On Generalization of 3D Generative Models

Arushika Bhansal*[1], Abhinanda Ranjit Punnakkal*[2], Dilip K. Prasad[2]

[1]IIT (ISM) Dhanbad, [2]UiT The Arcic University of Norway
abhinanda.r.punnakkal@uit.no, * *shows equal contribution*

Abstract

The generation of 3D data with single-shot or few-shot methods remains an unsolved challenge. In this study, we explore and analyze the generalization capabilities of a generative 3D shape model, 3DShape2VecSet. 3DShape2VecSet is a two-stage model in which the first stage is an autoencoder that generates a latent embedding of the data and the second stage is a diffusion model . We chose this model as our baseline because the latent embeddings are produced using cross-attentional layers, which provide flexible positional encoding depending on the input data. Therefore, it learns representations that capture local features of objects, such as curves and edges, making it well suited for generalization across classes. We study generalization by training it on different sizes of datasets. We show that by focusing on the encoding of the learned representations, the model is able to produce more robust and flexible latent spaces, improving its performance when trained on limited data. Our results show that despite training on smaller datasets, 3DShape2VecSet can effectively generalize across shapes from different categories by exploiting its ability to map shapes to meaningful latent representations. This study highlights the advantages of using learned representations for generalization and contributes to the understanding of the role of latent space in 3D shape generation and recognition.

1 Introduction

The generation of high-quality 3D shapes has become an important area of research in computer vision and deep learning. Despite remarkable progress in 2D image generation, the transition to 3D remains challenging due to the inherent complexity of generating and representing 3D data. High-quality 3D datasets are significantly more difficult to generate and annotate than 2D images, resulting in limited availability of large-scale 3D data, which is a bottleneck for training robust and generalizable models.

Recent advances, such as 3D neural fields (e.g., Neural Radiance Fields), have improved the quality of 3D perception and generation and show promise for rendering detailed and realistic 3D objects. However, several key challenges remain, such as model generalization across multiple object classes, few-shot learning and providing consistent rendering performance. In this work, we investigate the ability of a neural field 3D latent diffusion model to generalize across multiple classes and to learn with less data. To this end, we aim to learn latent shape spaces that are not class-specific in order to increase the generalization capability of the model. We base our experiments on the 3D Shape2Vec model, a framework that integrates an autoencoder and a diffusion model, allowing us to systematically explore these challenges. The 3DShape2Vec model's autoencoder-based latent space, combined with diffusion processes, provides a suitable testbed for analyzing generalizability, especially through its image-conditioned generation capabilities. Our contributions are threefold:

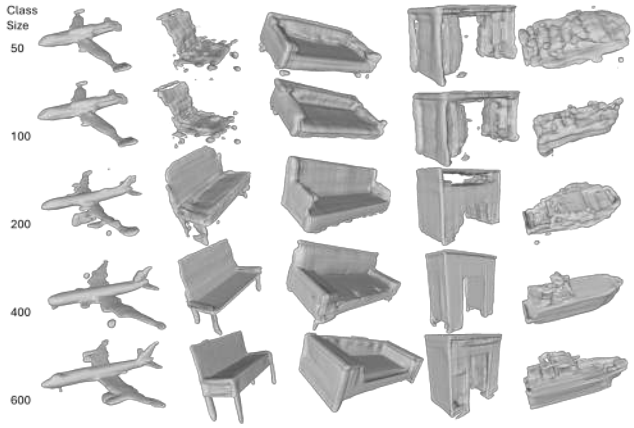

Figure 1: Qualitative results of the multi-class model trained on jet, armchair, couch, desk, and boat classes. One multi-class model is able to generate samples from five different classes.

1) Multi-class generalization: We investigate the generalization capabilities of the 3D Shape2Vec model for multi-class shape generation. While existing models demonstrate strong performance on single-class tasks, their ability to generalize to multiple object classes without performance degradation is underexplored. Our experiments aim to provide a comprehensive evaluation of how well this model can handle different object categories.

2) Effect of training data size: In this work, we systematically investigate the effect of training data size on the

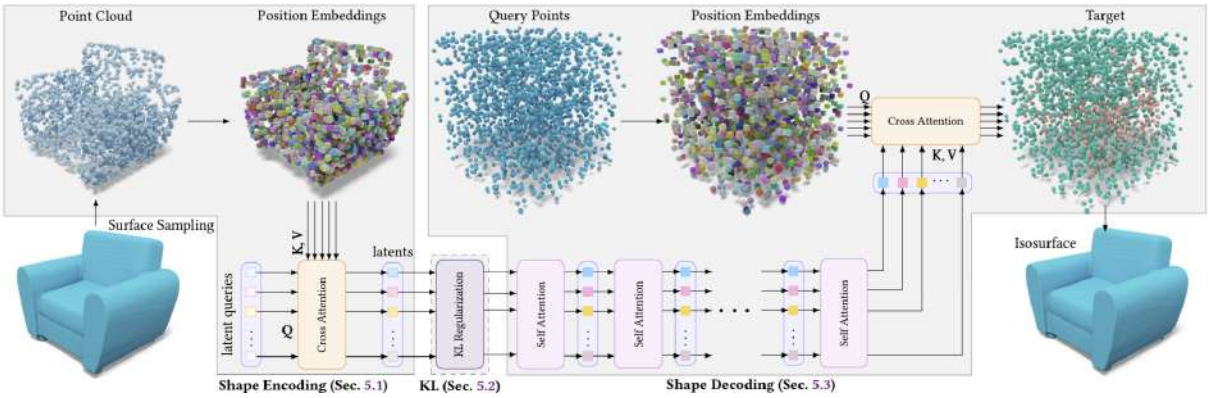

Figure 2: Architecture of Stage 1 autoencoder of 3DShape2VecSet used to create the latent representation (Figure from (Zhang et al. 2023))

performance of the 3DShape2VecSet model, hypothesizing that increasing data diversity, rather than sheer volume, improves generalization across object classes. We evaluate the model's performance on different data sizes to determine how data volume affects multi-class generation.

