Zero-Shot Point Cloud Segmentation by Semantic-Visual Aware Synthesis

Yuwei Yang[1] Munawar Hayat[2] Zhao Jin[1] Hongyuan Zhu[3] Yinjie Lei[1,✉]
[1]Sichuan University [2]Monash University [3]A*STAR
yuwei@stu.scu.edu.cn munawar.hayat@monash.edu jinzhao@stu.scu.edu.cn
hongyuanzhu.cn@gmail.com yinjie@scu.edu.cn

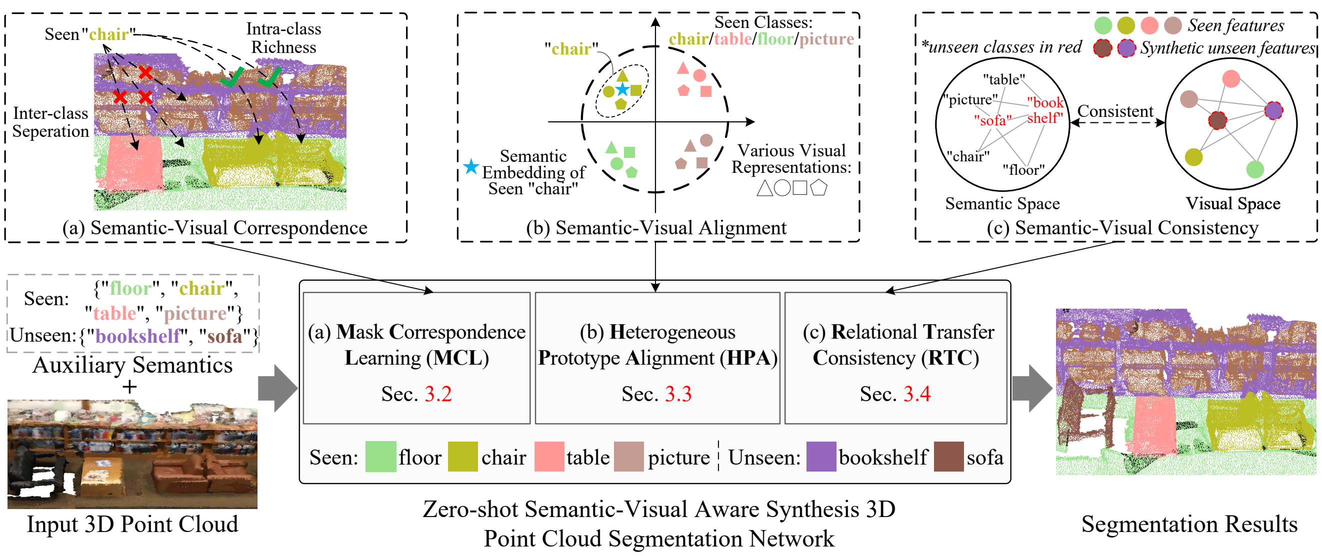

Figure 1. Our zero-shot synthesis approach for point cloud segmentation tackles multiple semantic-visual transfer issues, by enhancing correspondence (Sec. 3.2), alignment (Sec. 3.3) and consistency (Sec. 3.4) between the auxiliary-semantic and 3D-visual spaces.

Abstract

This paper proposes a feature synthesis approach for zero-shot semantic segmentation of 3D point clouds, enabling generalization to previously unseen categories. Given only the class-level semantic information for unseen objects, we strive to enhance the correspondence, alignment and consistency between the visual and semantic spaces, to synthesise diverse, generic and transferable visual features. We develop a masked learning strategy to promote diversity within the same class visual features and enhance the separation between different classes. We further cast the visual features into a prototypical space to model their distribution for alignment with the corresponding semantic space. Finally, we develop a consistency regularizer to preserve the semantic-visual relationships between the real-seen features and synthetic-unseen features. Our approach shows considerable semantic segmentation gains on ScanNet, S3DIS and SemanticKITTI benchmarks. Our code is available at: https://github.com/leolyj/3DPC-GZSL

1. Introduction

Semantic segmentation of 3D point clouds is mostly dominated by fully-supervised methods [37, 39, 49, 45, 21, 57] that require point-wise labelled data for training. While these methods perform well on previously seen objects, they lack scalability to novel and unseen classes for which no samples are available during training. Zero-Shot Learning (ZSL) provides a promising paradigm in such cases since it enables rapid generalization to unseen classes.

While ZSL from RGB images is well explored [15, 1, 17, 56, 19, 48, 24, 53, 7, 16, 41, 47, 6, 52, 18, 9, 20, 27, 58], ZSL for segmentation of point clouds is less investigated, due to unique challenges posed by 3D data (*e.g.* the lack of large-

✉ Corresponding Author: Yinjie Lei (yinjie@scu.edu.cn)

scale annotated datasets and pre-trained models [23] which are otherwise ubiquitous in 2D [34]). Most of the existing 3D ZSL methods tackle the relatively simpler classification problem [13, 10, 11, 12], with very few methods developed for segmentation [8, 30]. Chen *et al*. [8] learn shared geometric primitives to enable seen-to-unseen migration. However, their approach needs non-annotated unseen samples at training [8], which is restrictive and not suitable for practical scenarios where acquiring unseen data is not always feasible. [30] propose a feature synthesis-based approach for 3D segmentation that can simultaneously generalize to both seen and unseen, without requiring any data for unseen categories. Nevertheless, the features synthesized from the generator lack contextual diversity due to mode collapse [50], resulting in limited transfer to unseen classes.

To enable generalization to wider scenarios, we develop a feature synthesis framework, which doesn't require any samples (annotated or non-annotated) during training. Since semantics are the only common information available for seen and unseen, we need to ensure strong transfer capabilities from the semantic to the visual space. For this purpose, we consider the following semantic-visual transfer issues for ZSL: **1)** *Semantic-Visual Correspondence Mismatch.* The core of ZSL is to exploit and establish a mapping between semantics and vision, such that for a specific object, visual features can be uniquely identified from their corresponding semantics. **2)** *Heterogeneous Semantic-Visual Embedding.* The semantic vectors (embeddings of class-name words) and visual representations (from point cloud data) come from different modalities, and introduce inherent modality-specific heterogeneity that needs to be tackled in order to align the two data modalities. **3)** *Inconsistent Semantic-Visual Relationship.* The relationships between different classes, both seen and unseen, should be consistent in the semantic embedding space and visual feature space, so that the semantics for unseen can faithfully synthesize the unseen visual features.

