [38], a synthetic 3D scene dataset, enhances fine-tuning performance in 3D object detection through VoteNet [29] pre-training. However, the previous FDSL mainly focused on specific modalities and tasks.

Thus, we introduce FSVGP, which extends FDSL to visual-geometric representation learning, achieving supervised pre-training on a unified transformer model. FSVGP can effectively serve as a backbone network for a broad image and 3D object recognition spectrum through we develop vanilla transformers with minimal modification. Moreover, since FSVGP utilizes synthetic data, it circumvents the ethical issues of real data.

3 Formula-supervised visual-geometric pre-training

This section introduces FSVGP, a novel synthetic pre-training method designed to learn visual-geometric representations for image and 3D object recognition. Unlike previous visual-geometric representation learning methods that predominantly focus on image or 3D object recognition in isolation, FSVGP trains visual-geometric representations on a unified transformer model.

To implement FSVGP, we construct VG-FractalDB, which automatically generates fractal images and fractal point clouds based on fractal geometry, as shown in Figure 2. The generation process of fractal images and fractal point clouds refers to previous research [38, 39]. Our study’s key distinction from previous research [38, 39] is the implementation of supervised pre-training on a unified transformer model by utilizing formula-supervised consistency labels between visual and geometric modalities derived directly from a mathematical formula as cross-modality supervision. As a result, our approach supports visual-geometric learning within a shared label space. Furthermore, our study’s concept is based on the fact that FSVGP learns the natural law of visual-geometric relationships by using fractal geometry as a generation rule.

Furthermore, we do not develop special cross-modal modules and complex multi-task learning. As a result, by implementing minimal modifications to

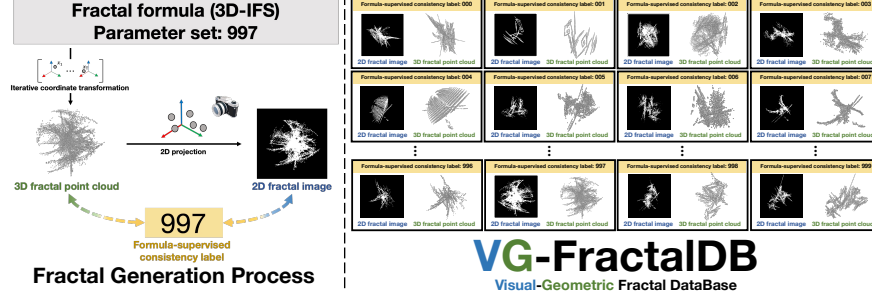


Fig. 2: Overview of the fractal generation process and VG-FractalDB. The fractal generation process creates paired fractal data and formula-supervised consistency labels. Initially, fractal point clouds are generated using the 3D Iterated Function System (3D-IFS). The fractal point clouds are then projected onto 2D planes to form fractal images. Simultaneously, formula-supervised consistency labels are automatically generated based on the variance of 3D coordinates, serving as cross-modality supervision. We construct the VG-FractalDB by repeating these generations.

the transformer model and employing a straightforward loss function for pre-training, we not only facilitate visual-geometric representation learning but also broaden the FSVGPs range of applicability.

3.1 Visual-geometric fractal database (VG-FractalDB)

VG-FractalDB is a pre-training dataset that consists of fractal images and fractal point clouds with formula-supervised consistency labels (see Figure 2). Specifically, the VG-FractalDB is defined by $\mathcal{D} = \{(X_j, I_j, y_j)\}_{j=1}^N$, where X_j represents a fractal point cloud, I_j indicates a fractal image, and y_j is a formula-supervised consistency label that relates to both visual and geometric modalities. Here, N signifies the total number of pre-training data. The formula-supervised consistency labels are categorized discretely within $\{1, 2, \dots, C\}$, where C denotes the number of categories. Note that the number of data in VG-FractalDB is N , since the fractal image is not generated by the other formulas but by simply projecting a fractal point cloud onto a 2D image plane.

Geometric modality – fractal point cloud. A fractal point cloud is generated using the 3D Iterated Function System (3D-IFS), a method rooted in fractal geometry for creating complex, self-similar structures. The 3D-IFS for generating a fractal point cloud of category c , denoted as Θ^c , is defined as $\Theta^c = \{\mathcal{X}; t_1^c, t_2^c, \dots, t_n^c; p_1^c, p_2^c, \dots, p_n^c\}$, where each $t_i^c : \mathcal{X} \rightarrow \mathcal{X}$ represents an affine transformation function within the space \mathcal{X} , and $\{p_i^c\}$ are the associated probabilities of each transformation, with $n = 7$ denoting the number of transformations. A fractal point cloud is represented as a set of coordinates, $\{\mathbf{x}_t\}_{t=1}^T$, within the complete metric space \mathcal{X} , here employing the 3D Euclidean space $\mathcal{X} = \mathbb{R}^3$. The point cloud is generated by a series of affine transformations t_i applied within this space. Each affine transformation t_i is defined by the equation $t_i(\mathbf{x}) = \mathbf{r}_i \mathbf{x} + \mathbf{b}_i$, where \mathbf{r}_i is a transformation matrix within $\mathbb{R}^{3 \times 3}$, \mathbf{x} represents

the coordinates within \mathbb{R}^3 , and \mathbf{b}_i is a bias vector. The initial point \mathbf{x}_1 is conventionally set to the origin (zero vector). The number of affine transformations, n , along with the elements of each rotation matrix \mathbf{r}_i and bias vector \mathbf{b}_i , are determined through random sampling from specified uniform distributions. The transformation probability, p_i , is calculated proportionally to the determinant of \mathbf{r}_i , normalized by the sum of determinants for all transformations. The final coordinate set, X , comprises the sequence of coordinates generated up to a predetermined limit $T = 8192$.