3) The role of latent representations: We analyze the role of the regularizer used to learn the latent representation in the generalization capability of the model.

Through this study, we aim to address some of the critical challenges faced by current 3D shape generation models and provide new insights into the factors that influence their generalizability and scalability.

2 Related Works

Implicit Shapes based 3D generation The field of 3D shape generation has rapidly advanced, driven by deep learning techniques capable of capturing complex geometric and structural details (Xu, Mu, and Yang 2023).

Of the different ways to represent 3D data such as voxels, point clouds, meshes, etc., implicit shape representations (Mescheder et al. 2019; Park et al. 2019; Chen and Zhang 2019; Michalkiewicz et al. 1901) are becoming more popular due to their ability to retain complex geometry without the loss of details and ease of modeling with deep learning.

Implicit shape-based 3D generative works like (Zheng et al. 2022; Chen and Zhang 2019) use GAN architecture or VAE architectures (Mescheder et al. 2019) or decoder only architecture (Park et al. 2019). As our work focuses on generating diverse shapes, we choose an architecture based on diffusion models. Due to the successful integration of diffusion architecture with implicit representations, there has been significant interest in this combination for 3D generative modeling (Shim, Kang, and Joo 2023; Zhang et al. 2023; Cheng et al. 2022; Chu et al. 2023) targeting problems of 3D reconstruction, shape competition, text-guided generation, etc. However, these techniques are designed and function exclusively for a single category of objects per model.

Few-shot, Zero-Shot, and generalization over classes in 3D Shape Reconstruction Despite the progress in 3D generative modeling, few-shot or zero-shot shape reconstruction is identified as one of the challenges encountered in 3D generation (Farshian et al. 2023). Works like (Wallace and Hariharan 2019; Michalkiewicz et al. 2021) attempted few-short learning for single-view reconstruction. For the case of zero-shot learning, (Zhang et al. 2018) learn priors for the reconstruction of novel categories, and (Thai et al. 2021) exploits image information for 3D reconstruction. (Bechtold et al. 2021; Rao, Nie, and Dai 2022) use a hierarchical approach to generalize over both local and global shapes prior. (Chu et al. 2023) shows generalization on objects of unseen classes by using a hierarchical feature aggregation mechanism. We hypothesize that having a latent space that learn the best possible spatial encoding depending on the data opens up possibilities for the generalization of the model across multiple classes and one-shot generation. Therefore, we choose 3DShape2VecSet (Zhang et al. 2023) as the baseline for our experiments.

Figure 3: Quantitative results of multi-class model with varying training size. The FID and KID scores are computed between the generated samples by the multi-class model and real samples of the classes used in the training set. Both metrics show a decreasing slope with increasing training size.

3 Baseline: 3DShape2VecSet

We summarize the baseline3DShape2VecSet in this section. There are two stages to this model, (i) the autoencoder that creates latent codes of the 3D shapes and (ii) a denoising diffusion probabilistic model that includes a forward diffusion process and an inverse denoising process. We choose this model for the following reasons: (i) 3DShape2VecSet

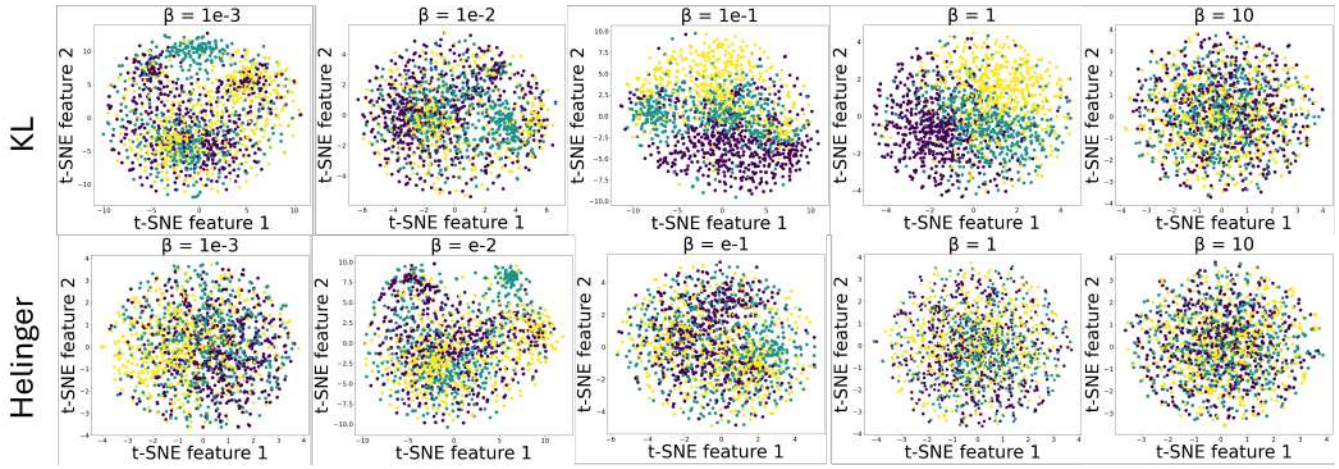

Figure 4: Latent space visualization of multi-class 3DShape2VecSet trained on three classes (airplane, chair, table). KL divergence better preserves class information in the latent embedding. Here yellow points represent airplanes, green points represent chairs and purple points represent tables.

