3D CoMPaT: Composition of Materials on Parts of 3D Things

Yuchen Li[1,*], Ujjwal Upadhyay[1,⋆], Habib Slim[1,*], Ahmed Abdelreheem[1], Arpit Prajapati[2], Suhail Pothigara[2], Peter Wonka[1], Mohamed Elhoseiny[1]

[1] KAUST, Thuwal, Saudi Arabia: {firstname.lastname}@kaust.edu.sa
[2] Poly9 Inc., San Francisco, California: {firstname}@polynine.com

Abstract. We present 3D CoMPaT, a richly annotated large-scale dataset of more than 7.19 million rendered compositions of Materials on Parts of 7262 unique 3D Models; 990 compositions per model on average. 3D CoMPaT covers 43 shape categories, 235 unique part names, and 167 unique material classes that can be applied to parts of 3D objects. Each object with the applied part-material compositions is rendered from four equally spaced views as well as four randomized views, leading to a total of 58 million renderings (7.19 million compositions ×8 views). This dataset primarily focuses on stylizing 3D shapes at part-level with compatible materials. We introduce a new task, called Grounded CoMPaT Recognition (GCR), to collectively recognize and ground compositions of materials on parts of 3D objects. We present two variations of this task and adapt state-of-art 2D/3D deep learning methods to solve the problem as baselines for future research. We hope our work will help ease future research on compositional 3D Vision. The dataset and code are publicly available at https://www.3dcompat-dataset.org/

1 Introduction

Various datasets have been proposed to facilitate 3D visual understanding including ShapeNet [4], ModelNet [33], and PartNet [26]. Recently, 3D-FUTURE [12] was proposed, which contains 9,992 industrial 3D CAD shapes of furniture with textures developed by professional designers. Despite these significant efforts to create 3D datasets, current 3D object datasets (e.g., [4,33,26]) and 3D Scene datasets (e.g., [8]) lack part-level material information. The availability of material information has multiple benefits. First, material information provides additional semantic information about an object. Second, material information enables more realistic renderings making the models more suitable for synthetic to real transfer. Third, the same geometric 3D shape can be rendered with different material assignments leading to more variability during training (see Fig. 1).

We introduce a richly annotated large-scale dataset, dubbed as *3D CoMPaT*, Compositions of Materials on Parts of 3D Things. The dataset contains more

[⋆] co-first authors

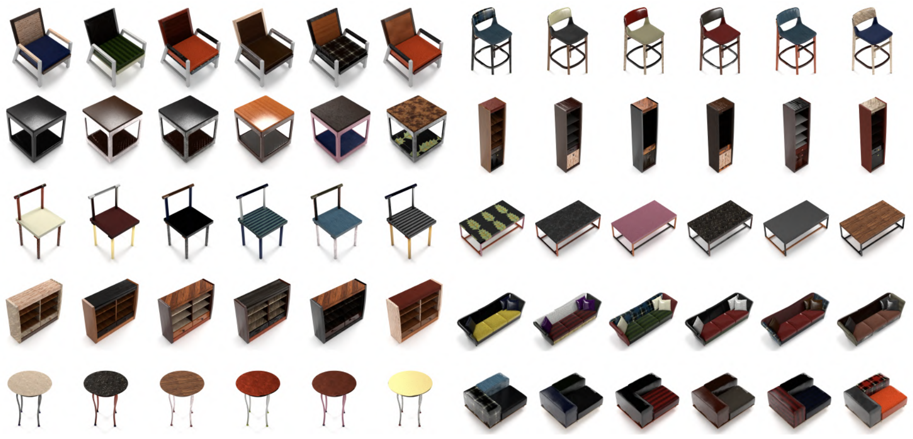


Fig. 1: Stylized models in the 3D CoMPaT dataset. We show several compositions of 8 selected models in the dataset, stylized with different materials.

than 7.19 million rendered model styles from 8 views, covers 43 shape categories, 235 unique and distinguishable part names, and 167 unique and distinguishable materials from 10 material classes that can be applied to parts of 3D objects. Each object with the applied part-material compositions is rendered from four equally spaced views, leading to 58 million (7.19 million compositions ×8 views) images in total. Examples of some rendered compositions and views can be seen in Fig. 1 and 2 respectively.

We start with 7262 unique shapes with a total of 37198 segmented parts (i.e., 5.12 segmented parts per shape on average), and we annotate the list of compatible/applicable materials for each part. Then, we sample a model by enumerating randomly over the compatible materials for each part with a limit of 1000 compositions per shape, leading to 7.19 million compositions of 3D objects.

Connection and differences to existing datasets. The proposed dataset is different from the currently available datasets in the literature in the following ways. First, the dataset contains a diverse set of high-quality materials beyond mere texture maps. Second, for each part found in every 3D model, the dataset defines a set of materials that may be applied to this part in that model, allowing us to generate multiple material combinations for a single model (we call each combination a *style*). The models in 3D-FUTURE [12] and ShapeNet [4] do not have multiple styles, and also, in the ShapeNet dataset, only a small portion of 3D shapes are stylized. The following four key aspects can characterize our 3D CoMPaT dataset in contrast to existing datasets.

–*(a) human-generated vs. 3D scanned geometry.* For example, ScanNet [8] and Matterport3D [3] datasets are scanned 3D geometry. Conversely, ShapeNet [4] and our 3D CoMPaT dataset are human-generated.

–*(b) part segmentation information.* For some datasets, none or only a subset of the shapes have segmented part information, which is an important aspect of datasets like PartNet [26] and is also a characteristic of our dataset.

