3D Compositional Zero-shot Learning with DeCompositional Consensus

Muhammad Ferjad Naeem*[1], Evin Pınar Örnek*[2], Yongqin Xian[1], Luc Van Gool[1], and Federico Tombari[2,3]

[1] ETH Zürich, [2] TUM, [3] Google

Abstract. Parts represent a basic unit of geometric and semantic similarity across different objects. We argue that part knowledge should be composable beyond the observed object classes. Towards this, we present 3D Compositional Zero-shot Learning as a problem of part generalization from seen to unseen object classes for semantic segmentation. We provide a structured study through benchmarking the task with the proposed Compositional-PartNet dataset. This dataset is created by processing the original PartNet to maximize part overlap across different objects. The existing point cloud part segmentation methods fail to generalize to unseen object classes in this setting. As a solution, we propose DeCompositional Consensus, which combines a part segmentation network with a part scoring network. The key intuition to our approach is that a segmentation mask over some parts should have a consensus with its part scores when each part is taken apart. The two networks reason over different part combinations defined in a per-object part prior to generate the most suitable segmentation mask. We demonstrate that our method allows compositional zero-shot segmentation and generalized zero-shot classification, and establishes the state of the art on both tasks.

Keywords: 3D Compositional Zero-shot Learning, Compositionality.

1 Introduction

A centaur is a mythological creature with the upper body of a human and the bottom body of a horse. This creature was never observed in our world, yet even a child can label its body parts from the human head to the horse legs. We humans can dissect the knowledge of basic concepts as primitives, like parts from human head to horse legs, to generalize to unseen objects. Cognitive studies have shown that humans learn part-whole relations in hippocampal memory to achieve object understanding through compositionality [18,47]. Compositionality has evolved as a survival need since every combination of every primitive cannot be observed.

Parts represent a basic primitive of geometric and semantic similarity across objects. Recently, PartNet dataset has been introduced to study fine-grained semantic segmentation of parts [35]. This has inspired several architectural works

[1] First and second author contributed equally.

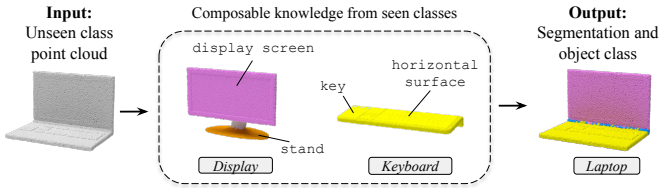

Fig. 1: We aim to compose parts (*e.g.,* `display screen, key, horizontal surface`) from seen (*e.g., Display, Keyboard*) to unseen object classes (*e.g., Laptop*) for semantic segmentation and classification in 3D point clouds.

towards improving supervised fine-grained segmentation in 3D models [54,5,57]. A parallel line of work uses the concept of parts to improve tasks like 3D reconstruction with hierarchical decomposition [39], unsupervised segmentation by finding repeated structural patterns [29], and instance segmentation in unseen objects [9]. However, these works do not predict semantic part classes.

With the availability of RGB-D sensors and the ease of acquiring 3D data in domains from augmented reality to robotic perception, the need for object understanding beyond seen object classes has emerged [38,2,51]. A model is unlikely to be trained for all possible existing objects [10,12], however, man-made environments consist of objects that share similarities through their parts. In this scenario, reasoning over learned parts can present an avenue for generalization to unseen objects classes. Zero-shot learning with 3D data has received far less attention compared to 2D domain. In this work, we introduce a new task, namely 3D Compositional Zero-Shot Learning (3D-CZSL), aiming at jointly segmenting and classifying 3D point clouds of both seen and unseen object classes (see Figure 1). 3D-CZSL is a challenging task as it requires generalizing parts from seen object classes to unseen classes that can be composed entirely of these parts.

Our contributions are as follows: (1) We formalize zero-shot compositionality for 3D object understanding with semantic parts and introduce the 3D-CZSL task. To the best of our knowledge, we present the first work for joint classification and semantic part labeling for compositional zero-shot learning in 3D. (2) We establish a novel benchmark through Compositional PartNet (C-PartNet), which enables research in 3D-CZSL through 16 seen and 8 unseen object classes. (3) We show that existing point cloud models fail to generalize beyond the seen object classes, whereas the performance of existing 2D zero-shot methods is severely limited in the 3D domain. (4) We propose a novel method, DeCompositional Consensus, which maximizes agreement between a segmentation hypothesis and its decomposed parts. Our method sets the state of the art for 3D-CZSL.

2 Related Work

Our work lies at the intersection of compositionality, zero-shot learning, and 3D point cloud part segmentation and discovery.

Compositionality is the notion of describing a whole through its parts, studied thoroughly in many disciplines such as mathematics, physics, and linguistics. Hoffman [19] and Biederman [4] suggested that human object recognition is based on compositionality. They heavily influenced both traditional and modern computer vision research, such as describing objects by their primitives in Deformable Part Models [15], images as a hierarchy of features in Convolutional Neural Networks [61,26], understanding a scene through its components as in Scene Graphs [22], events as a set of actions as in Space-Time Region Graphs [53]. Parts have been used as semantic and geometric object primitives, which were seen to be captured within CNN kernels implicitly [17,16].

Zero-shot learning (ZSL) addresses the task of recognizing object classes whose instances have not been seen during training [25,49,56]. This is attained through auxiliary information in the form of attributes (ALE [1]), word embeddings (SPNet [55]), or text descriptions [46]. ***Compositional zero-shot learning (CZSL)*** focuses on detecting unseen compositions of already observed primitives. The current literature on the topic focuses on state-object compositionality. Towards this, one line of research aims to learn a transformation between objects and states [34,37,28]. Another line proposes a joint compatibility function with respect to the image, state, and object [42,58,31]. Graph methods are also recently used in this direction including learning a causal graph of state object transformations [3] and using the dependency structure of state object compositions to learn graph embeddings [36,30]. There have been some preliminary works exploring ***zero-shot learning in 3D*** as an extension of 2D methods including projecting on word embeddings [12], using transductive approaches [10], along with some unlabelled data [11], and using generative models to learn the label distribution of unseen classes [32].

