We thank the reviewers for their constructive reviews which clearly show that all reviewers have thoroughly read the paper and are very familiar with related work. In our comments we address the reviews in the order of the reviewer number.

Reviewer #1

The reviewer notes that it would be great to include an analysis about how the pre-alignment of the objects in the ModelNet [35] and ShapeNet [5] dataset influences the quality of the learned representations. In previous literature, supervised tasks have consistently benefitted from pre-aligned objects by a small but significant margin (e.g. 0.3% classification accuracy on ModelNet40 in [16]) when compared to objects with random orientation. This very small gap was corroborated in our initial experiments, we therefore stuck to reporting pre-aligned objects in the paper. We are currently re-running the experiments with random orientation and will provide the numbers in the camera-ready version.

The reviewer also states that our paper could benefit from including PointNet++ [24] in the result tables to further highlight the architecture-agnosticity. We will include these results in the camera-ready version, as the required time for training PointNet++ for all provided experiments exceed the length of the author response period.

Reviewer #2

The reviewer notes that on page 4 line 152, the phrase "no limitation is needed on the receptive field size" would benefit from additional clarification in the context of our paper. In fact, the DGCNN and PointNet architectures used in our comparison use a max-pooling layer over all points in the input point cloud. The receptive field of the used architectures is therefore the entire input point cloud. In contrast, in [21] within the context of images, the receptive field is limited such that the neural network can only use the information from a single image patch. We will explicitly mention this in the camera-ready version.

The reviewer points out that there is a significant performance gap between using our method with PointNet and with DGCNN. The reviewer wonders how much of the performance gain of our method with regards to previous unsupervised methods (i.e. FoldingNet) really stems from the improved architecture of DGCNN or the proposed task. The reviewer argues that FoldingNet can outperform our method when used with PointNet even though it has a PointNet-like architecture, however we believe that FoldingNet has an architecture that can be better compared to DGCNN than to PointNet as it also uses graph convolutions. The FoldingNet decoder alone has 1.05M parameters. Unfortunately the parameters count of the encoder are not stated, but the code indicates around 0.65M parameters which results in 1.7M parameters in total whereas DGCNN only has 1.55M - nonetheless our method outperforms FoldingNet by a significant margin. For some previous unsupervised methods (e.g. VIP-GAN [12]), no detailed description of the architecture and no code is provided. We will do our best to provide the number of parameters and layers for each of the previous unsupervised methods in the camera-ready version by contacting the authors of these papers or re-implementing them to the best of our ability.

Reviewer #3

The reviewer points out that some additional ablation study might be beneficial. In terms of dependence on neural network architectures and training procedures, we did not modify those proposed in PointNet [23] and DGCNN [33]. With the inclusion of PointNet++ [24] in the camera-ready version, we will further highlight that the proposed method is architecture-agnostic. The reviewer also mentions that combining the proposed method with other ideas to further improve the performance of the neural network could be beneficial. We leave this as future work, as combining self-supervised tasks into a multi-task context is out of scope for this paper, whose main goal is instead to show in isolation that the proposed self-supervised learning task can be used flexibly in the context of deep learning on point clouds without particular fine-tuning to tasks or data-domains.