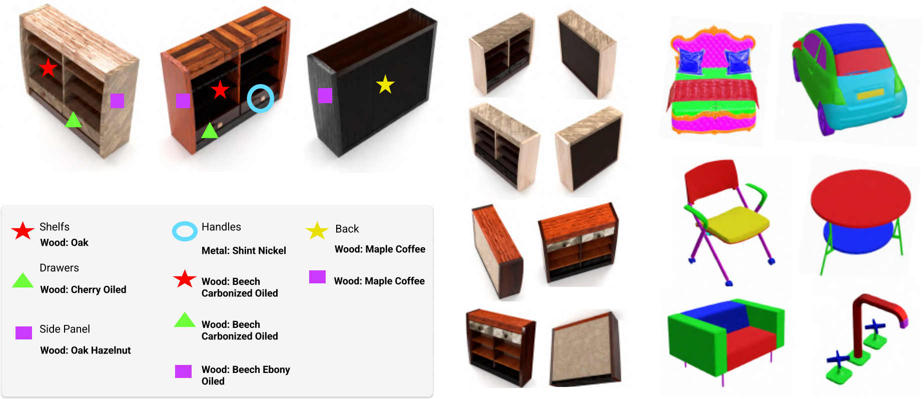


Fig. 2: 3D CoMPaT Dataset. **Left:** Examples of a stylized cabinet. The cabinet has five parts, shelves, drawers, handles, back and side panel, indicated as highlighted. The box below contains the material names for the indicated parts in different stylized cabinets. **Middle:** Each 3D object is rendered in four canonical and four randomized views. **Right:** Part segmentation masks for randomly selected shapes from our dataset.

–*(c) texture coordinates, textures, and materials.* Since stylizing the composition of 3D model parts is at the heart of our work, our models have texture coordinates and material compatibility information to enable high-quality rendering of hundreds of compositions of materials on each shape. This is the most important distinguishing characteristic of our 3D CoMPaT dataset compared to existing datasets. There was some earlier effort to augment a subset of ShapeNet with material information [24]. This dataset has fewer shapes (3080 vs 7262), parts, and materials compared to our work.
–*(d) automatically generated vs. human-generated information.* 3D CoMPaT part names are consistent and come from a list of allowable part names per model category. All models are manually segmented at a part level rather than being segmented with deep learning models like OpenRooms [23]. Furthermore, in 3D CoMPaT all texture coordinates are developed and verified by humans (refer to Sec. 3 for more details).

We validate our dataset with experiments covering main 3D recognition tasks, including 3D object classification, 3D part recognition, and material tagging.

Grounded CoMPaT Recognition (GCR) Task. Finally, we introduce a novel task, dubbed as CoMPaT recognition. It aims at recognizing and grounding the shape category collectively with the part-material pairs associated with the shape, e.g., recognizing that the example in Fig. 2 is a "Cabinet", with a handle made of "shiny nickel (metal)" and a back made of "maple coffee wood".

Contributions.

- We propose a new dataset of 7.19 million stylized models to study composition of Materials on Parts of 3D Things. Our dataset contains (a) a diverse set of 167 materials for 3D shapes. (b) The material assignment is done at part-level. (c) Segmentation masks in 2D and 3D are provided, alongside (d) human-verified texture coordinates. We hope this dataset may also facilitate future research on retrieving objects in 3D scenes (e.g., localizing a specific "chair" or "table" from Fig. 2).

- We validate our dataset by a set of experiments involving 2D/3D shape classification, part recognition (detection and segmentation), and material tagging.
- We also propose Grounded CoMPaT Recognition, a novel task of collectively recognizing and grounding compositions of materials on parts of 3D objects. We introduce two variants of this task, and adapt 2D/3D state-of-the-art methods as baselines for this problem.

2 Related Work

Datasets of 3D shapes and scenes. ModelNet [33] is a large-scale 3D CAD model dataset covering 40 categories. ShapeNet [4] is a richly-annotated, large-scale repository of shapes with semantic categories and organizes them under the WordNet taxonomy. PartNet [26] assigned rich fine-grained segmentation labels on the part level. Recently, 3D-FUTURE [12] was proposed, which contains 9,992 unique industrial 3D CAD shapes of furniture with high-resolution informative textures developed by professional designers. In contrast to 3D-FUTURE, PartNet, and ShapeNet, where only a small portion of 3D shapes can be assigned materials, our dataset contains 7262 models, all of which can be stylized with different textures. PhotoShape [27] is a dataset similar to ours. More specifically, it uses a technique to automatically apply materials to existing models, mainly from ShapeNet [4]. Some rendered models are not that realistic. The texture coordinates are generated automatically, while ours are human-generated and human-verified. PhotoShape only has a single shape category, i.e., chair, and has only five material classes (leather, fabric, wood, metal, plastic). In comparison, our dataset has 43 shape categories and thirteen material classes (wood, metal, fabric, marble, ceramic, glass, leather, paint, paper, plastic, rubber, granite, wax). OpenRooms [23] is a large dataset containing indoor scenes. The authors automatically segment CAD models of the scenes into parts based on a segmentation model trained on the PartNet dataset. Hence, OpenRooms parts are restricted by the part classes present in PartNet. This also may introduce some segmentation errors in part localization and naming since learned predictions are not as accurate as human annotations. In contrast, our models are manually annotated and verified. Furthermore, our dataset contains models of objects and not indoor scenes, which gives users more flexibility. The major difference between OpenRooms and 3D CoMPaT is that OpenRooms uses scanned geometry captured by sensors while our dataset is manually constructed. We also note that Lin et.al. [24] introduced 3080 stylized models for three categories with five material classes and part information. However, the scale of our dataset is much larger.

Texture Generation. TM-Net [13] is a novel deep generative model that generates meshes with detailed textures and synthesizes plausible textures for a given shape. Their work is inspired by SDM-NET [14]. Their method produces texture maps for each part, which means it works in a part-aware fashion. Each part is represented as a deformed box. They encode geometry and texture separately and learn the texture probability distribution conditioned on the geometry. This allows their method to be a generic framework for different application scenarios.

