RIVQ-VAE: Discrete Rotation-Invariant 3D Representation Learning

Mariem Mezghanni* Malika Boulkenafed Maks Ovsjanikov*

*LIX, Ecole Polytechnique, IP Paris
mezghanni,maks@lix.polytechnique.fr

Abstract

Building local surface representations has recently attracted significant attention in 3D vision, allowing to structure complex 3D shapes as sequences of simpler local geometries. Inspired by advances in 2D discrete representation learning, recent approaches have proposed to break up 3D shapes into regular grids, where each cell is associated with a discrete code sampled from a learnable codebook. Unfortunately, existing methods ignore both the local rigid self-similarities as well as the ambiguities inherent to 3D geometry related to possible changes in orientation. As a result, such techniques require very large codebooks to capture all possible variability in both geometry and pose. In this work, we propose a novel generative model that improves the generation quality by compactly embedding local geometries in a rotation- and translation-invariant manner. This strategy allows our codebook of discrete codes to express a larger range of geometric structures by avoiding local and global redundancies. Crucially, we demonstrate via a careful architecture design that our approach allows to recover meaningful shapes from local embeddings, while ensuring global consistency. The conducted experiments show that our approach outperforms baseline methods by a large margin under similar settings.

1. Introduction

Building 3D generative models that can generalize across different categories has seen rapid progress in recent years. Of particular interest are local-to-global approaches that express a 3D shape as a set of embeddings, each associated with a local surface patch [2, 3, 14, 17, 18, 25, 37, 46]. The motivating idea is that most 3D surfaces from different categories still tend to share *local* geometric details. In this respect, the goal is to learn 3D shape priors at part scale, consisting of meaningful shared abstractions for different shapes. For instance, a chair, a table and a lamp have globally different geometries, but still share local similarities at the chair leg, table leg and lamp pole level. Yet, there are other important similarities that are under-explored in the context of local-based embeddings, with *rotation*-induced similarities being a prime example. A table top and a chair back, for example, can be expected to share locally planar patches. However, their respective local patch embeddings might be inherently different due to the orientation bias when using rotation-dependent local features.

Figure 1. Our network RIVQ-VAE learns a local-based canonicalized (invariant up to rigid motions), discrete latent space of plausible shapes. By avoiding redundancies in the codebook, our formulation leads to higher shape reconstruction accuracy.

In this work, we focus on learning VAEs based on a local shape representation, combined with vector quantization (VQ) technique. In this approach, local embeddings are discrete latent vectors sampled from a finite learnable codebook. This technique, dubbed VQ-VAE [38], has recently led to several advances and proved beneficial not only to circumvent issues of "posterior collapse" [24, 38, 41], but also to facilitate the adoption of Transformer architectures [39] both for 2D [11] and 3D [25, 46] generation. However, existing generative 3D models that use this design [25, 46], build upon and extend 2D designs, and hence ignore vari-

ability inherent to 3D geometry. Particularity, within this context, state-of-the-art techniques discard geometric similarities such as equivalence up to rotation and translation, thus excessively allocating codebook capacity towards both changes in geometry and pose.

Motivated by this observation, we propose a novel compact vector-quantized representation that expresses a 3D shape by a sequence of discrete latent variables sampled from a learnable translation and rotation-invariant discrete codebook. Our formulation stems from the insight that most man-made shapes hold different types of symmetries (such as translational, rotational and reflective [27]) and often share similar geometric patterns (e.g., planar and cylindrical). Consequently, coupling a single local code with rigid transformations can contribute to different shape representations (cf. Figure 4), as opposed to representing each with a different code which depletes codebook capacity.

Towards this goal, we propsoe to encode shapes as sequences of independent discrete codes each representing the canonicalized content of a non-empty local patch, removing translation and rotation ambiguity. Such a latent space, in turn, enables the decoding of implicit function-based shapes by sampling plausible code sequences from the learned codebook.