To address these semantic-visual transfer challenges, as shown in Fig. 1, we design three modules. First, we propose a Mask Correspondence Learning (MCL) module (Sec. 3.2), to learn rich intra-class representations while enhancing inter-class boundary distribution. We believe promoting diversity between the same class features and ensuring separation between classes is critical to synthesize generalized features for unseen classes. Further, for better seen-to-unseen transfer, we align the seen visual prototypes with their corresponding semantics, using our proposed Heterogeneous Prototype Alignment (HPA) module (Sec. 3.3). Finally, while learning to synthesize the unseen visual features, we ensure that the inter-class structural relations of seen+unseen semantics are consistent with their corresponding visual features. For this purpose, we develop a Relational Transfer Consistency (RTC) module (Sec. 3.4) that transfers the seen+unseen semantic relationships with the corresponding real-seen+synthesized-unseen visual ones. Our proposed modules complement each other and constrain the generator to synthesize diverse, discriminative, and semantically relevant unseen visual features that generalize well for zero-shot segmentation.

We evaluate our model under the challenging Generalized Zero-Shot Learning (GZSL) in inductive setting, where training data contains no labelled or unlabelled unseen class samples, while the model is required to predict both seen and unseen classes at inference. We show significant gains over the current state-of-the-art on three public datasets ScanNet [14], S3DIS [2] and SemanticKITTI [3], by 7.7%, 3.8% and 3.0% respectively, according to the HmIoU metric. Our contributions can be summarized as follows:

- We propose an effective masked learning strategy, where visual features of the masked semantics are recovered via contrastive learning to enhance intra-class diversity and inter-class separation of the learned visual features, enhancing transfer to unseen classes.
- We propose cross-modality prototypical learning that aligns semantics with the visual space, thus promoting generalization to novel concepts.
- We develop consistency regularization that maintains relationships between the real+synthesized visual features with their corresponding semantics.

2. Related Works

By only using auxiliary class attributes or semantics, Zero-Shot Learning (ZSL) enables transfer of prior knowledge from seen to novel unseen classes. Here, we first review ZSL methods developed for RGB images. We then discuss existing 3D semantic segmentation techniques, followed by ZSL methods on 3D point clouds.

ZSL on RGB Images. The existing methods for zero-shot learning can be categorized as attribute-based, projection-based, knowledge-based and generative-based methods. The attribute-based methods [15, 26, 22, 1] recognize new objects using the semantic attributes of different classes. The projection-based approaches [17, 43, 51, 56] learn a mapping between visual representations and the auxiliary semantic prototypes (such as Word2Vec embeddings [31] or GloVe [36]). Knowledge-based models [19, 48, 24] employ graph networks to migrate structured knowledge from seen classes to unseen. Recently popular generative approaches [53, 7, 16, 41, 47] train generative models (*e.g*. conditional generative adversarial models [32] or variational autoencoder [46]), and then synthesize unseen latent features conditioned upon the corresponding class prototypes. The synthesized features are appiled to update the classifier to include unseen classes, which helps reduce the bias towards seen classes. The above mentioned ZSL methods are primarily developed for image classification. Some recent techniques [6, 52, 18, 9, 20, 27] extend them to ZSL for semantic segmentation in RGB images. Amongst these, generative approaches have shown most promise for RGB semantic segmentation in zero-shot setting [28, 18, 9].

3D Point Cloud Semantic Segmentation. Most of the existing methods on 3D segmentation are fully-supervised [44, 29, 38, 37, 39, 49, 45, 4, 21, 57], and project point cloud into multi-view 2D images [44, 38] or process them using voxel grids [29]. Since the seminal work PointNet [37], point clouds are encoded by using deep networks with MLPs [39], point-wise convolution [49, 45, 4], graph networks [21] or transformer [57]. While these deep models show impressive results in fully-supervised setting [35, 54], they require expensive point-wise annotations, and lack generalization to unseen classes in zero-shot setting.

ZSL on 3D Point Clouds. Compared with ZSL from RGB images, 3D ZSL is relatively less investigated. [13] adapts 2D ZSL to 3D, by learning a projection between the PointNet [37] features and the auxiliary semantics. Their work is further extended in [10, 12] to tackle the hubness problem [40], and in [11] assuming non-annotated unseen samples are available. For zero-shot segmentation, [8] learns shared geometric primitives between the seen and unseen classes by assuming that the samples of the unseen classes are available at training. Since their approach requires access to unlabelled unseen class data, *i.e.,* transductive setting, it limits their applicability to real-life scenarios where acquiring training samples for rare categories is not feasible. The closest to our approach is [30], where no training samples for unseen classes are used. While [30] synthesizes the features for unseen, they do not fully exploit the semantic-visual relationships, resulting in coarse visual features that lack effective transfer.

We can therefore conclude that while some progress has been made towards 3D zero-shot classification, 3D zero-shot semantic segmentation with no unseen training samples remains an open research problem. This work makes a progress towards this direction by learning diverse and discriminative visual features, that are well-aligned with the corresponding semantic space, thus enabling the synthesized features to generalize well to unseen classes.

3. Methodology

3.1. Problem Definition

Lets define a set of object categories as $\mathcal{C}$, with the seen $\mathcal{C}^S$ and the unseen $\mathcal{C}^U$ classes. Let $\mathcal{D}$ denote the dataset with the point cloud set $\mathcal{P}$, the corresponding label set $\mathcal{Y}$ and class prototypes set $\mathcal{T}$, where $\mathcal{T}$ contains the auxiliary D-dimensional semantic embedding vectors (*e.g.* given by Word2Vec [31] or GloVe [36]). Since we follow the challenging inductive *Generalized ZSL* setting instead of *vanilla ZSL*, we train the model using samples containing only $\mathcal{C}^S$ categories, and test on the scenes containing point cloud with classes both in $\mathcal{C}^S$ and $\mathcal{C}^U$. Thus, the training set $\mathcal{D}_{train}$ and test set $\mathcal{D}_{test}$ can be denoted as $\mathcal{D}_{train} = \left\{(p, y, t) \mid \forall i,\ y_i \in \mathcal{C}^S\right\}$ and $\mathcal{D}_{test} = \left\{(p, y, t) \mid \forall i,\ y_i \in \mathcal{C}^S \cup \mathcal{C}^U\right\}$, where $p \in \mathcal{P}$ has N points, $y \in \mathcal{Y}$, $t \in \mathcal{T}$ and y_i is the ground-truth label for point i.