3D part segmentation and discovery aims at parsing 3D objects into semantically and geometrically significant parts. The PartNet dataset [35] enabled studying fine-grained 3D semantic segmentation, hierarchical segmentation, and instance segmentation. The existing point cloud processing methods accomplish the task through conditioning the model over the known object class. PointNet models [43,44] provide multi-layer-perceptron (MLP) based solutions, DGCNN [54] uses graph convolutions for point clouds, ConvPoint [5] pre-processes points to define neighborhoods for convolutions, GDANet [57] uses attention in addition to MLP and currently holds state of the art for part semantic segmentation. Capsule Networks [48,62] propose architectural changes that implicitly model parts for tasks like object classification and segmentation.

An alternate line of work uses the idea of parts in objects for downstream tasks like instance segmentation and point cloud reconstruction. This includes discovering geometrically similar part prototypes, similar to superpixels [29], predicting category-agnostic segmentation through a clustering approach [52], finding repetitive structural patterns in instances of an object [9], modeling 3D objects as compositions of cuboids [50], superquadrics [21,40,39], convex functions [13], and binary space partitioning planes [8] through deep learning.

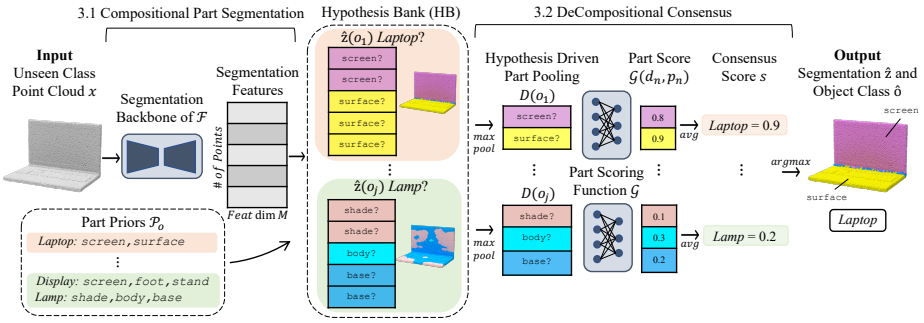

Fig. 2: **DeCompositional Consensus(DCC)** combines our compositional part segmentation function $\mathcal{F}$ with our part scoring function $\mathcal{G}$. We use the Part Prior $\mathcal{P}_o$ of which parts can exist in each object class to populate the Hypothesis Bank with multiple segmentation masks. These hypotheses are used in a Hypothesis Driven Part Pooling to get a part descriptor of each part as an input to the part scoring function $\mathcal{G}$ to calculate the DCC Score. This score measures the agreement of the segmentation mask with its part scores when each part is taken apart like lego blocks. Hypothesis with the maximum DCC score is selected for compositional zero-shot segmentation and zero-shot classification.

Our work lies at the intersection of these three areas. Similar to CZSL works [34,37,28,36,3], we study the compositionality of learned primitives, however, we are interested in parts of objects rather than state-object relations. Similar to ZSL works [25,49,56,1,55,46], we learn classification scores of unseen object classes, however, our method only uses parts as side information and does not rely on any pretrained models like word embeddings. Similar to part discovery in objects [29,9,39,23,62], we rely on parts as a basic unit of understanding an object. However, instead of geometric primitives, we use human-defined semantic parts, which tightly couple geometry, semantics, and affordances [14]. Our method further has parallels to ensemble learning, where a combination of learners solves the same downstream task [7], however, we use an agreement between different tasks to improve generalization.

3 Proposed Approach

In the following, we formalize the problem and explain the proposed solution.

Problem formulation. Let $\mathcal{T}$ define the training set with instances (x, o, z), where x is an input object point cloud described as a set of points in $\mathbb{R}^3$, o is the object class label from the set of seen object classes $\mathcal{O}_s$ and z is the part segmentation mask labelled with parts p from the set of all possible parts $\mathcal{P}$. We task a model to generalize to a set of unseen object classes $\mathcal{O}_u$, *i.e.*, $\mathcal{O}_s \cap \mathcal{O}_u = \emptyset$. We assume that $\mathcal{O}_u$ is labelled with the same part set P for part segmentation that was completely observed in seen object classes $\mathcal{O}_s$. This makes

the part segmentation task as a compositional zero-shot problem and the object classification task as a generalized zero-shot problem, *i.e.,* we predict over the full object set $\mathcal{O} = \mathcal{O}_s \cup \mathcal{O}_u$ at inference for object classification. We further assume that the model has access to a **part prior** for all object classes. For an object class o, this prior is defined as the set of parts $\mathcal{P}_o = \{p_1, ..., p_l\}$ that it can be labelled with for part segmentation.

Method overview. Part segmentation is a challenging task, as it requires one model to adapt to parts of varying scale, orientation, and geometry for all objects. Existing point cloud part segmentation methods simplify this by learning an object class conditioned model, either by training separate models specialized for each object class [35], or by feeding the object class label as an input to the model (one-hot class vector [43,54,57,62]). However, this requires an object class input at test time which is not available for unseen object classes. In this work, we refer to this case as **object prior**, *i.e.,* the model has access to the ground truth object class. The first step of our approach removes the object prior assumption and proposes Compositional Part Segmentation. In the second step, we propose our model DeCompositional Consensus which predicts the object class using the part segmentation from the previous step. It learns an agreement over a segmentation hypothesis and its part-based object classification score based on the idea of an object being taken apart like Lego blocks. The full model is depicted in Figure 2.