This formulation, however, raises additional challenges compared to conventional approaches due to the loss of rotation and translation attributes. Specifically, the decoding step must account for recovering *oriented* geometry while ensuring global consistency. We address these challenges by a careful decoder architecture design consisting of convolutional and attention blocks to capture the global structure from disjoint local codes, followed by two branches: the pose estimation branch that computes the rotation vector for each local code, and the geometry prediction branch that recovers the occupancy value at a given spatial point after being appropriately rotated (cf Figure 2). This approach thus allows the network to predict consistent local orientations and to cancel out the translation bias. We further promote global consistency and surface smoothness by employing an interpolation technique between occupancy estimations for neighboring regions. Jointly, this design enables our method to combine the merits of canonicalized local embeddings as a compact and generalizable shape representation, with global regularization, thereby ensuring shape consistency and plausibility.

We summarize our main contributions as follows: (1) We propose a novel latent representation based on sequences of discrete variables that approximate 3D shapes in a rotation and translation-invariant manner (2) We propose a novel RIVQ-VAE architecture tailored for this learned representation. (3) We demonstrate the performance of our model for shape auto-encoding, completion and single-view reconstruction, considerably outperforming existing methods.

2. Related Work

2.1. Distributed local shape representation

Using a single global latent code has been widely investigated for shape representation [1, 26, 28]. Though this line of work seems appealing for its simplicity, modeling 3D shapes in a holistic fashion may cause failure in capturing local details, scaling to complex surfaces and generalizing to unseen shape classes. To sidestep these challenges, research efforts turned towards representing a complex 3D surface as a sequence of smaller ones. Existing works typically partition 3D shapes into regular [2, 3, 7, 8, 17, 18] or irregular [37, 47] regions and process each separately. As such, the embedding process allocates more model capacity towards complex local details and allows scaling to large objects and scenes due to independent spatial geometric partitioning. However, most approaches [2, 3, 18, 47] are only suited for shape reconstruction since local surface locations are known in advance; each local code is associated with a selected geometric component of the input shape. To enable more challenging tasks such as random shape generation and editing, works in [17, 37] propose to infer local embeddings along with their locations by either defining the local surface position as a learnable parameter [37] or by predicting the complete grid-like local feature representation including empty space [17]. Our work falls within this context. We design a novel local-based 3D shape representation, which we further combine with vector quantization technique described below, to enable multiple complex shape generation applications.

2.2. Vector quantization

Vector quantization (VQ) refers to learning a latent space of discrete representations. Unlike the continuous setting, the shape encoder outputs *discrete vectors* sampled from a codebook. Besides its appealing simplicity, discrete representations allow to expand the success of architectures, such as the prominent Transformer [39], originally designed for other discretely-represented modalities, in the context of 2D [38] and 3D generation [25]. The seminal work in [38] proposed VQ-VAE that combines VAE learned on images and audio data with VQ, resulting in categorical posterior and prior distributions. This allows, in particular, to circumvent issues of "posterior collapse" that arises when learning a continuous latent space, where the variational distribution collapses towards the prior, making the decoder unable to benefit from all of the latent vectors dimensions [24, 41]. Recently, several related works have been built upon VQ-VAE [38] aiming to leverage VQ for image generation such as VQ-VAE-2 [33], RVQ-VAE [19], VQ-GAN [11] and VQ-GAN-CLIP [10], and for 3D shape generation such as P-VQ-VAE [25] and ShapeFormer [46]. The latter approach proposed a novel vector quantized deep implicit