For our approach, we define the generator as $G(\cdot)$, the feature embedding network as $\theta(\cdot)$, and the segmentor as $f(\cdot)$. As illustrated in Fig. 2, the overall training pipeline can be summarized as follow: **a)** Train a feature embedding network θ and a seen-class segmentor f_{seen} using only the seen class data; **b)** Train a generator $G(\cdot)$ on seen data using the auxiliary semantic vectors, so that the synthetic visual features generated by G are as similar as possible to the real point features extracted by frozen θ; **c)** Combine the synthetic unseen features of classes $\mathcal{C}^U$ generated by G together with the real extracted features on seen classes $\mathcal{C}^S$ to train the final segmentor f_{final}. The ultimate goal is that the resulting composite network θ with f_{final} can effectively segment point clouds for both seen and unseen classes. The challenge however is weak transfer from semantic to visual space. To promote generalization to unseen categories, we improve semantic-visual transfer by enhancing correspondence (Sec. 3.2), alignment (Sec. 3.3) and consistency (Sec. 3.4) between semantic and visual spaces, aiming to assist generator training and improve the synthesized features quality.

3.2. Mask Correspondence Learning

Given the auxiliary semantic embeddings t_s^c and random noise z_s^c of seen class c as input into $G(\cdot)$, we synthesize features $\hat{\mathbf{F}}_s^c$, such that they closely follow the distribution of the real features $\mathbf{F}_s^c = \theta(p_s^c)$ extracted from model θ. Unlike 2D counterparts, we lack large-scale 3D pre-trained backbones to train a generator that can synthesize diverse point-wise features. To promote transfer to unseen classes, we propose to establish a strong correspondence between the input semantic and output visual spaces. Such a correspondence should enhance intra-class richness and ensure features belonging to different classes are well separated. Therefore, while training generator $G(\cdot)$ on the seen classes, we ensure that the generated features have within-class diversity, and clear decision boundaries exist between different classes. Further, the class-wise features should be unique and follow the corresponding semantics. With these objectives in mind, we develop a masking strategy that learns by recovering the masked context.

As shown in Fig. 2, for the point cloud p_s^c containing seen class $c \in \mathcal{C}^S$, we randomly mask out part of the corresponding input auxiliary semantic embeddings t_s^c, then the

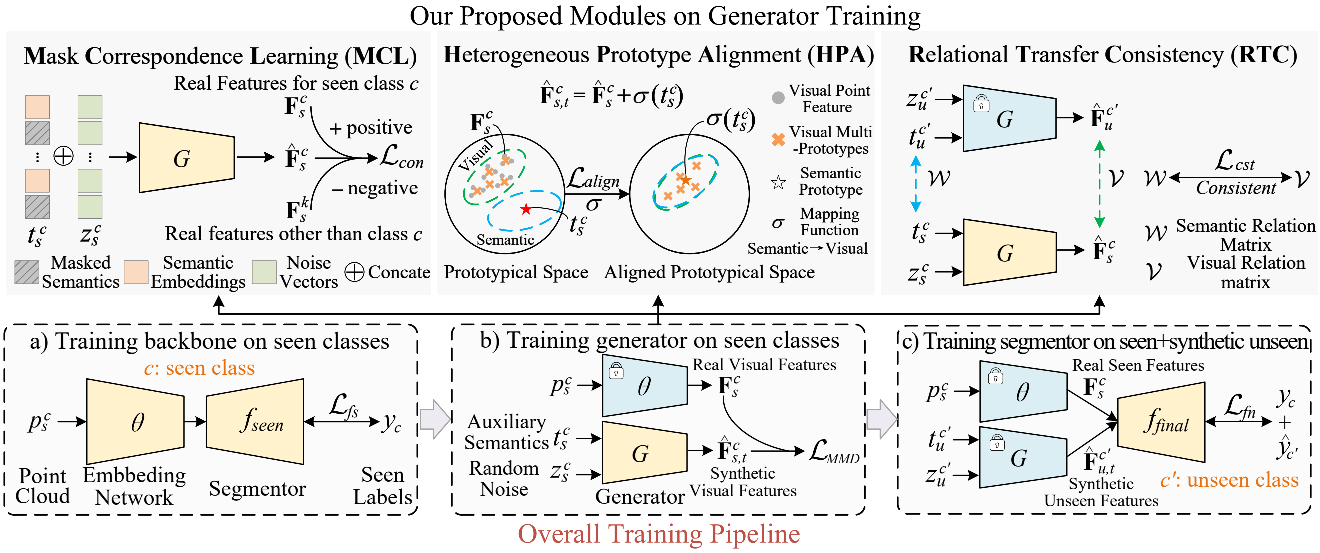

Figure 2. Schematics of our proposed framework for 3D generalized zero-shot semantic segmentation. The modules in blue are frozen while yellow are learnable. We develop 3 modules to tackle several semantic-visual transfer issues. In MCL module (Sec. 3.2), we recover the visual features corresponding to the masked semantics and develop contrastive learning to achieve the intra-class diversity and inter-class separation in visual features. In HPA module (Sec. 3.3), we align semantic and visual features in their prototypical space. In RTC module (Sec. 3.4), we ensure that the distance relations between seen and unseen classes are consistent in both visual and semantic spaces.

generated features by $G(\cdot)$ can be represented as:

$$\hat{\mathbf{F}}_s^c = G(\mathcal{H}(q)t_s^c \oplus z_s^c),\ c \in \mathcal{C}^S \tag{1}$$

where $\oplus$ indicates the concatenation operation, $\mathcal{H}(q)$ is initialized to 1 and masked with 0 with probability q. The size of t_s^c and z_s^c is set to match the number of points in class c of the current scene. During training, the generator $G(\cdot)$ recovers the visual features conditioned on the randomly masked semantics, which helps it learn the intra-class diversity. Semantic-conditioned visual synthesis is essentially one-to-many mapping, and masking the semantics introduces diversity in the semantic space, and thus promotes diversity and richness in the corresponding visual space. Besides, to promote discrimination between visual features of different classes, we consider the real feature $\mathbf{F}_s$ extracted by frozen θ of seen class c as the positive samples $\mathbf{F}_s^c$, and the features of other seen classes k in current p_s as the negative samples $\mathbf{F}_s^k$, and apply InfoNCE [33] loss:

$$\mathcal{L}_{con} = -log\frac{exp(\hat{\mathbf{F}}_s^c \cdot \mathbf{F}_s^c/\tau)}{\sum\limits_{k\in\mathcal{C}^S,\ k\neq c} exp(\hat{\mathbf{F}}_s^c \cdot \mathbf{F}_s^k/\tau) + exp(\hat{\mathbf{F}}_s^c \cdot \mathbf{F}_s^c/\tau)}, \tag{2}$$

where $\mathbf{F}_s^c = \theta(p_s^c)$, $\mathbf{F}_s^k = \theta(p_s^k)$, $p_s = p_s^c \cup p_s^k$, p_s represents the seen class point clouds and τ is the temperature parameter. Contrastive learning enhances the discrimination between different categories. Our proposed semantics masking and visual contrast learning strategies therefore ensure that the learned visual space is rich and discriminative.

