3DCoMPaT^{++}: An improved Large-scale 3D Vision Dataset for Compositional Recognition

Habib Slim, Xiang Li, Yuchen Li, Mahmoud Ahmed, Mohamed Ayman, Ujjwal Upadhyay, Ahmed Abdelreheem, Arpit Prajapati, Suhail Pothigara, Peter Wonka, *Senior Member, IEEE,* and Mohamed Elhoseiny, *Senior Member, IEEE*

***Abstract*—In this work, we present 3DCoMPaT^{++}, a multi-modal 2D/3D dataset with 160 million rendered views of more than 10 million stylized 3D shapes carefully annotated at the part-instance level, alongside matching RGB point clouds, 3D textured meshes, depth maps, and segmentation masks. 3DCoMPaT^{++} covers 42 shape categories, 275 fine-grained part categories, and 293 fine-grained material classes that can be compositionally applied to parts of 3D objects. We render a subset of one million stylized shapes from four equally spaced views as well as four randomized views, leading to a total of 160 million renderings. Parts are segmented at the instance level, with coarse-grained and fine-grained semantic levels. We introduce a new task, called Grounded CoMPaT Recognition (GCR), to collectively recognize and ground compositions of materials on parts of 3D objects. Additionally, we report the outcomes of a data challenge organized at the CVPR conference, showcasing the winning method's utilization of a modified PointNet++ model trained on 6D inputs, and exploring alternative techniques for GCR enhancement. We hope our work will help ease future research on compositional 3D Vision. The dataset and code have been made publicly available at https://3dcompat-dataset.org/v2/.**

***Index Terms*—3D vision, dataset, 3D modeling, multimodal learning, compositional learning.**

I. Introduction

MULTIPLE datasets have been proposed to facilitate 3D visual understanding including ShapeNet [1], ModelNet [2], and PartNet [3]. High-quality datasets like OmniObject3D [4] and ABO [5] were introduced in an attempt to provide 3D assets with high-resolution, realistic textures. 3D-Future [6] was also proposed and contains 10K industrial 3D CAD shapes of furniture with textures developed by professional designers. More recently, Objaverse [7] and its larger counterpart Objaverse-XL [8] were introduced, which contain more than 10 million artist-designed 3D objects with high-quality textures. Despite these notable efforts to advance 3D understanding, recent object-centric 3D datasets (e.g. [1], [2], [7], [8]) and 3D scene datasets (e.g. [9], [10]) lack part-level annotations. ShapeNet-Part [11] was proposed as an extension to ShapeNet [1] with part-level annotations, but only contains coarse-grained part segmentations extracted using a deep active learning framework. In contrast, PartNet [3] builds on ShapeNet [1] and provides fine-grained part segmentation labels, but similarly does not contain material information.

Material information offers several distinct advantages. First, it provides extra semantic information about an object, which enables a variety of important 3D object understanding tasks. Second, it helps create more realistic renderings, making the models better suited for transferring from synthetic to real scenarios. Finally, applying different materials to the same geometric 3D shape can be treated as a special form of training data augmentation. Current datasets lack part-level material information which underscores the need for a new resource. Our dataset fills this gap and invites researchers to explore new challenges and opportunities in 3D visual understanding. 3DCoMPaT^{++} is an extension of the 3DCoMPaT [12] dataset, which was previously published at a conference.

We introduce a new richly annotated multimodal 2D/3D large-scale dataset: **3DCoMPaT^{++}**, standing for **Co**mpositions of **M**aterials on **Par**ts of **3D T**hings. Our dataset comprises 10 million stylized 3D shapes rendered from 8 views, across 42 shape categories, 275 unique fine-grained part names and 43 coarse-grained part names, and 293 unique materials from 13 material classes. We sample object-compatible combinations of part-material pairs to create 1000 styles per shape. Each object with the applied part-material composition is rendered from 4 equally spaced views and 4 random views. We render images for a 1000 total compositions, leading to 160 million[1] total rendered views. Examples of some rendered compositions and views are provided in Figures 5 and 6 respectively.

Dataset. To create our dataset, we start with 10K unique geometries which we segment at a fine-grained part level into a total of 275 segmented parts (leading to 9.86 average part instances per shape). For each part of each shape, human experts determine a list of compatible/applicable materials. Then, we generate a stylized model by sampling over the compatible materials of each part with a limit of 1000 styles per shape, leading to 10 million stylized shapes.

Previous work. Our proposed dataset differs from previous work in numerous ways. Our dataset contains a diverse set of high-quality materials: for each part found in every 3D

Corresponding authors: H. Slim and M Elhoseiny with the Department of Computer Science, KAUST, Thuwal, Saudi Arabia.
E-mail: habib.slim@kaust.edu.sa; mohamed.elhoseiny@kaust.edu.sa
A. Prajapati, S. Pothigara are with Polynine, San Francisco, California.
X. Li, Y. Li, M. Ahmed, M. Ayman, U. Upadhyay, A. Abdelreheem, P. Wonka are with the Department of Computer Science, KAUST, Thuwal, Saudi Arabia.

[1]Figure detail: 10000 (#shapes) $\times$ 1000 (#styles) $\times$ 8 (#views) $\times$ 2 (#part semantic levels) $= 160M$ views

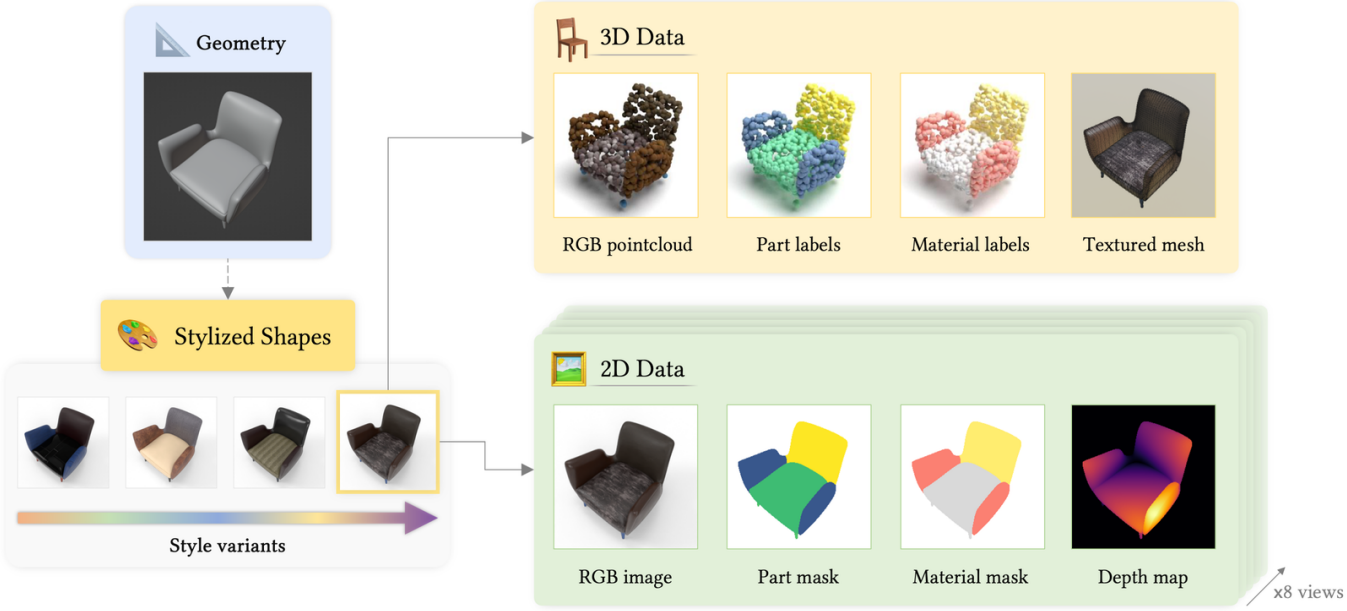


Fig. 1. **Data provided for each stylized shape.** For 3D: RGB pointclouds, textured shapes, and point-wise/triangle-wise part labels and material labels. For 2D: RGB images, depth maps, and corresponding part masks and material masks. Part and material annotations in 2D and 3D are provided in both *coarse* and *fine* semantic levels. In Figure 19 of the appendix, we show additional style variants for various shapes.

Fig. 2. **Grounded CoMPaT Recognition (GCR).** Given an input shape, here: a chair, the task consists of **(a)** recognizing the shape category and **(b)** segmenting the part-material pairs composing it.

shape, we annotate possible compatible materials that may be applied to each part, allowing us to generate multiple material combinations for a single shape (we refer to a combination of materials (*composition*) applied to a model as a *style*). We also enrich our dataset with 2D renders, depth maps, part masks, and material masks for each rendered view, and hierarchical part and material annotations in both the 2D and 3D modalities.

In summary, our 3DCoMPaT^{++} dataset can be distinguished from existing datasets by the following four key aspects:

(a) *Human-generated vs. 3D scanned geometry.* ScanNet [9] and Matterport3D [10] datasets are scanned 3D geometry datasets. Conversely, ShapeNet [1] and our 3DCoMPaT^{++} dataset are human-created, mostly by professional 3D modelers. Human-created geometry is generally of higher quality and has fewer artifacts, but is however more expensive and time-consuming to collect. For the Objaverse [7], [8] dataset, the authors thus propose to scrape 3D models from well-known web repositories which are mostly created by artists. However, the quality of these models is not guaranteed, and models are not annotated with part-level information. The realism of the collected objects is also not a given, as models in these repositories are not designed to be realistic, but rather to be visually appealing as they are typically targeted at the video game industry. Our dataset is human-generated and is designed to be realistic, and comprises high-quality textures and geometry.

(b) *Part segmentation information.* For some datasets, none or only a subset of the shapes have segmented part information, which is an important feature of datasets like PartNet [3] and is also a core characteristic of our dataset. We provide part segmentation information following two semantic levels, in both 2D and 3D modalities.

(c) *Texture coordinates, textures, and materials.* A key focus of our work lies in the stylization of 3D shapes with appropriate texture coordinates, textures, and materials. To achieve a superior level of quality when rendering numerous material compositions on each shape, our models are equipped with human-verified texture coordinates and part-wise material compatibility information. While previous attempts have been made to enhance a subset of ShapeNet with part-wise material information [13], it falls short in comparison to our work in terms of the number of shapes (3080 vs. 10000), shape classes (3 vs 42), and materials (6 vs. 13, and 293 fine-grained annotated classes).