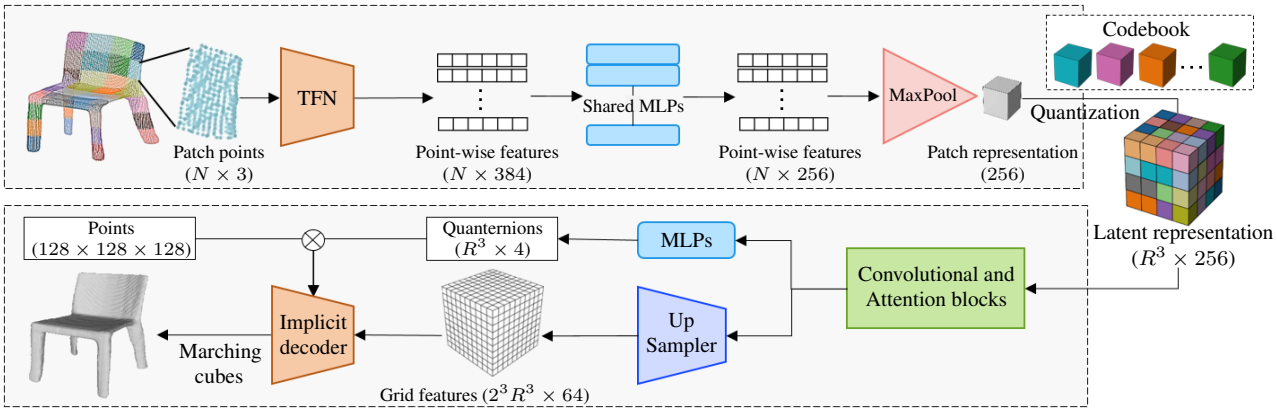

Figure 2. RIVQ-VAE architecture overview. Our model consists of (i) a local encoder that maps each surface patch into a rotation- and translation-invariant representation using a TFN [36] backbone, (ii) a learnable codebook, and (iii) an implicit surface-based decoder consisting of a pose estimation block (top branch) that recovers patch rotation, and a geometry prediction block (bottom branch) that estimates shape occupancy field.

function (VQDIF) reconstruction, using spatial sparsity to represent a 3D shape as a sequence of discrete variables. Meanwhile, 3DILG [49] proposed a novel spatial portioning of 3D shapes based on irregular grids, enabling shape abstractions to be sparse and adaptive.

Despite these advances, existing 3D methods typically constitute a natural extension of 2D designs, and hence ignore variability inherent to 3D geometry. In this work, we propose a novel discrete local representation that accounts for local 3D pose changes, allowing a more accurate and efficient formulation.

2.3. Rotation invariant point cloud embedding

There has been a steady stream of work aiming to efficiently process point cloud data such as PointNet [31], PointNet++ [32], PointCNN [23] and DGCNN [40] to name a few. While these methods improved the frontier of learning on point clouds, they lack robustness to 3D rotation, a desired property for many computer vision tasks such as shape classification and segmentation. Even though several methods rely on massive amounts of rotation-augmented data to tackle this limitation, they are, not only likely to have an increased learning time, but also still have no solid guarantee of the rotation invariance.

To address this challenge, a different line of work focuses on designing rotation invariant operations. Several approaches build rotation invariant point representation [5, 22, 30] and convolutional operations [50, 51] based on local relations between points and their neighbors such as distances and angles. Differently, steerable kernel bases have fueled the emergence of recent methods aiming to design rotation equivariant and invariant CNNs on non-Euclidean domains [12, 13, 20, 36, 42]. In this work, we build upon the TFN [36] architecture to compute a rotation-invariant embedding of local geometries that proves beneficial to guarantee the accuracy of our approach.

3. Method

3.1. Overview

In this work, we propose a novel vector-quantized representation that captures a 3D shape via a sequence of discrete variables. To this end, we design a novel RIVQ-VAE architecture illustrated in Figure 2. RIVQ-VAE consists of an encoder that maps local regions of input point cloud independently into a discrete rotation- and translation-invariant latent representations, sampled from a learnable codebook, while an implicit decoder jointly maps the region-based discrete representations into a watertight mesh which matches the input point cloud. With such a compact representation, we are able to use the Transformer [39] architecture to learn the autoregressive prior over the discrete representations enabling diverse generation tasks.

Our approach, while similar to baselines [25, 46] in using a patch-based encoder, differs in two key aspects (1) encoding local regions in a translation and rotation-invariant manner, and (2) designing a global decoder that promotes knowledge sharing across latent representations, thus enabling the recovery of consistent local orientations and the generation of smooth surfaces.

3.2. Network architecture

Tensor Field Networks To compute the canonicalized representation of local point cloud patches, we use the Tensor Field Networks (TFN) [36], a convolutional network that defines point convolution operation as the product of a learnable radial function and spherical harmonics. Importantly, this approach allows to process point cloud shapes in a permutation-invariant, translation- and rotation-equivariant way.

Given a point cloud $P \in \mathbb{R}^{M\times 3}$, TFN computes, for a spherical harmonics order l, an embedding $F^l(P) \in \mathbb{R}^{(2l+1)\times C}$ where C is a user-defined number of chan-

nels. $F^l(P)$ satisfies the rotation equivariance property $F^l(R_{ot}P) = D^l(R_{ot})F^l(P)$ where $R_{ot} \in SO(3)$ and $D^l : SO(3) \rightarrow SO(2l+1)$ is the so-called Wigner matrix of degree l [21, 34]. We refer the interested readers to [36] for a comprehensive overview.