introduces the encoding of 3D shapes into a set of latent vectors that can be used in a latent diffusion model architecture to create a generative model. (ii) 3DShape2VecSet uses neural fields for representation, providing a continuous representation of 3D surfaces and allowing learning through neural networks. (iii) The positional embedding used to encode the spatial information for the latent space is learned by the network. This means that the netorkdoes not rely on fixed spatial positions for its latent vectors, but instead allows the model to learn and infer the necessary spatial information. We hypothesize that this feature will increase flexibility and generalization capabilities of the model.

The architecture of the stage 1 encoder is shown in Figure. 2. The stage 1 encoder pipeline starts with a 3D ground truth surface mesh as input. A point cloud is first sampled from this mesh and then mapped to position embeddings. These embeddings are then encoded into a set of latent codes using a cross-attention module. Optional compression and KL regularization steps are then applied within the latent space to achieve structured and compact latent shape representations. The pipeline continues to the decoder which consists of self-attention layers to aggregate and share information within the latent set. A cross-attention module is used to compute the interpolation weights of query points.

The latent set representation of 3DShape2VecSet uses a cross-attention mechanism to relate the query coordinates to the anchored features based on a data-driven latent set grid. The grid in this context, refers to the latent set grid, which is not a traditional grid with fixed spacing such as a regular or irregular grid. Instead, it is determined by the data itself or what you are trying to encode, meaning that the structure of the grid is influenced by the underlying data features. The grid is adaptive, meaning it adjusts its structure based on the data you're working with, rather than following a predefined pattern. The structure of the grid and the relationships it encodes help the model process and understand the data efficiently.

The final step is to feed the interpolated feature vectors into a fully connected layer for occupancy prediction. This comprehensive pipeline integrates various attention mechanisms and regularization techniques to refine the shape coding process, thereby increasing the accuracy and efficiency of occupancy prediction. We refer to (Zhang et al. 2023) for more details on the baseline.

Thus, a shared encoder-decoder network represents all shapes, and each shape is represented by a latent code computed by the encoder. In addition, the latent representation does not rely on fixed spatial positions for its latent vectors, instead allowing the model to learn and infer the necessary spatial information. We want to investigate if this architecture can be used for a data set with class diversity, i.e. the data set is a combination of several classes of 3D shapes. We also investigate the role of the regularizer in this multi-class generation task and the possibility of using 3DShape2VecSet as a few-shot and single shot 3D generation model.

4 Experiments

4.1 Data

All our experiments use data from the ShapeNet dataset (Chang et al. 2015) (V2). The data is preprocessed according to the steps described in (Zheng et al. 2022). The rendered images from (Choy et al. 2016) is used in the image conditioning experiments. The exact details of classes used for the different experiments are provided in the specific experiments section.

4.2 Multi-class Generation

In general, the task of 3D generation has been limited to the extent of a model that learns from only one class. We investigate the ability of 3DShape2VecSet to generalize across multiple classes. A few-shot or multi-shot 3D model should be able to represent objects from different categories. To this end, we test the performance of 3DShape2VecSet when trained on multiple classes, and refer to the problem as