High-Level 3D Vision. Encouraging progress in 3D scene understanding, ScanNet [8] introduced a large-scale dataset of 1513 real-world scenes. More recently and at the intersection of 3D vision and natural language, ScanRefer [5] and Referit3D [2] datasets were recently introduced on top of ScanNet to study 3D object identification based on free-form natural language descriptions. The detailed composition of the shape category and part-material pairs provided in 3D CoMPaT can serve as a rich semantic description of shapes, and hence may facilitate more fine-grained visual grounding of language referring to 3D objects and scenes.

3 3D CoMPaT: Data Collection, Benchmark, and Validation

The 3D CoMPaT dataset collection pipeline comprises three main processes: 3D CAD models collection, materials collection, material assignment, and rendering.

3.1 3D CAD Models Collection

3D CoMPaT is based on a collection of 3D CAD models managed by Poly9 Inc.. The initial data has high-quality 3D models, but the part names, segmentation information, class information, and material information is often missing or faulty. Repetitions of the same CAD model may be present, and some CAD models contain a set of similar parts. The team for building 3D CoMPaT consisted of professional CAD modelers, researchers, and crowd-sourced workers from AMT. The process for creating 3D CoMPaT consists of frequent review meetings between researchers and professional modelers discussing issues with part names, shape categories, materials, and shape segmentation continuing for over one year. Based on these reviews, the professional modelers would adapt their processes, such as labeling instructions, or the allowable list of part names. While professional modelers did all the labeling and modeling work, researchers focused on automatic and manual quality control. While ultimately most of the class names, part names, and material assignments had to be changed during our effort, we only selected shapes that already had high geometric quality and texture coordinates so that little effort was needed to fix problems in the geometry. Due to our multi-stage verification process, each 3D shape was manually inspected more than once. Models that failed a stage of the quality control pipeline were sent back to the team of professional modelers.

- **(A) Shape and Part Category Labeling:** Each 3D CAD model is assigned a shape category label (e.g., chair, desk, table). All models are consistenly segmented into parts and every part in every model is assigned a part name (e.g., "seat, back, or legs"). Each part in each 3D CAD model is also assigned a list of compatible material types, e.g., for one particular chair a modeler could assign that the legs of the chair can be made of either metal or wood. Designing a consistent list of allowable part names for each shape category is a considerable effort. We sourced information from online retailers, other datasets such as PartNet [26], names used in 3D CAD models, crowdsourcing services and our own experience. In particular, we started with a smaller subset of shapes and some initial labeling of part names to verify these annotations using Amazon Mechanical Turk (AMT). Even though our goal to fix the list of allowable part names early in the process, we had to adapt the list over time in the review meetings as new shapes were being processed.
- **(B) Duplicates and Near-duplicates Removal:** Some 3D CAD models are repeated more than once or contain multiple instances of the same model (e.g., a 3D CAD model representing a set of vases with different sizes). we implemented an automatic procedure to detect duplicates and near-duplicates to remove them from the dataset.
- **(C) Part Segmentation:** Every CAD model should be part-segmented; i.e., every CAD model consists of a set of separated part meshes. We manually check the segmentation of each shape in a 3D viewer and correct them if they are not consistent with the defined part categories.
- **(D) Texture Coordinates Quality Check.** For a proper material assignment, the quality of texture coordinates was verified qualitatively. We followed two strategies. First, we overlay different materials over each part and check how it renders in different settings (we used an increasing level of light bounces to see how textures look). Second, we applied checkered textures to visually inspect the texture coordinates; as illustrated in Fig. 4 (left). For the evaluation of the 3D geometry, we checked that the models are watertight and that they have outward-pointing normals.

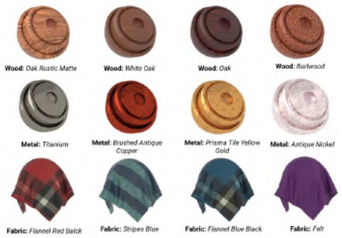


Fig. 3: Examples of materials found in the dataset. We show examples of wood, metal, and fabric materials in the first, second, and third rows respectively.

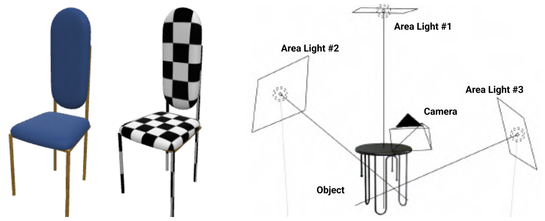


Fig. 4: Left: Examples of texture coordinate checks. Right: Blender rendering environment. The environment contains three light sources (a directional light source and three area lights) and a plane at the bottom. The 3D CAD model is normalized, centered on the origin, and placed on the plane.

Crowdsourcing the Verification of Shape and Category labels. To verify annotated class names of a given 3D model, we asked five MTurk participants to choose from the following four options: (1) ”yes, I would name the model the same,” ”(2) yes, but I would have given the model a different name,” (3) ”no, this is a wrong name (please specify a name)” or (4) ”no, the model cannot be given a specific name.”. We used a similar interface to verify the part class names; the part-annotation and model-annotation verification interface used in our AMT experiments is shown in the supplementary material [22].

3.2 Materials Collection.

For materials, we use the free and open-source Nvidia vMaterials library. The Nvidia vMaterials library has over 2150 real-world materials and continues to grow. These materials are defined by the Nvidia MDL specification, allowing PBR materials with higher visual quality than basic materials based on diffuse textures. Materials from this library have the infinite tiling feature, allowing textures to be spread across large areas without a clear repeating pattern. The library provides class labels for every material organized in a hierarchical tree (e.g., an antique oxidized aluminum is a rusted aluminum metal). Tab. 1 presents the count of distinct subtypes for each material class (10). In total, the number of materials in our dataset is 167.