Observing that the features of $F^l(P)$ have the same rotation equivariance property as the vectors of degree l spherical harmonics $Y^l(P) \in \mathbb{R}^{(2l+1)\times M}$ where $Y^l(.)$ denotes the spherical harmonic polynomials of degree l [29, 34], a *rotation-invariant* feature vector $S^l(P)$ can be computed as:

$$S^l(P) = Y^l(P)^T F^l(P) \in \mathbb{R}^{M\times C} \tag{1}$$

Key to this observation is the idea that the product of an equivariant signal by the transpose of an equivariant signal is rotation invariant [34]:

$$\begin{aligned} S^l(R_{ot}P) &= Y^l(R_{ot}P)^T F^l(R_{ot}P) \\ &= Y^l(P)^T D^l(R_{ot})^T D^l(R_{ot}) F^l(P) \\ &= S^l(P) \end{aligned} \tag{2}$$

for $R_{ot} \in SO(3)$. In what follows, we leverage $S^l(.)$ to define our latent space.

Encoder architecture Our network takes as input a point cloud shape $X \in \mathbb{R}^{N\times 3}$ split into regular voxel regions of resolution R, $\{X_i\}_{i=1..R^3}$. The point cloud patch within each region X_i is fed independently to our point cloud encoder E_ϕ with learnable parameters ϕ to compute latent representations as follows:

$$z_i = E_\phi(X_i) = concat_{l=1..L}\left[\max_{j=1..N_i} S^l_j(X_i - v_i)\right] \tag{3}$$

with $v_i \in \mathbb{R}^3$ the i^{th} grid center position and N_i the size of X_i. Note that z_i inherits the rotation invariance for $S^l(.)$. Besides, by construction, z_i is translation and permutation invariant due to the centering and max pooling operations respectively. Together, these properties justify the canonicalized criterion of our latent space.

Note that our independent encoding ensures that the learned discrete latent space only captures the local context, which benefits its generalization power across different categories. Besides, compared to global shape embeddings, our local and independent setting concentrates the impact of noisy and partial regions at the level of individual patch embeddings, since only the encodings of these regions is affected. This particularly enhances many tasks such as shape auto-encoding and completion.

Discrete latent representation Instead of working directly on the continuous embeddings $\{z_i\}_i$, we aim to transform the high-dimensional continuous representations to a more compact latent space greatly reducing the size and number of latent variables. Ultimately, this allows to efficiently learn the space of plausible distributions of discrete codes using the highly expressive Transformer architecture [39]. To this end, we follow the work in VQ-VAE [38]. Specifically, we define a discrete latent space of K vectors, each of dimension D: $e \in \mathbb{R}^{K\times D}$ stacked in a leanable codebook $\mathcal{D} = \{e_k \in \mathbb{R}^D\}_{k=1..K}$. Each computed continuous embedding z_i is thus vector-quantized into the closest vector in $\mathcal{D}$ in terms of the L_2-norm:

$$VQ(z_i) = z_i^q = \arg\min_{e\in\mathcal{D}} \|z_i - e\|_2 \tag{4}$$

Gradient backpropagation through the non-differentiable VQ operator is ensured via the straight-through gradient technique which copies the decoder gradient at z_i^q to z_i at the backward stage, enabling an end-to-end training of the model. It is important to highlight that, while exact rotation invariance may fall short due to noise or sampling changes in point cloud patches, the vector quantization step is expected to map close patches to the same codebook vector, thus counteracting this effect and leading to a rotation invariance up to discretization.

Decoder architecture The decoder part D_ψ, where ψ are learnable parameters, maps the quantized representations $Z^q = \{z_i^q\}_i \in \mathbb{R}^{R^3\times D}$ of shape X into the corresponding implicit field of occupancy values. Since input shape regions $\{X_i\}_i$ are encoded separately, the first part of D_ψ aims at capturing the global context from the separate region-based embeddings to recover pose and geometry estimations while ensuring global consistency.