(d) *Automatically generated vs. human-generated annotations.* 3DCoMPaT^{++} shapes are annotated manually by a team of trained humans. Part names are consistent across and within categories, and are defined in shape category-specific guidelines. Each guideline is defined by a team of researchers and professional modelers, and contains rigorous definitions

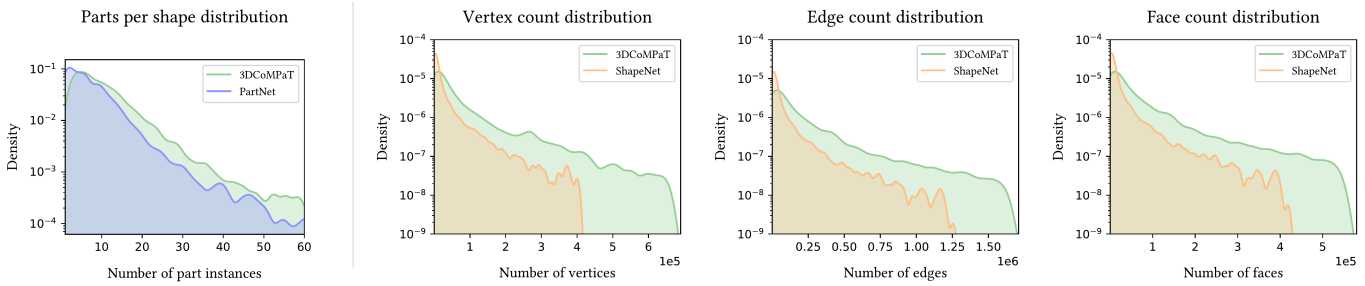


Fig. 3. **Comparison with PartNet.** Part instances per shape distributions compared to PartNet [3] (**left**), Density plots depicting the distribution of vertex counts, edge counts, and face counts across 3D shapes extracted from both the 3DCoMPaT^{++} and ShapeNet datasets (**right**). We show significantly higher numbers of vertices, edges, and faces in 3DCoMPaT^{++} shapes compared to ShapeNet. All annotated shapes in PartNet originate from ShapeNet.

and examples for each part that may occur within a given shape category. All models are manually segmented at a part level rather than with deep learning models like OpenRooms [14] or ShapeNet-Part [11].

Grounded CoMPaT Recognition. We introduce a novel task called CoMPaT recognition, which focuses on collectively recognizing and grounding shape categories along with the associated part-material pairs composing them. In Figure 2, we illustrate the task with an example. Given an input shape, the task aims to recognize both the shape category and all part-material pairs composing it. In the example shown, an agent first needs to identify the shape as a chair, and then all part-material pairs, such as a "seat" made of "leather" and a "backrest" made of "fabric". This novel task, compatible with both 2D and 3D modalities, goes beyond recognition with a grounded variant requiring the precise segmentation of parts alongside the recognition of associated materials.

Contributions. Our work introduces a new dataset, and introduces the GCR recognition task. The contributions of this work can be summarized in the following points:

- We propose a new dataset comprised of 10 million stylized models to study the composition of materials on parts of 3D objects. Our dataset contains (a) a diverse set of 293 materials for 3D shapes, where (b) material assignment is done at a coarse and fine-grained part-level; (c) segmentation masks in 2D and 3D, alongside (d) human-verified texture coordinates.
- We validate our dataset with experiments covering 2D and 3D vision tasks, including object classification, part recognition and segmentation, material tagging and shape generation.
- We also propose **G**rounded **C**oMPaT **R**ecognition (GCR), a novel task aiming at collectively recognizing and grounding compositions of materials on parts of 3D objects. We introduce two variants of this task, and leverage 2D/3D state-of-the-art methods as baselines for this problem.

II. Related Work

Early efforts. Several datasets have been initially proposed to facilitate 3D visual understanding, such as ShapeNet [1], ModelNet [2], and PartNet [3]. ModelNet [2] is one of the first datasets of 3D objects, and includes 40 shape categories and 12K unique 3D shapes. ShapeNet [1] is a large-scale dataset of 3D textured objects, with 55 shape categories and 51K unique 3D shapes. ShapeNet is annotated at the shape-level, and categories are extracted from WordNet [19]. It has emerged as an important benchmark for deep learning-based modeling, representation, and generation of 3D shapes. ObjectNet3D [17] is an object-centric dataset of 3D CAD models with 100 shape categories and 90K unique 3D shapes, and approximate 2D-3D image alignments. ModelNet, ShapeNet and ObjectNet3D are object-centric datasets, and do not contain part-level annotations.

Part-understanding. In an attempt to bridge this gap, ShapeNet-Part [11] was first proposed as an extension to ShapeNet [1] with part-level annotations. It contains 16 shape categories and 31K 3D shapes, but part annotations are only provided at a coarse-grained semantic level, and are extracted using a deep active learning framework instead of human annotation. PhotoShape [15] is one of the earliest efforts in gathering 3D shapes with high-quality textures. It contains 5.8K 3D shapes from 29 shape categories, and proposes to transfer material properties regressed from real images to untextured 3D shapes. PartNet [3] was built as a large-scale dataset of 3D shapes annotated with fine-grained, instance-level, and hierarchical part segmentations. PartNet is also created on top of ShapeNet [1] and contains 26K 3D shapes from 24 shape categories. PartNet is a valuable resource for advancing research in 3D shape analysis and understanding. Our work stands apart from PartNet in three main ways:

- **(a)** We provide both coarse-grained and fine-grained material information for each part of each shape.
- **(b)** We enrich 3D shapes with 2D renders, part masks, material masks, and depth maps.
- **(c)** We use a human verification process to ensure the compatibility of sampled materials with each part of each shape.

High-resolution datasets. In an effort to provide high-quality and realistic shapes and textures, OmniObject3D [4] and ABO [5] datasets introduced 3D assets with rich, high-quality

TABLE I

COMPARISON OF 3DCOMPAT++ WITH EXISTING 3D DATASETS. *Multi-level* INDICATES WHETHER THE DATASET CONTAINS HIERARCHICAL PART SEMANTICS. *Alignment* INDICATES WHETHER THE DATASET CONTAINS ALIGNED 2D/3D DATA. WE DENOTE BY ? MISSING ANNOTATIONS, *e.g.* CASES WHERE THE DATASET CONTAINS TEXTURED SHAPES, BUT DOES NOT INCLUDE MATERIAL ANNOTATIONS. AMONG DATASETS WITH PART-LEVEL ANNOTATIONS, 3DCOMPAT++ CONTAINS THE LARGEST NUMBER OF STYLIZED SHAPES, OBJECT CATEGORIES, MATERIALS AND ALSO PROVIDES A LARGE COLLECTION OF ALIGNED 2D IMAGES.

Dataset	Shapes				Materials		Parts			Images	
	Count	Stylized	Classes	Source	Count	Classes	#Instances/Shape	Multi-level	Instances	Count	Alignment
ModelNet [2]	128K	✗	662	modelled	✗	✗	✗	✗	✗	✗	✗
ShapeNet-Core [1]	51,3K	✗	55	modelled	?	?	✗	✗	✗	✗	✗
ShapeNet-Sem [1]	12K	✗	270	modelled	?	?	✗	✗	✗	✗	✗
PhotoShape [15]	5,8K	29K	1	modelled	658	8	✗	✗	✗	✗	✗
GSO [16]	1K	✗	17	scanned	?	?	✗	✗	✗	✗	✗
OmniObject3D [4]	6K	✗	190	scanned	?	?	✗	✗	✗	✗	✗
ObjectNet3D [17]	44,2K	✗	100	modelled	✗	✗	✗	✗	✗	90K	pseudo
3D-Future [6]	9,9K	✗	15	modelled	?	15	✗	✗	✗	20K	exact
ABO [5] + HAL3D [18]	8K	✗	63	modelled	?	?	?	?	?	398K	pseudo
Objaverse-XL [8]	10,2M	✗	?	modelled	?	?	✗	✗	✗	66M	exact
ShapeNet-Part [11]	31,9K	✗	16	modelled	✗	✗	2.99	✗	✗	✗	✗
ShapeNet-Mats [13]	3,2K	✗	3	modelled	?	6	6.20	✗	✗	✗	✗
PartNet [3]	26,7K	✗	24	modelled	✗	✗	7.21	✓	✓	✗	✗
3DCOMPAT++	10K	**10M**	**42**	modelled	**293**	**13**	**9.86**	✓	✓	**160M**	exact

textures. ABO [5] also provides unlabelled part instance segments for a subset of 3,4K shapes, in the form of unnamed connected shape pieces. HAL3D [18] builds on ABO using an active learning pipeline to provide semantic part instance names for ABO, but did not release the dataset to the public. Google Scanned Objects [16] is a scanned dataset of reconstructed 3D objects with high-quality textures and geometries, and contains 1K 3D shapes from 17 diverse categories of small objects. OmniObject3D [4] is a scanned dataset of 3D objects with high-quality textures, and contains 6K 3D shapes from 190 shape categories based on ImageNet [20] and LVIS [21]. ABO [5] is a dataset of 3D objects with high-quality textures and geometries, and contains 8K 3D shapes from 63 shape categories based on product catalogs extracted from Amazon.

3D-Future [6] presented a dataset comprising 10K industrial 3D CAD shapes of furniture developed by professional designers. More recently, Objaverse [7] and Objaverse-XL [8] expanded the horizon of 3D object datasets by releasing over 10 million artist-designed 3D objects with high-quality textures.

Despite these significant strides in advancing 3D understanding, these modern object-centric 3D datasets (e.g. [1], [2], [7], [8]) and scanned datasets (e.g. [9], [10], [4]) lack part-level annotations. PartNet [3], building on ShapeNet [1], offers fine-grained part segmentations of 3D meshes but does not include material information. The absence of such part-level annotations and material data points to the significance of a dataset like 3DCOMPAT++, which bridges these gaps and serves as a comprehensive resource for furthering research in 3D visual understanding.