3.1 Compositional Part Segmentation

We reformulate part segmentation to allow compositional reasoning by encoding the part prior into the optimization criterion and the model inference. Formally, given an input point cloud x, we define $\mathcal{F}(x, p)$ as the part segmentation function, with learnable parameters W, which returns a part score for each part p in the full part set $\mathcal{P}$. At training time, we compute the segmentation loss L_{Seg} from [43,44] as a cross entropy over parts in the part prior $\mathcal{P}_o$ of the ground truth object class o rather than the full part set $\mathcal{P}$. At inference, the predicted part segmentation mask $\hat{z}(o)$ of an object class o is computed over the scores of parts in its part prior:

$$\hat{z}(o) = \arg\max_{p \in \mathcal{P}_o}(\mathcal{F}(x, p)) \tag{1}$$

With the proposed changes, a part segmentation model such as PointNet[43] can now compositionally generate a part segmentation mask for any object class we have a part prior for. Furthermore, the proposed improvements also prevent unintended biases against similar parts in different object classes as shown experimentally later in Table 3b. Notably, when an object prior is available as ground truth class, it defines the upper bound of the part segmentation performance for a model (see Table 1a "Object Prior"). In the absence of an object prior (GT object class), we can predict $|\mathcal{O}|$ segmentation masks (hypotheses) for each object class we have a part prior for, generating the **Hypothesis Bank (HB)**. Next, we introduce our novel method which allows for selecting the most suitable segmentation hypothesis for an input point cloud.

3.2 DeCompositional Consensus

We propose a novel method, **DeCompositional Consensus (DCC)**, which learns an agreement (Consensus) over a segmentation hypothesis and its part-based object classification score when the object is taken apart (DeComposed) into parts like lego blocks as segmented in the hypothesis. DCC is based on the idea that we can learn what a valid part descriptor is from seen object classes to generalize to unseen object classes.

Hypothesis driven part pooling. We extract a part descriptor for each part in a segmentation hypothesis from the point-wise features of the segmentation backbone as shown in Figure 2. Part segmentation models like PointNet[43] generate this feature representation in the penultimate layer of the model, *i.e.,* before the final per-part segmentation scoring layer. We use the segmentation hypothesis as the pooling mask to pool over the point dimensions of the feature map for each part. This results in a permutation invariant part feature vector for each part, *i.e.,* part descriptor representing the features responsible for that part segmentation in this hypothesis. We choose maxpool as the pooling operation due to its wide adoption in point cloud literature [43,44]. For the segmentation hypothesis $\hat{z}(o)$ of an object class o, this operation returns a set $\mathcal{D}(o)$ with part descriptors d for each part p found in this hypothesis. Note that $|\mathcal{D}(o)|$ is not always equal to $|\mathcal{P}_o|$ as a segmentation hypothesis might not contain all parts defined in the $\mathcal{P}_o$, *e.g.,* an instance of a chair might or might not contain sidearms.

Learning DeCompositional Consensus. Our DCC model learns a part scoring function $\mathcal{G}$ with weights Θ. For a part descriptor d, the function returns a score $\mathcal{G}(d, p)$ which measures the likelihood of this part descriptor to belong to the part p. We define DeCompositional Consensus score as the agreement between the segmentation hypothesis and the part scores. For an object hypothesis $\hat{z}(o)$, the DCC score is defined as:

$$s(x, o) = \frac{1}{|\mathcal{D}(o)|} \sum_{n=1}^{|\mathcal{D}(o)|} \mathcal{G}(d_n, p_n) \tag{2}$$

Our novel DCC score measures the individual consensus of each part descriptor with the full segmentation mask to define an object classification score. We optimize DCC score for classification with a cross entropy loss over $\mathcal{O}_s$ as:

$$L_{DeComp} = -log(\frac{\exp s(x, o)}{\sum_{o' \in \mathcal{O}_s} \exp s(x, o')}) \tag{3}$$

Since L_{DeComp} is computed over the Hypothesis Bank generated by $\mathcal{F}$, an additional part classification loss L_{Part} is computed using the ground truth segmentation mask z of each input to prevent bias against parts that are hard to segment. L_{Part} uses the ground truth segmentation mask to extract part descriptor set $\mathcal{D}_{gt}$ and optimizes them for part classification over $\mathcal{P}$.

$$L_{Part} = \sum_{n=1}^{|\mathcal{D}_{gt}|} -log(\frac{\exp \mathcal{G}(d_n, p_n)}{\sum_{p' \in \mathcal{P}} \exp \mathcal{G}(d_n, p')}) \tag{4}$$

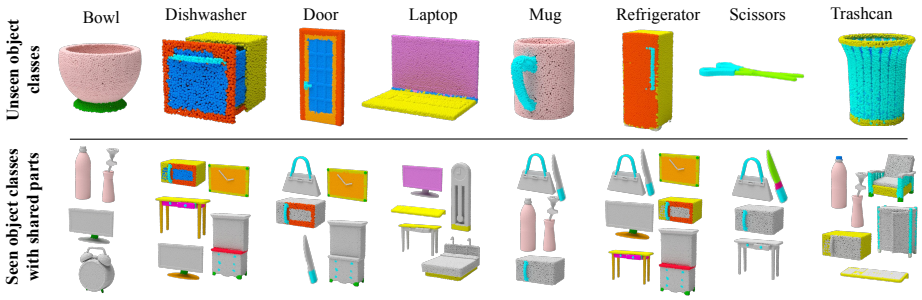

Fig. 3: **Compositional PartNet** refines the labels of PartNet dataset to maximize shared parts across different object classes, and enables studying 3D-CZSL task. The available 24 object classes are divided into 16 seen classes for training and 8 unseen classes for inference in zero-shot. We depict the shared labels between seen and unseen object classes in same colors.