3.3. Heterogeneous Prototype Alignment

The semantic embeddings and visual features are from different modalities, and directly using the semantics for visual synthesis, without any alignment, is sub-optimal. We therefore propose to align the cross-modality heterogeneous features before synthesis. Inspired by the prototypical learning [42], we cast the original features into the prototypical space to model their distribution for alignment. Since the semantic embedding vectors t_s^c corresponding to a seen class c can naturally be regarded as a prototype, we only need visual prototypes on the seen features $\mathbf{F}_s^c$.

To generate visual prototypes, instead of the simple average for point cloud visual features, we develop a neighbor-aware approach that reflects the intra-class fine-grained local structure. Specifically, we adopt the Farthest Point Sampling (FPS) algorithm to sample r-proportion $(0 < r < 1)$ point features $\{\mathbf{F}_s^{c,a}\}_{a=1}^{\lfloor n*r \rfloor}$ as anchors on the real seen features $\{\mathbf{F}_s^{c,i}\}_{i=1}^{n}$ embedded by θ, $n \leq N$ is the number of points for class c, $\lfloor\cdot\rfloor$ denotes the rounding operation. We calculate the ℓ_2 distance between n point features and $\lfloor n*r \rfloor$ anchors and assign the nearest anchor index to each point. We average the point features of the same anchor index to form $\lfloor n*r \rfloor$ (≥ 1) visual prototypes $\{\mathbf{H}_s^{c,b}\}_{b=1}^{\lfloor n*r \rfloor}$ as:

$$\mathbf{H}_s^{c,b} = \frac{1}{|\xi_s^{c,b}|}\sum_{\mathbf{F}_s^{c,i}\in\xi_s^{c,b}} \mathbf{F}_s^{c,i}, \tag{3}$$

where $\xi_s^{c,b}$ is the partition region composed of the point features assigned to anchor b. After getting the semantic and visual prototypes, we align them to enhance visual synthesis quality. We apply the linear $\sigma(\cdot)$ function to map the semantic embedding t_s^c to the same dimension as the visual prototypes $\mathbf{H}_s^{c,b}$, and minimize cosine distance $d(\cdot,\cdot)$:

$$\mathcal{L}_{align} = \frac{1}{\lfloor n*r \rfloor} \sum_{b=1}^{\lfloor n*r \rfloor} d\left(\mathbf{H}_s^{c,b}, \sigma(t_s^c)\right), \tag{4}$$

where $\sigma\colon \mathbb{R}^{D_1} \to \mathbb{R}^{D_2}$, D_1 and D_2 are the feature dimensions of semantic embeddings and visual prototypes respectively. We further add the aligned semantic vector $\sigma(t_s^c)$ with the synthesized feature $\hat{\mathbf{F}}_s^c$ to enhance representations $\hat{\mathbf{F}}_s^c + \sigma(t_s^c)$ for generator $G(\cdot)$ training. Besides, alignment on seen data helps to obtain a well-learned semantic-visual mapping which helps better synthesizes of unseen features $\hat{\mathbf{F}}_u^{c'} + \sigma(t_u^{c'})$, $c' \in \mathcal{C}^U$ for f_{final} segmentor.

3.4. Relational Transfer Consistency

Since the model is only optimized on the seen class data, and never encounters unseen data (as it is not available), the model becomes biased and confuses unseen classes as seen. To counter this, inspired by [27], we propose semantic-visual consistency regularization. We argue that even though the seen and unseen might have different semantic and visual structures, the inter-class relationships in their respective spaces should be preserved. Specifically, we employ the generator G to synthesize visual features for a specific unseen class c', denoted as $\hat{\mathbf{F}}_u^{c'}$, and its corresponding semantic prototype is $t_u^{c'}$. Similarly, for a seen class c, its synthetic visual features and semantic prototype can be represented as $\hat{\mathbf{F}}_s^c$ and t_s^c respectively. We construct unseen c' visual synthetic prototype $\hat{\mathbf{H}}_u^{c'}$ and seen c visual synthetic prototype $\hat{\mathbf{H}}_s^c$ as:

$$\hat{\mathbf{H}}_u^{c'},\ \hat{\mathbf{H}}_s^c = \frac{1}{|n'|}\sum_{i=1}^{n'} \hat{\mathbf{F}}_u^{c',i},\ \frac{1}{|n|}\sum_{i=1}^{n} \hat{\mathbf{F}}_s^{c,i}, \tag{5}$$

where n' and n denote the number of points belonging to unseen c' and seen classes c, respectively. we apply simple averaging to obtain the visual prototype, since the generated features lack fine-grained structure relative to the real features. We build sets $\{t_u^{c'}\}_{c' \in \mathcal{C}^U}$ and $\{\hat{\mathbf{H}}_u^{c'}\}_{c' \in \mathcal{C}^U}$ for semantic and visual prototypes in unseen classes. We further get the distance distribution relation matrices for semantic $\mathcal{W} \in \mathbb{R}^{m \times m}$ and visual $\mathcal{V} \in \mathbb{R}^{m \times m}$ between the prototypes of seen and unseen sets respectively,

$$\mathcal{W}_{ej} = ||t^e - t^j||^2,\ \mathcal{V}_{ej} = ||\hat{\mathbf{H}}^e - \hat{\mathbf{H}}^j||^2, \tag{6}$$

where m is the total number of elements in set $t^m = t_s^c \cup \{t_u^{c'}\}_{c' \in \mathcal{C}^U}$ or $\hat{\mathbf{H}}^m = \hat{\mathbf{H}}_s^c \cup \{\hat{\mathbf{H}}_u^{c'}\}_{c' \in \mathcal{C}^U}$. $e \leq m$ and $j \leq m$ denote the index of an element in the set t^m and $\hat{\mathbf{H}}^m$. We strive to keep the distance distribution in the two spaces consistent by minimizing the cosine distance $d(\cdot,\cdot)$ as:

$$\mathcal{L}_{cst} = \sum_{e=1}^{m} d(\mathcal{W}_{ej}, \mathcal{V}_{ej}). \tag{7}$$

Thus, we establish a consistency bridge between the visual and semantic space of seen and unseen classes, so the model can effectively tackle the bias towards the seen.

