We thank all the reviewers for their thorough reading of the paper, and we are happy to see their positive feedback. A common question was regarding Figure 4 and its example. We will apply suggested edits to make it more clear. Below we address other questions and comments.

Reviewer 1.

- Q: What are the assumptions of the vectorization of the input vector?
  - In the vectorization, the inner structure is repeated within the outer structure. We will revisit Figure 4 as suggested.
- Q: there is a similar characterization of equivariant layers in the general direct product setup in [Maron et al'20], Can the authors comment on that?
  - This is a very relevant paper. Both models are valid and each can be used in a different setting. In contrast to Maron et al'20 we note that hierarchical structures often have wreath product symmetry (i.e., substructures "move" independently) and focus on this type of group action. We plan to further discuss that paper.
- Q. Can the authors comment/discuss [the approximation power] as well?
  - We now have a proof of *maximality*: the proposed linear map of Eq (2) is the most expressive equivariant linear map for the given action, assuming input linear maps ($\mathbf{W}_{\mathcal{K}}$ and $\mathbf{W}_{\mathcal{H}}$) are also maximal. This will be added to the paper. The question of universality remains open.

Reviewer 2.

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

- Q: compare to the baselines in terms of model complexity (say, in terms of number of free parameters as a rough guide). In particular, it seems that the 4th order interactions introduced in the attention layers could greatly inflate the number of parameters in the model?
  - The following table compares the preprocessing and training time, as well as the number of parameters of our model and the competition for SEMANTIC3D dataset. Using attention indeed significantly increases the number of parameters. Note that we achieve SOTA even without using attention. We will add this table and more discussions on efficiency to the revised version.

Reviewer 3.

# Voxels Per Dim	Train (hrs)	Accuracy	mean IoU
2	42.60	81.7	62.7
3	48.77	85.9	67.3
4	56.12	90.6	70.5

- Q: [...] how is the wreath product used when the number of points are changing per voxel (rephrased)
  - This is a theoretically valid concern and we will clarify the text to avoid confusion. Since the number of parameters of the equivariant set layer does not change with the size of the set, we can have different number of points per voxel. The same logic allows DeepSets to be applied to point-clouds of different size, or a convolution filter to be applied to images of different size and so on.
- Q: [...] a brief discussion on how to choose the number of voxels D along each dimension. This choice will have various impact [...]
  - The reviewer is correct: increasing D increases the accuracy as well as the training time. The results in the following table shows this trend for coarser voxelizations (due to limited time). However, assuming all voxels remain (non-empty), changing D does not affect the number of parameters and the size of activations (which is proportional to the number of points), and so its effect on memory usage is minimal. We will add an extended version of this table to the paper.

Reviewer 4.

- Q: About wreath product imposing a "stronger inductive bias" compared to direct product.
  - Direct product "action" is in fact a subgroup of the imprimitive wreath product "action" as a permutation group. This means that their equivariant maps are directly comparable, and wreath product produces a more constrained layer. We will add this argument to support our statement.
- Q: Is it P+1 permutation matrices or P permutation matrices?
  - There are P permutations, one for each inner structure (blocks) and 1 permutation for the outer structure.
- Q: [...] the translation in each patch needs to be cyclic in order to be fully equivariant. Is that right?
  - Yes. However, while it is customary to work with cyclic groups, one could "assume" patches of input are zero padded by half the kernel width. This makes the theory applicable while having no effect in the actual implementation without cyclic assumption.