We thank the reviewers for their insightful comments. In this rebuttal, we respond to remarks from reviews. If accepted, we will extend the submission with discussions from below.

REVIEWER #1

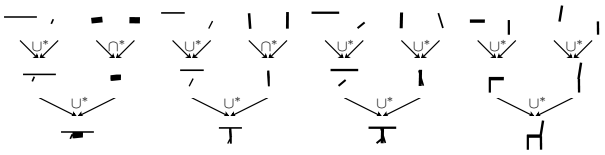
Remark 1 The work lacks discussion about the comparison of interpretability with BSP-Net.

Response BSP-Net generates multiple convex parts that are difficult to interpret due to an unbounded number of possible vertices to be used by each convex. These convexes are also problematic to modify from the perspective of a 3D designer. Moreover, their CSG structure is fixed by definition. It uses the intersection of hyperplanes first, and then the union of convexes.

REVIEWER #2

Remark 1 While the UCSG-NET can predict different CSG trees for different object instances, an empirical evaluation does not show that.

Response Our approach is capable of generating diverse CSG trees for different instances (see Figure on the right). We found empirically, that the diversity of CSG trees can be increased with a layer normalization applied for $\{l \in L | \mathbf{z}^{(l)}, h^{(l)}(\hat{\mathbf{V}}^{(l)})\}$. However, this operation slightly degrades quantitative results and we omitted it in our submission. We sacrificed the diversity of CSG trees for accuracy and reported the qualitative and quantitative results for the best model.



Remark 2 Only a single instance of CSG visualization for each class is shown.

Response We will provide more various CSG visualizations for each class in the supplementary (including Remark 1).

Remark 3 The proposed approach is inferior to the seemingly more straightforward superquadrics method. The structure of the ShapeNet dataset may cause it, and further experiments on the ABC dataset would be needed.

Response Superquadrics [36] method follows the evaluation methodology of Visual Primitives (VP) [37]. These two approaches use only a union operation. While our approach is inferior to the superquadrics, we focused on the applicability of our method in 3D design processes where the CSG is commonly used. Therefore, we argue that referenced approaches are not as versatile as ours. Each work was evaluated on ShapeNet as it is a standard benchmark. Presenting results on this dataset allows the reader to quickly grasp how the method performs in terms of the reconstruction. Referring to ABC, we will put effort into analyzing the dataset to use it in our approach.

REVIEWER #3

Remark 1 It is unclear, what low-level representation decisions allowed UCSG-NET, in contrast to CSG-NET, to be trained in an unsupervised fashion.

Response The CSG-NET predicts a 3D program. Since the program's instructions are discrete, the model is unable to be smoothly optimized towards a particular set of instructions. Therefore, the direct training in an unsupervised manner would be difficult. To solve this limitation, we proposed the following advancements over CSG-NET. Firstly, we applied smooth boolean operations that are pushed towards discrete forms during the training. Secondly, we used the SDF representation of shapes. It allowed us to convert shapes into occupancy values and apply differentiable CSG operations (Fig. 2). These operations are selected dynamically through the attention mechanism (Fig. 3.). Hence, UCSG-NET does not need any supervision to guide which CSG operations should be selected.

REVIEWER #4

Remark 1 Predicted CSG trees are often redundant and qualitatively dissimilar to human-created ones.

Response The network predicts such shapes that would not worsen the final reconstruction. Hence, it can predict highly overlapping boxes. At the same time, we do not force any particular structure of the tree to be learned. Since the task is ill-posed and there exist infinite CSG trees that reconstruct the same shape, we argue that the space solutions can be constrained by weak supervision that would enhance the fidelity of predicted CSG trees.

Remark 2 Could the authors explain the motivation of a bottom-up process that groups primitive to form final outputs?

Response To our knowledge, there are not many machine learning approaches that apply top-down decomposition of meshes that have proven high reconstruction quality. Moreover, our method was designed in such a way that it can be used as an extension for other bottom-up approaches in the literature. Therefore, we believe that our work will encourage future research further to investigate applicability of introduced CSG layers.

Remark 3 Can weak supervision be used in the introduced method?

Response Applying weak supervision is an interesting future direction. Possibly, it would require introducing a new term in $\mathcal{L}_{\texttt{total}}$. We would also limit the number of possible primitives to be predicted, match primitive predictions with the ground truth using the Hungarian method, and use cross-entropy loss to optimize predicted masks $\hat{\mathbf{V}}^{(l)}$ directly.

Remark 4 The sentence "contribute to learning disentangled representation of parts" from seems to be unsupported.

Response To clarify, we found empirically, that our method often separates particular semantic elements of the object during reconstruction (hence disentanglement) and merges them in the final layer, ex. wings and the hull of an airplane or legs and the counter of a desk.