Visual modality – fractal image. We detail transforming a fractal point cloud into a fractal image by projecting it onto an image plane. This conversion employs a mapping, \mathcal{F}_{RGB} , which interprets the coordinates in X_j as white dots against a black background, effectively rendering the fractal geometry visually in two dimensions. To facilitate this transformation, we introduce a virtual camera, \mathbf{c} , designed to map the 3D coordinate set into a fractal image, I_j . The mapping process is succinctly represented as $I_j = \mathcal{F}_{\text{RGB}}(X_j; \mathbf{c})$.

To ensure a precise alignment between the fractal images and their corresponding fractal point clouds—thus maintaining a coherent pairing of visual and geometric data—we allocate one virtual camera, \mathbf{c}_v , per fractal point cloud, where $v = 1$. The chosen method for projection is perspective projection, offering a realistic spatial representation. The positioning of the virtual camera is determined randomly but is strategically placed on a sphere that centers around the fractal object’s center of gravity, optimizing the view of the fractal’s intricate structures. The fractal image size is set to $(W, H) = (224, 224)$.

Formula-supervised consistency label. We introduce a new approach that assigns formula-based supervision, termed “formula-supervised consistency labels,” to fractal images and fractal point clouds, originating from a unified mathematical formula (see Figure 2 Left). These labels emerge from mathematical formulas, enabling the definition of common labels across modalities—a process traditionally requiring costly and specialized pre-processing to map images to point clouds and vice versa. Moreover, formula-supervised consistency labels facilitate the simultaneous input of fractal images and fractal point clouds into a unified transformer model, promoting learning within a shared label space.

As described above, a fractal category c is defined by 3D-IFS Θ^c . A 3D fractal point cloud is generated by the 3D-IFS and is projected onto a 2D image plane to generate a 2D fractal image. Therefore, the 2D fractal image and 3D fractal point cloud share the consistent label. To remove ineffective fractal categories, we employ a variance threshold criterion. This algorithm assesses whether the variance of a fractal category exceeds a predefined threshold (0.05) along each coordinate axis. We determined the threshold with reference to [38]. In other words, we only define fractal point clouds that are above the threshold as a fractal category. To enrich the diversity of instances within each fractal category, we implement a technique named FractalNoiseMix, as described in [38]. This method integrates an additional 20% of points generated from randomly selected fractal point clouds in VG-FractalDB, enhancing the dataset’s robustness and

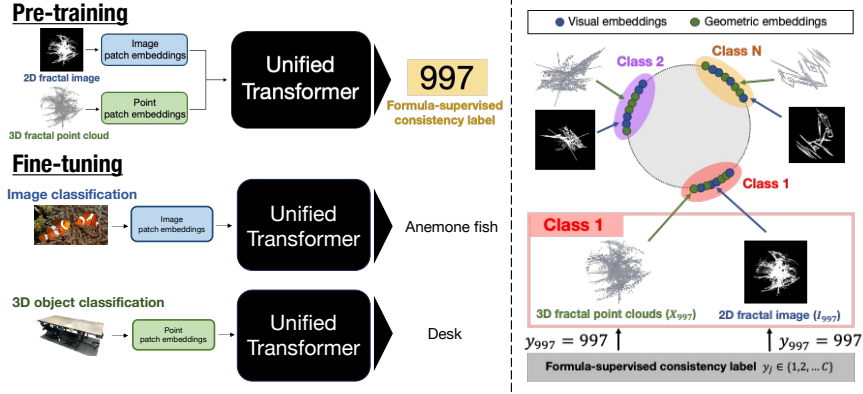


Fig. 3: VG-FractalDB pre-training. **Left:** We train VG-FractalDB on a unified transformer model. After pre-training, we can fine-tune the image and 3D object recognition by using the same unified transformer model. **Right:** FSVGP learns visual and geometric modalities by supervised pre-training based on a formula-supervised consistency label. Therefore, FSVGP can train different modalities within a common label space on a unified transformer model.

variability. Further details on VG-FractalDB, including examples and parameter specifications, are available in the [Supplementary Material](#).

3.2 VG-FractalDB Pre-training on a unified transformer model

To train VG-FractalDB with a unified transformer model, we modified the ViT and PointT to input a fractal image and a fractal point cloud simultaneously, as shown in Figure 3 Left. Our modifications were limited to the input processing, ensuring the transformer model remains as straightforward as possible. This approach maintains the model’s adaptability to various tasks without compromising its pre-training flexibility.

For embedding vectors specific to each modality, we utilized ViT for fractal images and PointT for fractal point clouds. Specifically, a fractal image and a fractal point cloud are divided and embedded into the image tokens $\mathbf{z}_i = [x_{\text{class}}, \mathbf{z}_i^1, \mathbf{z}_i^2, \dots, \mathbf{z}_i^{M_i}]$ and point cloud tokens $\mathbf{z}_p = [x_{\text{class}}, \mathbf{z}_p^1, \mathbf{z}_p^2, \dots, \mathbf{z}_p^{M_p}]$, where M_i and M_p are the numbers of image and point cloud tokens, respectively. Moreover, a class token x_{class} is added to the tokens of each modality. The fractal image and fractal point cloud tokens are then input into the transformer encoder. The class token x_{class} and the MLP layer used for classification are shared between the two modalities.