We manually inspected these materials, ignoring materials that may not be realistic when applied to specific parts. For example, a fabric material that has a mesh-like appearance is not suitable to be used for chair cushions (see Fig. 3).

Table 1: 3D CoMPaT material classes and number of materials per class.

Wood	Metal	Fabric	Marble	Ceramic	Glass
40	32	35	14	8	6
Leather	Plastic	Rubber	Granite	Wax	**Total**
13	10	5	3	1	**167**

3.3 Part-Material Assignment

One of the novel aspects of 3D CoMPaT is the presence of material compatibility information for parts present in each 3D model. Human workers conduct the process of material assignment. This process was realized at the instance level, i.e., the shape and parts of each 3D model were considered to compose them with appropriate materials. For example, the legs of one particular table could be assigned either metal or wood and the legs of another table could be assigned wood or plastic. The assignment space for this process is 7262 x 5.12 x 10 (where 10 is the number of material categories). We only assign material classes. For example, all 32 metals can be assigned to a shape part in the material sampling stage if a metal is a possible assignment. We also control compatibility to some extent through grouping information. For example, all table legs have to be assigned the same material. However, we do not explicitly control complex material combinations, as this is hard to integrate into the currently used 3D modeler. This has advantages and disadvantages. An advantage of the current solution is that we allow a greater variety of models which can be a good source of data augmentation. As a disadvantage, several sampled stylized models may not be aesthetically beautiful. Overall, we believe that the consistency of the material assignments is better controlled at sampling process that assigns materials. For example, the sampler could select the same material for chair back and legs significantly more often than different materials. We believe that is more efficient and compatible with our current approach while an explicit control of material combinations suffers from an exponential explosion of possible combinations that need to be controlled. To analyze this issue further will require a significant effort in synthetic to real transfer which we leave to future work.

3.4 Rendering Composition of Materials on Parts of the Collected CAD Models

Once material assignments for each part are available, this information is used to sample materials from the assigned categories. For example, if a tabletop is made of wood, we can sample one of the 40 wood types, like teak, oak, hazelnut, etc. In what follows, we refer to the combination of these materials assigned to parts of a given CAD model as a *composition*. The application of one such composition to the CAD model is called a *style* of the model.

For each 3D CAD model, we randomly select a material for each part from the list of its compatible materials. We sample at most 1000 styles per 3D model.

- **(A) Rendering.** We use Blender [1] to render each CAD model into RGB images from 8 different views, with the camera placed far enough for the entire object to be visible. For the lighting setup, we use three light sources; see Fig. 4 (right). We render each stylized model in 4 standard views (front and back with default model orientation and front left and back right with the model rotated with 30 degrees around the z (up) axis). Further, we also render each model from 4 random views. The camera for random views is parameterized with elevation angle θ_{cam} (in degree) $\in [0, 90]$, while keeping the x, y same. The model is rotated parameterized with random rotation angle θ_{model} (in degree) $\in [0, 360]$.
- **(B) Segmentation and Depth Maps.** The rendered images in the 3D CoMPaT dataset will be accompanied by corresponding segmentation maps and depth maps. These maps will be rendered with the same four fixed views and four random views.
- **(C) Stylized 3D models**. We plan to release the stylized 3D models, which will enable their use in many 3D computer vision applications, including retrieval, reconstruction, and 3D generation.

3.5 Dataset Statistics

3D CoMPaT contains 7262 unique 3D shapes covering 43 shape categories. The top 6 classes are (table, tray, bowl, chair, desk, cabinet). The dataset contains 37198 part instances covering 235 part classes, and 167 different materials from 10 material classes. The top 5 material classes are (metal, wood, fabric, paint, marble); please see the supplementary for more details about shape classes, the number of object instances per shape class, part, and material classes [22]. In

Table 2: Comparison of 3D CoMPaT with other datasets in the literature. To our knowledge, our dataset is the first one to have many different materials applied to different parts in the same 3D model. ✓*: only a subset of shapes are textured and the remaining shapes are with unidentified textures. ?: unknown. HVT stands for Human Verified Textures.

Benchmarks	**Shapes No.**	**Categories**	**Material Classes**	**Materials**	**Shape Source**	**Stylized Models**	**HVT**	**Parts per Shape**
3D-Future[12]	9992	34	14(+1)	?	industry	102972	×	10.3
3D Front[11]	13151	50	23	?	3D-Future	13151	×	6.5
PhotoShape[27]	5830	1	5	363	ShapeNet, industry	11000	×	1-3
ShapeNetCore[4]	51300	55	×	✓*	online, crowdsource	×	×	×
ShapeNetSem[4]	∼12000	270	×	✓*	online	×	×	×
ShapeNetPart[35]	31963	16	×	✓*	online	×	×	2.92
ModelNet[33]	151128	660	×	×	online	×	×	×
ObjectScans[6]	1,900	44	×	×	Scans	×	×	×
PartNet[26]	26671	24	×	×	ShapeNet	×	×	21.5
Lin et.al.,[24]	3080+115	3	5	×	online, ShapeNet	3080+115 ?	×	5.16
3D CoMPaT	**7262**	**43**	**10**	**167**	**industry**	**7.19 million**	✓	**5.12**

Table 2, we show how our proposed dataset has more variety in the number of materials and the number of materials assignments (styles) than the currently available datasets.