Comparison with existing work. In Table I, we compare 3DCOMPAT++ with existing prevalent 3D datasets. We distinguish between datasets originating from 3D artists (first group), scanned objects datasets (second group), datasets with aligned 2D images (third group), and datasets with part-level annotations (fourth group). We scrutinize fundamental aspects, including the number of shapes provided, whether or not stylized shapes are included, the number of classes represented and whether shapes come from scans or are designed from CAD tools. Additionally, we assess the availability of material annotations. We differentiate cases where textured shapes are provided but without material annotations (?) like in GSO [16] and Objaverse-XL [8], from cases where they are provided at a coarse-grained level only (e.g. 3D-Future [6] in which only coarse material annotations are available), or are provided at both coarse and fine-grained levels. Material annotations and part-wise material annotations are important as they provide essential contextual information about the surface properties and appearance of objects, facilitating compositional understanding and analysis of 3D shapes.

We also consider the inclusion of aligned 2D images, and differentiate between cases where images are pseudo-aligned or exactly aligned with matching 3D shapes. Pseudo-alignment includes using a manual 3D alignment pipeline with close candidate CAD models [17], or using an automatic 3D alignment strategy with exactly matching shapes (e.g. based on differentiable rendering [5]). Exact alignments are achieved by producing synthetic 2D images from 3D models using a rendering engine, and then projecting the 3D models into the 2D images using the camera pose ground truth (e.g. 3D-Future [6], this work). In contrast to other works, 3DCOMPAT++ emerges distinctively by offering a large collection of 10K stylized shapes, each accompanied by complete multi-level part-material information. With PartNet, 3DCOMPAT++ is the only dataset with instance-level part annotations, which are essential in tasks involving denumerating parts composing a shape.

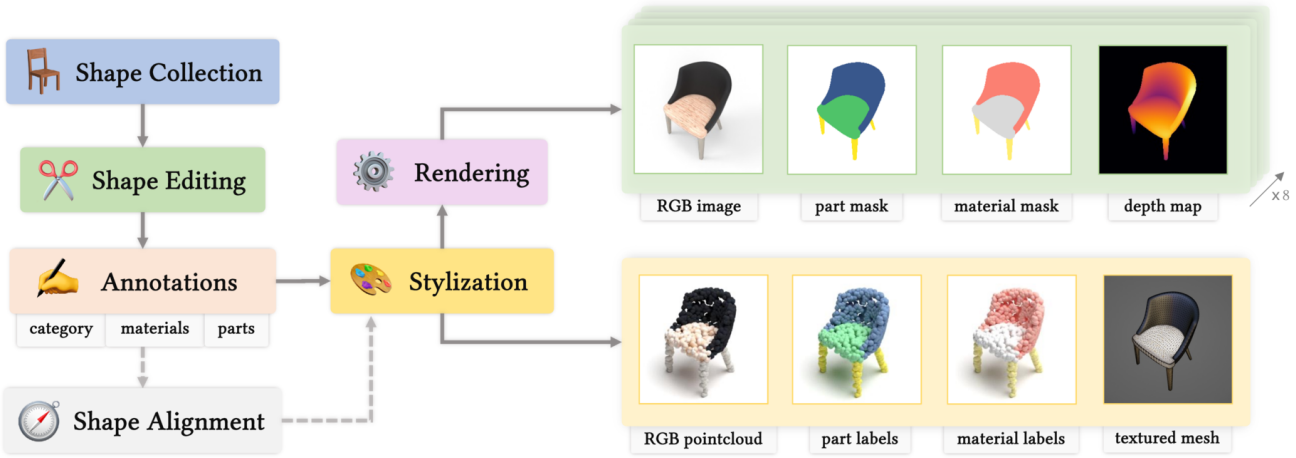


Fig. 4. **Detailing the data pipeline of 3DCoMPaT^{++}.** Starting from the collection of 3D shapes, we perform a first editing step consisting of model re-scaling, UV map correction and removal of undesirable meshes. Material compatibility information is also collected for each part of each shape, alongside the shape category. Shapes are then annotated at a fine-grained part-instance level, and part names are iteratively refined and uniformized using a web-based shape visualizer. Misaligned shapes are semi-automatically realigned using part annotations as a prior. Finally, we sample a set of materials for each part of each shape, and render each stylized shape from multiple viewpoints.

Notably, 3DCoMPaT^{++} also offers a large collection of aligned 2D/3D data with **160M images** and **10M** shapes, enabling its use in diverse multi-modal learning applications benefiting from scale like object classification, part recognition or novel view synthesis.

Mesh resolutions. In Figure 3, we compare part instances per shape distributions between 3DCoMPaT^{++} and PartNet [3] (left), and model resolution statistics for 3D CAD models from 3DCoMPaT^{++} and ShapeNet [1] (right). This comparison is important because ShapeNet [1] serves as the CAD model data source for PartNet [3] and ShapeNet-Part [11], which are two of the most prominent datasets for 3D part understanding. We provide density plots over vertex counts, edges counts, and faces counts for 3D CAD models from both datasets. We show that 3DCoMPaT^{++} exhibits both higher numbers of part instances per shape compared to PartNet and significantly higher average numbers of vertices, edges, and faces when compared to ShapeNet. While polygon count is not a perfect proxy for shape visual quality and realism, it is a useful metric for comparing the relative complexity of meshes in each dataset. This quantitative assessment underscores the richness of annotations and geometries within 3DCoMPaT^{++}, making it a valuable resource for advancing research in 3D shape analysis and understanding.

III. 3DCoMPaT^{++}

The 3DCoMPaT^{++} dataset is based on a collection of artist-designed 3D CAD models collected and annotated in collaboration with an industry partner. It contains 10K geometries annotated and segmented at a fine-grained part-instance level, with material compatibility information for each annotated part. For each shape, 8 rendered views are provided from canonical and random viewpoints (see Figure 6). For each rendered view, depth maps, part maps, and material maps are rendered (see Figures 7 and 21).
All annotations are provided by trained annotators following a rigorous multi-stage review process. 3DCoMPaT^{++} is a richly annotated, multimodal 2D/3D dataset: In Figure 1, we illustrate all data provided for a single stylized shape from our dataset.

A. Dataset

3D Data. Alongside each stylized shape, we provide a part-segmented textured 3D mesh, an RGB pointcloud, and point-wise and triangle-wise part and material annotations. All part segmentation information is provided in *coarse-grained* and *fine-grained* semantic levels. RGB pointclouds can be resampled at any resolution starting from the available textured 3D meshes. In Figure 1, we illustrate the 3D data provided for a single stylized shape.

2D Data. Each stylized shape is rendered from 8 viewpoints: 4 canonical viewpoints and 4 random viewpoints. Canonical viewpoints are equally spaced around the shape. Random viewpoints are sampled uniformly on the upper hemisphere centered on the center of the shape's bounding box. In Figure 7, we showcase the 2D data provided for the first *canonical* viewpoint across four different 3D shapes. Each 2D image is accompanied by part segmentation masks, material masks, and depth maps. For each image, camera parameters are also provided. Part segmentation masks and material masks are available in two semantic levels: *coarse-grained* and *fine-grained*.

B. Data collection pipeline

The complete data collection pipeline is depicted in Figure 4, and includes the following steps:

- **Collection and Editing.** 3D shapes are collected and edited by our industry partner.
- **Part annotations.** Annotators follow each category-level guideline when adding instance-level part annotations and segmentations to each shape.

Fig. 5. **Rendering of randomly sampled shapes from 3DCoMPaT^{++}.** The dataset comprises a rich collection of stylized 3D shapes annotated at the part-instance level. These renderings demonstrate the varying shapes, styles, and materials that are captured, enabling comprehensive exploration and analysis of compositional 3D vision tasks. Shapes are consistently aligned across classes and orientations are consistent for all 3D models. In the left circle, we illustrate the untextured 3D geometries we start from as a reference. We provide additional reference shapes from all 42 shape categories in Figure 17 of the appendix.

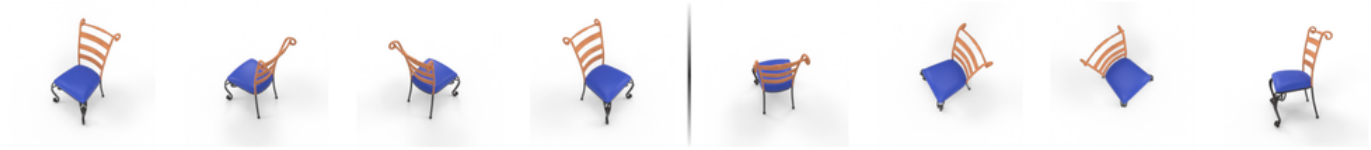


Fig. 6. **Canonical and random views.** Canonical (**left**) and random (**right**) views rendered for a single stylized 3D shape. Random viewpoints are sampled, while canonical viewpoints are equally spaced around the shape. In Figure 20 of the appendix, we provide additional examples of canonical and random views.

- **Material assignments.** Annotators select compatible materials for each part of each shape, from among 13 possible coarse classes.
- **Stylized shapes.** We sample a set of fine-grained materials for each part of each shape, which we refer to as a *style*.
- **Rendering.** We render each shape from multiple viewpoints with matching masks, depth maps and pointcloud data, as detailed in Section III-D.

Collection and Editing. All 3D shapes are collected by our industry partner. Editing steps include model scaling, the correction of UV maps, the removal of undesirable/invalid meshes in the shape (e.g., additional objects like a vase on top of a table), etc. Furthermore, as visible in Figure 5, all shapes are consistently aligned across classes and orientations are consistent for all 3D models. To align shapes, we use part annotations as a prior to automatically rotate a majority of misaligned shapes (for example, using the fact that the "`vertical_back_panel`" part should appear at the back of a shape). We then manually adjust the remaining misaligned shapes by using a web visualization tool[2] (see Figure 14). 3D shapes are also scaled to fit within a unit cube centered at the world origin.