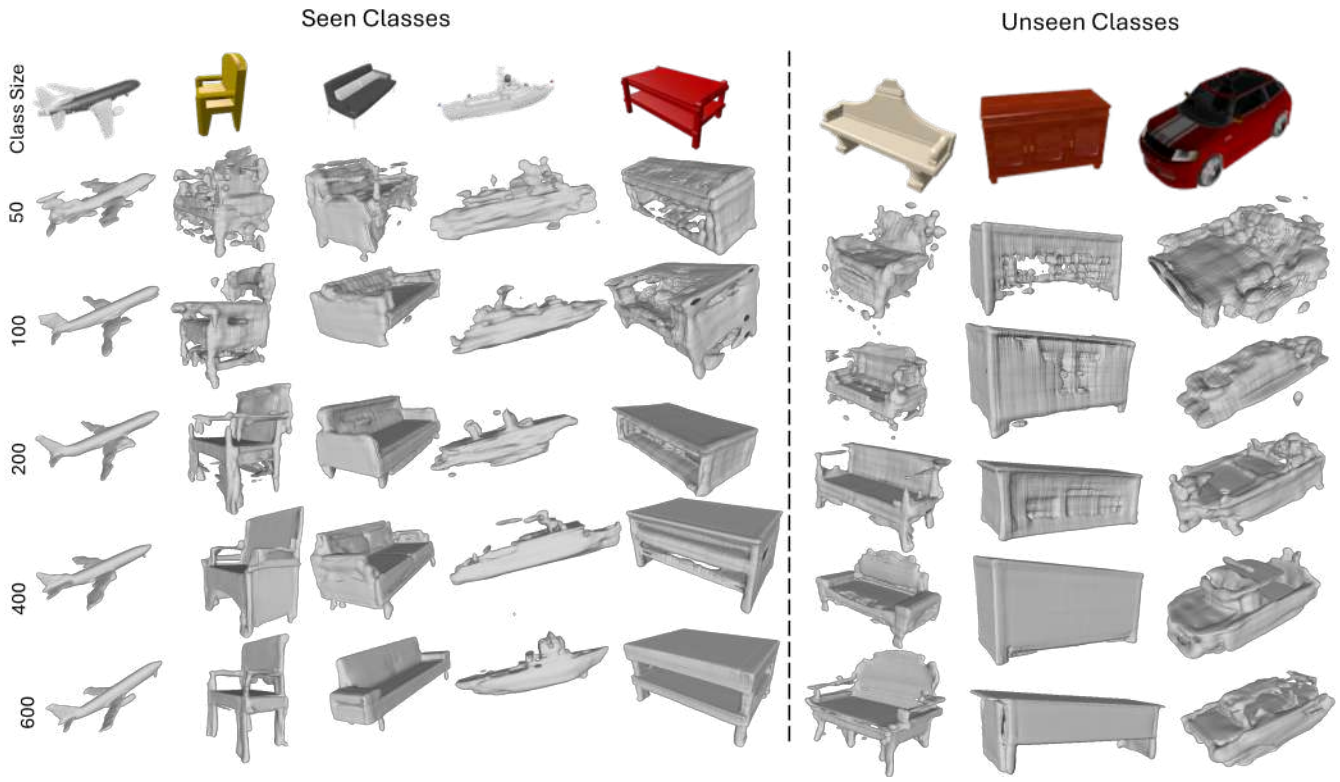

Figure 5: (Left) Qualitative results showing the effect of the size of training data in a multi-class model on seen classes. Top row shows the input image from the seen class provided as conditioning input. The results show that sample quality increases with training data size. (Right) Qualitative results showing the effect of the size of training data in a multi-class model on unseen classes. The results show that 3DShape2VecSet is suitable for single shot generation when trained in a multi-class setting.

"multi-class generation". We also see the variation in performance as the number of samples per class varies.

The multi-class generation problem is trained using the class-conditional variant of 3DShape2VecSet, where the class label is appended as a condition to the denoising network of 3DShape2VecSet (stage 2). The training set is a combination of samples from the jet, armchair, couch, desk, and boat classes of Shapenet. The training sizes are varied as 50, 100, 200, 400, and 600 samples from each class. we show qualitative results of the multi-class model of 3DShape2VecSet in Figure. 1.

Figure 1 shows that a single 3DShape2VecSet model is successful in retaining the features of the 5 different classes it was trained on. The quality of the generated samples improves as the number of training samples increase. We also compute the FID and KID scores for the different classes, as shown in the figure. 3.We follow [(Zhang, Nießner, and Wonka 2022),(Shue et al. 2023),(Ibing, Lim, and Kobbelt 2021)] to adapt the Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), which are commonly used to evaluate image generative models, to rendered images of 3D shapes. We used 150 generated samples per model versus 150 samples from the ground truth data for the FID and KID calculations. We observe that the slope of both the FID and KID scores have a decreasing slope as the number of samples increases. This indicates that the gain in sample quality decreases as the sample size increases.

4.3 Latent space vs Regularizers

In our second experiment, we test the effect of the regularizer on the latent encodings of the samples produced by the stage 1 autoencoder. The diffusion process is performed on these latent embeddings. Therefore, in a multi-class setting, it is desirable that the latent space of the data is meaningful and encodes some class-based clustering of the samples, i.e., intra-class distances are lower and inter-class distances are higher. We also use another regularizer to check its impact on the latent space distribution and the quality of the generated samples.

For this experiment, we train the 3DShape2VecSet model with a multi-class dataset containing samples from 3 classes (airplane, chair, table). We train the multi-class model with two regularizers, i.e., KL divergence (used in 3DShape2VecSet) and Hellinger distance, by varying the beta parameters for both (β parameter, i.e., the weight of the regularizer loss, takes values of 0.001, 0.01, 0.1, 1, 10). Hellinger distance(Hyun, Choi, and Kwak 2019) helps align embeddings from different classes into a unified latent space, reducing overfitting and promoting class-agnostic generalization. It preserves diversity, prevents mode col-

lapse, and ensures smooth transitions in the latent space, allowing for better interpolation and generation. This improves the ability of the latent diffusion model to handle diverse inputs while maintaining a meaningful representation. We adjust KL divergence as given in 3DShape2VecSet, which can be given as follows:

$$\mathcal{L}_{\text{reg}}\left(\{f_i\}_{i=1}^{M}\right) = \frac{1}{M \cdot C_0} \sum_{i=1}^{M} \sum_{j=1}^{C_0} \frac{1}{2} \left(\mu_{i,j}^2 + \sigma_{i,j}^2 - \log \sigma_{i,j}^2\right)$$

where , M is the number of data points or instances, C_0 is the size of the latent space, f_i is a set of features corresponding to each data point, $\mu_{i,j}$is the mean for the i-th data point and the j-th class. $\sigma_{i,j}^2$ is the variance for the i-th data point and the j-th class.

For Hellinger distance following equation is used:

$$H(P, Q) = 1 - \sqrt{\frac{2\sigma_p \sigma_q}{\sigma_p^2 + \sigma_q^2}} \exp\left(-\frac{(\mu_p - \mu_q)^2}{4(\sigma_p^2 + \sigma_q^2)}\right)$$

where, μ_p and σ_p are the mean and standard deviation of our data distribution P, and $\mu_q = 0$ and $\sigma_q = 1$ are the mean and standard deviation of distribution Q.