In addition, for the pre-training task of VG-FractalDB, we train our transformer model f using cross-entropy (CE) loss for the classification task (see Figure 3 Left), which is given by

$$\mathcal{L}_{\text{ce}}(f(\mathcal{D})) = -\frac{1}{N} \sum_{j=1}^N \sum_{c=1}^C y_{j,c} \log \hat{y}_{j,c} \quad (1)$$

Table 1: Details of fine-tuning datasets, fine-tuning task, model type, number of classes (#classes), training data (#train), validation data (#val), and evaluation metrics.

Fine-tuning dataset	Fine-tuning task	Table	Model type	#classes	#train	#val	Metrics
CIFAR10 (C10) [19]	Image CLS	3, 4	ViT-B	10	50k	10k	Acc.
CIFAR100 (C100) [19]	Image CLS	3, 4	ViT-B	100	50k	10k	Acc.
Stanford Cars (Cars) [18]	Image CLS	3, 4	ViT-B	196	8k	8k	Acc.
Oxford Flowers (Flowers) [27]	Image CLS	3, 4	ViT-B	102	6k	818	Acc.
PascalVOC 2012 (VOC) [9]	Image CLS	3, 4	ViT-B	20	13k	13k	Acc.
Places30 (P30) [43]	Image CLS	3, 4	ViT-B	30	150k	3k	Acc.
ImageNet100 (IN100) [7]	Image CLS	2, 3, 4, 10, 11, 12	ViT-B	100	120k	5k	Acc.
MS COCO 2017 (COCO) [34]	Image DET / SEG	2, 5	ViTDet-B	80	118k	5k	AP
ImageNet-1k [7]	Image CLS	6	ViT-B	1000	1.2M	50k	Acc.
ModelNet40 (M40) [37]	3D object CLS	3, 7, 8, 10, 11, 12	PointT-S	40	9.8k	2.4k	Acc.
ScanObjectNN (SONN) [35]	3D object CLS	2, 3, 7, 8	PointT-S	15	2.3k	581	Acc.
ShapeNet-Parts [35]	3D object (parts) SEG	2, 9	PointT-S	15	14k	2.8k	mIoU
ScanNet [4]	3D object DET	2, 9	3DETR	18	1.2k	312	mAP

where $\hat{y}_j = f(X_j, I_j)$ is the output vector. FSVGP trains a unified transformer model on VG-FractalDB to minimize CE loss using AdamW [24]. Our approach to supervised pre-training distinguishes itself from conventional visual-geometric learning methods in several key ways. Whereas the visual-geometric representation learning method [22] focused on pixel-point correspondence for training, FSVGP facilitates learning across different modalities within a common label space (see Figure 3 Right). This unified label space allows for optimizing a unified transformer model, the learning process across modalities.

4 Experiments

In this section, we evaluate the effectiveness of FSVGP by comparing it with previous pre-training methods. First, Section 4.1 outlines our experimental setup. In addition, Section 4.2 briefly describes the main experimental results in image recognition and 3D object recognition with FSVGP. Subsequently, Sections 4.3 and 4.4 compare FSVGP with established pre-training methods across six vision tasks, explicitly focusing on visual recognition and geometric recognition, respectively. Finally, Section 4.5 presents an ablation study to explore the fundamental components of FSVGP.

4.1 Experimental setting

Pre-training. We use VG-FractalDB-1k (1000 categories, 1000 instances per category), ensuring equitable comparison with the existing FDSL methods. Following the approach of previous SSL and FDSL methods, we conduct pre-training using a unified transformer (Base) for image recognition and a unified transformer (Small) for 3D object recognition. The Warm-up Cosine Scheduler is employed for scheduling during pre-training. The batch size is 64 for each GPU, the initial learning rate is 5e-4, the momentum is 0.9, the weight decay is 5e-2, and the number of epochs is 200. For example, training on VG-FractalDB-1k uses 16 NVIDIA V100 GPUs and requires about 60 hours.

Comparison methods. For image recognition, we evaluate FSVGP against transformer-based pre-training methods, including supervised pre-training on

Table 2: Comparison of the latest FDSL methods in image and 3D object classification (CLS), detection (DET), and segmentation (SEG). The best score is shown in **bold**.

Pre-training dataset	Visual recognition			Geometric recognition		
	CLS	DET	SEG	CLS	DET	SEG
	Acc.	AP ₅₀	AP ₅₀	Acc.	mAP ₂₅	mIoU _{cat}
VisualAtom-21k	91.3	66.3	63.3	X	X	X
PC-FractalDB-1k	X	X	X	83.3	63.0	83.7
VG-FractalDB-1k	92.0	68.3	65.6	83.7	63.7	84.1

ImageNet and SSL methods such as MAE [10] and DINO [2]. Additionally, we compare with FDSL methods, including ExFractalDB-21k [15], RCDB-21k [15], and VisualAtom-21k [32]. SAM [17] is also included in comparisons for image detection and segmentation. For 3D object recognition, we focus on transformer-based pre-training methods suited to point clouds, comparing SSL approaches such as PointBERT [41], PointMAE [28], and MaskPoint [21] to evaluate FSVGP’s effectiveness. For 3D object recognition, we specifically compare with the latest FDSL method, such as PC-FractalDB-1k [38]. However, since PC-FractalDB-1k proposed pre-trained on VoteNet [29], to ensure a fair comparison, we also pre-train PC-FractalDB-1k using 3DETR [25]. For fine-tuning in geometric classification and segmentation, we utilize the backbone network of the PC-FractalDB-1k pre-trained model on 3DETR.