Part annotations. By combining expert knowledge with the analysis of unannotated shapes, we define fine-grained part-level guidelines. A guideline is defined for each shape category and provides a non-ambiguous definition of each possible fine-grained part that can occur in shapes belonging to the category. Annotators follow each category-defined guideline when adding instance-level part annotations and segmentations to each shape (see Figure **??** for an example of a shape guideline for the `faucet, shower, sink` shape categories). Part segments and names are iteratively refined using a web-based shape visualizer (see Figure 14). This

[2]The 3DCoMPaT annotated shapes web-based browser is accessible here: https://3dcompat-dataset.org/browser/

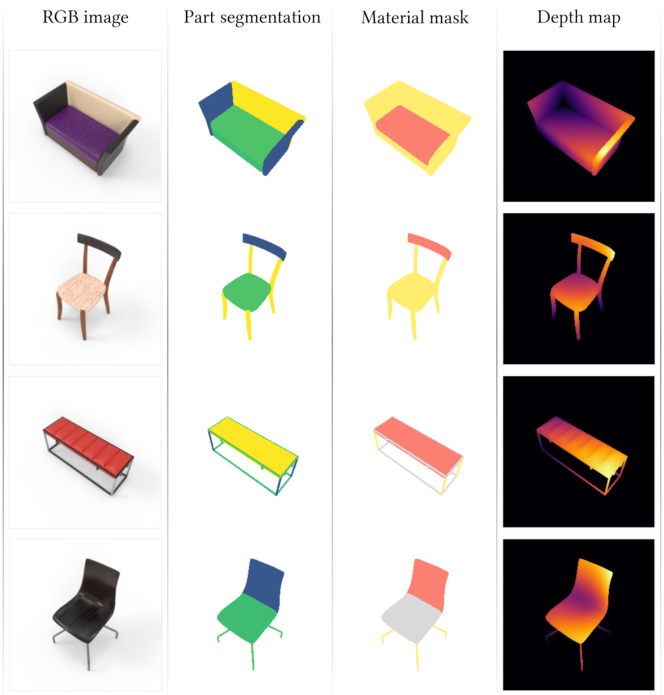


Fig. 7. **2D data associated with each viewpoint.** Each 2D image is complemented by corresponding part segmentation masks, material masks, and absolute depth maps, enabling various 2D/3D vision tasks. Camera parameters are also provided for each render. See Figure 21 in the appendix for additional examples.

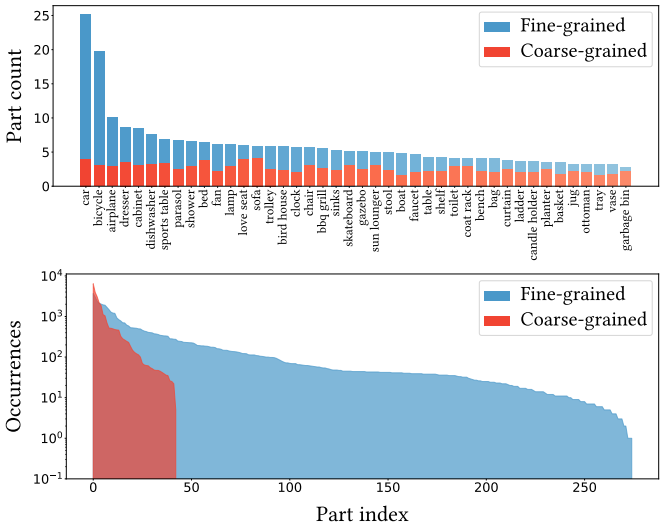


Fig. 8. **Distribution of part occurrences for fine/coarse levels.** We plot the average number of unique parts per object for fine and coarse semantic levels, across all shape categories (**top**). We also plot the sorted number of occurrences of each part (logscale, **bottom**).

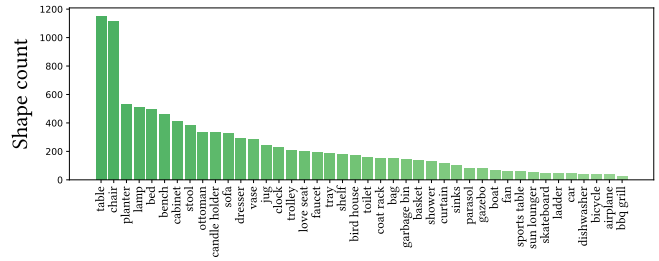


Fig. 9. **Distribution of shape category occurrences.** We plot the sorted number of occurrences of each shape category. The class distribution in 3DCoMPaT^{++} is significantly long-tailed.

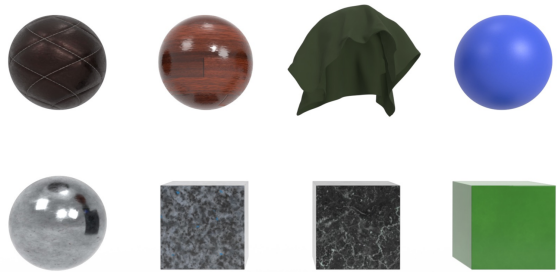


Fig. 10. **Example coarse materials in 3DCoMPaT^{++}.** We randomly sample a single material from 8 of the 13 materials provided in 3DCoMPaT^{++}, and render it on demo primitives. From left to right, top to bottom: `leather`, `wood`, `fabric`, `plastic`, `metal`, `granite`, `marble`, `ceramic`.

browser allows reviewers to visualize segments for a specific part class in a shape category, allowing to efficiently verify part semantics consistency across shapes and quickly identify annotation errors. Corner cases, when frequent enough, are identified and further refined into new meaningful part denominations for the category.

Material assignments. In Figure 10, we illustrate material categories in 3DCoMPaT^{++} with samples from our collection. We collect Physics-Based Rendering (PBR [22]) materials from various free-to-use repositories, including the NVIDIA vMaterials[3] library and the ambientCG[4] public domain material library. We filter collected PBR materials to ensure **1)** overall visual quality, **2)** compatibility with our rendering pipeline, **3)** visual affinity with our collected shapes. We collect a total of 13 coarse material categories, for 293 total PBR materials. With each segmented part, a set of compatible material categories is provided by the annotators (e.g. "metal, wood" for a leg in a chair.). The list of compatible materials for each part of each shape is first broadly defined at the shape category level and refined on a case-by-case basis for specific shapes.

Stylized shapes. Using the collected material compatibility information associated with each part, we randomly sample a material for each part of a shape to create a *style*. A *composition* is a combination of materials that could be applied to any shape, and a *style* is an instance of a *composition* applied to a specific shape. We detail the process of shape stylization in Figure 13. An average of 1000 styles are sampled per shape. The number of possible styles per shape S can be defined as:

$$\mathcal{N}(S) = \prod_{p \in \mathcal{P}(S)} |\mathcal{M}(S,p)|$$

where $\mathcal{P}(S)$ denotes the set of parts belonging to shape S, and $\mathcal{M}(S,p)$ the set of materials compatible with part p in shape S. For 14.6% of shapes, $\mathcal{N}(S) < 1000$, due to either a small number of parts or compatible materials per part. To

[3] vMaterials library: https://developer.nvidia.com/vmaterials
[4] ambientCG public domain repository: https://ambientcg.com/

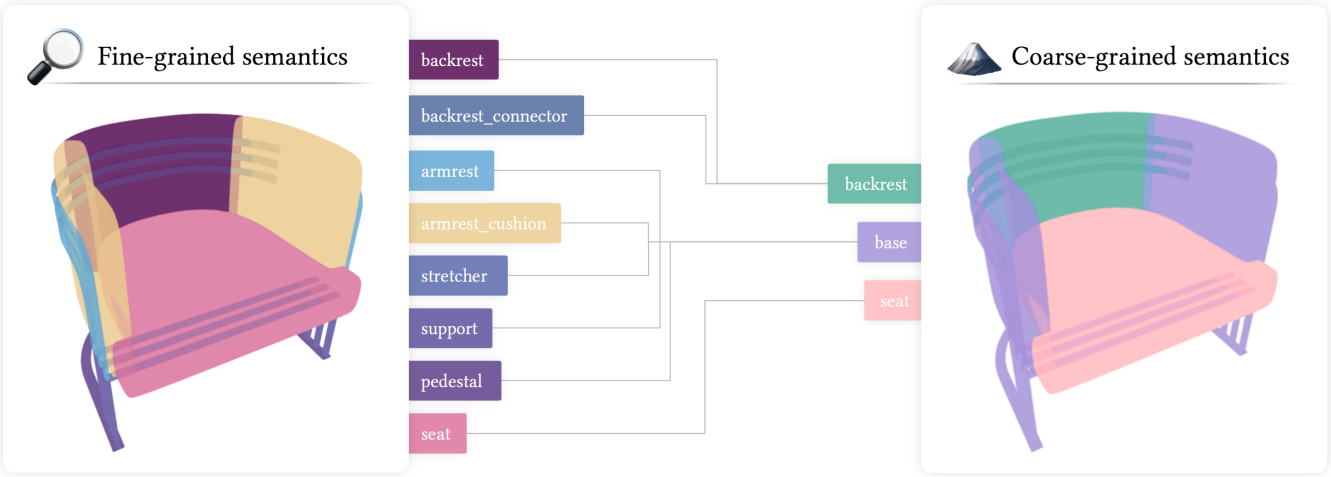


Fig. 11. **Going from fine-grained to coarse-grained segmentation.** Fine-grained classes are merged following a shape category-specific nomenclature to create coarse-grained classes. Resulting shapes are simplified and contain fewer parts.

compensate for this effect, we oversample from shapes where $\mathcal{N}(S) >> 1000$ to reach the desired average of 1000 styles per model.

C. Coarse/Fine-grained semantics

3DCoMPaT^{++} provides part and material annotations in two hierarchical semantic levels: *coarse* and *fine*.

Part hierarchies. Fine-grained part classes are defined from a hand-defined shape category-specific nomenclature. Coarse-grained part semantics are defined as shape category-specific groupings of fine-grained part categories. For example, in the "`airplane`" shape category, the `wheel`, `wheel_connector` and `wheel_cover` *fine-grained* parts are all merged into the `wheel` *coarse* part. Shape categories in our dataset can share part names by default. Parts that are category specific and relevant to the category are prefixed by the name of the category (for example: `airplane_wing`). We visualize these two semantic levels in Figure 11 for a single 3D shape, and highlight resulting grouped parts. The coarse-grained semantic level considerably simplifies the compositional structure of shapes, while the fine-grained semantic level provides a more detailed description of the composition of shapes. In the coarse-grained setting, the number of shape category-specific parts is also significantly reduced, while the number of parts shared across shape categories is increased. In Figure **??**, we provide additional examples of coarse-grained and fine-grained part semantics groupings for three distinct shape categories. Our coarse-level part semantics can be used for tasks that require a high-level understanding of shapes, while fine-grained part semantics can be used for tasks that require more detailed, shape category-specific understanding. We compare the average number of unique parts per object (**top**) for fine and coarse semantic levels across all shape categories in Figure 8. We also plot the sorted number of occurrences of each part (**bottom**) in both semantic levels. In the coarse-grained semantic level, parts occurrences are concentrated on a smaller number of parts, while the distribution for the fine-grained level is clearly long-tailed. The average number of parts per object is also equalized across shape categories in the coarse-grained level, while some shape categories present a significantly higher number of parts per object in the fine-grained level like cars and bicycles. Overall, the coarse-grained semantic level provides a more balanced distribution of parts across shape categories, while the fine-grained semantic level provides a more detailed description of the composition of shapes.

