Equivariant Maps for Hierarchical Structures

Renhao Wang, Marjan Albooyeh∗
University of British Columbia,
{renhaow,albooyeh}@cs.ubc.ca

Siamak Ravanbakhsh
McGill University & Mila
siamak@cs.mcgill.ca

Abstract

While using invariant and equivariant maps, it is possible to apply deep learning to a range of primitive data structures, a formalism for dealing with hierarchy is lacking. This is a significant issue because many practical structures are hierarchies of simple building blocks; some examples include sequences of sets, graphs of graphs, or multisresolution images. Observing that the symmetry of a hierarchical structure is the “wreath product” of symmetries of the building blocks, we express the equivariant map for the hierarchy using an intuitive combination of the equivariant linear layers of the building blocks. More generally, we show that any equivariant map for the hierarchy has this form. To demonstrate the effectiveness of this approach to model design, we consider its application in the semantic segmentation of point-cloud data. By voxelizing the point cloud, we impose a hierarchy of translation and permutation symmetries on the data and report state-of-the-art on SEMANTIC3D, S3DIS, and VKITTI, that include some of the largest real-world point-cloud benchmarks.

1 Introduction

In designing deep models for structured data, equivariance (invariance) of the model to transformation groups has proven to be a powerful inductive bias, which enables sample efficient learning. A widely used family of equivariant deep models constrain the feed-forward layer so that specific transformations of the input lead to the corresponding transformations of the output. A canonical example is the convolution layer, in which the constrained MLP is equivariant to translation operations. Many recent works have extended this idea to design equivariant networks for more exotic structures such as sets, exchangeable tensors and graphs, as well as relational and geometric structures.

This paper considers a nested hierarchy of such structures, or more generally, any hierarchical composition of transformation symmetries. These hierarchies naturally appear in many settings: for example, the interaction between nodes in a social graph may be a sequence or a set of events. Or in diffusion tensor imaging of the brain, each subject may be modeled as a set of sequences, where each sequence is a fibre bundle in the brain. The application we consider in this paper models point clouds as 3D images, where each voxel is a set of points with coordinates relative to the center of that voxel.

To get an intuition for a hierarchy of symmetry transformations, consider the example of a sequence of sequences – e.g., a text document can be viewed as a sequence of sentences, where each sentence is itself a sequence of words. Here, each inner sequence as well as the outer sequence is assumed to possess an “independent” translation symmetry. Contrast this with symmetries of an image (2D translation), where all inner sequences (say row pixels) translate together, so we have a total of two translations. This is the key difference between the wreath product of two translation groups (former) and their direct product (latter). It is the wreath product that often appears in nested structures. As is evident from this example, the wreath product results in a significantly larger set of transformations, and as we elaborate later, it provides a stronger inductive bias.

∗Currently at Borealis AI. Work done while at UBC.
Wreath product can express the symmetries of hierarchical structures: wreath product of three groups $U \rtimes H \rtimes K$ acting on the set of elements $P \times Q \times R$, can be seen as independent copies of groups, $H$, $K$ and $U$ at different level of hierarchy acting on copies of $P$, $Q$ and $R$. Intuitively, a linear map $W_U \rtimes H \rtimes K : \mathbb{R}^{PQR} \rightarrow \mathbb{R}^{PQR}$ equivariant to $U \rtimes H \rtimes K$, performs pooling over leaves under each inner node $(i.e., W_U, W_K$ and $W_H$ respectively), and broadcasting the output back to the leaves. A $U \rtimes H \rtimes K$-equivariant map can be constructed as the sum of these three contributions, from equivariant maps at each level of hierarchy.

We are interested in application of equivariant/invariant deep learning to this type of nested structure. The building blocks of equivariant and invariant MLPs are equivariant linear maps of the feedforward layer. We show that any equivariant linear map for the hierarchical structure is built using equivariant maps for the individual symmetry group at different levels of the hierarchy. Our construction only uses additional pooling and broadcasting operations; see Fig. 1.

In the following, after discussing related works in Section 2, we give a short background on equivariant MLPs in Section 3. Section 4 starts by giving the closed form of equivariant maps for direct product of groups before moving to the more difficult case of wreath product in Section 4.2. Finally, Section 5 applies this idea to impose a hierarchical structure on 3D point clouds. We show that the equivariant map for this hierarchical structure achieves state-of-the-art performance on the largest benchmark datasets for 3D semantic segmentation.

2 Related Works

Group theory has a long history in signal processing [22], where in particular Fourier transformation and group convolution for Abelian groups have found tremendous success over the past decades. However, among non-commutative groups, wreath product constructions have been the subject of few works. Rockmore [39] give efficient procedures for fast Fourier transforms for wreath products. In a series of related works Foote et al. [15], Mirchandani et al. [34] investigate the wreath product for multi-resolution signal processing. The focus of their work is on the wreath product of cyclic groups for compression and filtering of image data.

Group theory has also found many applications in machine learning [25], and in particular deep learning. Design of invariant MLPs for general groups goes back to Shawe-Taylor [41]. More recently, several works investigate the design of equivariant networks for general finite [38] and infinite groups [7, 10, 26]. In particular, use of the wreath product for design of networks equivariant to a hierarchy of symmetries is briefly discussed in [38]. Equivariant networks have found many applications in learning on various structures, from image and text [51], to sets [25, 50], exchangeable matrices [21], graphs [1, 28, 32], and relational data [18], to signals on spheres [9, 27]. A large body of works investigate equivariance to Euclidean isometries; e.g., [11, 44, 46].

When it comes to equivariant models for compositional structures, contributions have been sparse, with most theoretical work focusing on semidirect product (or more generally using induced representations from a subgroup) to model tensor-fields on homogeneous spaces [8, 10]. Direct product of symmetric groups have been used to model interactions across sets of entities [21] and in generalizations to relational data [18].

Relation to Maron et al. [33] Maron et al. [33] extensively study equivariant networks for direct product of groups; in contrast we focus on wreath product. Since both of these symmetry transformations operate on the Cartesian product of $G$-sets (see Fig. 1), the corresponding permutation groups are comparable. Indeed, direct product action is a sub-group of the imprimitive action of the
wreath product. The implication is that if sub-structures in the hierarchy transform independently (e.g., in sequence of sequences), then using the equivariant maps proposed in this paper gives a strictly stronger inductive bias. However, if sub-structures move together (e.g., in an image), then using a model equivariant to direct product of groups is preferable. One could also compare the learning bias of these models by noting that when using wreath product, the number of independent linear operators grows with the “sum” of independent operators for individual building blocks; in contrast, this number grows with the “product” of independent operators on blocks when using direct product. In the following we specifically contrast the use of direct product with wreath product for compositional structures.

3 Preliminaries

3.1 Group Action

A group \( G = \{g\} \) is a set equipped with a binary operation, such that the set is closed under this operation, \( gh \in G \), the operation is associative, \( g(\hat{h}h) = (g\hat{h})h \), there exists identity element \( e \in G \), and a unique inverse for each \( g \in G \) satisfying \( gg^{-1} = e \). The action of \( G \) on a finite set \( \mathbb{N} \) is a function \( \alpha : G \times \mathbb{N} \rightarrow \mathbb{N} \) that transforms the elements of \( \mathbb{N} \), for each choice of \( g \in G \); for short we write \( g \cdot n \) instead of \( \alpha(g, n) \). Group actions preserve the group structure, meaning that the transformation associated with the identity element is identity \( e \cdot n = n \), and composition of two actions is equal to the action of the composition of group elements \( (gh) \cdot n = g \cdot (\hat{h} \cdot n) \). Such a set, with a \( G \)-action defined on it is called a \( G \)-set. The group action on \( \mathbb{N} \) naturally extends to \( x \in \mathbb{R}^{[\mathbb{N}]} \), where it defines a permutation of indices \( g \cdot (x_1, \ldots, x_N) \equiv (x_{g_1}, \ldots, x_{g_N}) \). We often use a permutation matrix \( G(x) \in \{0, 1\}^{N \times N} \) to represent this action – that is \( G(x) = g \cdot x \).