Fine-tuning datasets and evaluation metrics. Table 1 describes the detailed settings of fine-tuning datasets used in the experimental section. For detailed information on each fine-tuning dataset, the hyperparameters employed in the pre-training and fine-tuning processes, and the comparison baselines, please refer to the [Supplementary Material](#).

4.2 FSVGP effects on image and 3D object recognition

In the beginning, we begin by presenting in Table 2 a comparison against the latest FDSL methods across all six tasks in both image and 3D object recognition classification (CLS), detection (DET) and segmentation (SEG). In image recognition, CLS signifies fine-tuning accuracy on ImageNet100, while DET and SEG refer to fine-tuning performance on MS COCO for detection and segmentation tasks. In 3D object recognition, CLS represents fine-tuning accuracy in ScanObjectNN (PB-T50-RS). DET involves fine-tuning performance on ScanNet, and SEG involves fine-tuning performance on ShapeNet-parts.

Table 2 demonstrates FSVGP (VG-FractalDB-1k) performs equally or better than VisualAtom-21k and PC-FractalDB-1k in all tasks of classification, detection, and segmentation for image and 3D object recognition. VisualAtom-21k or PC-FractalDB-1k pre-trained models are designed to fine-tune datasets of the same modality, and it is considered difficult to apply them to datasets of different modalities. However, our FSVGP performs better on both visual and geometric recognition. This result shows FSVGP’s capability to process image and 3D object recognition using a unified pre-trained model by supervised pre-training with the formula-supervised consistency label.

Table 3: Comparison of performance on pre-training on either visual or geometric modality with FSVGP (VG-FractalDB-1k, visual (V) + geometric (G)).

Dataset	Modal	#data	C10	C100	Cars	Flowers	VOC12	P30	IN100	Avg.	M40	SONN	Avg.
ImageNet	V	1M	99.0	89.6	81.9	99.1	86.5	82.1	93.1	90.2	92.2	82.0	87.1
ShapeNet	G	50k	82.1	65.4	8.25	74.8	53.1	79.3	79.7	63.2	92.7	83.3	88.0
VG-FractalDB-1k	V	1M	98.0	84.3	88.7	99.5	82.7	80.9	91.2	89.3	92.7	83.3	88.0
VG-FractalDB-1k	G	1M	87.5	68.4	11.2	82.1	57.6	80.6	82.8	67.2	92.6	83.3	88.0
VG-FractalDB-1k	V+G	1M	98.1	85.9	89.2	99.5	83.5	81.7	92.0	90.0	92.9	83.7	88.3

In addition, we investigate the pre-training effects of learning visual-geometric representations. We evaluate the performance of pre-trained models on either visual or geometric modality compared to pre-trained models on both modalities. Table 3 shows the results of pre-trained models on both real (ImageNet, ShapeNet) and synthetic (VG-FractalDB-1k) datasets. All pre-training utilized the same supervised learning to conduct a fair comparison. FSVGP (VG-FractalDB-1k, V + G), which uses both modalities, improves the recognition performance in both modalities and achieves similar performances to ImageNet and ShapeNet in visual and geometric recognition, respectively, even though VG-FractalDB consists of synthetic data. These results indicate the effect of FSVGP, which bridges the modality gap over single-modality pre-trained models.

In the following sections, we show the detailed analyses of FSVGP and comparisons with other pre-training on real datasets, such as SSL.

4.3 Comparative analysis in image recognition

Image classification. Table 4 compares the fine-tuning results with existing pre-training methods (SL, SSL, and FDSL) in image classification. Table 4 shows that the FSVGP (VG-FractalDB-1k) improvement compared with training from scratch (random initialization) in all datasets. In addition, the FSVGP (VG-FractalDB-1k) pre-trained model shows improvement compared with previous FDSL methods. In particular, FSVGP (VG-FractalDB-1k) improves performance by Avg. +0.3% from VisualAtom-21k despite the number of pre-training data is 1/21. However, FSVGP did not surpass the fine-tuning performance of SL, DINO, and MAE. Nevertheless, given real dataset issues associated with copyright, privacy, and social bias, we demonstrate the benefits of FSVGP.

Object detection and instance segmentation. We compare FSVGP with existing pre-training models for the fine-tuning results on average precision (AP) to COCO object detection and instance segmentation in Table 5. We employ ViT-B as the backbone network and use MaskR-CNN as the detection head, referring from ViTDeT. Table 5 reports that FSVGP has superior results to training from scratch (random initialization) for COCO object detection and instance segmentation. In comparison to VisualAtom-21k, FSVGP (VG-FractalDB-1k) provides 2.5% and 2.0% higher AP in visual detection and segmentation, respectively. FSVGP (VG-FractalDB-1k) is inferior to SAM and MAE. However, FSVGP (VG-FractalDB-1k) outperforms ImageNet supervised pre-training and DINO despite synthetic pre-training. From these results, ViT has a lower inductive bias, but because FSVGP is trained with visual and geometric modalities,

Table 4: Comparison of the latest supervised learning (SL), SSL, and FDSL methods in 2D image classification. ‘Modal’ indicates a modality with ‘V’isual and/or ‘G’eometric inputs. The best scores for each learning type are shown in **bold**.