One can understand that the canonicalized formulation of our codebook may compromise D_ψ performance compared to conventional setting. In fact, D_ψ needs to allocate its capacity not only for geometry prediction, but also for recovering patch orientations and ensuring global consistency. Motivated by this, we propose to decompose the decoder D_ψ into two branches: the pose estimation branch f_ρ that recovers for each local region i the rotation quaternion $r_i = [r_i^0, r_i^1, r_i^3, r_i^4]$, and a geometry prediction branch g_λ that computes the occupancy value at a given point x. Formally, $\psi = \{\rho, \lambda\}$ and:

$$\begin{aligned} f_\rho :&\quad Z^q \mapsto \{r_i\}_i \\ g_\lambda :&\quad (x, Z^q) \mapsto o_x \end{aligned} \tag{5}$$

As such, given a spatial point $x \in \mathbb{R}^3$ belonging to a region i of center v_i, its occupancy value o_x is predicted as:

$$D_\psi(x, Z^q) = g_\lambda\left(r_i^m(x - v_i), Z^q\right) = o_x, \tag{6}$$

where r_i^m is the rotation matrix in $SO(3)$ associated with r_i. We empirically observed that using this default design can lead to reconstructions that are discontinuous across patch boundaries. To remedy this effect, we interpolate the occupancy estimations from neighboring cells $\mathcal{N}_i$ for point x:

$$D_\psi(x, Z^q) = \sum_{j\in\mathcal{N}_i} \frac{w_j}{\sum_j w_j} g_\lambda\left(r_j^m(x - v_j), Z^q\right) \tag{7}$$

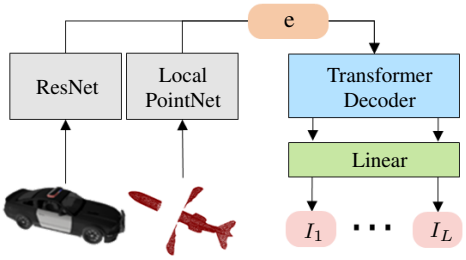

Figure 3. We leverage the Transformer [39] architecture to model the latent space distribution conditioned on image and point cloud observations, which are processed via ResNet [15] and Local PointNet [31] like architectures respectively.

where w_j is a weighting factor inversely proportional to the distance to the grid center v_j (cf our supplementary).

3.3. Model training

RIVQ-VAE Training Our model and codebook are trained end-to-end via the loss function:

$$\mathcal{L}(x;\phi,\psi,\mathcal{D}) = \mathcal{L}_r(o_x,\hat{o}_x) + \beta\mathcal{L}_{VQ}(Z,Z^q), \quad (8)$$

where $\hat{o}_x$ is the ground truth occupancy, β is a weighting parameter, $\mathcal{L}_r$ is the reconstruction binary cross-entropy loss and $\mathcal{L}_{VQ}$ denotes the vector quantization objective and the commitment loss [38].

Transformer Training Given an observation $\mathcal{O}$ such as partial point cloud shape or image view, we aim to model the distribution of plausible shape sequences to enable shape completion and single-reconstruction tasks respectively. Formally, given such an observation $\mathcal{O}$ of the target shape, the goal is to learn the distribution of the complete sequence $p(Z^q|\mathcal{O};\theta)$ where p designates the distribution of the discrete latent representation $Z^q \in \mathbb{R}^{R^3\times D}$ conditioned on $\mathcal{O}$, and θ denotes the learnable distribution parameters. We auto-regressively model $p(Z^q|\mathcal{O};\theta)$ such that the factorized sequence distribution can be written as follows:

$$p(Z^q|\mathcal{O};\theta) = \prod_{i=1}^{R^3} p(z_i^q|z_{<i}^q,\mathcal{O};\theta) \quad (9)$$

To learn $p(.;\theta)$, we adopt an encoder-decoder architecture illustrated in Figure 3. The observation $\mathcal{O}$ is first projected to an embedding space leading to a latent embedding e. Here we consider ResNet [15] and Local PointNet [31] like backbones to process image and point cloud observations respectively. Then, a Transformer [39] decoder takes e as a starting token to sequentially predict the complete target representation $Z^q = \{z_i^q\}_{1<i<R^3}$. The training objective maximizes the log-likelihood given $\mathcal{O}$ and Z^q:

$$\mathcal{L} = -log\ p(Z^q|\mathcal{O};\theta) \quad (10)$$

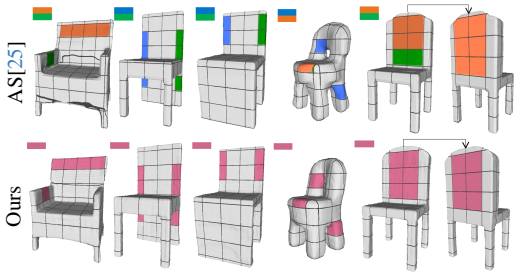

Figure 4. Embedding comparison. For each single shape, similarly colored patches are represented by the same codebook index. While baseline approach (first row) consumes multiple indices for representing similar rotated geometries, our method (second row) leverages these similarities for a more compact and efficient representation. In the supplementary, we illustrate the local rotation-invariance property of our approach across different categories.