3.2 Equivariant Multilayer Perceptrons

A function \( \phi : \mathbb{R}^N \rightarrow \mathbb{R}^M \) is equivariant to a given actions of group \( G \) iff \( \phi(Gx) = \hat{G}\phi(x) \) for any \( x \in \mathbb{R}^N \) and \( g \in G \). That is, a symmetry transformation of the input results in the corresponding symmetry transformation of the output. Note that the action on the input and output may in general be different. In particular, when \( \hat{G} = I_M \) for all \( g \) – that is, the action on the output is trivial – equivariance reduces to invariance. Here, \( I_M \) is the \( M \times M \) identity matrix. For simplicity and motivated by practical design choices, we assume the same action on the input and output.

For a feedforward layer \( \phi : x \rightarrow \sigma(Wx) \), where \( \sigma \) is a point-wise non-linearity and \( W \in \mathbb{R}^{N \times N} \), the equivariance condition above simplifies to commutativity condition \( G(x)W = WGx \forall g \in G \). This imposes a symmetry on \( W \) in the form of parameter-sharing [38][47]. While we can use computational means to solve this equation for any finite group, an efficient implementation requires a closed form solution. Several recent works derive the closed form solutions for interesting groups and structures. Note that in general the feedforward layer may have multiple input and output channels, with identical \( G \)-action on each channel. This only require replicating the parameter-sharing pattern in \( W \), for each combination of input and output channel. An Equivariant MLP is a stack of equivariant feed-forward layers, where the composition of equivariant layers is also equivariant. Therefore, our task in building MLPs equivariant to finite group actions is reduced to finding equivariant linear maps in the form of parameter-sharing matrices satisfying \( G(x)W = WGx \forall g \in G \); see Fig. 4a.1.a.2 for \( W \) that are equivariant to \( C_4 \), the group of circular translations of 4 objects (a.1), and symmetric group \( S_4 \), the group of all permutations of 4 objects (a.2).

4 Equivariant Map for Product Groups

In this section we formalize the imprimitive action of the wreath product, which is used in describing the symmetries of hierarchical structures. We then introduce the closed form of linear maps equivariant to the wreath product of two groups. With hierarchies of more than two levels, one only needs to iterate this construction. To put this approach in perspective and to make the distinction clear, first we present the simpler case of direct product; see [33] for an extensive discussion of this setting.
4.1 Equivariant Linear Maps for Direct Product of Groups

The easiest way to combine two groups is through their direct product \( G = \mathcal{H} \times \mathcal{K} \). Here, the underlying set is the Cartesian product of the input sets and group operation is \( (h, k)(h', k') = (hh', kk') \). If \( P \) is an \( \mathcal{H} \)-set and \( Q \) a \( \mathcal{K} \)-set then the group \( G = \mathcal{H} \times \mathcal{K} \) naturally acts on \( \mathbb{N} = P \times Q \) using \( (h, k) \cdot (p, q) = (hp, kq) \); see Fig. 2. This type of product is useful in modeling the Cartesian product of structures. The following claim characterises the equivariant map for direct product of two groups using the equivariant map for building blocks.

**Claim 1.** Let \( G^{(k)} \) represent \( \mathcal{H} \)-action, and let \( W_{\mathcal{H}} \in \mathbb{R}^{P \times P} \) be an equivariant linear map for this action. Similarly, let \( W_{\mathcal{K}} : \mathbb{R}^{Q \times Q} \) be equivariant to \( \mathcal{K} \)-action given by \( G^{(k)} \) for \( k \in \mathcal{K} \). Then, the product group \( G = \mathcal{H} \times \mathcal{K} \) naturally acts on \( \mathbb{R}^{N} = \mathbb{R}^{PQ} \) using \( G^{(k)} = G^{(k)} \otimes G^{(k)} \), and the Kronecker product \( W_{\mathcal{G}} = W_{\mathcal{H}} \otimes W_{\mathcal{K}} \) is a \( G \)-equivariant linear map.

Note that the claim is not restricted to permutation action\(^2\) and the proof follows from the mixed-product property of the Kronecker product\(^3\) and the equivariance of \( W_{\mathcal{H}} \) and \( W_{\mathcal{K}} \):

\[
G^{(g)} W_{\mathcal{G}} = (G^{(g)} \otimes G^{(g)}) (W_{\mathcal{H}} \otimes W_{\mathcal{K}}) = (G^{(g)} W_{\mathcal{H}}) \otimes (G^{(g)} W_{\mathcal{K}}) = (W_{\mathcal{H}} G^{(g)}) \otimes (W_{\mathcal{K}} G^{(g)}) = (W_{\mathcal{H}} \otimes W_{\mathcal{K}})(G^{(g)} \otimes G^{(g)}) = W_{\mathcal{G}} G^{(g)} \quad \forall g \in G.
\]

An implication of the tensor product form of \( W_{\mathcal{H} \times \mathcal{K}} \) is that the number of independent linear operators of the product map (free parameters in the parameter-sharing) is the product of the independent operators of the building blocks. Note that whenever dealing with product of inputs belonging to \( \mathbb{R}^{P} \) and \( \mathbb{R}^{Q} \), the result which belongs to \( \mathbb{R}^{PQ} \) and therefore it is vectorized.

**Example 1** (Convolution in Higher Dimensions). \( D \)-dimensional convolution is a Kronecker (tensor) product of one-dimensional convolutions. The number of parameters grows with the product of kernel width across all dimensions; Fig. 2(a.1) shows the parameter-sharing for circular 1D convolution \( W_{C_4} \), and (b.1) shows the parameter-sharing for the direct product \( W_{C_4 \times C_4} \).

**Example 2** (Exchangeable Tensors). Hartford et al.\(^{21}\) introduce a layer for modeling interactions across multiple sets of entities, e.g., a user-movie rating matrix. Their model can be derived as the Kronecker product of 2-parameter equivariant layer for sets \( \mathbb{S}_n \). The number of parameters is therefore \( 2^D \) for a rank \( D \) exchangeable tensor. Fig. 2(b.1) shows the 2-parameter model \( W_{\mathbb{S}_4} \) for sets, and (c.1) shows the parameter-sharing for the direct product \( W_{\mathbb{S}_4 \times \mathbb{S}_4} \).

4.2 Wreath Product Action and Equivariance to a Hierarchy of Symmetries

Let us start with an informal definition. Suppose as before \( P \) and \( Q \) are respectively an \( \mathcal{H} \)-set and a \( \mathcal{K} \)-set. We can attach one copy of \( Q \) to each element of \( P \). Each of these inner sets or fibers have their own copy of \( \mathcal{K} \) acting on them. Action of \( \mathcal{H} \) on \( P \) simply permutes these fibers. Therefore the combination of all \( \mathcal{K} \) actions on all inner sets combined with \( \mathcal{H} \)-action on the outer set defines the action of the wreath product on \( P \times Q \). Fig. 3 demonstrates how one point \( (p, q) \) moves under this action. Next few paragraphs formalize this.

**Semidirect Product.** Formally, wreath product is defined using semidirect product which is a generalization of direct product. In part due to its use in building networks equivariant to Euclidean isometries, application of semidirect product in building equivariant networks is explored in several recent works; see \(^{10}\) and citations therein. In semidirect product, the underlying set (of group members) is again the product set. However the group operation is more involved. The (external) semi-direct product \( G' = \mathcal{H} \ltimes \mathcal{K} \), requires a homomorphism \( \gamma : \mathcal{H} \to \text{Aut}(\mathcal{K}) \) that for each choice of

\[\text{(A} \otimes \text{B})(\text{C} \otimes \text{D}) = (\text{AC}) \otimes (\text{BD})\]
As before, let $H$ be a group, and let $Q$ be a set. Then $H \times Q$ is a group where the operation is defined as $(h, q) \cdot (h', q') = (hh', q + q')$. If in addition to $Q$, $P$ is a second group, then $H \times (Q \times P)$ is a group where the operation is defined as $(h, q, p) \cdot (h', q', p') = (hh', q + q', pp')$. Each member of this group is a tuple $(h, q, p)$, and index these copies by an integer $k$.