3.5. Network Training and Inference

For the backbone $\theta(\cdot)$ and $f_{seen}(\cdot)$ training, we apply the cross-entropy loss between network output and labels y_c on only seen point clouds $p_s^c, c \in \mathcal{C}^S$,

$$\mathcal{L}_{fs} = -\sum_c y_c log(f_{seen}(\theta(p_s^c))). \tag{8}$$

To train the generator G, we apply the *Maximum Mean Discrepancy* (MMD) loss [28] to narrow the distribution mismatch between the synthesised $\hat{\mathbf{F}}_s^c + \sigma(t_s^c)$ as $\hat{\mathbf{F}}_{s,t}^c$ and the real features $\mathbf{F}_s^c$ on seen c, and combine $\mathcal{L}_{con}$, $\mathcal{L}_{align}$, $\mathcal{L}_{cst}$ losses to form the joint loss:

$$\begin{aligned}\mathcal{L}_{MMD} = \sum_{x,x' \in \mathbf{F}_s^c} \mu(x,x') + \sum_{\hat{x},\hat{x}' \in \hat{\mathbf{F}}_{s,t}^c} \mu(\hat{x},\hat{x}') \\ -2\sum_{x \in \mathbf{F}_s^c} \sum_{\hat{x} \in \hat{\mathbf{F}}_{s,t}^c} \mu(x,\hat{x}),\end{aligned} \tag{9}$$

$$\mathcal{L}_G = \sum_c \left(\mathcal{L}_{MMD} + \mathcal{L}_{con} + \mathcal{L}_{align} + \alpha\mathcal{L}_{cst}\right), \tag{10}$$

where $\mu(\cdot,\cdot)$ is the Gaussian kernel function, $\mu(x,x') = exp(-\frac{1}{2}||x - x'||^2)$, α is a hyper-parameter for loss balance. It should be noted that we do not use discriminator to make the features more realistic, which is demonstrated in [30] that it may be harmful for 3D point clouds. The well-trained generator G synthesizes unseen features $\hat{\mathbf{F}}_u^{c'} + \sigma(t_u^{c'})$ as $\hat{\mathbf{F}}_{u,t}^{c'}$, $c' \in \mathcal{C}^U$ which will combine with the real seen features $\mathbf{F}_s^c$ on c to train the final segmentor f_{final} using,

$$\mathcal{L}_{fn} = -\sum_c y_c log(f_{final}(\mathbf{F}_s^c)) - \sum_{c'} \hat{y}_{c'} log(f_{final}(\hat{\mathbf{F}}_{u,t}^{c'})), \tag{11}$$

where $\hat{y}_{c'}$ denotes the synthetic unseen labels. At inference time, we combine the θ and f_{final} to jointly predict both seen $\mathcal{C}^S$ and unseen $\mathcal{C}^U$ categories.

4. Experiments

4.1. Datasets and Settings

Datasets. We follow [30] to conduct experiments based on three public 3D semantic segmentation datasets ScanNet

Table 1. Generalized 3D zero-shot semantic segmentation results on three benchmarks. All methods use GloVe+Word2Vec embbeddings. The evaluation metric are mIoU and HmIoU (%). $\hat{\mathcal{C}}^U$ stands for pseudo generated unseen data. The results of all comparison methods are derived from [30]. Our approach shows impressive gains of 7.7%, 3.8%, 3.0% based on HmIoU in ScanNet, S3DIS and SemanticKITII datasets respectively.

	Training set		ScanNet				S3DIS				SemanticKITTI			
	Backbone	segmentor	mIoU $\mathcal{C}^S$	mIoU $\mathcal{C}^U$	mIoU All	HmIoU	mIoU $\mathcal{C}^S$	mIoU $\mathcal{C}^U$	mIoU All	HmIoU	mIoU $\mathcal{C}^S$	mIoU $\mathcal{C}^U$	mIoU All	HmIoU
Supervised methods with different levels of supervision														
Full supervision	$\mathcal{C}^S \cup \mathcal{C}^U$	$\mathcal{C}^S \cup \mathcal{C}^U$	43.3	51.9	45.1	47.2	74.0	50.0	66.6	59.6	59.4	50.3	57.5	54.5
ZSL backbone	$\mathcal{C}^S$	$\mathcal{C}^S \cup \mathcal{C}^U$	41.5	39.2	40.3	40.3	60.9	21.5	48.7	31.8	52.9	13.2	42.3	21.2
ZSL-trivial	$\mathcal{C}^S$	$\mathcal{C}^S$	39.2	0.0	31.3	0.0	70.2	0.0	48.6	0.0	55.8	0.0	44.0	0.0
Generalized zero-shot-learning methods														
ZSLPC-Seg* [13]	$\mathcal{C}^S$	$\mathcal{C}^U$	28.2	0.0	22.6	0.0	65.5	0.0	45.3	0.0	49.1	0.0	34.8	0.0
DeViSe-3DSeg* [17]	$\mathcal{C}^S$	$\mathcal{C}^U$	20.0	0.0	16.0	0.0	70.2	0.0	48.6	0.0	49.7	0.0	36.6	0.0
ZSLPC-Seg [13]	$\mathcal{C}^S$	$\mathcal{C}^U$	16.4	4.2	13.9	6.7	5.2	1.3	4.0	2.1	26.4	10.2	21.8	14.7
DeviSe-3DSeg [17]	$\mathcal{C}^S$	$\mathcal{C}^U$	12.8	3.0	10.9	4.8	3.6	1.4	3.0	2.0	42.9	4.2	27.6	7.5
3DGenZ [30]	$\mathcal{C}^S$	$\mathcal{C}^S \cup \hat{\mathcal{C}}^U$	32.8	7.7	27.8	12.5	53.1	7.3	39.0	12.9	41.4	10.8	35.0	17.1
Ours	$\mathcal{C}^S$	$\mathcal{C}^S \cup \hat{\mathcal{C}}^U$	**34.5**	**14.3**	**30.4**	**20.2**	**58.9**	**9.7**	**43.8**	**16.7**	**46.4**	**12.8**	**39.4**	**20.1**

[14], S3DIS [2] and SemanticKITTI [3]. (a) ScanNet is an RGB-D video dataset having 1201 training scans, 312 validation scans and 100 test scans belonging with points annotated with 20 classes. (b) S3DIS is an indoor scene dataset containing 272 rooms in 6 areas, with each point labeled as one of 13 classes. (c) SemanticKITTI contains 21 sequences of 43,552 annotated streetscape LiDAR scans, the points of each object are annotated in 19 semantic classes. According to the original division, sequences 00∼07 and 09∼10 are used for training, sequence 08 for validation, and sequence 11∼21 for online testing. Since the test set of ScanNet and SemanticKITTI are not available, we chose their validation for ZSL testing. For S3DIS, we select area 1 as the test set and the other areas are used for training [30].

Settings. Following [30], we divide each dataset into two non-overlapping parts, *i.e.,* seen and unseen classes. For each dataset, we consider 4 unseen classes, desk, bookshelf, sofa, toilet for ScanNet; beam, column, window, sofa for S3DIS; motorcycle, truck, bicyclist, traffic-sign for SemanticKITTI. There exist semantic similarities between seen and unseen classes *e.g.* unseen sofa and seen chair in ScanNet and S3DIS. Note that we discard any point clouds in the training set that contain unseen points and their labels. We use the mean Intersection-over-Union (mIoU) as the evaluation metric. In addition, we use Harmonic Mean (HM) to report the combined seen+unseen results.