Material hierarchies. Coarse-grained materials correspond to a high-level set of 13 material categories (e.g. "`wood`, `metal`, `ceramic`, etc.). Each high-level material category is composed of fine-grained specific materials belonging to that category (e.g. "`pine_wood`" in "`wood`".). In Table II, we detail the number of fine-grained materials within each coarse category in 3DCoMPaT^{++}.

TABLE II

MATERIALS IN THE 3DCOMPAT^{++} DATASET. WE SHOW THE NUMBER OF FINE-GRAINED MATERIALS WITHIN EACH OF THE 13 COARSE-GRAINED CATEGORY.

Material	ceramic	fabric	glass	granite	leather	marble	metal
# **Count**	6	32	6	11	34	39	66

Material	plant	plastic	rubber	soil	wax	wood	**Total**
# **Count**	1	21	5	4	4	64	**293**

D. Rendering

Scene. We render each shape in the same scene with a single directional light and three area lights positioned around the shape. In Figure 12, we detail our rendering scene setup with an example shape. The stylized shape is placed inside an ovoid surface with a white color, to ensure that the shape is always rendered on a uniformly white background. Projected shadows only appear on the $z = 0$ plane on which the shape is placed. When rendering depth maps and masks, the background surface is removed from the scene.

All images are rendered in a 256x256 resolution, and 2D images are encoded in the PNG format. Depth maps are

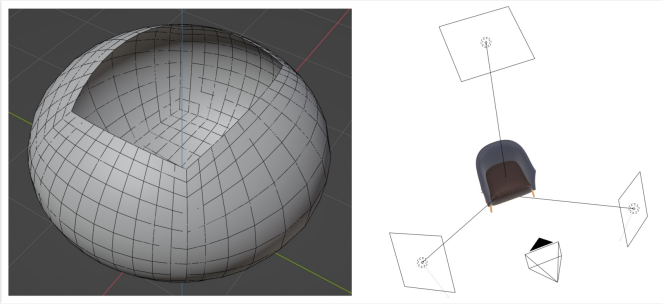


Fig. 12. **Blender rendering scene setup.** We depict the ovoid surface in the middle of which shapes are placed (**left**). We also show the scene setup used with three area lights and the camera (**right**). The directional light is positioned above the whole scene and centered on the shape.

stored in the OpenEXR format to accommodate absolute distances to the image plane, which are represented using floating-point values.

Viewpoints. Each stylized shape is rendered from multiple perspectives: 4 canonical viewpoints and 4 random viewpoints. We first translate each shape above the $z = 0$ plane. Camera viewpoints are defined in spherical coordinates (ϕ, θ) where the origin is set to the center of the shape's bounding box, which we note $\mathbf{o}_c$. The camera is rotated around $\mathbf{o}_c$ by ϕ and θ. Canonical viewpoints are distributed evenly around the shape with a fixed elevation θ. We set the base spherical angle ϕ to 40 degrees and then increment it by 90 degrees for each of the four viewpoints, while keeping θ fixed at 0 degrees. Random viewpoints are sampled uniformly from an upper hemisphere above the plane. We randomly sample ϕ from the range $[0, 2\pi]$ and θ from the range $[-\frac{1}{3}\pi, \frac{1}{3}\pi]$. Using the obtained ϕ and θ angles, we define the position and orientation of the camera. The camera's initial position denoted as $\mathbf{c_0}$ is rotated around $\mathbf{o_c}$. The orientation of the camera is then adjusted to ensure that the image plane is centered on $\mathbf{o_c}$. Extrinsic and intrisic camera parameters are recorded for each view and are provided alongside the rendered images. The sampling procedure of camera parameters is detailed in Algorithm 1.

E. Toolbox

To support the use of 3DCoMPaT^{++}, we provide a toolbox for easily loading and visualizing the data. Mainly, we provide the following elements:

- **Python API** for easily loading the data, based on PyTorch [24] and WebDataset [25].
- **Web-based browser** for easily exploring 3D shapes and part annotations in both coarse and fine-grained semantic levels (see Figure 14).
- **Documentation and notebooks** to facilitate the use of the dataset.

All of these elements are available on the 3DCoMPaT^{++} website[5].

[5]3DCoMPaT^{++} website: https://3dcompat-dataset.org/doc

Algorithm 1 Sampling camera parameters

Input:
- $\mathbf{obj} \in \mathbb{R}^{N\times 3}$: Object to render.
- $\mathbf{c_0} \in \mathbb{R}^3$: Base position of the camera.
- $v_k \in [\![0, 3]\!]$: View identifier.
- `is_random_view` $\in \{0, 1\}$: Sample a random view.

Output:
- $\mathbf{m_w} \in \mathbb{R}^{4\times 4}$: Camera transformation matrix.

function SAMPLE_CAMERA($\mathbf{obj}, \mathbf{c_0}, v_k$, `is_random`)
  $\mathbf{o_c} \leftarrow$ bounding_box_center($\mathbf{obj}$)
  if `is_random_view` **then**
    $\theta \sim \mathcal{U}\left[-\frac{1}{3}\pi, \frac{1}{3}\pi\right],\ \phi \sim \mathcal{U}[0, 2\pi]$
  else
    $\theta \leftarrow 0,\ \phi \leftarrow \frac{2}{9}\pi + \frac{\pi}{2}v_k$
  $\mathbf{c_1} \leftarrow$ rotate_point($\mathbf{c_0};\ \mathbf{o_c}, \phi, \theta$)
  $\mathbf{m_w} \leftarrow$ look_at($\mathbf{c_1}, \mathbf{o_c}$)
  return $\mathbf{m_w}$ ▷ Camera transformation matrix

IV. Experiments

A. Classification and Segmentation

Shape classification. As illustrated in Figure 9, the shape class distribution of our dataset is significantly long-tailed. We conduct shape classification experiments on 2D renders and 3D XYZ pointclouds to assess the difficulty of this task on our dataset. All pointclouds are sampled with a resolution of 2048 points, and all methods are trained from scratch for 200 epochs. For 2D classification, we fine-tune ResNet models [26] pretrained on ImageNet [20] for 30 epochs. We report 2D and 3D shape classification results in Table V. We reach a maximum top-1 accuracy of 90.20% on 2D renders with ResNet-50, and 85.14% on 3D pointclouds with CurveNet [27].

Part segmentation. We conduct 3D part segmentation experiments on pointclouds and 2D renders to assess the difficulty of this task on our dataset. We provide results for both fine-grained and coarse-grained 3D part segmentation in Table III. We report pointwise accuracy (*shape-agnostic*) and mIOU for each model. For mIOU, we consider the shape-informed version of the metric where we restrict the set of predicted parts to the parts that are present in the ground-truth shape category, and the shape-agnostic version where all possible parts are considered. We also report results with and without using a shape prior during training and inference for PCT [28], PointNet++ [29] and CurveNet [27]. We note that getting accurate part segmentations without RGB information is challenging but remains possible. Without using a shape prior, CurveNet [27] reaches a shape-agnostic mIOU of 53.09% on fine-grained part segmentation. In this setting, the model has to perform the challenging task of point-wise part classification from a set of 275 possible parts.

Overall, a large gap exists (around 30 accuracy points across models) between shape-informed and shape-agnostic mIOU, highlighting the difficulty of the task over the full space of possible parts. The task of coarse-grained part segmentation

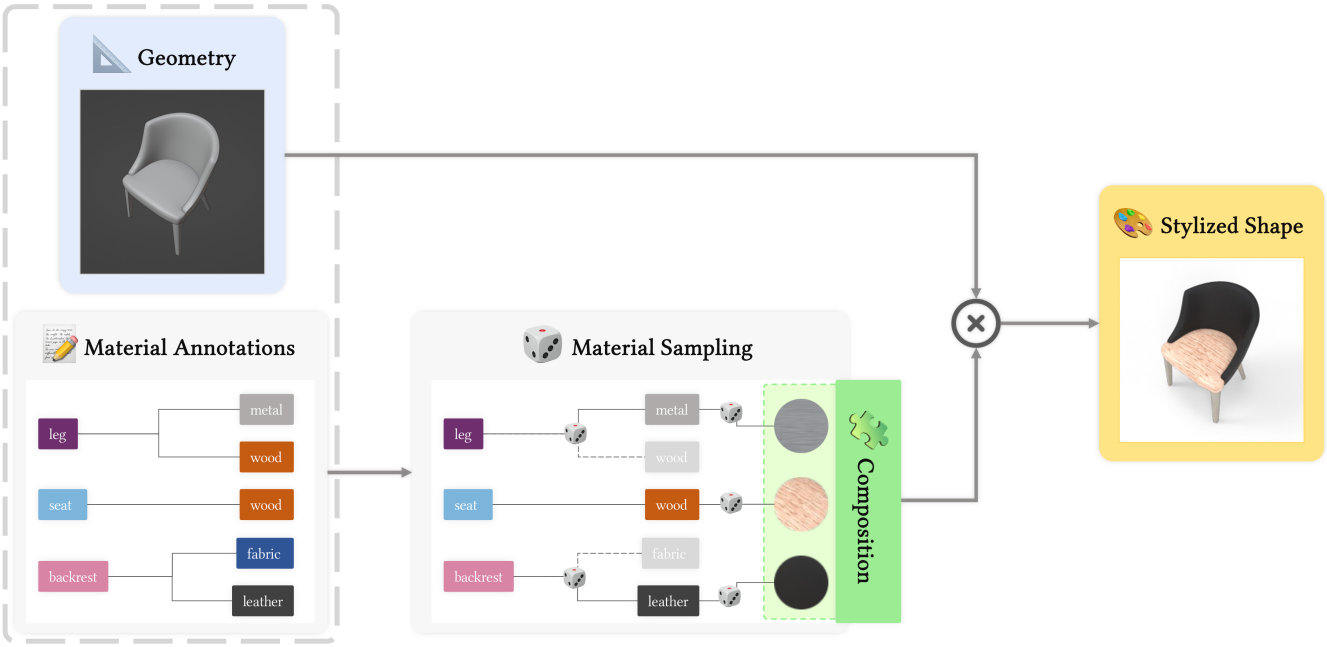


Fig. 13. **Sampling a shape style.** Starting from part-material annotations, we first sample a coarse material class from a set of compatible material classes for each part. We then sample fine-grained materials within each sampled coarse material class. The combination of fine-grained materials for each part defines a *composition*. Finally, we apply the sampled part-materials pairs to the shape to obtain a stylized shape.