Since $P$ is an $H$-set, $H$ acts on $P$ by permuting the fibers $P_k$ in $P$. The semidirect product $P \rtimes H$ defined using this action of $H$ on $P$ is called the wreath product, and written as $H \ltimes P$. Each member of the product group can be identified by the pair $(h, \delta)$, where $\delta$ is a member from the base group $P$ itself.

This shows that the order of the wreath product group is $|H|^{|P|}$, which can be much larger than the direct product group $P \times H$, whose order is $|H|$. Each member of the product group can be identified by the pair $(h, \delta)$, where $\delta$ is a member from the base group $P$ itself.

Now we are ready to define the wreath product of two groups. As before, let $\mathcal{H}$ and $\mathcal{K}$ denote two finite groups, and let $\mathcal{P}$ be an $\mathcal{H}$-set. Define $\mathcal{B}$ as the direct product of $\mathcal{P} = |\mathcal{P}|$ copies of $\mathcal{K}$, and index these copies by $p \in \mathcal{P}$: $\mathcal{B} = \mathcal{K} \times \mathcal{K} \times \cdots \times \mathcal{K}_{p} \times \cdots \times \mathcal{K}_{\mathcal{P}}$. Each member of this group is a tuple $\delta = (\delta_1, \delta_2, \ldots, \delta_p)$, where $\delta_i$ is an element of $\mathcal{K}$.

Since $\mathcal{P}$ is an $\mathcal{H}$-set, $\mathcal{H}$ also naturally acts on $\mathcal{B}$ by permuting the fibers $\mathcal{K}_p$. The semidirect product $\mathcal{B} \rtimes \mathcal{H}$ defined using this action of $\mathcal{H}$ on $\mathcal{B}$ is called the wreath product, and written as $\mathcal{K} \ltimes \mathcal{B}$. Each member of the product group can be identified by the pair $(h, \delta)$, where $\delta$ is a member from the base group $P$ itself.

This shows that the order of the wreath product group is $|\mathcal{K}|^{|\mathcal{P}|}$, which can be much larger than the direct product group $\mathcal{K} \times \mathcal{P}$, whose order is $|\mathcal{K}|$. Now we are ready to define the wreath product of two groups.

Now we are ready to define the wreath product of two groups. As before, let $\mathcal{H}$ and $\mathcal{K}$ denote two finite groups, and let $\mathcal{P}$ be an $\mathcal{H}$-set. Define $\mathcal{B}$ as the direct product of $\mathcal{P} = |\mathcal{P}|$ copies of $\mathcal{K}$, and index these copies by $p \in \mathcal{P}$: $\mathcal{B} = \mathcal{K} \times \mathcal{K} \times \cdots \times \mathcal{K}_{p} \times \cdots \times \mathcal{K}_{\mathcal{P}}$. Each member of this group is a tuple $\delta = (\delta_1, \delta_2, \ldots, \delta_p)$, where $\delta_i$ is an element of $\mathcal{K}$.

Since $\mathcal{P}$ is an $\mathcal{H}$-set, $\mathcal{H}$ also naturally acts on $\mathcal{B}$ by permuting the fibers $\mathcal{K}_p$. The semidirect product $\mathcal{B} \rtimes \mathcal{H}$ defined using this action of $\mathcal{H}$ on $\mathcal{B}$ is called the wreath product, and written as $\mathcal{K} \ltimes \mathcal{B}$. Each member of the product group can be identified by the pair $(h, \delta)$, where $\delta$ is a member from the base group $P$ itself.

This shows that the order of the wreath product group is $|\mathcal{K}|^{|\mathcal{P}|}$, which can be much larger than the direct product group $\mathcal{K} \times \mathcal{P}$, whose order is $|\mathcal{K}|$.

### 4.2.1 Imprimitive Action of Wreath Product

If in addition to $\mathcal{P}$ being an $\mathcal{H}$-set, $\mathcal{K}$ action on $\mathcal{Q}$ is also defined, the wreath product group acts on $\mathcal{P} \times \mathcal{Q}$ (making it comparable to the direct product). Specifically, $(h, h_1, \ldots, h_p) \in \mathcal{K} \ltimes \mathcal{B}$ acts on $(p, q) \in \mathcal{P} \times \mathcal{Q}$ as follows:

$$(h, h_1, \ldots, h_p) \cdot (p, q) = (h \cdot h_1 \cdot \cdots \cdot h_p, q).$$

Intuitively, $\mathcal{H}$ permutes the copies of $\mathcal{K}$ acting on each $\mathcal{Q}$, and itself acts on $\mathcal{P}$. We can think of $\mathcal{P}$ as the outer structure and $\mathcal{Q}$ as the inner structure; see Fig. 3.

**Example 3** (Sequence of Sequences). Consider our early example where both $\mathcal{H} = \mathcal{C}_P$ and $\mathcal{K} = \mathcal{C}_Q$ are cyclic groups with regular action on $\mathcal{P} \cong \mathcal{H}$, $\mathcal{Q} \cong \mathcal{K}$. Each member of the wreath product $\mathcal{C}_Q \ltimes \mathcal{C}_P$, acts by translating each inner sequence using some $c \in \mathcal{C}_Q$, while $c' \in \mathcal{C}_P$ translates the outer sequence by $c'$.
Let the $P \times P$ permutation matrix $G^{(k)}$ represent the action of $k \in \mathcal{H}$, and the $Q \times Q$ matrices $G^{(k_1)}, \ldots, G^{(k_P)}$ represent the action of $k_p$ on $Q$. Then the action of $g \in \mathcal{H} \times \mathcal{H}$ on (the vectorized) $P \times Q$ is the following $PQ \times PQ$ permutation matrix:

$$G^{(g)} = \begin{bmatrix}
G^{(k_1)}_{1,1} & \cdots & G^{(k_1)}_{1,P} \\
\vdots & \ddots & \vdots \\
G^{(k_P)}_{P,1} & \cdots & G^{(k_P)}_{P,P}
\end{bmatrix} = \sum_{p=1}^{P} 1_{p,h} \otimes G^{(k_p)} (1)
$$

where $1_{p,h} \otimes G^{(k_p)}$ is a $P \times P$ matrix whose only nonzero element is at row $p$ and column $h_p$ with the value of 1; and $G^{(k_p)}_{p,p}$ is the element at row $p$ and column $p'$ of the permutation matrix $G^{(k_p)}$. In the summation formulation, the Kronecker product $1_{p,h} \otimes G^{(k_p)}$ puts a copy of permutation matrix $G^{(k_p)}$ as a block of $G^{(g)}$. Note that the resulting permutation matrix is different from the Kronecker product; in this case we have $P + 1$ (permutation) matrices participating in creating $G^{(g)}$, compared with two matrices in the vanilla Kronecker product; see Fig. 5.

### 4.2.2 Equivariant Map

Consider a hierarchical structure, potentially with more than two levels of hierarchy, such as a set of sets $\{W_i\}$ or a sequence of sets $\{S_i\}$. The question answered by the following theorem is: how to use the equivariant map for the entire hierarchy? For now we only consider two levels of hierarchy; extension to more levels follows naturally, and is discussed later. Note that one instance of the hierarchy (e.g., a sequence of sets) belongs to $\mathbb{R}^{PQ}$, as each instance of the structure is in $\mathbb{R}^Q$ and the outer structure is of size $P$ — that is vectorization of the input follows the hierarchy.

**Theorem 4.1.** Let $W_{\mathcal{H} \times \mathcal{H}} \in \mathbb{R}^{PQ \times PQ}$ be the matrix of some linear map equivariant to the imprimitivity action of $\mathcal{H} \times \mathcal{H}$ on $P \times Q$. Any such map can be written as follows

$$W_{\mathcal{H} \times \mathcal{H}} = W_{\mathcal{H}} (1_Q 1_Q^T) + I_P \otimes W_{\mathcal{H}}, \quad \text{where} \quad 1_Q = [1, \ldots, 1]^T, (2)$$

$I_P$ is the $P \times P$ identity matrix, $W_{\mathcal{H}} \in \mathbb{R}^{P \times P}$ is $\mathcal{H}$-equivariant, and $W_{\mathcal{H}} \in \mathbb{R}^{Q \times Q}$ is equivariant to $\mathcal{H}$-action.