Fig. 5a and Fig. 5b show the frequency with which different shape classes and parts occur in the dataset. We observe that the dataset has an uneven distribution as some parts, model classes, and materials are more frequent than others. Fig. 5c shows the distribution of subsets of parts in different model classes. The size of the bubble represents the occurrences of a part in a certain model class, which we further categorized as very frequent, frequent, or less frequent. It can be inferred from Fig. 5c that some parts are centered around a single "model class"; for example, the "top" is mostly centered around the "table" class, indicating the very frequent tabletop part in the dataset. Some parts have high variability across different models. From Fig. 5c, we can see that "leg" is a part that frequently occurs in "table" models but also in several other models like "chair", "cabinet", and "desk".

Table 3 shows the statistics of 3D CoMPaT in different aspects including parts, model classes, material and styles. The scale of material compositions on model parts is a key difference of 3D CoMPaT when compared to existing datasets, enabling styling of all existing model classes with different part-material combinations that are compatible. As we pointed out earlier, some materials cannot be applied to some parts (e.g., wood in exchange for glass).

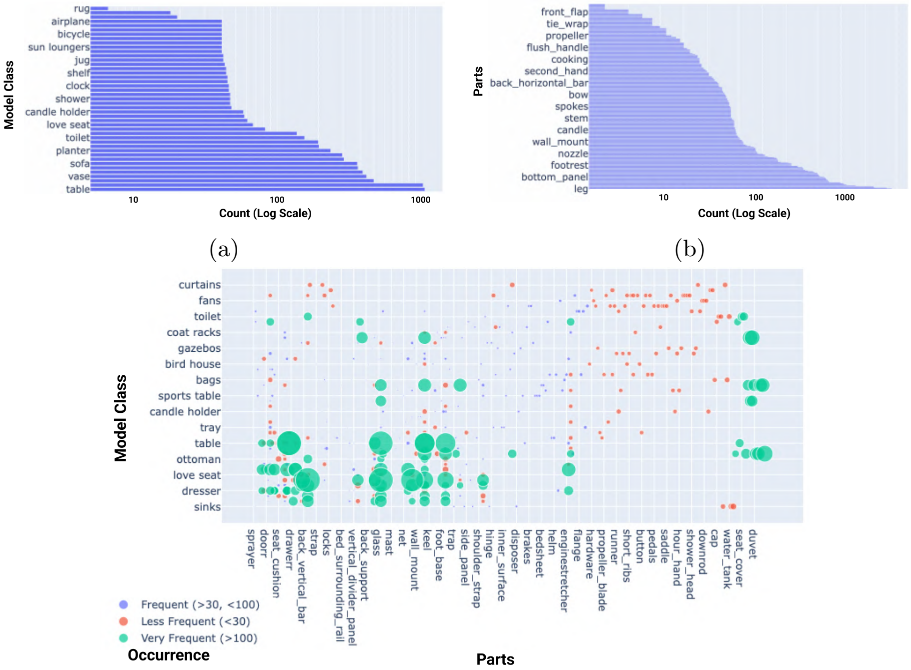
Table 3: Dataset statistics.

Total Number of Models	7262
Total Number of Model Classes	43
Total Number of Parts	37198
Total Number of Parts Classes	235
Minimum Number of Parts Per Model	1
Maximum Number of Parts Per Model	17
Average Number of Parts Per Model	5.12
Average Compositions Per Model	990
Total Number of Materials	167
Total Number of Stylized Models	7.19 million

3.6 Dataset Split and Non-Compositional Validation Experiments

We create training, validation, and test splits for the renderings. All renderings and stylized versions of a 3D shape have to be assigned together to either training, validation, or test. Therefore, the splits are defined on shapes to prevent data leak. The training set has 5597 shapes, the test set 924 shapes and the validation set 477 shapes. Despite compositional recognition being the focus of our work, a variety of standard tasks can benefit from our proposed dataset, including 3D object classification, 3D semantic segmentation, shape classification, image shape retrieval, shape reconstruction from single/multiple images. We conducted experiments on some of these tasks to validate the properties of our proposed 3D CoMPaT dataset.

3D Shape and Part Classification. Our dataset has an uneven distribution, so some models and parts have more examples and hence help in better generalization. Some parts and models resemble their more frequently occurring counterparts (e.g. jar and container), making classification challenging. Note that there is some intersection between model class and part names, namely



(a)

(b)

(c)

Fig. 5: 3D CoMPaT. (a) Number of samples per shape category. (b) Number of samples per part class. (c) Frequency of occurrence of part and model pair. Note that both visualizations do not cover all shape and part labels because of space constraints; more details are provided in the supplementary [22].

basket, bowl, candle holder, glass, shelf, table, tray, vase. This is because some models are parts of other larger models (e.g. shelf as part of shelf structure or cabinet). The presence of various materials to style the 3D object gives our dataset an edge over other existing datasets. We conduct shape classification experiments for the 3D models and the 3D parts. Results are reported in Table 4, where we benchmark Point Cloud Transformer (PCT) [16], DGCNN [32] , and Pointnet++ [29] on shape classification and Pointnet++ and PCT on part classification; see results in Table 4.

Table 4: 3D CoMPaT non-compositional validation experiments.

Architecture	Task	Test Performance
Pointnet++[29]	3D Shape Classification	57.95% Accuracy
DGCNN[32]		68.32% Accuracy
PCT[15]		69.09% Accuracy
Pointnet++[29]	3D Part Classification	24.18% Accuracy
PCT[15]		37.37% Accuracy
BPNet 2D [19]	2D Material Segmentation	35.75% mIOU
BPNet 3D [19]	3D Material Segmentation	17.03% mIOU
ResNet50 [18]	2D Material Tagging	0.53 F1, 0.67 AP
ResNet50 [18]	2D Shape Classification	76.82% Accuracy

2D and 3D Material Segmentation. We benchmark BPNet [19], a 2D and 3D joint UNet, for our 2D and 3D Material Segmentation in Table 4.