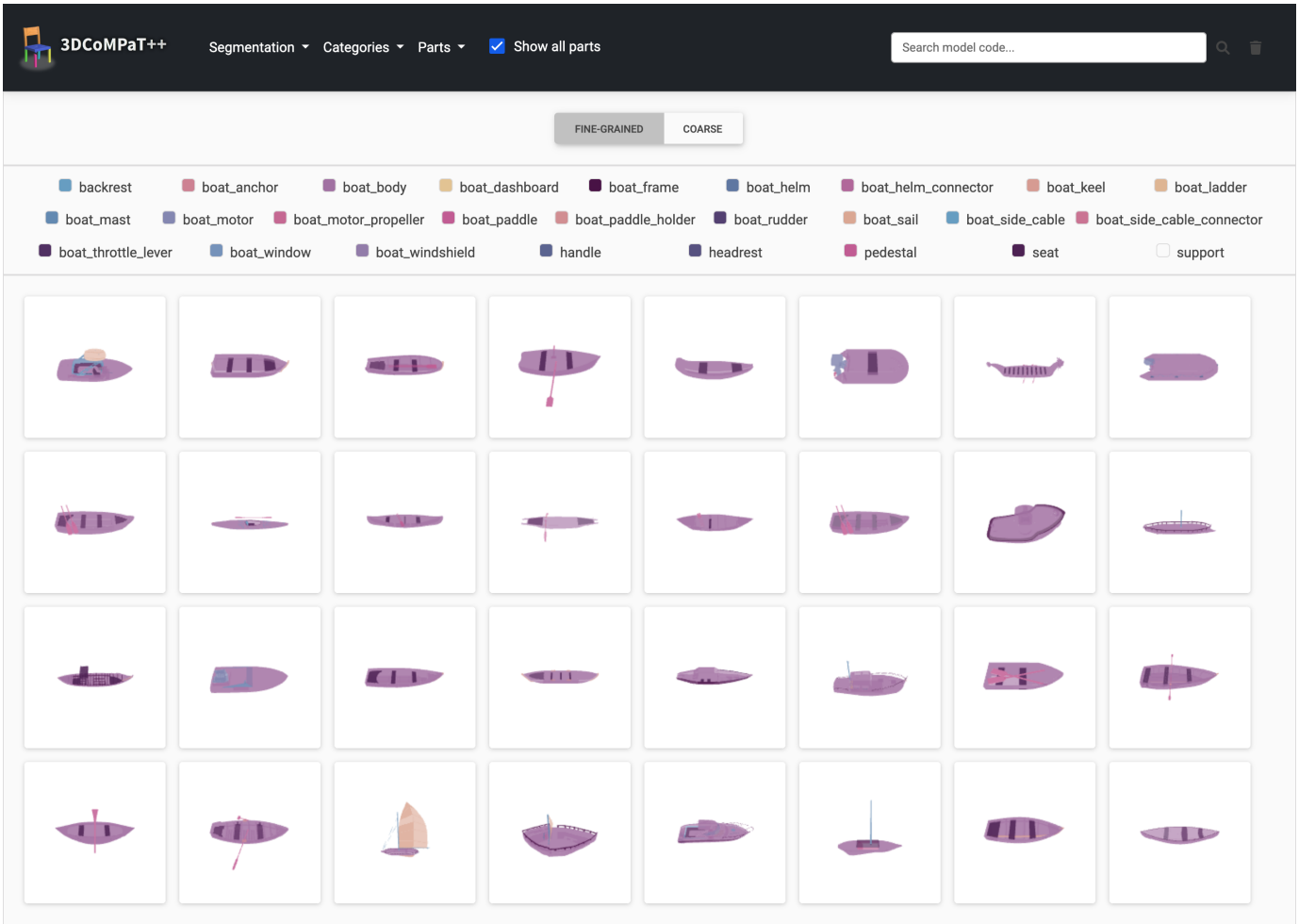


Fig. 14. **Interactive web-based browser for 3DCoMPaT^{++}.** We provide an interactive 3D shape browser built using Three.JS [23] to visualize the 3D shapes in our dataset. The browser allows to easily visualize instance and semantic part annotations in both coarse and fine-grained levels.

is easier, as the model only has to perform part classification from a set of 43 possible parts. In this setting, CurveNet [27] reaches a shape-agnostic mIOU of 76.32%. For 2D fine and coarse part segmentation, we report results for SegFormer [30] in Table IV, alongside material segmentation results. We obtain a mIOU of 52.24% for fine-grained part segmentation, 73.35% for coarse-grained part segmentation, and 82.45% for material segmentation.

B. Grounded Compositional Recognition (GCR)

Task. One key property of our 3DCoMPaT^{++} dataset is that it enables understanding of the complete part-material compositions of a given 3D shape. This involves predicting the category of the object, all part categories, and the associated materials for each of those parts within the 3D model. In Figure 2, we detail all information that has to be predicted for a given shape in this proposed GCR task. Grounded Compositional Recognition can be related to Zero-Shot Recognition, which aims at predicting the category of an object from a set of unseen categories, where the unseen categories are defined by unseen compositions of visual attributes [31], [32], [33]. The GCR task can also be related to situation recognition [34], [35] which can be defined as the identification of role - entity pairs in a given scene [34].

Metrics. Drawing inspiration from the metrics introduced in [34], [35] initially designed for the compositional recognition of activities in images, we define the GCR compositional metrics in 2D/3D as follows:

- **Shape accuracy.** Proportion of correctly predicted shape categories.
- **Value.** Proportion of correctly predicted part-material pairs.
- **Value-all.** Accuracy of predicting all part-material pairs of a shape correctly.

We extend these metrics to the segmentation of parts and materials in 2D/3D by defining *grounded* variants of **value** and **value-all** metrics:

- **Grounded-value.** Proportion of correctly predicted part-material pairs, where the part is correctly segmented.
- **Grounded-value-all.** Accuracy of predicting all part-material pairs of a shape correctly, where all parts are correctly segmented.

We consider a part to be correctly segmented if the predicted part segmentation mask has an intersection over union (IoU) of at least 0.5 with the ground-truth part segmentation mask. In 2D, we use the pixel-wise definition of IoU. For the 3D modality, we use the point-wise definition.

Formally, for a test set with N shapes, let y_i and $\hat{y}_i$ denote the true and predicted shape categories for shape i, respectively. Additionally, let $\mathcal{C}_i = \{(p_1, m_1), (p_2, m_2), \ldots, (p_{K_i}, m_{K_i})\}$ represent the set of true part-material pairs for shape i, and $\hat{\mathcal{C}}_i$ the corresponding predicted pairs. Let $\mathcal{P}_i$ denote the set of all parts in shape i. We define the metrics as follows:

Shape Accuracy measures the proportion of correctly identified shape categories:

$$\text{SHAPE ACCURACY} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(y_i = \hat{y}_i)$$

Value computes the average proportion of correctly predicted part-material pairs per shape:

$$\text{V} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\mathcal{C}_i \cap \hat{\mathcal{C}}_i|}{|\mathcal{C}_i|}$$

Value-all requires perfect prediction of all part-material pairs:

$$\text{VALL} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\mathcal{C}_i = \hat{\mathcal{C}}_i)$$

With $\mathbb{1}(\cdot)$ being the indicator function. Note that performing well on these metrics only requires part-material labels to be predicted, and does not require part segmentation masks. For the grounded variants, we define $\mathcal{S}_i$ as the set of correctly segmented parts with IoU ≥ 0.5. This allows us to define:

Grounded-value which measures correctly predicted part-material pairs that are also correctly segmented:

$$\text{GV} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\{(p,m) \in \mathcal{C}_i : p \in \mathcal{S}_i \wedge (p,m) \in \hat{\mathcal{C}}_i\}|}{|\mathcal{C}_i|}$$

Grounded-value-all which requires all part-material pairs to be correctly predicted and all parts to be correctly segmented:

$$\text{GVALL} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left((\mathcal{C}_i = \hat{\mathcal{C}}_i) \wedge (\forall p \in \mathcal{P}_i : \text{IoU}(p, \hat{p}) \geq 0.5)\right)$$

Note that *Value* and *Grounded-value* are both evaluated at the shape level: we divide the number of correctly identified (resp. grounded) part-material pairs by the total number of parts appearing in each shape, and then average across all samples. *Value* is thus upper bounded by *Value-all*, and *Grounded-value* by *Grounded-value-all*.

Baselines. We experiment with two fusion-based baselines to assess the performance of the GCR task on 3DCoMPaT^{++}.

- **"PointNeXt+SegFormer"**. This baseline employs separate 2D/3D models and fuses predictions at evaluation time. We use PointNeXt [36] for 3D shape classification and SegFormer [30] for 2D material segmentation and 2D part segmentation. 2D dense predictions are projected to the 3D space using the depth maps and camera parameters. We use this baseline to assess the feasibility of the GCR task on 3DCoMPaT^{++} when all part-pair predictions are performed on the 2D space.

- **BPNet.** We adapt the BPNet 2D/3D multimodal method to the GCR task. BPNet leverages complementary information from 2D and 3D modalities by fusing features from both modalities using a bidirectional projection module for feature fusion. We detail the BPNet architecture we employ in Figure 24 in the appendix.

TABLE III

3D PART SEGMENTATION. WE REPORT MEAN INTERSECTION OVER UNION (MIOU) AND POINTWISE ACCURACY FOR VARIOUS MODELS ON THE 3DCOMPAT^{++} DATASET FOR BOTH FINE-GRAINED (**LEFT**) AND COARSE-GRAINED (**RIGHT**) 3D PART SEGMENTATION. FOR MIOU, WE DIFFERENTIATE BETWEEN SHAPE-INFORMED (WHERE THE GROUND-TRUTH SHAPE CATEGORY IS PROVIDED AS INPUT) AND SHAPE-AGNOSTIC EVALUATION. FOR PCT, POINTNET++ AND CURVENET, WE ALSO REPORT RESULTS WITH AND WITHOUT USING A SHAPE PRIOR DURING TRAINING AND INFERENCE. ALL MODELS ARE TRAINED ON 3D XYZ POINTCLOUDS.