Implementation details: For a direct comparison with existing research [30], we use the same backbones, *i.e.,* FKAConv [5] for ScanNet, ConvPoint [4] for S3DIS, and KPConv [45] for SemanticKITTI. All backbones are pretrained on the seen data and labels with recommended parameters in respective papers. After pre-training, we freeze the backbone, and extract features on seen classes to train the generator. We chose a generative Moment Matching Network (GMMN) [28] as the generator. Similar to [30], we use 600-dimensional GloVe+Word2Vec as auxiliary semantic vectors. We use the Adam [25] optimizer with initial learning rate of 2e-4, and empirically set the mask probability q to 0.2 and the ratio r of visual prototypes to 0.04 for all three datasets. α is set to 0.4. The backbone features are fed to the final segmentor for fine-tuning, using the initial learning rate of 7e-3 and 7e-2, respectively. A poly learning rate scheduler is applied for final training [52]. We train our zero-shot model for 20 epochs. We follow [30] to apply class-dependent weighting and calibrated stacking to reduce the bias towards seen classes, and construct cross-validation sets with randomly selected 20% or at least 2 seen classes of training data. We also report results on 3 fully supervised models as upper-bound.

4.2. Experimental Results

Quantitative comparison with the state-of-art methods. Tab. 1 compares different approaches in terms of mIoU. We observe that our method achieves consistently superior performance on the three datasets. We outperform the current state-of-the-art method 3DGenZ [30] by a large margin of 7.7%, 3.8% and 3.0% of HmIoU metric on ScanNet, S3DIS and SemanticKITTI, respectively. The results suggest that our proposed modules can effectively migrate seen-to-unseen knowledge and generalize to novel categories. Moreover, it should be noted that our approach also retains performance on seen classes. We believe that our generator synthesizes realistic features that are distinguishable between different classes, while the seen and unseen

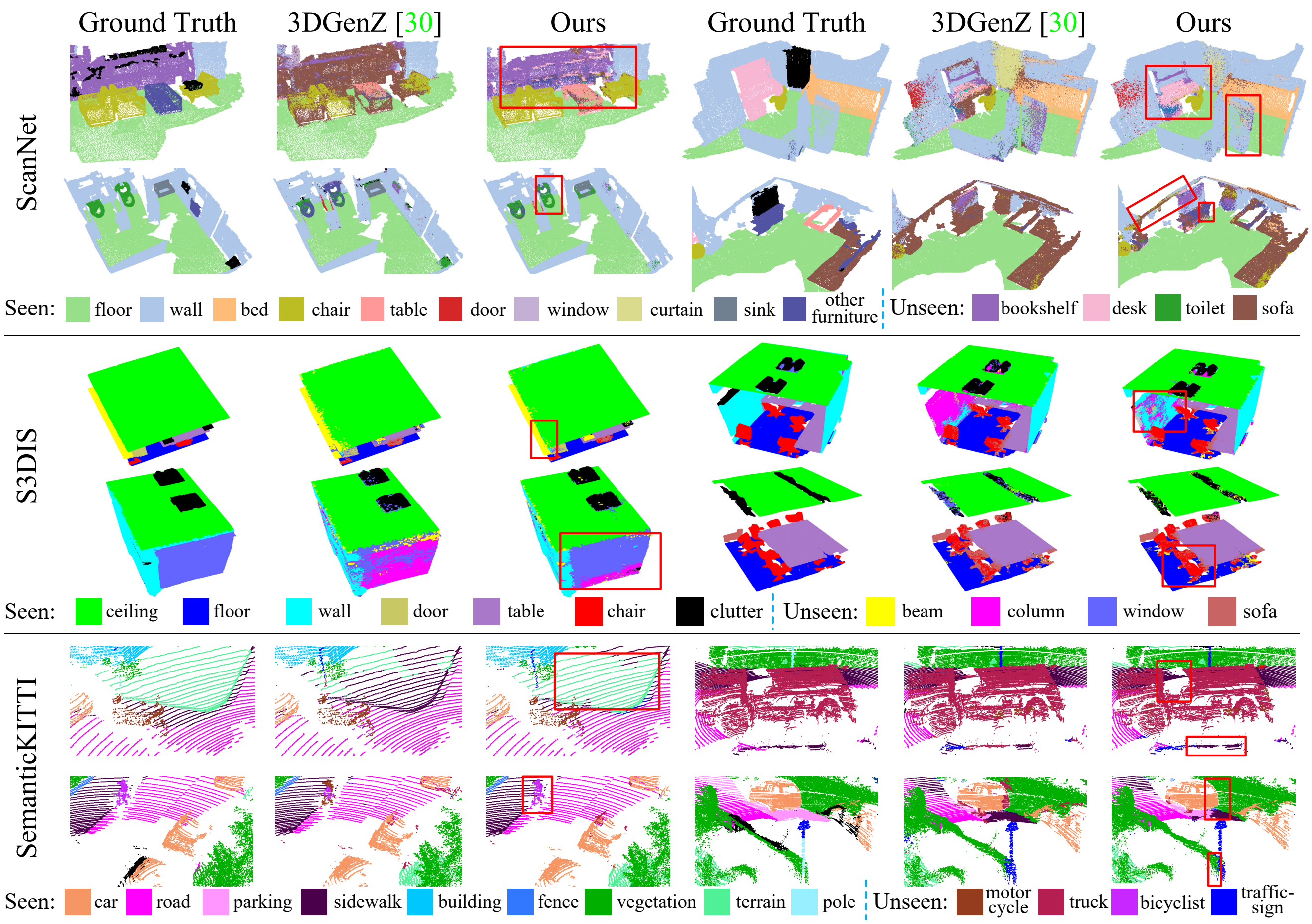

Figure 3. Qualitative comparison with 3DGenZ [30] under inductive generalized zero-shot setting. The results in black on the ScanNet and SemanticKITTI datasets represent unlabeled data. The regions in red boxes highlight the effectiveness of our method.

visual features are well aligned with the corresponding semantic space, thus benefiting the training of the final segmentor. Compared with the performance on SematicKITTI, our method shows a higher improvement than other methods on the ScanNet and S3DIS datasets. The reason for this might be that the categories in the large outdoor scenes involved in SemanticKITTI are more complex, making it challenging to generalize to novel unseen classes. We further report the IoU of the individual seen and unseen categories for various datasets in the supplementary material.

More comparisons with adapted 2D methods: In Tab. 2, we adapt five 2D generalized zero-shot semantic segmentation methods [52, 6, 27, 18, 55] to 3D point cloud. We evaluate these methods in our inductive setting using the same 3D backbone (FKAConv [5]) on ScanNet dataset. Results suggest that existing classical 2D methods are not directly suitable for 3D point cloud data.