Method	#Data	Modal	Supervision	C10	C100	Cars	Flowers	VOC12	P30	IN100	Avg.
From scratch	–	–	–	78.3	57.7	16.1	77.1	64.8	75.7	73.2	63.3
ImageNet-1k	1.2M	V	SL	99.0	89.6	81.9	99.1	86.5	82.1	93.1	90.2
ImageNet-1k	1.2M	V	SSL (DINO)	98.9	88.9	92.5	99.6	89.4	82.3	93.2	92.1
ImageNet-1k	1.2M	V	SSL (MAE)	99.1	90.1	91.3	99.8	90.2	82.8	94.1	92.5
ExFractalDB-21k [15]	21M	V	FDSL	97.8	85.2	88.1	99.5	82.7	81.6	90.1	89.3
RCDB-21k [15]	21M	V	FDSL	96.8	82.9	85.9	99.0	81.2	81.2	90.2	88.2
VisualAtom-21k [32]	21M	V	FDSL	97.7	86.7	89.2	99.0	82.4	81.6	91.3	89.7
VG-FractalDB-1k	1.0M	V + G	FDSL (FSVGP)	98.1	85.9	89.2	99.5	83.5	81.7	92.0	90.0

Table 5: Comparison of representative pre-trained models in image object detection and instance segmentation. The best values for each learning type are in **bold**.

Method	COCO Det			COCO Ins Seg		
	AP ₅₀	AP	AP ₇₅	AP ₅₀	AP	AP ₇₅
From scratch	65.7	45.5	49.3	62.8	40.4	43.7
ImageNet-1k (SL)	63.9	43.1	47.4	60.9	38.9	41.7
ImageNet-1k (DINO)	65.0	44.6	48.8	62.3	39.9	42.8
ImageNet-1k (MAE)	70.7	50.5	55.4	68.1	44.8	48.6
SAM-1B (SAM)	70.7	50.5	55.3	68.4	45.0	48.5
ExFractalDB-21k	66.8	46.1	50.3	63.8	40.7	43.4
RCDB-21k	64.5	44.1	48.1	61.7	39.1	41.5
VisualAtom-21k	66.3	45.4	49.8	63.3	40.4	42.9
VG-FractalDB-1k (FSVGP)	68.3	47.9	51.6	65.6	42.4	45.3

Table 6: Comparison of fine-tuning accuracy on ImageNet-1k. The best value for each image resolution is in **bold**.

Method	Res.	Image	ImageNet-1k
From scratch	224 ²	Real	80.5
ImageNet-1k (DINO) [2]	224 ²	Real	82.8
ImageNet-1k (MAE) [10]	224 ²	Real	83.6
ExFractalDB-21k [15]	224 ²	Synthetic	82.7
RCDB-21k [15]	224 ²	Synthetic	82.4
VisualAtom-21k [32]	224 ²	Synthetic	82.7
VG-FractalDB-1k (FSVGP)	224 ²	Synthetic	82.7
From scratch	384 ²	Real	81.2
ImageNet-21k (SL) [8]	384 ²	Real	83.0
JFT-300M (Dosovitskiy et al.) [8]	384 ²	Real	84.2
VisualAtom-21k [32]	384 ²	Synthetic	83.7
VG-FractalDB-1k (FSVGP)	384 ²	Synthetic	83.6
VG-FractalDB-21k (FSVGP)	384 ²	Synthetic	83.8

it has the potential to give ViT a stronger inductive bias regarding spatial information and object shape than ImageNet supervised pre-training.

ImageNet-1k classification. Table 6 presents the fine-tuning accuracy on ImageNet-1k with different image resolutions (224×224 or 384×384), compared with representative conventional approaches. Table 6 reveals that FSVGP (VG-FractalDB-1k) performs comparably to the VisualAtom-21k pre-trained model in image resolution, respectively. Moreover, it is noteworthy that the large-scale FSVGP (VG-FractalDB-21k; 21000 categories, 1000 instances per category) achieves a similar performance of JFT-300M [8] pre-training (83.8% vs. 84.2%), even though FSVGP using about 1/14 of the pre-training data when fine-tuning with image resolutions of 384×384 . We consider that the result is worthwhile because JFT-300M is a non-public dataset. However, VG-FractalDB is more transparent and has fewer copyright, privacy, and social bias issues.

4.4 Comparative analysis in geometric recognition

3D object classification. In Table 7, we evaluate the fine-tuning accuracy on ModelNet40 and three subsets of ScanObjectNN, namely {OBJ-BG, OBJ-ONLY, PB-T50-RS} in 3D object classification. Table 7 shows that FSVGP (VG-FractalDB-1k) yields more accurate performance than training from scratch (random initialization), similar to the experimental results in image classification. Furthermore, FSVGP (VG-FractalDB-1k) improves the average accuracy

Table 7: Comparison of the latest supervised learning (SL), SSL, and FDSL methods in 3D object classification. ‘Modal’ indicates a modality with ‘V’isual and/or ‘G’eometric inputs. The best score for each learning type is in **bold**.