Model	Shape Prior	Pointwise Acc. (%)	mIOU (%)	
			Shape-Informed	Shape-Agnostic
PCT [28]	✗	70.49	81.31	49.09
	✓	78.51	82.84	56.14
PointNet++ [29]	✗	71.09	80.01	50.39
	✓	78.61	81.19	56.46
CurveNet [27]	✗	72.49	81.37	53.09
	✓	79.90	82.15	59.61
PointNeXt [36]	✓	82.07	83.92	63.73
PointStack [37]	✓	78.51	81.98	56.20

Fine-grained segmentation

Model	Shape Prior	Pointwise Acc. (%)	mIOU (%)	
			Shape-Informed	Shape-Agnostic
PCT [28]	✗	80.64	75.49	66.95
	✓	92.15	82.57	80.49
PointNet++ [29]	✗	84.73	77.98	73.79
	✓	92.02	81.82	81.03
CurveNet [27]	✗	86.02	80.64	76.32
	✓	93.40	84.62	83.85
PointNeXt [36]	✓	94.18	86.80	85.46
PointStack [37]	✓	93.49	84.97	83.67

Coarse-grained segmentation

TABLE IV

2D PART SEGMENTATION. WE REPORT MEAN INTERSECTION OVER UNION (MIOU) FOR SEGFORMER ON THREE 2D PART SEGMENTATION TASKS: FINE-GRAINED PART SEGMENTATION, COARSE-GRAINED PART SEGMENTATION AND MATERIAL SEGMENTATION.

Model	Task	mIOU (%)
SegFormer [30]	Fine-grained part segmentation	52.24
	Coarse-grained part segmentation	73.35
	Material segmentation	82.45

TABLE V

SHAPE CLASSIFICATION. WE REPORT THE ACCURACY OF VARIOUS MODELS ON THE 3DCOMPAT^{++} DATASET FOR BOTH 3D AND 2D SHAPE CLASSIFICATION. 3D MODELS ARE TRAINED ON POINTCLOUDS PROVIDED WITH EACH SHAPE, WHILE 2D MODELS ARE TRAINED AND TESTED ON RENDERS WITH CANONICAL AND RANDOM VIEWPOINTS.

Model	2D Classification
	Accuracy (top-1, %)
ResNet-18 [26]	76.27
ResNet-50 [26]	90.20

Model	3D Classification
	Accuracy (top-1, %)
PCT [28]	68.88
DGCNN [38]	78.85
PointNet++ [29]	84.10
PointStack [37]	83.04
CurveNet [27]	85.14
PointNeXt [36]	82.21
PointMLP [39]	83.71

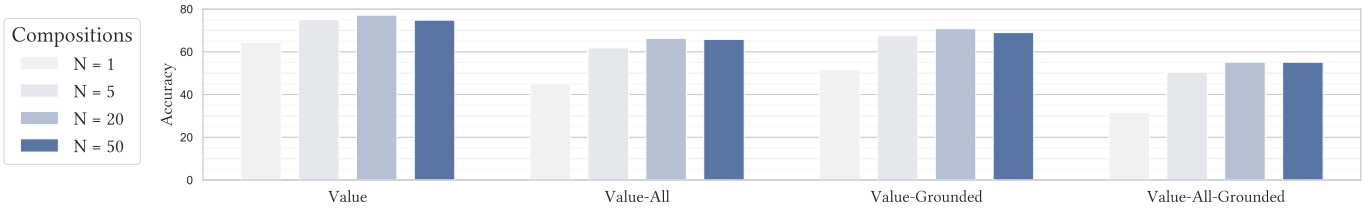


Fig. 15. **Analysis of the performance with different numbers of sampled styles.** We train the BPNet [40] model using 1/5/20/50 style compositions and report all the compositional metrics defined in Figure 2. Overall, we observe a clear trend of improvement with the number of styles, especially for the **Value-All** and **Value-All-Grounded** metrics. We do notice however the start of a saturation effect when training with $N = 50$ styles.

TABLE VI

GROUNDED COMPOSITIONAL RECOGNITION (GCR). WE EVALUATE THE PERFORMANCE OF VARIOUS BASELINES ON THE GCR TASK. WHILE MODALITY FUSION-BASED METHODS LIKE BPNET [40] AND POINTNEXT+SEGFORMER [30], [36] PERFORM WELL ON THE SHAPE CLASSIFICATION TASK THEY STILL UNDERPERFORMED COMPARED TO THE RGB POINTCLOUD-BASED BASELINE. WE REPORT THE GCR METRICS UNDER BOTH FINE-GRAINED AND COARSE-GRAINED SETTINGS, USING 10 COMPOSITIONS PER SHAPE. OVERALL, RECOGNIZING AND GROUDING ALL PART-MATERIAL PAIRS OF A SHAPE IS PARTICULARLY CHALLENGING, ESPECIALLY IN THE FINE-GRAINED SETTING.

Semantic level	Model	Shape Acc.	Value	Value-all	Grounded-value	Grounded-value-all
Fine-grained	PointNeXt+SegFormer [30], [36]	**84.18**	42.57	9.05	26.68	3.84
	BPNet [40]	79.57	**59.98**	**27.74**	45.46	15.41
	PointNet++$^{\text{RGB}}$ [29]	83.70	57.78	25.36	**49.34**	**17.55**
Coarse-grained	PointNeXt+SegFormer [30], [36]	84.27	65.61	44.82	52.82	29.74
	BPNet [40]	84.72	75.04	61.81	67.49	50.44
	PointNet++$^{\text{RGB}}$ [29]	**85.19**	**75.66**	**63.88**	**72.14**	**58.99**

Challenge. We organized a compositional 3D visual understanding challenge on the GCR task of 3DCoMPaT^{++}, with the goal of benchmarking the performance of various methods, in the context of the C3DV workshop at CVPR 2023 [6]. The best-performing method (**PointNet++$^{\mathbf{RGB}}$** in Table VI) on the GCR task consisted of an unimodal 3D model based on a modified PointNet++ [29] trained on 6D inputs (XYZ coordinates and RGB color) [7]. One important design choice is the point grouping method employed which relies on spatial proximity only. The winning method achieved a **Grounded-value-all** accuracy of 58.99% in the coarse-grained setting and 17.55% in the fine-grained setting. Other solutions included a late fusion of 2D and 3D features by averaging logits of part and material segmentation and training a PointNet++ model with additional 2D segmentation features. More information about the challenge submissions can be found on the workshop website.

Results. Table VI summarizes GCR results of baseline methods and challenge winners. The PointNeXt+SegFormer 2D-based baseline is markedly outperformed by the BPNet multimodal baseline, which can be imputed to the absence of explicit 3D-aware modality fusion during training. The winning method **PointNet++$^{\mathbf{RGB}}$**, which takes only 3D point clouds as inputs and leverages a powerful point grouping module, performs best in the coarse-grained setting, reaching 58.99% on the **Grounded-value-all** metric. BPNet on the other hand, performs best in the fine-grained setting, and outperforms PointNet++$^{\mathbf{RGB}}$ on the classification metrics (Value and Value-all), but is overtaken by PointNet++$^{\mathbf{RGB}}$ on the grounded metrics. More importantly, we notice that even our best performing models struggle to perform on the fine-grained GCR task, where we reach most 17.55% on the **Grounded-value-all** metric. This suggests that creating a single model able to achieve strong performance across GCR metrics poses great challenges, especially in the fine-grained setting.

In this sense, Grounded Compositional Recognition is a challenging task that can be used to benchmark the compositional understanding of future multimodal models.

Number of compositions. We conduct further ablation analysis to investigate the impact of varying the number of compositions during the training of the BPNet [40] model. We focus our analysis on the compositional metrics outlined in Figure 15, related to the GCR task (2D/3D material mIoU, 2D shape accuracy, and 3D part mIoU). We train multiple instance of the BPNet model with 1/5/20/50 compositions from each shape and report the performance obtained for each epoch, for each GCR compositional metric. Our findings reveal a clear and consistent improvement in the performance of all metrics as the number of compositions utilized in training is increased, specifically when going from $N_c = 1$ to $N_c = 5$ and $N_c = 20$. However, the observed trend becomes less discernible when transitioning from $N_c = 20$ to $N_c = 50$ compositions. This highlights the need for further investigation into efficient ways of leveraging a large number of compositions during training.

[6] C3DV Workshop: https://3dcompat-dataset.org/workshop/C3DV23/

[7] Winning solution repository: https://github.com/Cattalyya/3DCoMPaT-challenge

V. Conclusion

We introduce 3DCoMPaT^{++}, a large-scale dataset of Compositions of Materials on Parts of 3D Things, which contains 10M styled models stemming from 10000 3D shapes from 42 object categories. 3DCoMPaT^{++} contains 3D shapes, part segmentation information in fine-grained and coarse-grained semantic levels and material compatibility information, so that multiple high-quality PBR materials can be assigned to the same shape part. We also propose a new task, dubbed as Grounded CoMPaT Recognition (GCR), that our dataset enables and introduces baseline methods to solve them. Future directions may include additional tasks such as 3D part-aware shape synthesis, 3D part-aware reconstruction from 2D views, and 3D part-style transfer, which all can be enabled by the rich data provided in 3DCoMPaT^{++}.

Acknowledgements

For computing support, this research used the resources of the Supercomputing Laboratory at King Abdullah University of Science & Technology (KAUST). We extend our sincere gratitude to the KAUST HPC Team for their invaluable assistance and support during the course of this research project. We also thank the Amazon Open Data program for providing us with free storage of our large-scale data on their servers, and the Polynine team for their relentless effort in collecting and annotating the data.