Qualitative results. We visualize the results of our method compared with 3DGenZ [30] on three different datasets in Fig. 3. Our method performs better than 3DGenZ on all classes, especially on unseen classes *e.g*. On the ScanNet dataset our method is more successful in segmenting unseen bookshelf, whereas 3DGenZ is confused on unseen sofa. The same phenomenon occurs in window on the S3DIS dataset and bicyclist on the SemanticKITTI dataset. Our method can more effectively help the network to transfer knowledge from the seen to the unseen situation by synthesizing the unseen features being semantic-visual aware.

Table 2. Comparisons with adapted 2D Generalized ZSL methods on ScanNet. Asterisk (*) denotes the methods in our reproduction.

Methods	Publication	mIoU $\mathcal{C}^S$	mIoU $\mathcal{C}^U$	mIoU All	HmIoU
SPNet [52]*	CVPR 2019	16.2	1.6	13.3	2.9
ZS3Net [6]*	NeurIPS 2019	33.6	4.1	27.7	7.3
CSRL [27]*	NeurIPS 2020	34.2	4.6	28.3	8.2
GaGNet [18]*	ACM MM 2020	33.8	5.2	28.1	8.9
PMOSR [55]*	ICCV 2021	32.5	5.4	27.1	9.3
Ours	ICCV 2023	**34.5**	**14.3**	**30.4**	**20.2**

4.3. Ablation study

Ablation study of different modules. We progressively integrate different modules to study their contribution

Table 3. Ablation study of MCL (Sec. 3.2), HPA (Sec. 3.3) and RTC (Sec. 3.4) modules on ScanNet dataset. We observe that all the three proposed modules contribute to the performance.

MCL	HPA	RTC	mIoU $\mathcal{C}^S$	mIoU $\mathcal{C}^U$	mIoU All	HmIoU
×	×	×	34.2	7.7	29.1	12.5
×	✓	×	33.5	9.7	28.8	15.0
×	×	✓	33.5	10.0	28.8	15.4
×	✓	✓	34.5	10.6	29.8	16.3
✓	×	×	34.0	13.0	29.8	18.8
✓	✓	×	33.9	13.9	29.9	19.7
✓	×	✓	33.7	14.0	29.7	19.8
✓	✓	✓	**34.5**	**14.3**	**30.4**	**20.2**

in Tab. 3. We can notice that MCL, HPA and RTC modules show complementary gains. The most pronounced gain comes from the MCL module, suggesting that the separable and diverse representations generated by G are more conducive for the final segmentor training. We notice that while all three modules show individual gains, their combination achieves the best results, improving the baseline by 7.7% in terms of HmIoU, suggesting that these modules complement each other for enhanced generalization to unseen categories. From these empirical evaluations, we can conclude that the modules proposed in our method can effectively generalize to the recognition of unseen classes on the basis of seen knowledge.

Dissecting MCL module. The MCL module has two components: masking and contrastive learning. We study the impact of these two components in Tab. 4. We observe that only using contrastive learning improves the baseline's HmIoU by 4.3, but its mIoU decreases on seen classes. The combination of both masking and contrastive learning retains performance on seen categories, while simultaneously promoting generalization to unseen classes.

Table 4. Contributions of Contrastive Learning (CL) and Masking Strategy (MS) in MCL module (Sec. 3.2).

Methods	mIoU $\mathcal{C}^S$	mIoU $\mathcal{C}^U$	mIoU All	HmIoU
Baseline	34.2	7.7	29.1	12.5
+ CL	31.7	11.4	27.7	16.8
+ CL + MS	**34.0**	**13.0**	**29.8**	**18.8**

Prototypes in HPA and RTC module. Tab. 5 compares three different prototype construction strategies in Sec. 3.3 and Sec. 3.4. The HPA module equipped with our proposed neighbor-aware approach achieves the best result, since it models the rich local point cloud structure that is well-aligned with the corresponding semantics. Unlike HPA, RTC module using simple averaging is better than other strategies, since averaging helps remove any noise in the synthesized visual features by smoothing.

Table 5. Effects of different prototype generation strategies. For HPA, the neighbor-aware prototype generation works best, while for simple averaging performs better for RTC.

Section	Prototype Construction Methods	mIoU $\mathcal{C}^S$	mIoU $\mathcal{C}^U$	mIoU All	HmIoU
Sec. 3.3 HPA	Simple Averaging	34.1	13.8	30.0	19.6
	K-Means Clustering	32.7	13.8	28.9	19.4
	Neighbor-Aware	**34.5**	**14.3**	**30.4**	**20.2**
Sec. 3.4 RTC	Simple Averaging	**34.5**	**14.3**	**30.4**	**20.2**
	K-Means Clustering	32.3	13.3	28.5	18.9
	Neighbor-Aware	33.4	12.1	29.2	17.7

Hyper-parameters. Fig. 4 shows the impact of two critical hyper-parameters (*i.e.,* mask probability q and visual prototypes ratio r). We observe that with a gradual increase in q, the models performance improves, indicating that the model is able to complete the visual features of the missing semantic embeddings according to the contextual information, so as to obtain better representations. However, the higher mask probability will result in a lack of sufficient semantics to assist generator training, resulting in performance degradation. The highest performance is achieved when the mask probability is 0.2 for all three datasets. In addition, we observe a low performance when the ratio is small for the visual prototypes. It is due to the insufficient prototypical representations in the visual space, which leads to the deviation in alignment with semantic vectors. Beyond $r > 0.04$, the performance starts to to decline, probably caused by over-fitting that leads to adverse impact.

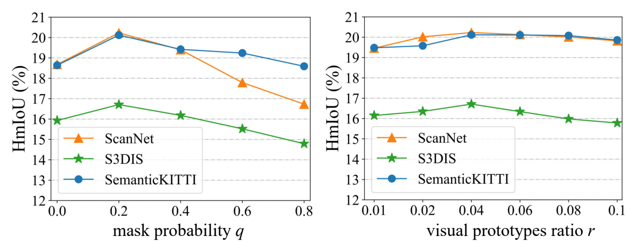

Figure 4. Effect of hyper-parameters: mask probability q and visual prototypes ratio r on three datasets, $q = 0.2$ and $r = 0.04$ show the best results across all the datasets.