Model	CD	EMD	F1
IMNet[6]	0.436	3.574	0.618
AutoSDF[25]	0.182	2.852	0.615
ShapeFormer-8[46]	0.195	2.924	0.564
Ours-PN	0.166	2.792	0.633
Ours*	0.126	2.640	0.692
Ours	**0.120**	**2.602**	**0.707**
ShapeFormer-16[46]	0.104	2.300	0.705
Ours-16	**0.072**	**2.162**	**0.768**

Table 1. Quantitative results for shape auto-encoding. Our approach outperforms baseline approaches by a large margin under similar settings. CD and EMD are multiplied by 10^2.

At inference time, given a partial point cloud or a single image view $\mathcal{O}$, the corresponding shape representation Z^q is sequentially sampled following the top-p [16] technique where the code index is selected from the set of indices whose probability sum exceeds a threshold p.

4. Experiments

All models are trained on 13 categories from ShapeNet [4] dataset (airplane, bench, cabinet, car, chair, display, lamp, speaker, rifle, sofa, table, phone and watercraft) using the train/test splits provided by [45].

4.1. Shape reconstruction

As a first experiment, we measure the reconstruction performance of our model to emphasize how our latent space can incorporate different geometric configurations to accurately represent test shapes. We compare our RIVQ-VAE to state-of-the-art global-based approach IM-Net [6], and local based approaches ShapeFormer-8 [46] and AutoSDF [25]. We use the common reconstruction metrics Chamfer distance (CD), Earth moving distance (EMD) and F-score%1 (F1) [35]. For a relevant evaluation, we compare the different methods using the baselines' settings for codebook size K, latent space resolution R and latent code size D. To this end, we report quantitative results in Ta-

Figure 5. Qualitative results for shape auto-encoding. Our method allows higher detailed reconstructions.

ble 1 for AutoSDF [25], ShapeFormer-8 [46] and RIVQ-VAE (ours) with K=512, R=8 and D=256, as well as for ShapeFormer-16 [46] and RIVQ-VAE-16 (ours-16) with K=4096, R=16 and D=128. As an ablation study, we train RIVQ-VAE without the pose estimation block (ours*) and with PointNet[31]-like encoder leading to non rotation-invariant embedding (ours-PN). For implicit function based approaches, we sample shapes at 128^3 resolution.

First, the comparison to the baseline IM-Net [6] confirms the higher-quality afforded by local-based representations due to their higher capacity in capturing surface details and their superior generalization power when handling different shape categories. Second, under similar settings, RIVQ-VAE outperforms baseline local-based methods across different metrics. This is further supported by qualitative results in Figures 1 and 5 (We provide qualitative results for $R = 16$ experiments in our supplementary). The obtained evaluations demonstrate that our canonicalized formulation allows the codebook to capture more diverse and high-detailed local geometries. For illustration, Figure 4 shows how our strategy allows to compactly represent similar geometries up to rigid transformations, which inherently leads to promote codebook diversity by better allocating its capacity. In the supplementary, we show additional illustrations of the discrete rotation invariance property of our approach across different categories. Moreover, we further highlight the superior expressive power of our codebook by evaluating the performance under a smaller codebook size.

Model	CD	EMD	F1
IMSVR[6]	0.761	4.433	0.487
AutoSDF [25]	0.629	4.481	0.430
ResNet2Occ	0.912	4.796	0.463
Ours	**0.569**	**4.000**	**0.506**

Table 2. Quantitative results for single-view reconstruction. Our method provides higher quality reconstructions. CD and EMD are multiplied by 10^2.