Method	#Data	Modal	Supervision	ModelNet40	ScanObjectNN			Avg.
					OBJ-BG /	OBJ-ONLY /	PB-T50-RS	
From scratch	–	–	–	92.1	86.6 /	86.9 /	81.1	86.7
ShapeNet	50k	G	SSL (Point-BERT)	93.1	90.5 /	89.5 /	85.0	89.5
ShapeNet	50k	G	SSL (Point-MAE)	93.1	90.4 /	88.1 /	85.8	89.4
ShapeNet	50k	G	SSL (MaskPoint)	92.8	89.5 /	88.1 /	83.8	88.6
PC-FractalDB-1k	1.0M	G	FDSL	92.6	88.3 /	88.3 /	83.3	88.1
VG-FractalDB-1k	1.0M	V + G	FDSL (FSVGP)	92.9	88.9 /	88.5 /	83.7	88.5

Table 8: Comparison of few-shot learning. The best averaged accuracy for each learning type is in **bold**.

Method	ModelNet40 Classification			
	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
Scratch	94.8	96.2	91.2	92.8
PointBERT	94.6	96.3	91.0	92.7
PointMAE	96.3	97.8	92.6	95.0
MaskPoint	95.0	97.2	91.4	93.4
PC-FDB-1k	95.6	96.9	91.4	93.2
VG-FDB-1k	96.4	96.8	92.2	93.3

Table 9: Comparison of pre-training methods in 3D object detection and parts segmentation. The best value for each learning type is in **bold**.

Method	ScanNetV2 Det		ShapeNet PartsSeg	
	mAP ₂₅ /	mAP ₅₀	mIoU _{cat} /	mIoU _{ins}
Scratch	62.7 /	37.5	83.3 /	85.4
PointBERT	61.0 /	38.3	84.1 /	86.0
PointMAE	– /	–	84.1 /	86.1
MaskPoint	63.4 /	40.6	84.4 /	86.0
PC-FDB-1k	63.0 /	42.5	83.7 /	85.7
VG-FDB-1k	63.7 /	42.0	84.1 /	85.7

in fine-tuning performance by +0.4% compared to the latest FDSL method (PC-FractalDB). In contrast, FSVGP (VG-FractalDB-1k) has lower fine-tuning accuracy than ShapeNet self-supervised pre-training methods (PointBERT, PointMAE, and MaskPoint). One reason for this result may be that there are many overlapping categories, such as chairs, desks, etc., in the pre-training dataset (ShapeNet) and the fine-tuning datasets (ModelNet40 and ScanObjectNN). However, the goal of FSVGP is not only to perform the highest result on a specific task but also to achieve a superior pre-training effect for various tasks.

Few-shot learning. We compare and verify the performance of few-shot learning on ModelNet40 in Table 8. For few-shot learning, we randomly sample K categories from ModelNet40 and N shots of training samples from each category, following the experimental setting in [28]. FSVGP (VG-FractalDB-1k) achieves inspiring performance, outperforming training from scratch (random initialization) by a large margin in all few-shot settings despite different domain ModelNet40. In addition, FSVGP (VG-FractalDB-1k) achieves equal or better performance improvement over PC-FractalDB-1k in all few-shot settings. This result shows the FSVGP can achieve few-shot learning.

3D object detection. Table 9 reports fine-tuning performance for 3D bounding box mAP in 3D object detection. We fine-tune 3DETR on ScanNet. Our FSVGP (VG-FractalDB-1k) performs better under mAP₂₅ and mAP₅₀ than previous SSL methods. The FSVGP (VG-FractalDB-1k) is 1.4 points higher than MaskPoint [21], the latest SSL method (42.0 vs 40.6, mAP₅₀). More significantly, FSVGP (VG-FractalDB-1k) outperforms the PC-FractalDB-1k pre-

Table 10: ShapeNet vs. VG-FractalDB (VG-FDB) in visual geometric pre-training. The datasets were assigned visual and geometric modalities. **Table 11:** Effect of generation rules. We compared Perlin noise (VG-PN-1k) with fractal (VG-FDB-1k) in FSVGP. **Table 12:** Effect of supervisions in SSL (MAE) and our FSVGP. In both supervision settings, we utilized VG-FractalDB-1k (VG-FDB-1k) with visual and geometric modalities.

Method	#data	IN100	M40
ShapeNet	50k	87.3	92.7
VG-FDB	50k	87.9	92.8

Method	IN100	M40
VG-PN-1k	90.7	92.6
VG-FDB-1k	92.0	92.9

Method	IN100	M40
VG-FDB-1k (MAE)	80.3	92.8
VG-FDB-1k (FSVGP)	92.0	92.9

trained model [38] by 0.7 points on mAP_{25} , but falls short by 0.5 points on mAP_{50} . This result suggests that the potential of FSVGP works well when applied to backbone networks of 3DETR.