REFERENCES

[1] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An Information-Rich 3D Model Repository," in *arXiv*, 2015. 1, 2, 3, 4, 5

[2] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A Deep Representation for Volumetric Shapes," in *CVPR*, 2015. 1, 3, 4

[3] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A Large-scale Benchmark for Fine-grained and Hierarchical Part-level 3D Object Understanding," in *CVPR*, 2019. 1, 2, 3, 4, 5

[4] T. Wu, J. Zhang, X. Fu, Y. Wang, J. Ren, L. Pan, W. Wu, L. Yang, J. Wang, C. Qian, D. Lin, and Z. Liu, "OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation," in *CVPR*, 2023. 1, 3, 4

[5] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Y. Vicente, T. Dideriksen, H. Arora, M. Guillaumin, and J. Malik, "ABO: Dataset and Benchmarks for Real-World 3D Object Understanding," in *CVPR*, 2022. 1, 3, 4

[6] H. Fu, R. Jia, L. Gao, M. Gong, B. Zhao, S. Maybank, and D. Tao, "3D-FUTURE: 3D Furniture shape with TextURE," in *IJCV*, 2021. 1, 4

[7] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A Universe of Annotated 3D Objects," in *CVPR*, 2023. 1, 2, 4

[8] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, E. VanderBilt, A. Kembhavi, C. Vondrick, G. Gkioxari, K. Ehsani, L. Schmidt, and A. Farhadi, "Objaverse-XL: A Universe of 10M+ 3D Objects," in *arXiv*, 2023. 1, 2, 4

[9] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes," in *CVPR*, 2017. 1, 2, 4

[10] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D Data in Indoor Environments," in *3DV*, 2017. 1, 2, 4

[11] L. Yi, V. G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, and L. Guibas, "A scalable active framework for region annotation in 3D shape collections," in *SIGGRAPH*, 2016. 1, 3, 4, 5

[12] Y. Li, U. Upadhyay, H. Slim, A. Abdelreheem, A. Prajapati, S. Pothigara, P. Wonka, and M. Elhoseiny, "3D CoMPaT: Composition of Materials on Parts of 3D Things," in *ECCV*, 2022. 1

[13] H. Lin, M. Averkiou, E. Kalogerakis, B. Kovacs, S. Ranade, V. G. Kim, S. Chaudhuri, and K. Bala, "Learning Material-Aware Local Descriptors for 3D Shapes," in *3DV*, 2018. 2, 4

[14] Z. Li, T.-W. Yu, S. Sang, S. Wang, M. Song, Y. Liu, Y.-Y. Yeh, R. Zhu, N. Gundavarapu, J. Shi, S. Bi, H.-X. Yu, Z. Xu, K. Sunkavalli, M. Hasan, R. Ramamoorthi, and M. Chandraker, "OpenRooms: An Open Framework for Photorealistic Indoor Scene Datasets," in *CVPR*, 2021. 3

[15] K. Park, K. Rematas, A. Farhadi, and S. M. Seitz, "PhotoShape: Photorealistic Materials for Large-Scale Shape Collections," in *SIGGRAPH Asia*, 2018. 3, 4

[16] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google Scanned Objects: A High-Quality Dataset of 3D Scanned Household Items," in *ICRA*, 2022. 4

[17] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese, "ObjectNet3D: A Large Scale Database for 3D Object Recognition," in *ECCV*, 2016. 3, 4

[18] F. Yu, Y. Qian, F. Gil-Ureta, B. Jackson, E. Bennett, and H. Zhang, "Hal3d: Hierarchical active learning for fine-grained 3d part labeling," in *ICCV*, 2023. 4

[19] G. A. Miller, "WordNet: a lexical database for English," in *ACM Communications*, 1995. 3

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009. 4, 9

[21] A. Gupta, P. Dollár, and R. Girshick, "LVIS: A Dataset for Large Vocabulary Instance Segmentation," in *CVPR*, 2019. 4

[22] M. Pharr, W. Jakob, and G. Humphreys, *Physically Based Rendering: From Theory to Implementation (3rd ed.)*. Morgan Kaufmann Publishers Inc., 3rd ed., 2016. 7

[23] three.js, "three.js: A 3D Javascript graphics library." https://threejs.org/, 2023. 10

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019. 9

[25] FastML, "WebDataset: A PyTorch Dataset for Large-Scale and High-Resolution Data." https://github.com/webdataset/webdataset, 2023. 9

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016. 9, 12

[27] T. Xiang, C. Zhang, Y. Song, J. Yu, and W. Cai, "Walk in the Cloud: Learning Curves for Point Clouds Shape Analysis," in *ICCV*, 2021. 9, 11, 12

[28] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Computational Visual Media*, 2021. 9, 12

[29] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *CVPR*, 2017. 9, 12, 13

[30] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *NeurIPS*, 2021. 11, 12

[31] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*, 2009. 11

[32] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009. 11

[33] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a Classifier: Zero-Shot Learning Using Purely Textual Descriptions," in *CVPR*, 2013. 11

[34] S. Pratt, M. Yatskar, L. Weihs, A. Farhadi, and A. Kembhavi, "Grounded Situation Recognition," in *ECCV*, 2020. 11

[35] M. Yatskar, L. Zettlemoyer, and A. Farhadi, "Situation Recognition: Visual Semantic Role Labeling for Image Understanding," in *CVPR*, 2016. 11

[36] G. Qian, Y. Li, H. Peng, J. Mai, H. A. A. K. Hammoud, M. Elhoseiny, and B. Ghanem, "PointNeXt: Revisiting PointNet++ with Improved Training and Scaling Strategies," in *NeurIPS*, 2022. 11, 12

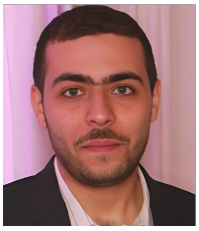
[37] K. T. Wijaya, D.-H. Paek, and S.-H. Kong, "Advanced Feature Learning on Point Clouds using Multi-resolution Features and Learnable Pooling," in *arXiv*, 2022. 12

[38] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic Graph CNN for Learning on Point Clouds," in *SIGGRAPH*, 2019. 12

[39] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking Network Design and Local Geometry in Point Cloud: A Simple Residual MLP Framework," in *ICLR*, 2022. 12

[40] W. Hu, H. Zhao, L. Jiang, J. Jia, and T.-T. Wong, "Bidirectional Projection Network for Cross Dimension Scene Understanding," in *CVPR*, 2021. 12, 13

BIOGRAPHY SECTION



Habib Slim is a Ph.D. student at KAUST, Saudi Arabia. He earned a M.Res. in Data Science from Université Grenoble Alpes (UGA), France, during which he worked on class-incremental learning for image classification at Université Paris-Saclay. He received a M.Eng. in Computer Science from École Nationale Supérieure d'Informatique et de Mathématiques Appliquées de Grenoble (ENSIMAG) in 2020. He is interested in continual/compositional 2D/3D vision.

Xiang Li is a postdoctoral researcher in computer vision at KAUST, Saudi Arabia. He received a B.S. degree in Remote Sensing Science and Technology from Wuhan University, Wuhan, China, in 2014. He received a Ph.D. in Cartography and GIS from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, China, in 2019. His research interests include computer vision, deep learning, and remote sensing.

Yuchen Li is a PhD Student at KAUST, Saudi Arabia. Before joining KAUST, Yuchen was a research intern at iFLYTEK, an intelligent speech and artificial intelligence company in Hefei, China, for three months. He was a Rocket MQ open source contributor and Alibaba summer of code student developer with rich research and engineering experience. He is interested in meta-learning, few-shot learning and 3D object recognition.

Mahmoud Ahmed is an M.S student at KAUST, Saudi Arabia, and received his B.S degree from the American University in Cairo (AUC), Egypt, in 2022. Prior to that, he worked as a Data Science intern at Dell Technologies, then as a 5G Software Engineer. His research interests include computer vision, graphics, and deep learning.

Mohamed Ayman is a M.S. student at University of Alberta and received his B.S. degree from the American University in Cairo (AUC), Egypt, in 2023. He worked as an Applied Science intern at Microsoft. Currently, he is a research intern at KAUST, Saudi Arabia. His research interests are focused on computer vision, NLP, and software optimization.

Ujjwal Upadhyay is an AI Scientist at qure.ai where he works on applying novel deep learning methods to medical data. His research interests include computer vision, adversarial machine learning, and representation learning. He has been involved in cutting-edge research in 3D vision, scene understanding, and neuroscience.

Ahmed Abdelreheem is a Ph.D. student at KAUST, Saudi Arabia. He received a BSc in Computer Engineering from Cairo University, Egypt, in 2019. He attained his MSc degree in Computer Science from KAUST, Saudi Arabia, in 2022. His research interests lie in the intersection of 3D vision, computer graphics, and natural language. More specifically, he is interested in linking 3D object-centric representations to natural language.

Arpit Prajapati is the Director of Technology at Poly9, where his responsibilities encompass developing solutions for product sampling in the Home and Lifestyle industry. Prior to this role, he was the owner of Lanover Solutions for nearly 8 years, managing company operations. He holds a Bachelor of Engineering (BE) degree in Computer from Gujarat University, completed between 2005 and 2009.

Suhail Pothigara is the CEO of Poly9, a recognized Product Management leader in ecommerce and digital transformation. With 15 years of experience at luxury brands and retailers in fashion, home, and consumer electronics, his accomplishments include driving over $3 billion in revenue cumulatively through his roles at e-commerce and cloud businesses at Macy's, LVMH, and HP.

Peter Wonka is Full Professor in Computer Science at KAUST, Saudi Arabia, and Interim Director of the Visual Computing Center (VCC). Peter Wonka received his Ph.D. from the Technical University of Vienna in computer science. Additionally, he received a M.Sc. in Urban Planning from the same institution. After his Ph.D., he worked as a postdoctoral researcher at the Georgia Institute of Technology and as faculty at Arizona State University. His research publications tackle various topics in computer vision, computer graphics, remote sensing, image processing, visualization, and machine learning. The current research focus is on deep learning, generative models, and 3D shape analysis and reconstruction.

Mohamed Elhoseiny is an Assistant Professor of Computer Science at KAUST, Saudi Arabia, and a Senior Member of IEEE and AAAI. He was a Visiting Faculty at Stanford Computer Science Department (2019-2020), a Visiting Faculty at Baidu Research Silicon Valley Lab (2019), and a Postdoc Researcher at Facebook AI Research (2016-2019). Dr. Elhoseiny completed his Ph.D. in 2016 at Rutgers University, during which he spent time at Adobe Research (2015-2016) for more than a year and at SRI International in 2014. He received an NSF Fellowship in 2014 and the Doctoral Consortium Award at CVPR 2016. His primary research interest is in computer vision, especially in efficient multimodal learning with limited data in areas like zero/few-shot learning, vision and language, and language-guided visual perception. He is also interested in affective AI and particularly in producing novel art and fashion with AI. His creative AI work was featured in MIT Tech Review, New Scientist Magazine, and the HBO show Silicon Valley.