Inference. For generalized zero-shot inference, the HB is populated over all object classes $\mathcal{O} = \mathcal{O}_s + \mathcal{O}_u$. The object class prediction $\hat{o}$ for an input point cloud x is retrieved by selecting the object class with the highest DeCompositional Consensus score:

$$\hat{o} = \arg\max_{o' \in \mathcal{O}} (s(x, o')) \tag{5}$$

The corresponding hypothesis of the predicted class $\hat{o}$ becomes the final part segmentation output, *i.e.,* $\hat{z}(\hat{o})$. Our technical novelty lies in defining part descriptors as features responsible for part segmentation in a hypothesis; and using their likelihood to define an object class level consensus score to achieve zero-shot compositionality. In contrast to several zero-shot baselines [55,36], our method does not require any supervised calibration step over the unseen classes.

4 Compositional PartNet Benchmark

Zero-shot compositionality in machine learning algorithms has mainly been studied for state-object relations in image datasets like MIT-States [20], UT-Zappos [60], AO-CLEVr [3], and more recent C-GQA [36]. These datasets have several limitations such as including label noise [20,3], lacking visual cues [60,36], being too simple [3], or missing multilabel information [36].

We believe that 3D part object relations provide an ideal avenue to study zero-shot compositionality, as they tend to be more well-defined albeit challenging. There have been several attempts in a part-based benchmark [6,59] for 3D object understanding. Recently, ShapeNet has been extended with fine-grained part labels to form the new dataset PartNet [35]. PartNet provides 24 distinct object classes, annotated with fine-grained, instance-level, and hierarchical 3D part information, consisting of around 26K 3D models with over 500K part instances and 128 part classes. However, these part class labels are not unified

across different object categories, preventing a study into zero-shot compositionality. We refine PartNet into **Compositional PartNet (C-PartNet)** with a new labeling scheme that relates the compositional knowledge between objects by merging and renaming the repeated labels as shown in Figure 3.

Unifying part labels. While PartNet provides three levels of hierarchical part labels, not all objects are labeled at the deepest level. We take the **deepest level** available for each object. We find similar parts within and across different objects by training a supervised segmentation model and compute pairwise similarities between parts across PartNet. Parts that share a high similarity and have the same semantic meaning (*e.g.,* `bed horizontal surface` in object Bed and `horizontal surface` in Storage Furniture) are merged into a single general part label (`horizontal surface`). Furthermore, parts with a similar function but different name (*e.g.,* `screen side` of Laptop and `display screen` of Display) are merged together. The relabelled C-PartNet consists of 96 parts compared to 128 distinct part labels in the original PartNet. Details in the supplementary.

Selecting test time unseen object classes. Objects that share a similar function tend to have similar parts [4]. We divide PartNet objects into several functional categories. Details of this categorization and the dataset statistics can be found in the supplementary. We identify three easy to compose unseen object classes (*i.e.,* Mug, Bowl and TrashCan), that share large similarities with seen object classes (Bottle and Vase). Furthermore, we choose three object classes of medium difficulty that require generalizing parts beyond the context they were observed in (*i.e.,* Dishwasher, Refrigerator, and Laptop). Finally, Scissors and Door present two hard-to-compose object classes that require generalizing beyond scale, context, and number instances of parts compared to seen object classes. The validation set contains all seen and 2 unseen object classes (Bowl and Dishwasher). The test set consists of 16 seen O_s and 8 unseen classes O_u.

5 Experiments

Since our proposed benchmark lies at the intersection of point cloud processing, attribute learning, zero-shot learning, and its specialized sub-domain compositional zero-shot learning, we adapt baselines representing these lines of works.

Baselines. *Object Prior* uses a point cloud part segmentation model trained with our framework and evaluates the segmentation performance on the ground truth object. This is the oracle upper bound for the zero-shot models. *Direct Seg* trains a point cloud part segmentation model $\mathcal{F}$ without a part prior to predict over all parts $\mathcal{P}$ in the dataset. *PartPred* is inspired from a classic zero-shot baseline DAP [24] and trains a part prediction network from the global feature of each point cloud. The predicted parts are used as P_o for equation 1 to condition the compositional part segmentation network [24]. For zero-shot classification baselines, we use the predicted class to select the corresponding segmentation mask from the Hypothesis Bank. Among these, *SPNet* [55] learns classification by projecting the global feature of an input on a pretrained distribution where both seen and unseen objects lie *e.g.,* word embeddings. *CGE*

[36] proposes to model compositional relations using a graph consisting of parts connected to objects they occur in. We reformulate CGE to a multitask setup and use part nodes for segmentation and object nodes for classification. *PartPred DCC* uses the part prediction network's scores for parts found in each segmentation hypothesis to calculate the consensus score from Equation 2. Finally, *3D Capsule Networks* [62] aim to discover part prototypes through unsupervised reconstruction. Segmentation is subsequently learned by a linear mapper from capsules to part labels. We give additional details about these baselines in the supplementary and also compare with the current SOTA for part class agnostic segmentation method, Learning to Group [29], on unseen object classes.

Metrics. The proposed benchmark consists of two jointly learned tasks. For the **compositional zero-shot segmentation**, we report the mean object class-wise Intersection-over-Union (mIoU) for part labels over seen and unseen object classes. We also report the harmonic mean over seen and unseen object classes to study the best generalized zero-shot performance. In addition, we report a per-object mIoU to study model performance on each unseen object across the three difficulty levels. For **generalized zero-shot classification**, we report for the per-object class top-1 classification accuracy over unseen classes, mean accuracy over seen classes, unseen classes and their harmonic mean. For models that apply joint classification and segmentation, we choose the checkpoint with the best segmentation performance to encourage compositional part understanding. Part based classification baselines can give the same scores across two objects if an instance does not have all parts, *e.g.,* an empty Vase has the same parts in Vase Hypothesis and Bowl. This is counted as an accurate classification, since the ground truth object still receives the highest score and achieves compositional segmentation.