Proof is in Appendix A. Pictorially, the first term of Eq. (2) scales up the matrix $W_{\mathcal{H}}$, while the second term adds copies of $W_{\mathcal{H}}$ on the diagonal blocks of the scaled matrix. Assuming sets of independent equivariant linear operators $\{W_{\mathcal{H}}\}$, it is evident that the number of independent linear operators equivariant to $\mathcal{H} \times \mathcal{H}$ grows with the sum of those of the building blocks.

**Example 4** (Various Hierarchies of Sets and Sequences). Consider two of the most widely used equivariant maps, translation equivariant convolution $W_{\mathcal{C}_p}$, and $\mathcal{S}_Q$-equivariant map which has the form $W_{\mathcal{S}_Q} = w_1 1_Q + w_2 1_Q^T$. There are four combinations of these structures in a two level hierarchical structure: 1) set of sets $\mathcal{S}_Q \times \mathcal{S}_P$; 2) sequence of sequences $\mathcal{C}_Q \times \mathcal{C}_P$; 3) set of sequences $\mathcal{C}_Q \times \mathcal{S}_P$; 4) sequence of sets $\mathcal{S}_Q \times \mathcal{C}_P$. Fig. 6(a,b,c,d) show the parameter-sharing matrix for these hierarchical structures, assuming a full kernel in 1D circular convolution.

---

*In the proof we show that the number of parameters is the sum of those of the building blocks minus one.*
The reader is invited to contrast the three parameter layer for set of sets, with the four parameter layer for interactions across sets in Fig. 4(c.1,c.2). Similarly, a model for sequence of sequences has far fewer parameters than a model for an image, as seen in Fig. 4(b1, b.2).

4.2.3 Deeper Hierarchies and Combinations with Direct Product

With more than two levels, the symmetry group involves more than one wreath product, which means that the equivariant map for the hierarchy is given by a recursive application of Theorem 4.1. For example, the equivariant map for \((\mathcal{K} \bowtie \mathcal{H}) \bowtie \mathcal{U}\), in which \(\mathcal{U}\) acts on some set \(\mathbb{R}\), and \(W_\mathcal{U}\) is the corresponding equivariant map, is given by \(W_{(\mathcal{K} \bowtie \mathcal{H}) \bowtie \mathcal{U}} = W_\mathcal{U} \otimes (1_{PQ} 1_{PQ}) + I_\mathcal{R} \otimes W_{\mathcal{K} \bowtie \mathcal{H}}\), where \(W_{\mathcal{K} \bowtie \mathcal{H}}\) is in turn given by Eq. (2). Note that wreath product is associative, and so the iterative construction above leads the same equivariant map as \(W_{\mathcal{K} \bowtie \mathcal{H} \bowtie \mathcal{U}}\).

We can also mix and match this construction with that of direct product in Claim 1 for example, to produce the map for exchangeable tensors (product of sets), where each interaction is in the form of an image (hierarchy) – i.e., the group is \((C_\mathcal{R} \times C_\mathcal{V}) \bowtie (\delta_p \times \delta_q)\).

4.3 Efficient Implementation

With equivariant maps for direct product of groups, efficient implementation of \(W_{\mathcal{K} \bowtie \mathcal{H}}\) using efficient implementation for individual blocks \(W_{\mathcal{K}}\), and \(W_{\mathcal{H}}\) is non-trivial – e.g., consider 2D convolution. In contrast, it is possible to use a black-box implementation of the parameter-sharing layers for \(W_{\mathcal{K}}\), and \(W_{\mathcal{H}}\) to construct \(W_{\mathcal{K} \bowtie \mathcal{H}}\). To this end, let \(x \in \mathbb{R}^{PQ}\) be the input signal, and \(\text{mat}(x) \in \mathbb{R}^{P \times Q}\) denote its matrix form; for example in a set of sets, each column is an inner set in this matrix form. Throughout, we are assuming a single input-output channel, relying on the idea that having multiple channels simply corresponds to replicating the linear map for each input-output channel combination. Then we can rewrite Eq. (2) as

\[
W_{\mathcal{K} \bowtie \mathcal{H}} x = \text{vec} \left( (W_{\mathcal{H}} (\text{mat}(x) 1_Q)) 1_Q^T + \text{mat}(x) W_{\mathcal{K}} \right),
\]

where the first multiplication \(\text{mat}(x) 1_Q\) pools over columns (inner structures), and after application of the \(\mathcal{H}\)-equivariant map \(W_{\mathcal{H}}\) to the pooled value, the result is broadcasted back using right-multiplication by \(1_Q^T\). The second term simply transforms each inner structure using \(W_{\mathcal{K}}\). The overall operation turns out to be simple and intuitive: the inner equivariant map is applied to individual inner structures, and the outer equivariant map is applied to pooled values and broadcasted back.

**Example 5 (Equivariant Map for Multiresolution Image).** Consider a coarse pixelization of an image into small patches. Eq. (3) gives the following recipe for a layer equivariant to independent translations within each patch as well as global translation of the coarse image: 1) apply convolution to each patch independently; 2) pool over each patch, apply convolution to the coarse image, and broadcast back to individual pixels in each patch. 3) add the contribution of these two operations. Notice how pooling over regions, a widely used operation in image processing, appears naturally in this approach. One could also easily extend this to more levels of hierarchy for larger images.

5 Application: Point-Cloud Segmentation

In this section we consider a simple application of the theory above to large-scale point-cloud segmentation. The layer is the 3D version of equivariant linear map for sequence of sets, which is visualized in Fig. 4(d.2) – that is we combine translation and permutation symmetry by voxelizing the point-cloud; see Fig. 6. Using a hierarchical structure is beneficial compared to both the set model and 3D convolution. In particular, the set model lacks any prior of the Euclidean nature of the data, while the 3D convolution, in order to preserve resolution requires a fine-grained voxelization where each point appears in one voxel. The reader may ask: “how is the wreath product used to characterize the setting when the number of points is changing per voxel?” While our derivation using wreath...
Table 1: Performance of various models on point cloud segmentation benchmarks.

<table>
<thead>
<tr>
<th>Model</th>
<th>SEMANTIC-8 OA</th>
<th>mIoU</th>
<th>3DIS OA</th>
<th>mIoU</th>
<th>VKITTI OA</th>
<th>mIoU</th>
<th>mAcc</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEEPSETS [50]</td>
<td>89.3</td>
<td>60.5</td>
<td>67.3</td>
<td>42.7</td>
<td>74.2</td>
<td>42.9</td>
<td>36.8</td>
</tr>
<tr>
<td>POINTNET++ [36]</td>
<td>85.7</td>
<td>63.1</td>
<td>81.0</td>
<td>54.5</td>
<td>79.7</td>
<td>34.4</td>
<td>47.0</td>
</tr>
<tr>
<td>SP3 [40]</td>
<td>92.9</td>
<td>76.2</td>
<td>85.5</td>
<td>62.1</td>
<td>84.3</td>
<td>67.3</td>
<td>52.0</td>
</tr>
<tr>
<td>CONVPOINT [41]</td>
<td>93.4</td>
<td>76.5</td>
<td>88.8</td>
<td>68.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KP-FCNN [43]</td>
<td>-</td>
<td>-</td>
<td>70.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RSNET [23]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>56.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PCCN [45]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>58.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SNAPNET [5]</td>
<td>91.0</td>
<td>67.4</td>
<td>-</td>
<td>-</td>
<td>79.7</td>
<td>57.6</td>
<td>35.6</td>
</tr>
<tr>
<td>ENGELMANN ET AL. 2018 [13]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>80.6</td>
<td>54.1</td>
<td>36.2</td>
</tr>
<tr>
<td>ENGELMANN ET AL. 2017 [12]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87.8</td>
<td>54.1</td>
<td>41.6</td>
</tr>
<tr>
<td>3P-RNN [49]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>WREATH PRODUCT NET. (ours)</td>
<td>93.9</td>
<td>75.4</td>
<td>90.6</td>
<td>71.2</td>
<td>88.4</td>
<td>68.9</td>
<td>58.6</td>
</tr>
<tr>
<td>WREATH PRODUCT NET. + ATTN (ours)</td>
<td>95.2</td>
<td>77.4</td>
<td>95.8</td>
<td>80.1</td>
<td>90.7</td>
<td>69.5</td>
<td>59.2</td>
</tr>
</tbody>
</table>

product assumes a fixed number of points, since the resulting model for the set is independent of the number of data points, we can operate on various points per voxel. This is the same reason why we can apply the same convolution filter to images of different sizes.