Parts segmentation. Table 9 also compares the performance of FSVGP (VG-FractalDB-1k) with existing pre-training methods for parts segmentation. We fine-tune PointT-S on ShapeNet-parts and evaluate the performance using the mIoU for all categories (mIoU_{cat}) and all instances (mIoU_{ins}). Table 9 shows that our FSVGP (VG-FractalDB-1k) improves results over training from scratch (random initialization), for example, by 0.8 points (84.1 vs. 83.3, mIoU_{cat}). In addition, even though SSL duplicates pre-training and fine-tuning data, our method achieved performance comparable to SSL. These results suggest that FSVGP is even more effective for recognizing more detailed 3D object structures.

4.5 Ablation study

In this section, we conduct additional experiments to explore essential factors of FSVGP guiding visual-geometric representation learning. Specifically, we investigate to answer the following questions: (i) Which is more effective, fractal point clouds or CAD models? in FSVGP (ii) Can other generation rules be effective in FSVGP? Moreover, (iii) what is the effect of the pre-training task for VG-FractalDB?

(i) Which is more effective, fractal point clouds or CAD models in FSVGP? We investigate the pre-training effects in FSVGP using VG-FractalDB and an existing 3D dataset, ShapeNet. Under the same conditions as VG-FractalDB, we project point clouds from ShapeNet onto images to generate visual-geometric data. Subsequently, the generated visual-geometric data from ShapeNet undergoes pre-training under the same conditions as FSVGP. Please refer to the **Supplementary Material** for instances of image and point cloud data generated from ShapeNet. To equalize the number of instances for ShapeNet and pre-training data, VG-FractalDB undergoes random sampling of data, with 50 instances per category. Table 10 shows that the VG-FractalDB pre-trained model is more accurate than the ShapeNet pre-trained model in ImageNet100 and ModelNet40 despite the same number of data.

(ii) Can other generation rules be effective in FSVGP? We verify the pre-training effect regarding which generation rules (fractal and Perlin noise) are more effective in FSVGP. We extended Perlin Noise, effective as a dataset

generation function [12, 14], to point clouds and constructed a Visual Geometric Perlin Noise (VG-PN) dataset. Please consult the [Supplementary Material](#) for additional details on the generation process of the VG-PN dataset and examples of both point clouds and images. We pre-train VG-PN with the same config and fine-tune it for image and 3D object classification. Table 11 shows that the VG-FractalDB outperforms the VG-PN in ImageNet100 and ModelNet40.

(iii) Effect of pre-training tasks for VG-FractalDB. We explore the supervision types of VG-FractalDB by comparing the formula-supervised consistency label with self-supervision adopted by MAE, a representative SSL method. The implementations of decoders for the SSL are the same as MAE [10] and PointMAE [28], and the decoders reconstruct masked patches of 2D fractal images and 3D point clouds for pre-training based on the self-supervision. Table 12 shows that utilizing the formula-supervised consistency label is more effective than utilizing the self-supervision based on MAE in ImageNet100 and ModelNet40.

5 Discussion and Conclusion

This paper proposes FSVGP, which blends visual and geometric representations to achieve image and 3D object recognition on a unified transform model. FSVGP automatically generates fractal images, fractal point clouds, and their formula-supervised consistency labels based on fractal geometry.

In the beginning, contrary to previous visual-geometric representation learning transfer the one-way knowledge using visual and geometric modalities, we show that FSVGP effectively achieved the pre-training effects in both image and 3D object classification, detection, and segmentation. Furthermore, Table 10 and Table 11 show that VG-FractalDB pre-training is more effective than the pre-training of ShapeNet and VG-PN dataset. We consider that these results are attributed to VG-FractalDB being generated based on more parameters of generation function, allowing it to pre-train on geometric shapes that are even more complex than those in the ShapeNet and VG-PN dataset. Finally, Table 12 shows the effectiveness of the pre-training by simple classification task for VG-FractalDB and suggests that FSVGP is sufficient to learn visual-geometric representation in VG-FractalDB pre-training. These observations show that FSVGP effectively improves fine-tuning performance in image recognition and 3D object recognition despite synthetic pre-training. Furthermore, FSVGP can reduce real dataset issues such as copyright, personal information, and social bias.

Technical limitations and future work. FSVGP is a pre-training method using VG-FractalDB (*i.e.*, synthetic data). A previous FDSL study [26] reported that when FDSL is fine-tuning, a certain amount of real data regarding the domain gap between real and synthetic data is necessary. Therefore, compared to MAE (ImageNet), FSVGP (VG-FractalDB) is more challenging to work well with linear probing (Please consult the [Supplementary Material](#) for more details). Designing an efficient fine-tuning approach using FSVGP will be essential. We consider it important to validate FSVGP in 3D shape retrieval and multi-modal recognition involving bird’s-eye view images and point clouds for future applications such as autonomous driving and search systems.

6 Acknowledgments

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). A computational resource, AI Bridging Cloud Infrastructure (ABCI), provided by the National Institute of Advanced Industrial Science and Technology (AIST), was used. We want to thank Ryota Suzuki, Yoshihiro Fukuhara, Naoya Chiba, Ryo Nakamura, Kodai Nakashima, Sora Takashima, and Ryu Tadokoro for their helpful comments in the research discussions.