2D Material Tagging/ Shape Classification. We use a ResNet50 [18] backbone for encoding the rendered images, to train a multi-label classifier over the 167 materials in the rendered images. The F1 score and average precision were 0.53 and 0.67 respectively. We also report a 2D shape classification performance of 76.82% using ResNet50, on 50 canonical compositions; see Table 4.

Sim2Real 3D shape recognition. We trained a PointMLP [25] model on ModelNet40 and 3D CoMPaT (with only one sampled composition/shape) and evaluated on the hardest variant of ScanObjectNN [31] (*without finetuning*). Table 5 shows 3D shape classification results for 9 classes. Results show that pretraining on 3DCoMPaT shapes leads to better generalization to real-world data than pretraining on ModelNet40.

Table 5: Sim2Real transfer: Accuracy results for PointMLP [25] trained on ModelNet40 and 3DCoMPaT (1 random composition), on ScanObjectNN's hardest variant.

Dataset	Acc. (%)
ModelNet40	24.33
3DCoMPaT	**29.21**

4 2D/3D Grounded CoMPaT Recognition (GCR) Task, Baselines, and Results

The goal of compositional modeling on 3D CoMPaT is to recognize the entire composition of materials on parts of a given 3D model. More specifically, we aim at correctly predicting the object category, part categories and the associated material for every part in the 3D model. Fig. 6 visualizes some ground truth and prediction examples for this task. This task is challenging because 96.31% of the compositional frames at test time are unseen. We define a 3D CoMPaT compositional frame as a shape category and a set of part-material categorical pairs. Two compositional frames are different if they differ in a single part or material assignment. In standard recognition settings the model only has to select

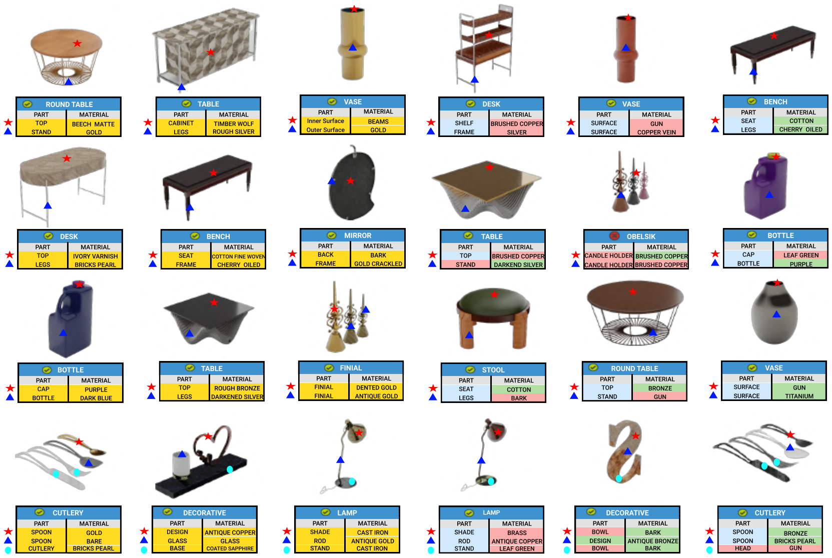


Fig. 6: Example realized materials over parts of certain model classes. Below each image is a table where the first row is the model class, the left column is part names, and the right column is material name for those parts. On the left outlined in gold are ground truth. The output from our material recognition, part recognition, and model recognition model is on the right. Incorrect part names and material names are highlighted in red, whereas correct ones are green.

the correct shape class and examples of all shape classes have been seen before. By contrast, several proposed metrics require the correct recognition of compositional frames. The number of compositional frames is much higher than the number of shape classes and for most compositional frames in the test set there are no examples in the training set. This can be related to zero-shot recognition, which aims at recognizing unseen categories that are defined by unseen compositions of visual attributes (e.g., [21,10,9,17]). This is also connected to existing yet different compositional 2D computer vision tasks, including situation recognition [28,34], which aims at identifying an activity like "surfing" in an image, the engaged entities with their roles (e.g., "agent: woman", "tool: surfboard", and "place: ocean"), and bounding-box groundings of entities.

Metrics. Inspired by the metrics proposed in [34,28] for compositional situation recognition of activities in images, we define the compositional metrics of the 2D/3D **G**rounded **C**oMPaT **R**ecognition (**GCR**) task as follows:

(a) Shape Accuracy: accuracy of the predicted shape category. **(b) Value:** accuracy of predicting both part category and the material of a given part correctly. **(c) Value-all**: accuracy of predicting all the (part, material) pairs of a shape correctly. We similarly define grounding metrics to check segmentation masks. A grounding is correct if the IoU of a predicted part and ground truth part is more than 0.5; it can be IoU on segmentation masks. **(d) Grounded-**

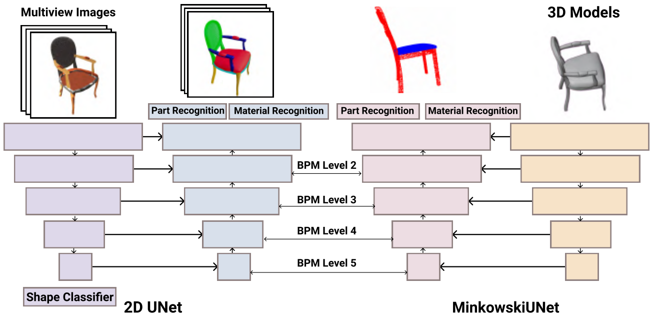


Fig. 7: 2D/3D GCR-SEG: Modified 2D/3D BPNet segmentation [19] architecture with 2D UNet [30] on the left and 3D MinkowskiUNet [7] on the right with same number of pyramid levels.

value: accuracy of predicting both part category and the material of a given part as well as correctly grounding it. **(e) Grounded-value-all**: accuracy of predicting all the (part, material) pairs of a given shape correctly and grounding all of them correctly. All these metrics are calculated for each shape and then averaged across them to avoid bias toward shapes with more parts.