The past few years have seen a growing body of work on learning with point cloud data; see [19] for a survey. Many methods use hierarchical aggregation and pooling; this includes the use of furthest point clustering for pooling in POINNET++ [37], use of concentric spheres for pooling in SHELLNET, or KD-tree guided pooling in [24]. Several works extend convolution operation to maintain both translation and permutation equivariance in one way or another [3, 4, 45]; see also [6, 17]. Here, objective is not to introduce a radically new procedure, but to show the effectiveness of the approach discussed in previous sections in deep model design from first principles. Indeed, we are able to achieve state-of-the-art in several benchmarks for large point-clouds.

5.1 Equivariance to a Hierarchy of Translations and Permutations

Let $X \in \mathbb{R}^{N \times C}$ be the input point cloud, where $N$ is the number of points and $C$ is the number of input channels (for concreteness, in this section are including the channel dimension in our formulae.) In addition to 3D coordinates, these channels may include RGB values, normal vectors or any other auxiliary information. Consider a voxelization of the point cloud with a resolution $D$ voxels per dimension, and consider the hierarchy of translation symmetry across voxels and permutation symmetry within each voxel. We may also replace 3D coordinates with relative coordinates within each voxel. Let $\Pi \in \{0, 1\}^{D^3 \times N}$ with one non-zero per column identify the voxel membership. Then the combination of equivariant set layer $XW_1 + \Pi^T \Pi XW_3$ [50] with 3D convolution using the pool-broadcast interpretation given in Eq. (3), results in the following wreath-product equivariant linear layer

$$\phi(X) = XW_1 + \Pi^T (\Pi X W_3 (\Pi X)),$$  

where $W_1 \in \mathbb{R}^{C \times C'}$, $*$ denotes the convolution operation, $W_3 \in \mathbb{R}^{K^3 \times C \times C'}$ is the convolution kernel, with kernel width $K$, and $C'$ output channels. Here, multiplication with $\Pi$ and $\Pi^T$ performs the pooling and broadcasting from/to points within voxels, respectively. Note that we have dropped the “set” operation $\Pi^T (\Pi X) W_2$ that pools over each voxel, multiplies by the weight and broadcasts back to the points. This is because it can be absorbed in the convolution operation, and therefore it is redundant. In the equation above pooling operation can replace summation (implicit in matrix multiplication) with any other commutative operation; see [35, 48, 51], we use mean-pooling for the experiments. While the layer of Eq. (4) already achieves state-of-the-art in the experiments, we also consider adding an (equivariant) attention mechanism for further improvement; see Appendix B for details.
5.2 Empirical Results

We evaluate our model on two of the largest real world point cloud segmentation benchmarks, **SEMANTIC3D** \[20\] and the Stanford Large-Scale 3D Indoor Spaces (s3DIS) \[2\], as well as a dataset of virtual point cloud scenes, the \textsc{vkitti} benchmark \[16\]. As shown in Table \[1\] in all cases we report new state-of-the-art. Table 2 compares the processing time and the size of different models. The architecture of \textsc{wreath product net.} is a stack of equivariant layer Eq. (4) plus ReLU nonlinearity and residual connections. \textsc{wreath product net.}+\textsc{attn} also adds the attention mechanism. \[6\] Details on architecture and training, as well as further analysis of our results appear in Appendix C.

### Table 2: Comparison of pre-processing and training time and the number of parameters for **SEMANTIC**-8.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-Proc. (hrs)</th>
<th>Train (hrs)</th>
<th># Params. $\times 10^6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>POINTNET</td>
<td>8.82</td>
<td>3.54</td>
<td>3.50</td>
</tr>
<tr>
<td>POINTNET++</td>
<td>8.84</td>
<td>7.46</td>
<td>12.40</td>
</tr>
<tr>
<td>SNAPNET</td>
<td>13.42</td>
<td>53.44</td>
<td>30.76</td>
</tr>
<tr>
<td>SPG</td>
<td>17.43</td>
<td>1.50</td>
<td>0.25</td>
</tr>
<tr>
<td>CONVPOINT</td>
<td>13.42</td>
<td>48.74</td>
<td>2.76</td>
</tr>
<tr>
<td>OURS</td>
<td>4.39</td>
<td>53.76</td>
<td>5.27</td>
</tr>
<tr>
<td>OURS + ATTN</td>
<td>4.39</td>
<td>91.68</td>
<td>47.01</td>
</tr>
</tbody>
</table>

Outdoor Scene Segmentation - **SEMANTIC3D** \[20\] is the largest LiDAR benchmark dataset, consisting of 15 training point clouds and 15 tests point clouds with withheld labels, amassing altogether over 4 billion labeled points from a variety of urban and rural scenes. In particular, rather than working with the smaller \textsc{reduced-8} variation, we run our experiments on the full dataset (**SEMANTIC-8** \[19\]) Table \[1\] reports mIoU unweighted mean intersection over union metric (mIoU), as well as the overall accuracy (OA) for various methods. In Table \[3\] we report these measures for different voxelization resolution. At highest resolutions the performance degrades due to overfitting.

Indoor Scene Segmentation - Stanford Large-Scale 3D Indoor Spaces (s3DIS) \[2\] consists of various 3D RGB point cloud scans from an assortment of room types on six different floor in three buildings on the Stanford campus, totaling almost 600 million points. Table \[1\] show that we achieve the best overall accuracy as well as mean intersection over union. This is in spite of the fact that our competition use extensive data-augmentation also for this dataset. In addition to random jittering and subsampling employed by both \textsc{kpcov} and \textsc{convpoint}, \textsc{kpcov} also uses random dropout of RGB data.

Virtual Scene Segmentation - Virtual KITTI (\textsc{vkitti}) \[16\] contains 35 monocular photo-realistic synthetic videos with fully annotated pixel-level labels for each frame and 13 semantic classes in total. Following \[12\], we project the 2D depth information within these synthetic frames into 3D space, thereby obtaining semantically annotated 3D point clouds. Note that \textsc{vkitti} is significantly smaller than either **SEMANTIC3D** or s3DIS, containing only 15 millions points in total.

### Table 3: Effect of voxelization resolution for **SEMANTIC**-8.

<table>
<thead>
<tr>
<th># Voxel/s Per Dim</th>
<th>OA</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>81.7</td>
<td>62.7</td>
</tr>
<tr>
<td>3</td>
<td>85.9</td>
<td>67.3</td>
</tr>
<tr>
<td>4</td>
<td>90.6</td>
<td>70.5</td>
</tr>
<tr>
<td>5</td>
<td>92.3</td>
<td>72.2</td>
</tr>
<tr>
<td>6</td>
<td>93.7</td>
<td>72.8</td>
</tr>
<tr>
<td>7</td>
<td>94.4</td>
<td>74.1</td>
</tr>
<tr>
<td>8</td>
<td>95.1</td>
<td>75.6</td>
</tr>
<tr>
<td>9</td>
<td>94.6</td>
<td>77.1</td>
</tr>
<tr>
<td>10</td>
<td>95.2</td>
<td>77.4</td>
</tr>
<tr>
<td>12</td>
<td>94.8</td>
<td>73.0</td>
</tr>
<tr>
<td>16</td>
<td>90.6</td>
<td>71.5</td>
</tr>
</tbody>
</table>

Conclusion

This paper presents a procedure to design neural networks equivariant to hierarchical symmetries and nested structures. We describe how the imprimitive action of wreath product can formulate the symmetries of the hierarchy, contrast its use case with direct product, and characterize linear maps equivariant to the wreath product of groups. This analysis showed that we can always build an equivariant layer for the hierarchy using equivariant maps for the building blocks, and additional pool-and-broadcast operations. We consider one of the many use cases of this approach to design a deep model for large-scale semantic segmentation of point cloud data, where we are able to achieve state-of-the-art using a simple architecture.