Effects of different auxiliary semantic embeddings. Tab. 6 compares different choices of auxiliary semantic embeddings (Word2Vec, GloVe and GloVe+Word2Vec for 300, 300, 600-dimensional semantic embeddings respectively). In general, a higher dimensional semantic embeddings produce richer feature representations, but we observe that more dimensional embeddings on different datasets may not always lead to the best results. Using the GloVe embeddings produces better performance in unseen mIoU and HM for ScanNet datasets, but not the other two, and similar phenomenon appears in the Word2Vec embeddings for SemanticKITTI. Only for S3DIS dataset, we achieve the best gain using a combination of Word2Vec and GloVe. We argue that the higher dimensional embedding space may bring more complexity and information redundancy for our model, especially for ScanNet and SemanticKITTI, which contains relatively more classes than S3DIS.

Table 6. Effects of different auxiliary semantic embeddings on various datasets. "SN", "S3", "SK" represent ScanNet, S3DIS and SemanticKITTI dataset, respectively. HM denotes harmonic mean (%).

	Word2Vec				GloVe				GloVe + Word2Vec			
	mIoU			HM	mIoU			HM	mIoU			HM
	$\mathcal{C}^S$	$\mathcal{C}^U$	All		$\mathcal{C}^S$	$\mathcal{C}^U$	All		$\mathcal{C}^S$	$\mathcal{C}^U$	All	
SN	33.3	11.7	29.0	17.3	33.1	**14.9**	29.5	**20.6**	**34.5**	14.3	**30.4**	20.2
S3	58.9	9.1	43.6	15.8	58.7	6.2	42.5	11.2	**58.9**	**9.7**	**43.8**	**16.7**
SK	45.8	**14.4**	39.2	**21.9**	46.0	6.1	37.6	10.8	**46.4**	12.8	**39.4**	20.1

5. Conclusion

In this paper, we propose a feature synthesis-based approach for Generalized Zero-Shot Semantic Segmentation of 3D point clouds. Our goal is to enhance semantic-visual correspondence, alignment and consistency, to learn generic representations that can transfer across novel unseen classes. Through our developed strategies, we promote the intra-class diversity in the visual features, while enhancing separation between classes. We further align the visual features and their semantics in the prototypical space, and preserve semantic-visual relationships through consistency regularization. Our empirical evaluations suggest that the proposed method can effectively segment point clouds for both seen and unseen classes at inference time, and achieve significant gains over the current state-of-the-art. For future work, we plan to extend our current approach for open-vocabulary zero-shot point cloud semantic segmentation.

Acknowledgement: This work was supported by the National Natural Science Foundation of China (No.62276176).

References

[1] Ziad Al-Halah, Makarand Tapaswi, and Rainer Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5975–5984, 2016. 1, 2

[2] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1534–1543, 2016. 2, 6

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9297–9307, 2019. 2, 6

[4] Alexandre Boulch. Convpoint: Continuous convolutions for point cloud processing. *Computers & Graphics*, 88:24–34, 2020. 3, 6

[5] Alexandre Boulch, Gilles Puy, and Renaud Marlet. Fkaconv: Feature-kernel alignment for point cloud convolution. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 6, 7

[6] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 1, 3, 7

[7] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1043–1052, 2018. 1, 2

[8] Runnan Chen, Xinge Zhu, Nenglun chen, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. Zero-shot point cloud segmentation by transferring geometric primitives. *arXiv preprint arXiv:2210.09923*, 2022. 2, 3

[9] Jiaxin Cheng, Soumyaroop Nandi, Prem Natarajan, and Wael Abd-Almageed. Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9556–9566, 2021. 1, 3

[10] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Mitigating the hubness problem for zero-shot learning of 3d objects. *arXiv preprint arXiv:1907.06371*, 2019. 2, 3

[11] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Transductive zero-shot learning for 3d point cloud classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 923–933, 2020. 2, 3

[12] Ali Cheraghian, Shafin Rahman, Townim F Chowdhury, Dylan Campbell, and Lars Petersson. Zero-shot learning on 3d point cloud objects and beyond. *International Journal of Computer Vision (IJCV)*, 130(10):2364–2384, 2022. 2, 3

[13] Ali Cheraghian, Shafin Rahman, and Lars Petersson. Zero-shot learning of 3d point cloud objects. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE, 2019. 2, 3, 6

[14] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 2, 6

[15] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785. IEEE, 2009. 1, 2

[16] Rafael Felix, Ian Reid, Gustavo Carneiro, et al. Multimodal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2018. 1, 2

[17] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013. 1, 2, 6

[18] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pages 1921–1929, 2020. 1, 3, 7

[19] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Zero-shot learning with transferred samples. *IEEE Transactions on Image Processing (TIP)*, 26(7):3277–3290, 2017. 1, 2

[20] Ping Hu, Stan Sclaroff, and Kate Saenko. Uncertainty-aware learning for zero-shot semantic segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21713–21724, 2020. 1, 3

[21] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11108–11117, 2020. 1, 3

[22] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014. 2

[23] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10984–10994, 2023. 2

[24] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11487–11496, 2019. 1, 2

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 951–958. IEEE, 2009. 2

[27] Peike Li, Yunchao Wei, and Yi Yang. Consistent structural relation learning for zero-shot segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:10317–10327, 2020. 1, 3, 5, 7

[28] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning (ICML)*, pages 1718–1727. PMLR, 2015. 3, 5, 6

[29] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928. IEEE, 2015. 3

[30] Björn Michele, Alexandre Boulch, Gilles Puy, Maxime Bucher, and Renaud Marlet. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *2021 International Conference on 3D Vision (3DV)*, pages 992–1002. IEEE, 2021. 2, 3, 5, 6, 7

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013. 2, 3

[32] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning (ICML)*, pages 2642–2651. PMLR, 2017. 2

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[34] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 2

[35] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7108–7117, 2021. 3

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 2, 3

[37] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 1, 3

[38] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5648–5656, 2016. 3

[39] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5099–5108, 2017. 1, 3

[40] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531, 2010. 3

[41] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8247–8255, 2019. 1, 2

[42] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 4

[43] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013. 2

[44] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 945–953, 2015. 3

[45] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6411–6420, 2019. 1, 3, 6

[46] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4281–4289, 2018. 2

[47] Maunil R Vyas, Hemanth Venkateswara, and Sethuraman Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–86. Springer, 2020. 1, 2

[48] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6857–6866, 2018. 1, 2

[49] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9621–9630, 2019. 1, 3

[50] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8456–8465, 2018. 2

[51] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 69–77, 2016. 2

[52] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8256–8265, 2019. 1, 3, 6, 7

[53] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5542–5551, 2018. 1, 2

[54] Yuwei Yang, Munawar Hayat, Zhao Jin, Chao Ren, and Yinjie Lei. Geometry and uncertainty-aware 3d point cloud class-incremental semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21759–21768, 2023. 3

[55] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6974–6983, 2021. 7

[56] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2021–2030, 2017. 1, 2

[57] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021. 1, 3

[58] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11175–11185, 2023. 1