We visualize the latent encoding of 600 random samples using T-SNE(Van der Maaten and Hinton 2008) Visualization of the latent embedding, 2D T-SNE projections shown in Fig. 4) show that, in general, the KL divergence is better at clustering samples from the same class together for each of the values of the β parameter. The samples from different classes appear to be evenly or randomly distributed in the latent space generated by the Hellinger distance. The highest class-based clustering occurs for KL $\beta = 1$, while β values of 1e-1 and 1e-2 also show decreasing degrees of class-based clustering. We also provide qualitative results in 7 (left) to test the effect of the regularizer on the latent space for the image-conditioned multi-class model of 3DShape2VecSet in section. 4.3.

i. Effect of training data size We train the multi-class models with data from 5 leaf node classes of the Shapenet taxonomy. These leaf classes are chosen to minimize the variation within each class. The training dataset consists of the classes jet, armchair, couch, desk, and boat, which are child classes of the parent classes airplane, chair, sofa table, and watercraft, respectively.

Figure 5 (left) shows representative examples of results generated by a multi-class version of an image-conditioned 3DShape2VecSet model for varying class size of the training set. We can see from Figure. 5 (left) that multi-class 3DShape2VecSet is successful in producing good 3D shape results from the 5 classes it was trained with. We also see that the sample quality improves with increasing training size for all classes.

Next, we use the same model that was trained on the leaf classes (jet, chair, couch, desk, and boat) to generate results given images of unseen classes (classes not used in the training dataset). The figure. 5 (right) shows representative examples of results for image inputs from unseen classes with the different class sizes of the training data.

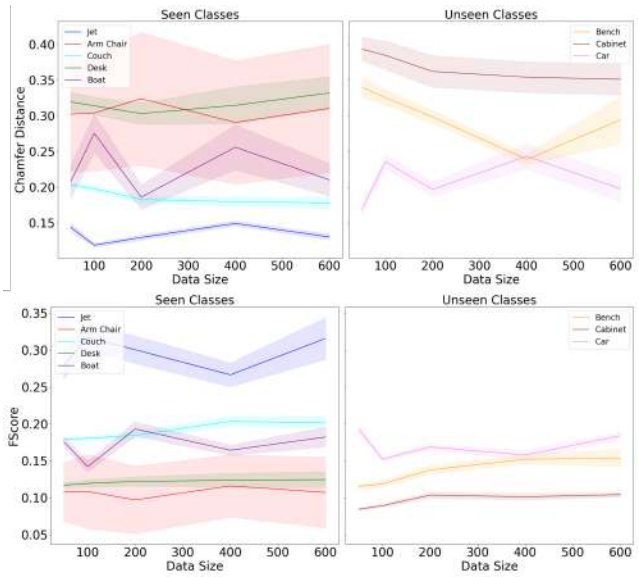

Figure 6: Quantitative metrics of F-score and Chamfer distance of the multi-class model for seen and unseen classes.

We believe that local and global features from different classes of the training data are mixed and matched to produce results closer to the input image of the unseen class. For example, the leftmost image in Figure. 5 (right) is from the bench class. The multiclass image-conditioned 3DShape2VecSet model combines features from the chair, desk, and couch leaf classes present in the training dataset to generate the shape of the bench sample. The second column shows the results of an input image from the cabinet class of ShapeNet and the multi-class 3DShape2VecSet combines features of the couch and desk leaf classes present in the training set to match the input condition at test time. The third column has an input image of the car (unseen class), which is generated by the model by using the leaf class, boat, to give the structure of the car and the 4 legs features of the couch, armchair, and desk leaf classes to form the wheels (can be seen in the 200 training size model). Thus, training with multiple classes has enabled the generation of good quality 3D samples that match input images from unseen classes.

We also ask whether a smaller number of classes in a multi-class training paradigm is better at preserving details in the geometry. Generated samples from the cabinet and bench input images (Figure. 5(right part) , first and second column) show finer geometry (the intended line at the junction of the backrest and seat surface of the first image and the surface indentation for the cabinet door) retained in the output of models trained with 100 and 200 samples than in 400 and 600, which lack these details. Increasing the class size produces smoother surfaces that are more consistent with the overall topology and global shape of the object in the input image. This suggests that training with fewer samples in a multi-class setting results in more robust learning of local features in a single-shot setting.