---

6 We have released a \textsc{pytorch} implementation of our models at \url{https://github.com/rw435/wreathProdNet}.

7 The leader-board can be viewed at \url{http://www.semantic3d.net/view_results.php?chl=1}.
Broader Impact

As deep learning finds its way in various real-world applications, the practitioners are finding more constrains in representing their data in formats and structures amenable to existing deep architectures. The list of basic structures such as images, sets, and graphs that we can approach using deep models has been growing over the past few years. The theoretical contribution of this paper substantially expands this list by enabling deep learning on a hierarchy of structures. This could potentially unlock new applications in data-poor and structure-rich settings. The task we consider in our experiments is deep learning with large point-cloud data, which is finding growing applications, from autonomous vehicles to geographical surveys. While this is not a new task, our empirical results demonstrate the effectiveness of the proposed methodology in dealing with hierarchy in data structure.

Acknowledgments and Disclosure of Funding

This research was in part supported by CIFAR AI Chairs program and NSERC Discovery Grant. Computational resources were provided by Mila and Compute Canada.

References


A Proof of Theorem 4.1

Proof. We show that both of the two terms in Eq. (2) are equivariant to $\mathcal{K} \cdot \mathcal{H}$-action as expressed by the permutation matrix of Eq. (1).

Part 1. We show that $G^{(g)}$ commutes with the first term in Eq. (2) for any $g \in \mathcal{K} \cdot \mathcal{H}$:

\[
(W_{\mathcal{H}} \otimes (1_Q 1_Q^\top)) G^{(g)} = (W_{\mathcal{H}} \otimes (1_Q 1_Q^\top)) \left( \sum_{p=1}^{P} 1_{p,h,p} \otimes G^{(h_p)} \right)
\]

\[
= \sum_{p=1}^{P} (W_{\mathcal{H}} 1_{p,h,p}) \otimes (1_Q 1_Q^\top) G^{(h_p)}
\]

\[
= W_{\mathcal{H}} \left( \sum_{p=1}^{P} 1_{p,h,p} \right) \otimes (1_Q 1_Q^\top)
\]

\[
= (W_{\mathcal{H}} G^{(h)}) \otimes (1_Q 1_Q^\top) = (G^{(h)} W_{\mathcal{H}}) \otimes (1_Q 1_Q^\top)
\]

\[
= \sum_{p=1}^{P} (1_{p,h,p} W_{\mathcal{H}}) \otimes (1_Q 1_Q^\top) G^{(h_p)}
\]

\[
= \left( \sum_{p=1}^{P} 1_{p,h,p} \otimes G^{(h_p)} \right) (W_{\mathcal{H}} \otimes (1_Q 1_Q^\top))
\]

\[
= \left( \sum_{p=1}^{P} 1_{p,h,p} \otimes G^{(h_p)} \right) (W_{\mathcal{H}} \otimes (1_Q 1_Q^\top)) = G^{(g)} (W_{\mathcal{H}} \otimes (1_Q 1_Q^\top)).
\]

In Eq. (5) we substitute from Eq. (1). In Eq. (6) we pulled the summation out, and then used the mixed product property of Kronecker product. We also note that a permutation of the uniformly constant matrix $1_Q 1_Q^\top$ is constant. In Eq. (7), since the summation can be restricted to the first term we pull out $W_{\mathcal{H}}$. In Eq. (8) we use the fact that the summation of the previous line is the permutation matrix $G^{(h)}$ and apply the key assumption that $W_{\mathcal{H}}$ and $W_{\mathcal{H}}$ are equivariant to $\mathcal{H}$ and $\mathcal{K}$-action respectively. The following lines repeat the procedure so far in reverse order to commute $G^{(g)}$ and $(W_{\mathcal{H}} \otimes (1_Q 1_Q^\top))$.

Part 2. For the second term, again we use the mixed-product property and equivariance of the input maps

\[
(I_p \otimes W_{\mathcal{H}}) G^{(g)} = (I_p \otimes W_{\mathcal{H}}) \left( \sum_{p=1}^{P} 1_{p,h,p} \otimes G^{(h_p)} \right)
\]

\[
= \sum_{p=1}^{P} (I_p 1_{p,h,p}) \otimes (W_{\mathcal{H}} G^{(h_p)})
\]

\[
= \sum_{p=1}^{P} (1_{p,h,p} I_p) \otimes (G^{(h_p)} W_{\mathcal{H}})
\]

\[
= \sum_{p=1}^{P} (1_{p,h,p} \otimes G^{(h_p)}) (I_p \otimes W_{\mathcal{H}}) = G^{(g)} (I_p \otimes W_{\mathcal{H}})
\]

where in Eq. (13) we used the fact that the identity matrix commutes with any matrix, as well as the equivariance of $G^{(h_p)}$. Eqs. (13) and (15) simply use the mixed product property. Putting the two parts together shows the equivariance of the first and second term in Eq. (2), completing the proof.
A.1 Proof of Maximality

Previously we proved that the linear maps of Eq. (2) is equivariant to imprimitive action of wreath product $\mathcal{R} \rtimes \mathcal{H}$ given by Eq. (1). Here, we prove that any $\mathcal{R} \rtimes \mathcal{H}$ equivariant linear map has this form.

Claim 2. Assuming that any $\mathcal{R}$-equivariant ($\mathcal{H}$-equivariant) linear map $W_{\mathcal{R}}(W_{\mathcal{H}})$ can be written as a linear combination of $K$ ($H$) independent linear operators, then any $W_{\mathcal{R} \rtimes \mathcal{H}}$ as defined in Eq. (2) is a linear combination of $K \cdot H - 1$ independent linear bases.

Proof. Consider the two terms (I) $W_{\mathcal{R}} \otimes (1Q_1Q_q^T)$ and (II) $I_p \otimes W_{\mathcal{H}}$ in Eq. (2), where (I) has $H$ independent bases and (II) has $K$ independent bases. If they had independent bases, their summation would have $K \cdot H$ linear bases. However, (I) and (II) have exactly one shared basis: $W_{\mathcal{R}} = I_p$ is $\mathcal{H}$-equivariant, and $W_{\mathcal{R}} = 1Q_1Q_q^T$ is $\mathcal{R}$-equivariant. Therefore, $I_p \otimes (1Q_1Q_q^T)$ is a basis for both (I) and (II). \(\square\)

Next, we show that any $\mathcal{R} \rtimes \mathcal{H}$-equivariant linear map can be written using as a weighted sum of $K \cdot H - 1$ independent linear operators. Together with proof of equivariance of $W_{\mathcal{R} \rtimes \mathcal{H}}$ these two facts prove that any $\mathcal{R} \rtimes \mathcal{H}$ equivariant linear map has the form of Eq. (3).

Claim 3. Any linear map that is equivariant to the imprimitive action of $\mathcal{R} \rtimes \mathcal{H}$ can a linear combination of $K + H - 1$ equivariant linear bases.

Proof. Let $W \in \mathbb{R}^{N \times N}$ be the matrix representing this equivariant map, satisfying $WG^g = G^gW \quad \forall g \in \mathcal{R} \rtimes \mathcal{H}$. Since $G^gW = \mathbf{W} \quad \forall g \in \mathcal{R} \rtimes \mathcal{H}$, all elements of the matrix $W$ that are in the same orbit are tied together. This constraint means that the number of unique values in $W$ are equal to the number of orbits in this simultaneous action on rows and columns of $W$. Using a property of Kronecker product, we can also write the equation above as $\text{vec}(G^gWG^{g^T}) = (G^g \otimes G^g) \text{vec}(W)$. Therefore, the diagonal group action is given by $G^g \otimes G^g$. Using Burnside’s lemma we get the number of orbits $L$ in this action to be

\[
L = \frac{1}{|G|} \sum_{g \in \mathcal{R} \rtimes \mathcal{H}} \text{Tr}(G^g \otimes G^g) = \sum_{g \in \mathcal{R} \rtimes \mathcal{H}} \text{Tr}(G^g)^2 \quad (16)
\]

Our objective is to show that $L = H + K - 1$. Our strategy is to replace the definition of $G^g$ from Eq. (1) into Eq. (16) and simplify. This simplification involves numerous steps.