Given the shape dependence of metrics, we define three settings: **(a) Ground Truth Shape**: the ground truth shape is assumed to be correct. **(b) Top-1 Shape**: Shape category is predicted correctly. **(C) Top-5 Shape**: Shape category is in the top-5 predictions. For (b) and (c), part-material pairs and their groundings are considered incorrect if shape is not in top-1 or top-5 predictions, respectively. We investigate two variants of the GCR task:

– **(A) Joint 2D/3D GCR-SEG Setting and Baseline:** 2D/3D GCR-SEG aims at solving the GCR Task in 2D or 3D, where grounding of parts is measured at the pixel precision for 2D and mesh triangle precision for 3D. Hence, we adopted a segmentation approach to solve it, specifically the joint 2D/3D BPNet segmentation model [19]; see Fig. 7. We adapted [19] to jointly predict shape, part, and material recognition in the Grounded CoMPat recognition task. As shown in Fig. 7, our adapted network consists of two branches: according to their functional domains, we denote the left one as the 2D UNet branch and the right as the 3D MinkowskiUNet branch. A Bidirectional Projection Module (BPM) bidirectionally fuses the multi-view 2D and 3D features between two branches. Features from the encoder of the U-Net are fed to a fully connected layer for shape classification. Images and voxels can be aggregated in a coarse-to-fine manner. BPNet can collect low-level and advanced complementary information; see more details in the supplementary [22].

– **(B) 3D GCR-SEG Setting and Baseline:** Similar to (A), but in 3D only, where part label and material labels are predicted at the point precision for grounding. [20]. We built on PointGroup [20], a point cloud based method for 3D segmentation; see PointGroup adaptation details in the supplementary [22].

Results. Table 6 shows the results for 2D GCR-SEG using BPNet, 3D GCR-SEG using our joint 2D/3D BPNet-based baseline. We report the "Standard"

Table 6: 2D/3D Grounded CoMPaT recognition (GCR) Results.

Exp	Top-1 predicted shape					Top-5 predicted shape					Ground Truth Shape			
	Shape Acc.	Value	Value-all	Value-grnd	Value-all grnd	Shape Acc.	Value	Value-all	Value-grnd	Value-all grnd	Value	Value-all	Value-grnd	Value-all grnd
2D GCR-SEG (BPNet)														
Standard	36.91	6.81	3.54	3.29	0.07	39.07	7.29	3.74	3.48	0.07	65.07	27.72	36.15	2.89
GT Material	36.91	7.10	3.54	4.30	0.60	39.07	7.59	3.74	4.51	0.60	69.54	27.72	40.46	4.80
GT Part	36.91	7.52	5.77	4.94	1.20	39.07	8.07	6.28	5.37	1.49	86.54	71.73	71.83	44.80
2D GCR-SEG (BPNet) + Separate 2D Shape Classifier														
Standard	67.86	40.39	16.57	23.23	1.78	77.60	49.28	20.67	27.50	1.89	65.07	27.72	36.15	2.89
GT Material	67.86	42.98	16.57	26.40	3.46	77.60	52.33	20.67	31.18	3.64	69.54	27.72	40.46	4.80
GT Part	67.86	55.12	46.58	46.02	30.08	77.60	66.89	57.36	55.81	37.30	86.54	71.73	71.83	44.80
3D GCR-SEG (BPNet)														
Standard	36.91	5.39	2.02	0.65	0.03	39.07	5.68	2.07	0.70	0.03	41.35	11.48	4.62	0.03
GT Material	36.91	6.10	2.02	1.34	0.15	39.07	6.43	2.07	1.40	0.15	46.39	11.48	7.80	0.27
GT Part	36.91	6.30	4.13	2.05	0.82	39.07	6.72	4.40	2.29	0.92	77.40	66.47	46.73	39.03
3D GCR-SEG (BPNet) + Separate 3D Shape Classifier														
Standard	67.53	29.83	8.12	3.25	0.46	87.23	38.34	11.07	4.30	0.47	44.59	12.52	4.92	0.47
GT Material	67.53	33.72	8.12	6.07	0.63	87.23	43.15	11.07	7.38	0.66	50.71	12.52	8.23	0.66
GT Part	67.53	50.08	42.42	27.86	22.65	87.23	63.53	54.06	36.44	29.31	77.27	65.30	44.83	36.57

compositional metrics described earlier in this section, on ten compositions. To demonstrate how perfect prediction of either material or part influence the performance, we also report results where we use ground truth materials as the predicted material labels ("GT Material"), and ground truth parts as predicted parts ("GT Part"). It is not surprising that compared to "Standard", some metrics improve under "GT Material" and "GT Part" evaluation, especially for the value and value-all metrics that depend on the predicted part and the material labels. Note that all these baselines are composed of one model that jointly predicts shape and part material pairs in 2D or 3D. These models have a shape recognition performance ranging between 15.64% and 38.29% top-1 accuracy., and between 62.32% and 85.2% top-5 accuracy.; see Table 6. This limits the compositional performance, especially as we showed earlier in Table 4 that separate 2D and 3D Shape classifiers can reach 76.8% and 69.1% Top-1 Acc respectively. Hence, we also evaluated our BPNet-based 3D GCR-SEG approach where shape classes are predicted with a separate 3D PCT [15] classifier, leading to improved compositional performance. We observe similar behavior with PointGroup-based 3D GCR-SEG solution; see supplementary for materials for details [22]. The results suggest that designing a single model capable of performing well on GCR metrics is a challenge, and we hope that our 3D CoMPaT dataset and GCR baselines help ease future research.