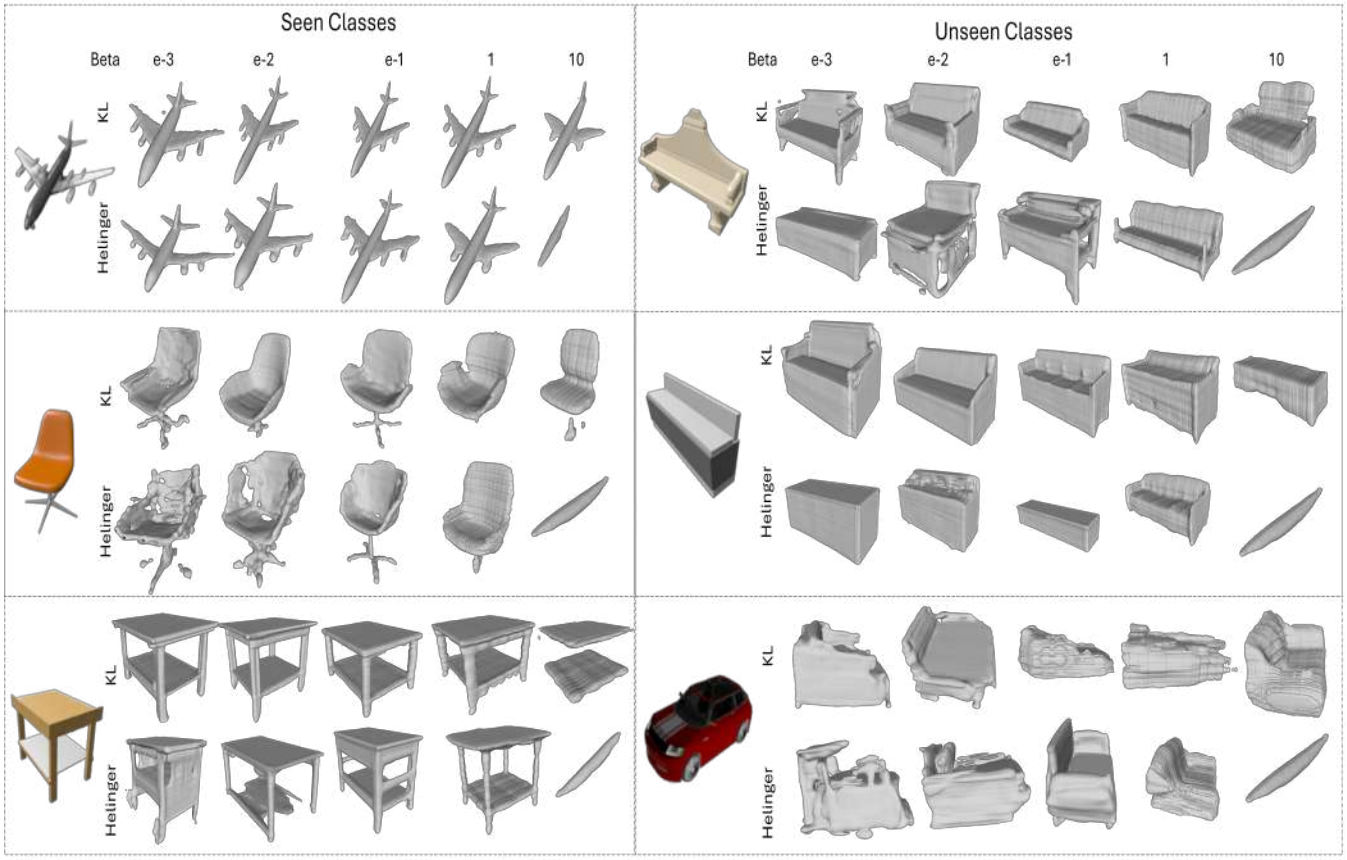

Figure 7: (Left) Qualitative result showing the effect of different regularizers with different β values on the sample generation for seen classes by the model. (Right) Qualitative result showing the effect of different regularizers with different β values on the sample generation for unseen classes by the model.

We compute quantitative metrics of chamfer distance and F-score (Mescheder et al. 2019) on all the meshes in the test set using our multi-class image conditioned models Figure. 6, for the seen classes and unseen classes. Similar to 3DShape2VecSet, Chamfer distance and F-score are calculated between two sampled point clouds, each consisting of 50k points, taken from the reconstructed and ground truth surfaces, respectively. Lower values for Chamfer and higher values for F-score are desirable. The arm chair class has the highest variance in the both the chamfer distance and F-score The chamfer distance of jet, couch, and boat classes in the seen set and the cabinet and car classes decreases with increasing training data size. We observe that both F-score for all 3 unseen classes are within the range of F-scores of the seen class test set and also shows little variation from mean values compared to that of seen classes. While the F-score of seen classes increases with increasing data size of the model, the rate of increase in the unseen classes is lower

Comparison to singleton models: We compare the performance of our image conditioned models against models trained on only one class. We choose the classes of couch and arm chair for this comparison as representatives of best and worst scores for the F-score and chamfer distance metric (Figure. 6). We consider the couch class as the simpler class to learn and the armchair class as the more difficult class to learn among the 5 classes used in the multi-class image condition experiments dues to their F-score and Chamfer values in Figure 6. The F-score and chamfer distance metrics are compared between multi-class and single-class versions of models in Figure. 8. The chamfer distance and F-score metric plots for the couch class shows that the single and multi-class versions show similar trends as the data size increases. The variance in single-class couch model for both Chamfer and F-score is greater than multiclass indicating more robustness in the multi-class results. It is also interesting to note that in the few shot regime (100 - 200 training data size), the multi-class mean metrics are better (F-score) or at par (Chamfer distance) with the single class metric.

ii. Effect of Regularizer We train the multi-class models with data from 3 classes of the Shapenet taxonomy to see the generalization of the multi-class 3DShape2VecSet to unseen classes with different regularizer settings. The training dataset consisting of the airplane, chair, and table classes and the experiment are run for the KL and Hellinger regularizers with β parameter values of 0.001, 0.01, 0.1, 1, 10.

Figure. 7(left) shows representative examples of qualitative results generated by the multi-class version of an image-conditioned 3DShape2VecSet model for varying weight of the regularizer in the total loss of the autoencoder. These are examples taken from the test set of the classes used to train the models. We can see from Figure. 7(left) that multi-class 3DShape2VecSet is successful in generating 3D shapes that match the input image for all beta values except $\beta = 10$ for both KL and Hellinger models. The pattern generated by models with β set to 0.001 is noisier, and the pattern quality improves as we move to higher values of β for both Hellinger and KL, up to $\beta = 1$. We can also see that finer details (local features) are preserved for β values of 0.01 and 0.1 (the lower part of the chair is missing as the β value increases, details of the wing in the aircraft class). More quantitative results are needed to conclude whether Hellinger or KL is more suitable for multi-class 3D generation.