\[
L = \frac{1}{|\mathcal{R}||\mathcal{H}|} \sum_{p \in \mathcal{R}, h_1, \ldots, h_p \in \mathcal{H}} \text{Tr}\left( \sum_{p=1}^{P} 1_{p,h_1} \otimes G^{(h_p)} \right)^2 \quad (17)
\]

\[
= \frac{1}{|\mathcal{R}||\mathcal{H}|} \sum_{p \in \mathcal{R}, h_1, \ldots, h_p \in \mathcal{H}} \text{Tr}\left( \sum_{p=1}^{P} G^{(h_1)}_{p,h_1} G^{(h_p)} \right)^2 \quad (18)
\]

\[
= \frac{1}{|\mathcal{R}||\mathcal{H}|} \sum_{p \in \mathcal{R}, h_1, \ldots, h_p \in \mathcal{H}} \left( \sum_{p=1}^{P} G^{(h_1)}_{p,h_1} \text{Tr}(G^{(h_p)}) \left( \sum_{p'=1}^{P} G^{(h_1)}_{p',p'} \text{Tr}(G^{(h_{p'})}) \right) \right) \quad (19)
\]

\[
= \frac{1}{|\mathcal{R}||\mathcal{H}|} \sum_{p_1 \neq p_2, p_1, p_2 \in \mathcal{R}} \left( \sum_{p=1}^{P} G^{(h_1)}_{p,h_1} \text{Tr}(G^{(h_p)}) \text{Tr}(G^{(h_{p'})}) \right) \quad (20)
\]

where in Eq. (18) we observed that the contribution from $G^{(h_p)}$ to the trace of $G^{(h_p)}$ is non-zero only if $G^{(h_p)}_{p,p} = 1$. Note that all the $G$ matrices are permutation matrices and in the following we will continue to use this fact without elaborating. In Eq. (19) we used linearity of trace, and in Eq. (20) we divided the product of the previous line into $p = p'$ and $p \neq p'$. As we show below, the first term simplifies to $K$ and the second terms simplifies to $H - 1$, showing that $L = K + H - 1$. We start by
We then use a linear function that models the interaction between latent classes, for each pair of

\[ W \sim \Pi \]

while the layer of Eq. (4) performs competitively, to improve its performance we add a novel attention

B Adaptive Pooling and Attention

While the layer of Eq. (4) performs competitively, to improve its performance we add a novel attention mechanism. To this end we complement the pool-and-broadcast scheme of Eq. (4) with an adaptive pooling mechanism using a learned \( \tilde{\Pi} : R^N \rightarrow \Delta_L \) in Eq. (4), where \( \Delta_L \) is the \( L \)-dimensional probability simplex. Here, one may interpret \( L \) as the number of latent classes, analogous to \( D^3 \) voxels. The pooling matrix \( \tilde{\Pi} \) is a function of input through a linear map \( W_{3}^{(t)} \in R_{C \times L} \) followed by \( \text{Softmax} : R^L \rightarrow (0, 1)^L \), so that the probability of each point belonging to different classes sums to one – that is

\[ \tilde{\Pi}_n = \text{Softmax} ((XW_3)_n) \forall n \in [N]. \]
back using $\overline{\Pi}_n$, giving us the adaptive pooling layer, in which the output for channel $c$ is given by

$$
\phi(X)_c = \sum_{c'=1}^{C'} \overline{\Pi} W_{4,c,c'} \overline{\Pi} X_c.
$$

(32)

Intuitively this layer can decide which subset of nodes to pool, and therefore acts like an equivariant attention mechanism. In experiments, we see further improvement in results when adding this nonlinear layer to the linear layer of Eq. (4).

C Details on Experiments

C.1 Architecture and Data Processing

Our results are produced using minor architectural variations and limited hyperparameter search. Concretely, our best-performing models on semantic3D and s3dis involve 56 residual-style blocks, where each block comprises of 2 equivariant linear maps of Eq. (4). Each map involves a single 3D convolution operator with periodic padding and a $3 \times 3 \times 3$ kernel. We use an identity map to facilitate skip connections within these blocks. The number channels is fixed 64, except the final block, which maps to the number of segmentation classes (8 for semantic3D and 13 for s3dis). When incorporating adaptive pooling, we utilize 5, 10, 15, 25 and 50 latent classes $L$ for blocks 1-19, 19-29, 30-39, 40-49 and 50-56, respectively and inclusively. Our best-performing model on vKITTI uses 22 such residual-style blocks with identical architectural hyperparameters. For adaptive pooling, we then use 5, 10, 15, and 20 latent classes for blocks 1-9, 10-14, 15-19, and 20-22, respectively and inclusively.

Additionally, while previous works rely on stochastic point cloud subsampling during pre-processing and reprojection during post-processing, we simply split large point clouds into subsets of 1 million points each to ameliorate memory issues. We then compute a $9 \times 9 \times 9$ voxelization over each sample, and train on mini-batches of 4 such samples. To generate predictions, we stitch predictions on these smaller samples together.

C.2 Datasets: Training, Validation and Test

Outdoor Scene Segmentation - semantic3D semantic3D [20] We hold out 4 of the available 15 training point clouds to use as a validation set, as in [30].

Table 4 reports the overall accuracy (OA), and mean intersection over union (mIoU) for our method and the competition. We achieve both the best overall accuracy as well as the best mean intersection over union.

<table>
<thead>
<tr>
<th>Method</th>
<th>OA</th>
<th>mIoU</th>
<th>Man-Made Terrain</th>
<th>Natural Vegetation</th>
<th>High Vegetation</th>
<th>Low Vegetation</th>
<th>Buildings</th>
<th>Hardscape</th>
<th>Scanning Artifacts</th>
<th>Cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEEPSET [5]</td>
<td>89.3</td>
<td>60.5</td>
<td>90.8</td>
<td>76.3</td>
<td>41.9</td>
<td>22.1</td>
<td>94.0</td>
<td>46.5</td>
<td>26.8</td>
<td>85.4</td>
</tr>
<tr>
<td>POINTNET [36]</td>
<td>85.7</td>
<td>63.1</td>
<td>81.9</td>
<td>78.1</td>
<td>64.3</td>
<td>51.7</td>
<td>75.9</td>
<td>36.4</td>
<td>43.7</td>
<td>72.6</td>
</tr>
<tr>
<td>SNAPNET [5]</td>
<td>91.0</td>
<td>67.4</td>
<td>89.6</td>
<td>79.5</td>
<td>74.8</td>
<td>56.1</td>
<td>90.9</td>
<td>36.5</td>
<td>34.3</td>
<td>77.2</td>
</tr>
<tr>
<td>SPG [30]</td>
<td>92.9</td>
<td>76.2</td>
<td>91.5</td>
<td>75.6</td>
<td>78.3</td>
<td>71.7</td>
<td>94.4</td>
<td>56.8</td>
<td>52.9</td>
<td>88.4</td>
</tr>
<tr>
<td>CONVPOINT [4]</td>
<td>93.4</td>
<td>76.5</td>
<td>92.1</td>
<td>80.6</td>
<td>76.0</td>
<td>71.9</td>
<td>95.6</td>
<td>47.3</td>
<td>61.1</td>
<td>87.7</td>
</tr>
<tr>
<td>WREATH PRODUCT NET. (ours)</td>
<td>93.9</td>
<td>75.4</td>
<td>93.4</td>
<td>84.0</td>
<td>76.4</td>
<td>68.2</td>
<td>96.8</td>
<td>45.6</td>
<td>47.4</td>
<td>91.9</td>
</tr>
<tr>
<td>WREATH PRODUCT NET. + ATTN (ours)</td>
<td>94.6</td>
<td>77.1</td>
<td>95.2</td>
<td>87.1</td>
<td>75.3</td>
<td>67.1</td>
<td>96.4</td>
<td>51.3</td>
<td>51.0</td>
<td>93.4</td>
</tr>
</tbody>
</table>

Table 4: Performance of various models on the full semantic3D dataset (semantic-8). Higher is better, bolded is best. mIoU is unweighted mean intersection over union metric. OA is overall accuracy. Per class splits show mIoU.