5 Conclusion

We introduce 3D CoMPaT, a large-scale dataset of Compositions of Materials on Parts of 3D Things. It contains 7.19 million styled models stemming from 7262 CAD models from 43 object categories. The unique aspect of 3D CoMPaT is that it contains 3D shapes, part segmentation information, texture coordinates, 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 2D/3D Grounded CoMPaT Recognition (GCR), that the dataset enables and introduce baseline methods to solve them.

Acknowledgments. The authors wish to thank Poly9 Inc. participants for all the hard work, without whom this work would not be possible. This research is supported by King Abdullah University of Science and Technology (KAUST).

References

1. Blender foundation, blender.org - home of the blender project - free and open 3d creation software (2021) 8
2. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. 16th European Conference on Computer Vision (ECCV) (2020) 5
3. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. International Conference on 3D Vision (3DV) (2017) 2
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015) 1, 2, 4, 9
5. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. arXiv preprint arXiv:1912.08830 (2019) 5
6. Choi, S., Zhou, Q.Y., Miller, S., Koltun, V.: A large dataset of object scans (2016) 9
7. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3075–3084 (2019) 13
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 1, 2, 5
9. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: Zero-shot learning using purely textual descriptions. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2584–2591 (2013) 12
10. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR 2009. pp. 1778–1785. IEEE (2009) 12
11. Fu, H., Cai, B., Gao, L., Zhang, L., Li, J.W.C., Xun, Z., Sun, C., Jia, R., Zhao, B., Zhang, H.: 3d-front: 3d furnished rooms with layouts and semantics (2021) 9
12. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. arXiv preprint arXiv:2009.09633 (2020) 1, 2, 4, 9
13. Gao, L., Wu, T., Yuan, Y., Lin, M., Lai, Y., Zhang, H.: TM-NET: deep generative networks for textured meshes. CoRR **abs/2010.06217** (2020), https://arxiv.org/abs/2010.06217 5
14. Gao, L., Yang, J., Wu, T., Yuan, Y., Fu, H., Lai, Y., Zhang, H.: SDM-NET: deep generative network for structured deformable mesh. CoRR **abs/1908.04520** (2019), http://arxiv.org/abs/1908.04520 5
15. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: Pct: Point cloud transformer (2021) 11, 14
16. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: Pct: Point cloud transformer. Computational Visual Media **7**(2), 187–199 (Apr 2021). https://doi.org/10.1007/s41095-021-0229-5, http://dx.doi.org/10.1007/s41095-021-0229-5 11
17. Guo, Y., Ding, G., Han, J., Gao, Y.: Synthesizing samples for zero-shot learning. In: IJCAI (2017) 12
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015) 11

19. Hu, W., Zhao, H., Jiang, L., Jia, J., Wong, T.T.: Bidirectional projection network for cross dimension scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14373–14382 (2021) 11, 13
20. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: Pointgroup: Dual-set point grouping for 3d instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4867–4876 (2020) 13
21. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR. pp. 951–958. IEEE (2009) 12
22. Li*, Y., Upadhyay*, U., Slim*, H., Abdelreheem, A., Prajapati, A., Pothigara, S., Wonka, P., Elhoseiny, M.: Supplementary Material for 3D CoMPaT: Composition of Materials on Parts of 3D Things (2022), available at https://3dcompat-dataset.org/pdf/supplementary.pdf, version 1.0 7, 8, 10, 13, 14
23. Li, Z., Yu, T.W., Sang, S., Wang, S., Song, M., Liu, Y., Yeh, Y.Y., Zhu, R., Gundavarapu, N., Shi, J., Bi, S., Xu, Z., Yu, H.X., Sunkavalli, K., Hašan, M., Ramamoorthi, R., Chandraker, M.: Openrooms: An end-to-end open framework for photorealistic indoor scene datasets (2021) 3, 4
24. Lin, H., Averkiou, M., Kalogerakis, E., Kovacs, B., Ranade, S., Kim, V., Chaudhuri, S., Bala, K.: Learning material-aware local descriptors for 3d shapes. 2018 International Conference on 3D Vision (3DV) (Sep 2018). https://doi.org/10.1109/3dv.2018.00027, http://dx.doi.org/10.1109/3DV.2018.00027 3, 4, 9
25. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual mlp framework. arXiv preprint arXiv:2202.07123 (2022) 11
26. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 1, 2, 4, 6, 9
27. Park, K., Rematas, K., Farhadi, A., Seitz, S.M.: Photoshape: Photorealistic materials for large-scale shape collections. ACM Trans. Graph. **37**(6) (Nov 2018) 4, 9
28. Pratt, S., Yatskar, M., Weihs, L., Farhadi, A., Kembhavi, A.: Grounded situation recognition. In: ECCV (2020) 12
29. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems. pp. 5099–5108 (2017) 11
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 13
31. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, D.T., Yeung, S.K.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: International Conference on Computer Vision (ICCV) (2019) 11
32. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (TOG) **38**, 1 – 12 (2019) 11
33. 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 conference on computer vision and pattern recognition. pp. 1912–1920 (2015) 1, 4, 9

34. Yatskar, M., Zettlemoyer, L., Farhadi, A.: Situation recognition: Visual semantic role labeling for image understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5534–5542 (2016) 12

35. Yi, L., Kim, V.G., Ceylan, D., Shen, I.C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.: A scalable active framework for region annotation in 3d shape collections. ACM Transactions on Graphics (ToG) **35**(6), 1–12 (2016) 9