Figure. 7(Right) shows representative examples of qualitative results on unseen classes generated by a multi-class version of an image-conditioned 3DShape2VecSet model for different settings of the regularizer. We observe that at higher values of β, i.e. (1 and 10 for KL and 1 for Hellinger, the results get closer to the local and global features of the training data set, while at lower values of β, such as 0.001, 0.01, and 0.1, the model is more flexible and tries to shape the pre-learned local features according to the input image.

As seen in Figure 4, the combination of regularizer and β that produces clear class clustering (e.g., KL with $\beta = 1$) reduces the fidelity of the representation of the input image features. For example, in Figure 7 (right, row 2, KL with $\beta = 1$) generates legs for an image that lacks them. Similarly, Hellinger with higher β values emphasizes learned features over input fidelity. On the other hand, smaller β values (e.g., $\beta < 0.1$) cause overlap in the latent space due to shared local features (e.g.table and chair classes both have legs). In Figure 7(right, row 3), KL with $\beta < 1$ mixes airplane and table features to reconstruct the frontal structure of the input, consistent with their overlapping clusters in latent space. However, KL with $\beta = 1$, which has distinct clusters, fail to produce the same.

5 Conclusion

This study demonstrates that generative models with learned latent space can be used effectively in a multi-class problem. It also sheds light on the role of regularization and training size in achieving class-agnostic representations with the 3D latent generative models. Our experiments show that while increasing the weight of the regularizer can promote a more uniform latent space, it can also lead to a decrease in generalization ability. Additionally, varying the training size revealed the ability of the model to maintain class-agnostic properties. Here, a smaller training size per class leads to the retention of finer details in the generation. Comparing quanitative metric sof Chamfer distance and F-score indicated that multi-class models performed better in the few-shot regime than their single-shot counter parts. Future work can explore alternative regularization techniques and larger and more diverse datasets to further improve class-agnostic capabilities in 3D shape modeling. We believe that as VR, AR, and gaming become more popular, efficiently creating 3D content with less training data and for new categories is essential to making these technologies scalable and accessible. This study provides valuable insights into understanding 3D content creation in the face of these challenges.

References

Bechtold, J.; Tatarchenko, M.; Fischer, V.; and Brox, T. 2021. Fostering generalization in single-view 3d reconstruction by learning a hierarchy of local and global shape priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15880–15889.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, Z.; and Zhang, H. 2019. Learning Implicit Fields for Generative Shape Modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cheng, Y.-C.; Lee, H.-Y.; Tuyakov, S.; Schwing, A.; and Gui, L. 2022. SDFusion: Multimodal 3D Shape Completion, Reconstruction, and Generation. *arXiv*.

Choy, C. B.; Xu, D.; Gwak, J.; Chen, K.; and Savarese, S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 628–644. Springer.

Chu, R.; Xie, E.; Mo, S.; Li, Z.; Nießner, M.; Fu, C.-W.; and Jia, J. 2023. Diffcomplete: Diffusion-based generative 3d shape completion. *Advances in Neural Information Processing Systems*.

Farshian, A.; Götz, M.; Cavallaro, G.; Debus, C.; Nießner, M.; Benediktsson, J. A.; and Streit, A. 2023. Deep-Learning-Based 3-D Surface Reconstruction—A Survey. *Proceedings of the IEEE*, 111(11): 1464–1501.

Hyun, M.; Choi, J.; and Kwak, N. 2019. Disentangling Options with Hellinger Distance Regularizer. *arXiv preprint arXiv:1904.06887*.

Ibing, M.; Lim, I.; and Kobbelt, L. 2021. 3d shape generation with grid-based implicit functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13559–13568.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Michalkiewicz, M.; Pontes, J.; Jack, D.; Baktashmotlagh, M.; and Eriksson, A. 1901. Deep level sets: Implicit surface representations for 3d shape inference. CoRR abs/1901.06802 (2019).

Michalkiewicz, M.; Tsogkas, S.; Parisot, S.; Baktashmotlagh, M.; Eriksson, A.; and Belilovsky, E. 2021. Learning Compositional Shape Priors for Few-Shot 3D Reconstruction. *arXiv preprint arXiv:2106.06440*.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rao, Y.; Nie, Y.; and Dai, A. 2022. Patchcomplete: Learning multi-resolution patch priors for 3d shape completion on unseen categories. *Advances in Neural Information Processing Systems*, 35: 34436–34450.

Shim, J.; Kang, C.; and Joo, K. 2023. Diffusion-Based Signed Distance Fields for 3D Shape Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20887–20897.

Shue, J. R.; Chan, E. R.; Po, R.; Ankner, Z.; Wu, J.; and Wetzstein, G. 2023. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20875–20886.

Thai, A.; Stojanov, S.; Upadhya, V.; and Rehg, J. M. 2021. 3d reconstruction of novel object shapes from single images. In *2021 International Conference on 3D Vision (3DV)*, 85–95. IEEE.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wallace, B.; and Hariharan, B. 2019. Few-shot generalization for single-image 3d reconstruction via priors. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3818–3827.