References

1. Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R.: 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 (CVPR). pp. 9902–9912 (2022) [2](#), [4](#)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021) [3](#), [9](#), [11](#)
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) [3](#)
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5828–5839 (2017) [8](#)
5. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 7480–7512 (2023) [3](#)
6. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-XL: A universe of 10m+ 3d objects. In: Proceedings of the Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023) [3](#)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009) [8](#)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representation (ICLR) (2021) [3](#), [11](#)
9. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision (IJCV) **111**(1), 98–136 (2015) [8](#)

10. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022) [3](#), [9](#), [11](#), [14](#)
11. Hou, J., Xie, S., Graham, B., Dai, A., Nießner, M.: Pri3d: Can 3d priors help 2d representation learning? In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR). pp. 5693–5702 (2021) [2](#), [4](#)
12. Inoue, N., Yamagata, E., Kataoka, H.: Initialization using perlin noise for training networks with a limited amount of data. In: Proceedings of the International Conference on Pattern Recognition (ICPR). pp. 1023–1028 (2021) [14](#)
13. Jain, A., Katara, P., Gkanatsios, N., Harley, A.W., Sarch, G., Aggarwal, K., Chaudhary, V., Fragkiadaki, K.: Odin: A single model for 2d and 3d segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3564–3574 (2024) [2](#)
14. Kataoka, H., Hara, K., Hayashi, R., Yamagata, E., Inoue, N.: Spatiotemporal initialization for 3d cnns with generated motion patterns. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1279–1288 (2022) [14](#)
15. Kataoka, H., Hayamizu, R., Yamada, R., Nakashima, K., Takashima, S., Zhang, X., Martinez-Noriega, E.J., Inoue, N., Yokota, R.: Replacing labeled real-image datasets with auto-generated contours. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21232–21241 (June 2022) [4](#), [9](#), [11](#)
16. Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., Satoh, Y.: Pre-training without natural images. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2020) [2](#), [4](#)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) [9](#)
18. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the International IEEE Workshop on 3D Representation and Recognition (3DRR-13). pp. 554–561 (2013) [8](#)
19. Krizhevsky, A., Hinton, G.: Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto (2009) [8](#)
20. Kundu, A., Yin, X., Fathi, A., Ross, D., Brewington, B., Funkhouser, T., Pantofaru, C.: Virtual multi-view fusion for 3d semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 518–535 (2020) [2](#), [4](#)
21. Liu, H., Cai, M., Lee, Y.J.: Masked discrimination for self-supervised learning on point clouds. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 657–675 (2022) [3](#), [9](#), [12](#)
22. Liu, Y.C., Huang, Y.K., Chiang, H.Y., Su, H.T., Liu, Z.Y., Chen, C.T., Tseng, C.Y., Hsu, W.H.: Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. arXiv preprint arXiv:2104.04687 (2021) [2](#), [4](#), [8](#)
23. Liu, Z., Qi, X., Fu, C.W.: 3d-to-2d distillation for indoor scene parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4464–4474 (2021) [2](#), [4](#)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) [8](#)
25. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2906–2917 (2021) [9](#)

26. Nakashima, K., Kataoka, H., Satoh, Y.: Does formula-driven supervised learning work on small datasets? *IEEE Access* (2023) [14](#)
27. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*. pp. 722–729 (2008) [8](#)
28. Pang, Y., Wang, W., Tay, F.E., Liu, W., Tian, Y., Yuan, L.: Masked autoencoders for point cloud self-supervised learning. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 604–621 (2022) [3](#), [9](#), [12](#), [14](#)
29. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 9277–9286 (2019) [4](#), [9](#)
30. Robert, D., Vallet, B., Landrieu, L.: Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 5575–5584 (2022) [2](#), [4](#)
31. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. pp. 843–852 (2017) [3](#)
32. Takashima, S., Hayamizu, R., Inoue, N., Kataoka, H., Yokota, R.: Visual atoms: Pre-training vision transformers with sinusoidal waves. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 18579–18588 (2023) [4](#), [9](#), [11](#)
33. Tang, P., Xu, H.M., Ma, C.: Prototransfer: Cross-modal prototype transfer for point cloud segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 3337–3347 (2023) [2](#), [4](#)
34. Tsung-Yi Lin, Michael Maire, S.B.L.B.R.G.J.H.P.P.D.R.C.L.Z.P.D.: Microsoft COCO: Common objects in context. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 740–755 (2014) [8](#)
35. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., Yeung, S.K.: 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 (ICCV)*. pp. 1588–1597 (2019) [8](#)
36. Wang, Z., Yu, X., Rao, Y., Zhou, J., Lu, J.: P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* pp. 14388–14402 (2022) [2](#), [4](#)
37. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1912–1920 (2015) [8](#)
38. Yamada, R., Kataoka, H., Chiba, N., Domae, Y., Ogata, T.: Point cloud pre-training with natural 3d structures. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 21283–21293 (2022) [4](#), [6](#), [9](#), [13](#)
39. Yamada, R., Takahashi, R., Suzuki, R., Nakamura, A., Yoshiyasu, Y., Sagawa, R., Kataoka, H.: Mv-fractaldb: formula-driven supervised learning for multi-view image recognition. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 2076–2083 (2021) [4](#)
40. Yi, X., Deng, J., Sun, Q., Hua, X.S., Lim, J.H., Zhang, H.: Invariant training 2d-3d joint hard samples for few-shot point cloud recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*. pp. 14463–14474 (2023) [2](#), [4](#)

41. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: 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 (CVPR). pp. 19313–19322 (June 2022) [3](#), [9](#)
42. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16259–16268 (2021) [3](#)
43. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **40**(6), 1452–1464 (2017) [8](#)