Training details. For its simplicity and competitive performance in our ablations (see Table 3), we choose PointNet [43] as the backbone model for $\mathcal{F}$ in our baseline comparisons in Table 1a, 1b. We also report further results on DGCNN [54], ConvPoint [5] and GDANet [57] in Table 3 . All backbones are pretrained with the author's implementations extended by our framework. The pretrained models are then used as initialization for the zero-shot models and are finetuned. For our model DCC, we use a 2-layer MLP with 512 hidden dimensions, ReLU, and dropout followed by a linear layer as function $\mathcal{G}$. We use a step size learning rate scheduler between $1e^{-3}$ and $1e^{-5}$ with Adam optimizer. We use cross entropy as segmentation loss L_{Seg} for part segmentation similar to [43,44,54,57]. L_{Seg}, L_{DeComp} and L_{Part} are equally weighted and the network is trained until convergence on the validation set. We use Word2Vec [33] for models that rely on word embeddings [55,36]. For CGE, we choose the graph configuration that achieved the best result on the validation set at 2 layers of GCN with a hidden dimension of 1024. Our framework is implemented in PyTorch [41] and all experiments are conducted using Nvidia A100 GPUs. The dataset and experimental framework will be released upon acceptance.

Method	Unseen object classes										
	HM	S	U	Bowl	Dish	Door	Lap	Mug	Refr	Scis	Trash
Object Prior [43]	47.9	52.8	43.8	77.0	40.2	25.1	72.4	47.1	31.9	22.5	34.2
Direct Seg [43]	28.5	**48.7**	20.1	62.9	4.0	1.6	19.9	35.7	0.9	0.0	33.9
SPNet* [55]	8.5	28.5	5.0	12.6	2.5	0.5	0.0	2.6	2.6	0.0	15.8
CGE* [36]	30.8	37.0	26.4	67.0	19.5	0.3	35.1	39.6	11.2	0.0	33.6
3D-PointCapsNet [62]	4.4	9.4	2.9	4.3	0.0	0.2	1.2	11.2	0.1	0.1	6.5
PartPred [24]	26.3	33.6	21.6	66.2	2.3	7.2	19.4	43.1	0.5	0.0	32.5
PartPred DCC [24]	20.9	41.3	14.0	35.5	2.1	**7.2**	17.2	29.2	0.7	0.0	20.0
DCC (ours)	**35.2**	38.0	**32.7**	**66.1**	**30.9**	5.3	**56.3**	**40.4**	**28.4**	0.0	**34.2**

(a) **Compositional Zero-shot Segmentation**

Method	Unseen object classes										
	HM	S	U	Bowl	Dish	Door	Lap	Mug	Refr	Scis	Trash
SPNet* [55]	3.8	46.7	2.0	12.0	0.0	3.1	0.0	0.0	0.0	0.0	0.0
CGE* [36]	33.1	54.3	23.8	31.9	0.0	0.0	52.0	1.0	33.7	0.0	**71.9**
PartPred DCC [24]	19.9	**74.0**	11.5	4.3	3.3	**25.8**	0.0	13.5	0.0	0.0	45.2
DCC(ours)	**55.9**	73.2	**45.2**	**79.8**	**57.1**	5.3	**55.4**	**71.9**	**55.6**	0.0	36.8

(b) **Generalized Zero-shot Classification**

Table 1: **Baseline comparison.** We compare our proposed method, DeCompositional Consensus (DCC), against baseline and report results for the two tasks. * marks baselines that require supervised calibration. For (a), we report mIoU % over part labels per object class over seen objects, unseen objects, and their harmonic mean. We also report the mIoU over each unseen object class. For (b), we report the top-1 classification accuracy. DCC achieves SOTA on both tasks.

5.1 Comparing with State of the Art

We compare our method with baselines on compositional zero-shot segmentation in Table 1a and generalized zero-shot classification in Table 1b. Our method outperforms all baselines on almost all metrics and establishes state of the art on both tasks.

Compositional zero-shot segmentation performance. Our method demonstrates remarkable performance gains on all unseen classes and achieves the best harmonic mean on compositional zero-shot segmentation in Table 1a. We achieve a 50% improvement over the direct segmentation demonstrating that the introduction of object class conditioned inference with DCC can improve compositional zero-shot segmentation in point cloud models. This improvement is observed most in unseen object classes that have large variations in parts from

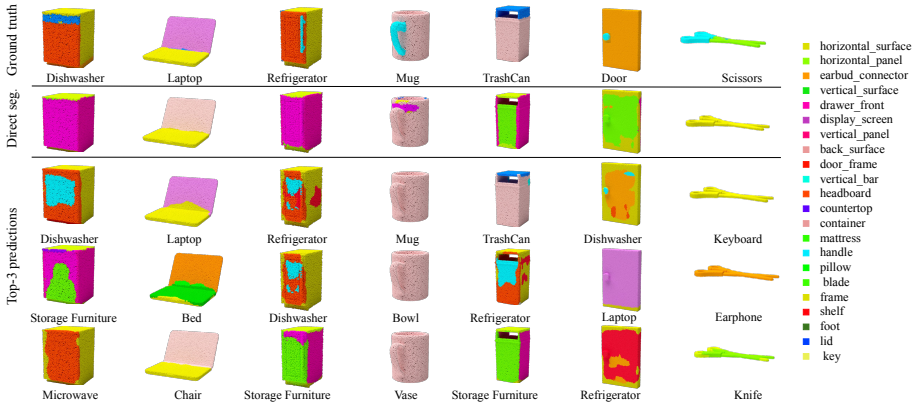

Fig. 4: **Qualitative results.** Direct segmentation tends to segment an input point cloud to parts from seen objects with large geometric similarities. While this works for the objects from Container category, it fails in more complex objects that share similarity with Furniture while being composed of parts from other categories. In contrast, DeCompositional Consensus builds an implicit understanding of what parts can occur together in different categories, and achieves meaningful segmentations for all object classes but Door and Scissors.