Xu, Q.-C.; Mu, T.-J.; and Yang, Y.-L. 2023. A survey of deep learning-based 3D shape generation. *Computational Visual Media*, 9(3): 407–442.

Zhang, B.; Nießner, M.; and Wonka, P. 2022. 3dilg: Irregular latent grids for 3d generative modeling. *Advances in Neural Information Processing Systems*, 35: 21871–21885.

Zhang, B.; Tang, J.; Nießner, M.; and Wonka, P. 2023. 3DShape2VecSet: A 3D Shape Representation for Neural Fields and Generative Diffusion Models. *ACM Trans. Graph.*, 42(4).

Zhang, X.; Zhang, Z.; Zhang, C.; Tenenbaum, J.; Freeman, B.; and Wu, J. 2018. Learning to reconstruct shapes from unseen classes. *Advances in neural information processing systems*, 31.

Zheng, X.; Liu, Y.; Wang, P.; and Tong, X. 2022. SDF-StyleGAN: Implicit SDF-Based StyleGAN for 3D Shape Generation. *Computer Graphics Forum*.

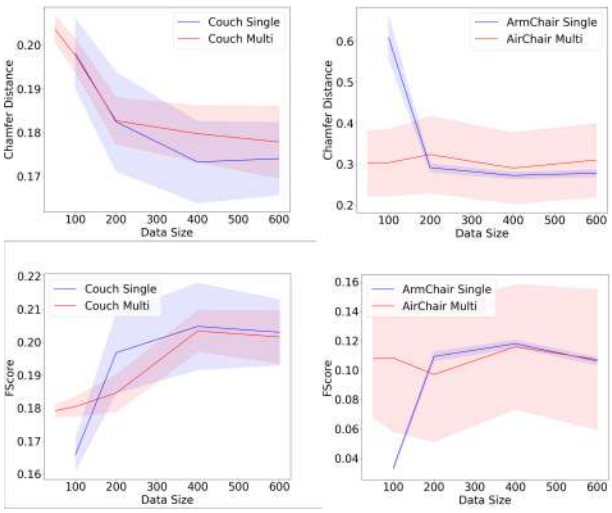

Figure 8: Quantitative comparison of the couch (highest F-score and lowest Chamfer distance among the 5 classes in Figure 7) and arm-chair (lowest F-score, highest Chamfer in Figure 7) classes in the multi-class model against their singleton versions with varying training data size.

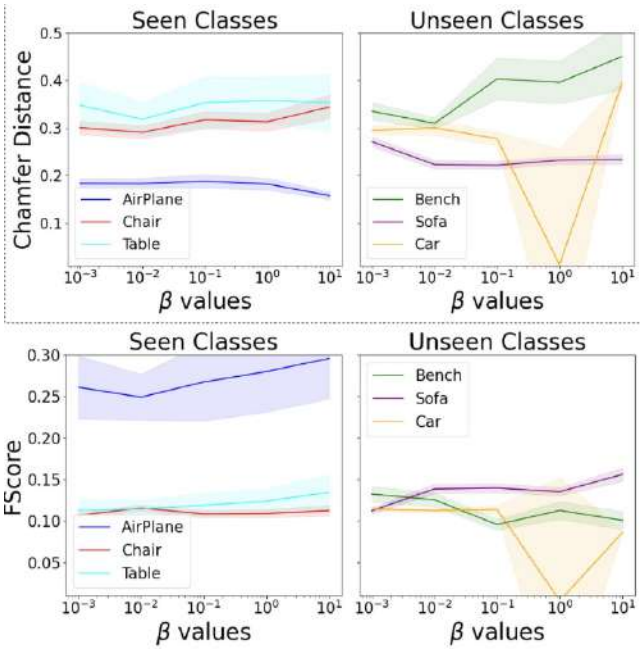

Figure 9: Chamfer Distance and Fscore for the seen and unseen Classes with varying β values of the KL.

Supplementary

Comparison to singleton models: We compare the performance of our image conditioned models against models trained on only one class. We choose the classes of couch and arm chair for this comparison as representatives of best and worst scores for the F-score and chamfer distance metric (Figure. 7). We consider the couch class as the simpler class to learn and the armchair class as the more difficult class to learn among the 5 classes used in the multi-class image condition experiments dues to their F-score and Chamfer values in Figure 7. The F-score and chamfer distance metrics are compared between multi-class and single-class versions of models in Figure. 8. The chamfer distance and F-score metric plots for the couch class shows that the single and multi-class versions show similar trends as the data size increases. The variance in single-class couch model for both Chamfer and F-score is greater than multiclass indicating more robustness in the multi-class results. It is also interesting to note that in the few shot regime (100 - 200 training data size), the multi-class mean metrics are better (F-score) or at par (Chamfer distance) with the single class metric.

In the case of the Armchair (class with lower metric scores in the 5 classes used for training, or the more difficult class), the mean value of both metrics of single and multi- models follow similar trend and have almost equal values for the models of training data size higher that 200. The higher variance in the multi-class could be attributed to the niche details of geometry like wheels, rounded seats and varying topology (holes introduced by the arm-rest), that are present in only the archair class and not the other 4 classes of the multi-class model and is missed by the model. This implies that the singleton model that is trained on just one class is able to pick up these details better. However, we see again in the difficult class of arm-chair, the same observation as the couch class that the multi-class versions are performing better in the few short regime than the singleton models. Between the easy couch class and the more difficult arm chair class, a noticeable difference is that the variance in arm-chair metrics is consistently much higher for the multi-class model than for the single-class model.