Indoor Scene Segmentation - Stanford Large-Scale 3D Indoor Spaces (s3dis) The s3dis dataset [2] consists of various 3D RGB point cloud scans from an assortment of room types on six different floor in three buildings on the Stanford campus, totaling almost 600 million points. Following previous works by [12, 30, 36, 37, 42], we perform 6-fold cross validation with micro-averaging, computing all metrics once over the merged predictions of all test folds.

Here, we achieve the best overall accuracy as well as mean intersection over union. This is in spite of the fact that our competition use extensive data-augmentation also for this dataset. In particular, both KPConv and CONVPOINT use random jittering and subsampling.
Figure 7: (top left) RGB scan of a point cloud with 280 994 028 points from the SEMANTIC3D dataset; (middle left) segmentation results from our method; (bottom left) ground truth segmentation; (top right) RGB scan of a point cloud with 19 767 991 points from the SEMANTIC3D dataset. (middle right) segmentation results from our method; (bottom right) ground truth segmentation;

Table 5: Performance of various models on the S3DIS dataset (micro-averaged over all 6 folds). Higher is better, bolded is best. mIoU is unweighted mean intersection over union metric. OA is overall accuracy. Per class splits show mIoU.

<table>
<thead>
<tr>
<th>Method</th>
<th>OA</th>
<th>mIoU</th>
<th>Ceiling</th>
<th>Floor</th>
<th>Wall</th>
<th>Beam</th>
<th>Column</th>
<th>Window</th>
<th>Door</th>
<th>Chair</th>
<th>Table</th>
<th>Bookcase</th>
<th>Sofa</th>
<th>Board</th>
<th>Clutter</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEEPSETS [50]</td>
<td>67.3</td>
<td>42.7</td>
<td>81.1</td>
<td>72.4</td>
<td>85.2</td>
<td>16.9</td>
<td>25.8</td>
<td>44.2</td>
<td>48.5</td>
<td>51.0</td>
<td>49.8</td>
<td>21.7</td>
<td>24.4</td>
<td>17.2</td>
<td>34.6</td>
</tr>
<tr>
<td>POINTNET [36]</td>
<td>78.5</td>
<td>47.6</td>
<td>88.0</td>
<td>88.7</td>
<td>69.3</td>
<td>42.4</td>
<td>23.1</td>
<td>47.5</td>
<td>51.6</td>
<td>42.0</td>
<td>54.1</td>
<td>38.2</td>
<td>9.6</td>
<td>29.4</td>
<td>35.2</td>
</tr>
<tr>
<td>RSNET [23]</td>
<td>N/A</td>
<td>56.5</td>
<td>92.8</td>
<td>78.6</td>
<td>32.8</td>
<td>34.4</td>
<td>51.6</td>
<td>48.1</td>
<td>60.1</td>
<td>59.7</td>
<td>50.2</td>
<td>16.4</td>
<td>44.9</td>
<td>52.0</td>
<td></td>
</tr>
<tr>
<td>PCCN [45]</td>
<td>N/A</td>
<td>58.3</td>
<td>92.3</td>
<td>75.9</td>
<td>0.27</td>
<td>6.0</td>
<td>69.5</td>
<td>63.5</td>
<td>65.6</td>
<td>66.9</td>
<td>47.3</td>
<td>39.1</td>
<td>46.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPG [30]</td>
<td>85.3</td>
<td>62.1</td>
<td>95.1</td>
<td>75.6</td>
<td>62.9</td>
<td>47.8</td>
<td>55.3</td>
<td>68.4</td>
<td>73.9</td>
<td>69.2</td>
<td>63.2</td>
<td>45.9</td>
<td>5.7</td>
<td>52.9</td>
<td></td>
</tr>
<tr>
<td>CONVPOINT [4]</td>
<td>88.8</td>
<td>68.2</td>
<td>95.0</td>
<td>97.8</td>
<td>81.7</td>
<td>47.1</td>
<td>34.6</td>
<td>63.2</td>
<td>73.2</td>
<td>75.8</td>
<td>71.8</td>
<td>64.9</td>
<td>59.2</td>
<td>57.6</td>
<td>65.0</td>
</tr>
<tr>
<td>KP-FCNN [43]</td>
<td>N/A</td>
<td>70.6</td>
<td>93.6</td>
<td>82.4</td>
<td>88.3</td>
<td>63.0</td>
<td>54.3</td>
<td>66.1</td>
<td>76.8</td>
<td>57.8</td>
<td>64.0</td>
<td>69.3</td>
<td>74.9</td>
<td>61.3</td>
<td>60.3</td>
</tr>
<tr>
<td>WREATH PRODUCT NET (ours)</td>
<td>90.6</td>
<td>71.2</td>
<td>94.3</td>
<td>95.7</td>
<td>80.6</td>
<td>65.0</td>
<td>76.2</td>
<td>82.1</td>
<td>73.9</td>
<td>64.5</td>
<td>52.7</td>
<td>49.2</td>
<td>68.8</td>
<td>73.5</td>
<td>59.9</td>
</tr>
<tr>
<td>WREATH PRODUCT NET + ATTN (ours)</td>
<td>95.8</td>
<td>80.1</td>
<td>97.2</td>
<td>97.8</td>
<td>88.6</td>
<td>72.3</td>
<td>82.0</td>
<td>73.8</td>
<td>78.6</td>
<td>75.1</td>
<td>72.0</td>
<td>73.1</td>
<td>71.4</td>
<td>82.1</td>
<td>77.2</td>
</tr>
</tbody>
</table>

C.3 Virtual Scene Segmentation - Virtual KITTI

The vKITTI dataset [16] contains 35 monocular photo-realistic synthetic videos with fully annotated pixel-level labels for each frame and 13 semantic classes in total. Following [12], we project the 2D depth information within these synthetic frames into 3D space, thereby obtaining semantically annotated 3D point clouds. Similar to the training and evaluation scheme in [12, 29], we separate the original set of sequences into 6 non-overlapping subsequences and use a 6 fold cross-validation protocol (with micro-averaging similar to the methodology on S3DIS).

Note that vKITTI is significantly smaller than either SEMANTIC3D or S3DIS, containing only 15 millions points in total. We hypothesize that these smaller, and sparser point clouds provide little geometric signal outside vegetation and road structure. This partially explains our only incremental improvement.
<table>
<thead>
<tr>
<th>Method</th>
<th>OA</th>
<th>mIoU</th>
<th>mAcc</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEEPSETS \cite{50}</td>
<td>74.2</td>
<td>42.9</td>
<td>36.8</td>
</tr>
<tr>
<td>POINTNET \cite{36}</td>
<td>79.7</td>
<td>57.6</td>
<td>35.6</td>
</tr>
<tr>
<td>ENGELMANN ET AL. 2018 \cite{13}</td>
<td>79.7</td>
<td>54.1</td>
<td>36.2</td>
</tr>
<tr>
<td>ENGELMANN ET AL. 2017 \cite{12}</td>
<td>80.6</td>
<td>54.1</td>
<td>41.6</td>
</tr>
<tr>
<td>3P-ENN \cite{49}</td>
<td>87.8</td>
<td>67.3</td>
<td>52.0</td>
</tr>
<tr>
<td>SPG \cite{29}</td>
<td>84.3</td>
<td>67.3</td>
<td>52.0</td>
</tr>
<tr>
<td>WREATH PRODUCT NET. (Ours)</td>
<td>88.4</td>
<td>68.9</td>
<td>58.6</td>
</tr>
<tr>
<td>WREATH PRODUCT NET. + ATTN (Ours)</td>
<td>90.7</td>
<td>69.5</td>
<td>59.2</td>
</tr>
</tbody>
</table>

Table 6: Performance of various models on the VKITTI dataset (micro-averaged over all 6 folds). Higher is better, best results are in bold. mIoU is unweighted mean intersection over union metric. OA is overall accuracy. mAcc is mean accuracy.

over the state of the art (see Table 6). However, we expect future simulations of point cloud scenes to become increasingly dense, in line with increasingly powerful LiDAR scanners for real world applications \cite{14,40}, where our simple baselines can potentially produce significant gain both in accuracy and computation.