the seen object classes like Dishwasher (7.5×), Laptop (2.5×) and Refrigerator (28×) as shown in Fig. 4. Unseen object classes that share large geometric and semantic similarities with respect to parts to seen object classes also have significant improvements. This includes improvements in Bowl (4%), Mug (14%), and TrashCan (1.4%) that have very similar parts with seen Bottle and Vase.

Comparing with zero-shot learning baselines, we observe that our method achieves the best performance in 6 out of 8 classes and establishes a state of the art in overall harmonic mean and unseen mIoU while achieving competitive seen IoU. PartPred [24] learns to dynamically predict parts and generalizes to unseen objects that share part and geometric similarities with seen objects in the Container category but fails in other objects. As SPNet [55] does not use any part information, it fails to generalize to unseen objects by projecting on word embeddings alone. Compared to SPNet, CGE [36] performs much better as it uses the part prior and refines the word embeddings by using the dependency structure defined in the graph. Although being competitive on Bowl, Dishwasher, Mug and TrashCan, it performs much poorer on other unseen objects. 3D-PointCapsNet [62], while conceptually engineered for part-whole relations, fails to generalize to unseen objects, likewise having very low performance on seen objects. We relate this performance to the capsules' inability to generalize without object prior as further shown in the supplementary material. Finally, PartPred DCC, achieves impressive performance on seen objects but fails on the unseen objects showing the importance of learning consensus over the features responsible for a hypothesis. We observe that while some methods have almost zero classification

accuracy, they can still achieve some segmentation performance due to confusion with objects that share some parts with the ground truth object. All methods fail to generalize to challenging object classes Door and Scissors. We discuss that in qualitative analysis in Section 5.3.

Zero-shot classification performance. Our method also achieves significant gains on generalized zero-shot classification as seen in Table 1b. DCC attains the best harmonic mean and unseen classification accuracy while maintaining a competitive seen performance. In fact, the best seen performance is achieved by PartPred DCC which extends our DCC score to a simple attribute (part) prediction model. This shows the power of enforcing consensus in different decisions of a model. Specifically, DCC is able to classify 6 out of the 8 unseen object classes with an outstanding accuracy. SPNet is only able to classify Bowl with a low accuracy of 12%. CGE is again a competitive baseline here. However, it is only able to receive reasonable classification scores on 4 of the 8 unseen object classes while maintaining a competitive seen class performance.

5.2 Ablations

We ablate our design choices and compare performance against different point cloud backbones.

Optimization criteria. We ablate over the two optimization criteria for DCC in Table 2. As seen from row a) that only training for L_{DeComp} is unable to attain high performance as it can introduce bias against hard to predict parts to increase classification performance. Similarly, only training for L_{Part} in row b) achieves low performance as the model is not optimized for the downstream classification task of predicting the consensus score. Row c) and d) combine both of these losses and see a big performance gain. In row c) we replace the predicted segmentation mask corresponding to the ground truth object class in HB with the ground truth segmentation mask. Comparing row c) and d) in Table 2, we see that when we learn DCC score exclusively on the model's predicted segmentation instead of using ground truth segmentation mask, we see a large improvement in seen and unseen performance. We conjecture that the part scoring function $\mathcal{G}$ learns the segmentation network's limitations in this setting, *i.e.,* if a part is not predicted well by $\mathcal{F}$, $\mathcal{G}$ can look for cues from other parts.

Comparing point cloud backbones. We compare point cloud backbones under Direct Segmentation and DCC in Table 3a. We see that all models are unable to achieve competitive performance over unseen object classes with direct segmentation. In fact, ConvPoint [5] completely fails under this setting. The introduction of DCC to every backbone leads to a major increase in performance on the unseen object while being competitive over seen classes. This shows that our model can be readily extended to various families of point cloud backbones. In Table 3b, we compare the oracle segmentation performance over the ground truth object class when trained with and without our part prior optimization criterion (L_{Seg} over $\mathcal{P}_o$ or over $\mathcal{P}$) . In absence of our criterion, we observe a large difference between the performance on seen and unseen object classes. We

	Hyperparameters			**Classification**			**Segmentation**		
	L_{DeComp}	L_{Part}	*Segonly*	HM	S	U	HM	S	U
a)	✓		✓	29.0	38.1	23.4	23.3	24.2	22.5
b)		✓		14.9	34.8	9.4	24.8	22.1	28.3
c)	✓	✓		52.6	54.4	50.9	39.1	35.9	42.9
d)	✓	✓	✓	**72.8**	**76.6**	**69.3**	**45.1**	**40.9**	**50.2**

Table 2: **Ablating over L_{DeComp} and L_{Part}**, we see that both the criterion complement each other to achieve the best performance.

Backbone	**Direct Seg**			**DCC**		
	HM	S	U	HM	S	U
PointNet [43]	28.5	48.7	20.1	35.2	38.0	**32.7**
DGCNN [54]	**29.5**	**50.0**	**20.9**	**36.2**	45.1	30.2
ConvPoint [5]	2.9	5.2	2.0	29.5	35.0	25.5
GDANet [57]	28.7	47.7	20.5	33.5	**46.2**	26.4

(a) **CZSL Segmentation**

Backbone	L_{Seg} **over** $\mathcal{P}$			L_{Seg} **over** $\mathcal{P}_o$		
	HM	S	U	HM	S	U
PointNet [43]	43.2	51.7	37.1	47.9	52.8	43.8
DGCNN [54]	**44.6**	52.4	**38.8**	**50.0**	**55.0**	**46.3**
ConvPoint [5]	29.1	28.7	29.5	43.5	42.4	43.0
GDANet [57]	43.4	**53.1**	36.8	48.0	53.7	43.4

(b) **Oracle Performance**

Table 3: **Backbone ablation**. (a) We see DCC results in a large improvement compared to direct segmentation across all ablated point cloud models (b) We further see that our Part Prior optimization criterion greatly benefits all backbones under oracle evaluation especially on unseen objects.

conjecture that the model overfits to seen object classes, limiting compositionality to unseen object classes. With our criterion, *e.g.,* PointNet segmentation network improves up to 19% on unseen and 1% on seen classes.

5.3 Qualitative and Model Limitation

In Figure 4, we show some qualitative results for direct segmentation versus top-3 results of our model across unseen objects. We further validate our results from Table 1a, and see that for the easy object Mug, the direct segmentation can give a meaningful result. However, it fails for other relatively harder objects, which can be attributed to the lack of affordance, *i.e.,* how an agent interacts with an object. Our method, despite not having access to affordances, builds an implicit understanding of what parts occur in each object category and is thus able to learn a reasonable consensus score. This brings about an increased generalization and meaningful results for all objects. We see a correct prediction for even Dishwasher and Refrigerator, which are closer geometrically to Furniture than Microwave, their closest functional seen object. However, although less, DCC also suffers from lack of affordances. For example, among the top-3 result for a Laptop in Figure 4 is a Chair which shares geometrical similarity to an open Laptop. This indicates an upper bound to the performance that can be

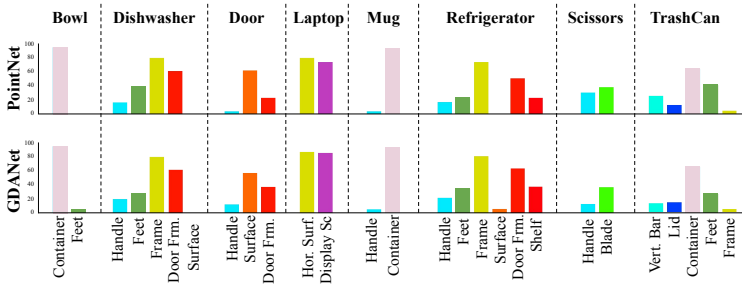

Fig. 5: **Error plots.** We find that PointNet[43] is comparable in mIoU to a much newer model, GDANet[57], across unseen objects.

achieved from visual data alone [27,45]. An affordance prior can help address this limitation for part-object relations.

Another aspect that limits our model performance is the generalization limitations of point cloud backbones. In Figure 5, we compare the object prior per part performance on the unseen objects between PointNet [43] and GDANet [57], which were released five years apart. A surprising insight we observe is that years of progress in point cloud processing, while making a significant advance on seen object classes, does not translate to improvement on unseen object classes. We see that there is no clear consensus on which model is better for unseen object class generalization. Even using the right part prior, some parts are unlikely to be segmented in unseen classes. An example of this is handle, which is unable to be reasonably segmented for Mug, Dishwasher, and Refrigerator. A more extreme case of this is observed in Door and Scissors, where the segmentation fails completely as shown in last two columns of Figure 4. These objects have a large variation with respect to parts from the seen objects in scale, the number of instances (two blades in Scissors vs one in Knife), and orientation.

6 Conclusion and Future work

We introduce 3D-CZSL as a joint compositional zero-shot segmentation and generalized zero-shot classification task. We provide a structured study into zero-shot compositionality through a novel benchmark on the proposed C-PartNet dataset and show that previous models do not generalize beyond the training object classes. Towards this, our novel approach, DeCompositional Consensus, maximizes the agreement between a segmentation hypothesis and its parts when taken apart, and sets a new SOTA. We also show that while there has been a lot of progress in part segmentation in a supervised setting, simple models like PointNet are still competitive in unseen object classes, arguably because the current research has not focused on this task. There are several future directions that can stem from this work, including introducing affordance priors and extension of the capsules paradigm for part reasoning on unseen object classes. We also hope to inspire future research into compositional point cloud models.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for attribute-based classification. In: CVPR (2013)
2. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: CVPR (2016)
3. Atzmon, Y., Kreuk, F., Shalit, U., Chechik, G.: A causal view of compositional zero-shot recognition. In: NeurIPS (2020)
4. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review **94**, 115–147 (1987)
5. Boulch, A.: Convpoint: Continuous convolutions for point cloud processing. Computers and Graphics **88**, 24–34 (2020)
6. Chen, X., Golovinskiy, A., Funkhouser, T.: A benchmark for 3D mesh segmentation. SIGGRAPH (2009)
7. Chen, Z., Wang, S., Li, J., Huang, Z.: Rethinking generative zero-shot learning: An ensemble learning perspective for recognising visual patches. In: Proceedings of the 28th ACM International Conference on Multimedia (2020)
8. Chen, Z., Tagliasacchi, A., Zhang, H.: Bsp-net: Generating compact meshes via binary space partitioning. CVPR (2020)
9. Chen, Z., Yin, K., Fisher, M., Chaudhuri, S., Zhang, H.: Bae-net: Branched autoencoder for shape co-segmentation. ICCV (2019)
10. Cheraghian, A., Rahman, S., Campbell, D., Petersson, L.: Transductive zero-shot learning for 3d point cloud classification. WACV (2020)
11. Cheraghian, A., Rahman, S., Campbell, D., Petersson, L.: Mitigating the hubness problem for zero-shot learning of 3d objects. In: BMVC (2019), `https://bmvc2019.org/wp-content/uploads/papers/0233-paper.pdf`
12. Cheraghian, A., Rahman, S., Petersson, L.: Zero-shot learning of 3d point cloud objects. MVA (2019)
13. Deng, B., Genova, K., Yazdani, S., Bouaziz, S., Hinton, G., Tagliasacchi, A.: Cvxnet: Learnable convex decomposition. In: CVPR (2020)
14. Deng, S., Xu, X., Wu, C., Chen, K., Jia, K.: 3d affordancenet: A benchmark for visual object affordance understanding. In: CVPR (2021)
15. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
16. Gonzalez-Garcia, A., Modolo, D., Ferrari, V.: Do semantic parts emerge in convolutional neural networks? In: Int. J. Comput. Vis. vol. 126, pp. 476 – 494 (2018)
17. Gonzalez-Garcia, A., Modolo, D., Ferrari, V.: Objects as context for detecting their semantic parts. In: CVPR (2018)
18. Hinton, G.: Some demonstrations of the effects of structural descriptions in mental imagery. Cognitive Science **3**(3), 231–250 (1979)
19. Hoffman, D.D., Richards, W.A.: Parts of recognition. Cognition **18**(1-3), 65–96 (1984)
20. Isola, P., Lim, J.J., Adelson, E.H.: Discovering states and transformations in image collections. In: CVPR (2015)
21. Jaklic, A., Leonardis, A., Solina, F., Solina, F.: Segmentation and recovery of superquadrics, vol. 20. Springer Science & Business Media (2000)
22. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: CVPR (2015)
23. Kawana, Y., Mukuta, Y., Harada, T.: Unsupervised pose-aware part decomposition for 3d articulated objects (2021)

24. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
25. Larochelle, H., Erhan, D., Bengio, Y.: Zero-data learning of new tasks. In: AAAI (2008)
26. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
27. Li, X., Liu, S., Kim, K., Wang, X., Yang, M.H., Kautz, J.: Putting humans in a scene: Learning affordance in 3d indoor environments. In: CVPR (2019)
28. Li, Y.L., Xu, Y., Mao, X., Lu, C.: Symmetry and group in attribute-object compositions. In: CVPR (2020)
29. Luo, T., Mo, K., Huang, Z., Xu, J., Hu, S., Wang, L., Su, H.: Learning to group: A bottom-up framework for 3d part discovery in unseen categories. ICLR (2020)
30. Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Learning graph embeddings for open world compositional zero-shot learning. In: arXiv (2021)
31. Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Open world compositional zero-shot learning. In: CVPR (2021)
32. Michele, B., Boulch, A., Puy, G., Bucher, M., Marlet, R.: Generative zero-shot learning for semantic segmentation of 3d point cloud. CoRR **abs/2108.06230** (2021), `https://arxiv.org/abs/2108.06230`
33. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NeurIPS (2013)
34. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: Composition with context. In: CVPR (2017)
35. 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: CVPR (2019)
36. Naeem, M.F., Xian, Y., Tombari, F., Akata, Z.: Learning graph embeddings for compositional zero-shot learning. In: CVPR (2021)
37. Nagarajan, T., Grauman, K.: Attributes as operators: factorizing unseen attribute-object compositions. In: ECCV (2018)
38. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: ISMAR (2011)
39. Paschalidou, D., Gool, L.V., Geiger, A.: Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In: CVPR (2020)
40. Paschalidou, D., Ulusoy, A.O., Geiger, A.: Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In: CVPR (2019)
41. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS (2019)
42. Purushwalkam, S., Nickel, M., Gupta, A., Ranzato, M.: Task-driven modular networks for zero-shot compositional learning. In: ICCV (2019)
43. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. CVPR (2017)
44. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. NeurIPS (2017)
45. Qi, W., Mullapudi, R.T., Gupta, S., Ramanan, D.: Learning to move with affordance maps. In: ICLR (2020)
46. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: CVPR (2016)

47. Rolls, E.T., Treves, A.: Neural networks in the brain involved in memory and recall. In: Van Pelt, J., Corner, M., Uylings, H., Lopes Da Silva, F. (eds.) The Self-Organizing Brain: From Growth Cones to Functional Networks, Progress in Brain Research, vol. 102, pp. 335–341. Elsevier (1994)
48. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: ICLR (2017)
49. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: NeurIPS (2013)
50. Tulsiani, S., Su, H., Guibas, L.J., Efros, A.A., Malik, J.: Learning shape abstractions by assembling volumetric primitives. In: CVPR (2017)
51. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Niessner, M.: Rio: 3d object instance re-localization in changing indoor environments. In: ICCV (2019)
52. Wang, X., Sun, X., Cao, X., Xu, K., Zhou, B.: Learning fine-grained segmentation of 3d shapes without part labels. In: CVPR (2021)
53. Wang, X., Gupta, A.: Videos as space-time region graphs. In: ECCV (2018)
54. 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) (2019)
55. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Semantic projection network for zero-and few-label semantic segmentation. In: CVPR (2019)
56. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE TPAMI **41**(9), 2251–2265 (2019). https://doi.org/10.1109/TPAMI.2018.2857768
57. Xu, M., Zhang, J., Zhou, Z., Xu, M., Qi, X., Qiao, Y.: Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. AAAI (2021)
58. Yang, M., Deng, C., Yan, J., Liu, X., Tao, D.: Learning unseen concepts via hierarchical decomposition and composition. In: CVPR (2020)
59. 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. SIGGRAPH Asia (2016)
60. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: CVPR (2014)
61. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)
62. Zhao, Y., Birdal, T., Deng, H., Tombari, F.: 3d point capsule networks. In: CVPR (2019)