Finally, note that, as shown in Table 1 and Figure 5, adding the pose estimation block slightly enhances the performance of our approach, revealing that the convolutional and attention blocks efficiently handle the pose estimation. In what follows, we use RIVQ-VAE (ours).

4.2. Single-view reconstruction

We compare our approach on single view reconstruction of ShapeNet rendered images [9] with baseline methods: IMSVR [6], AutoSDF [25] and ResNet2Occ which refers to our approach where the ResNet [15] model is used to directly predict the shape sequence without using any shape prior. Since AutoSDF [25] and our method allow to sample multiple solutions for a single input image, we sam-

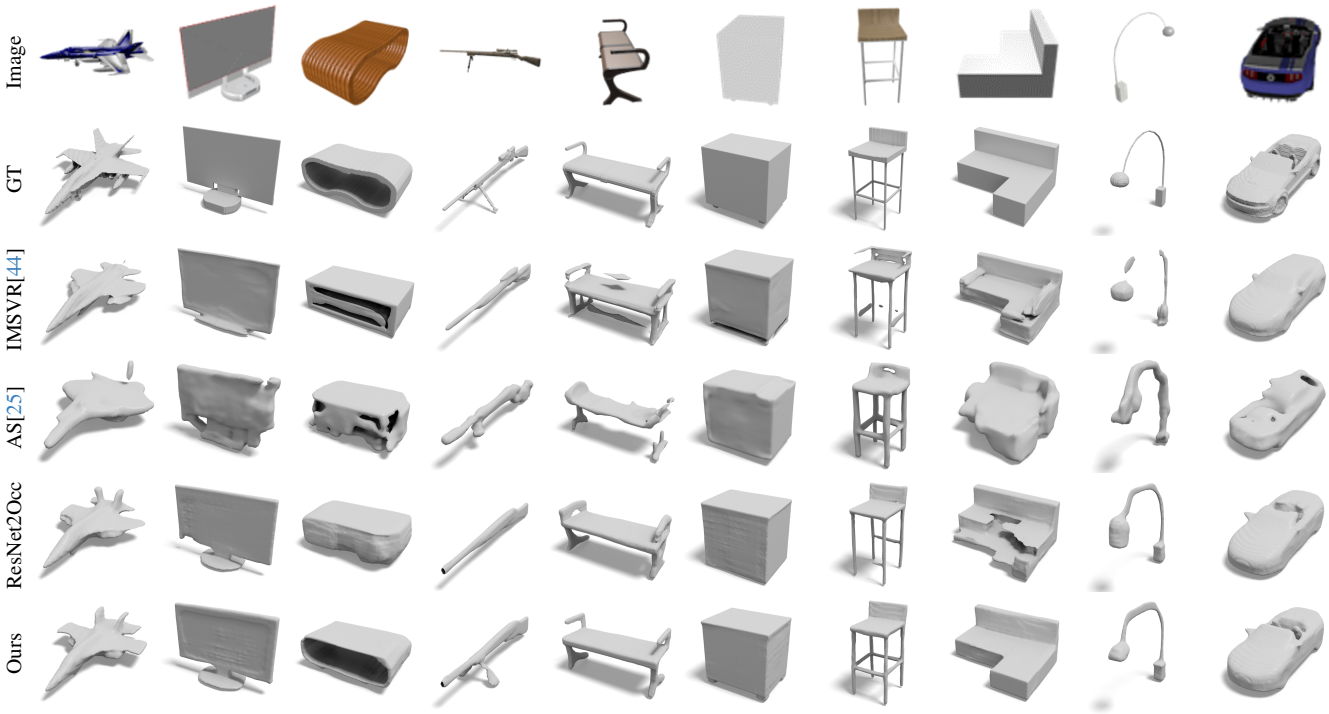

Figure 6. Qualitative results for single-view reconstruction. Our approach achieves higher accuracy compared to baselines' reconstructions.

Ambiguity	Low			High		
Model	TMD↑	UHD↓	MMD↓	TMD↑	UHD↓	MMD↓
PoinTr[48]	-	2.543	7.143	-	2.297	7.523
SF-16[46]	2.339	**1.152**	5.205	**4.194**	**1.425**	5.389
Ours	**3.199**	1.996	**4.858**	3.715	2.188	**4.892**

Table 3. Quantitative results for shape completion. Our method produces diverse shape completion candidates while ensuring higher geometric quality and plausibility. TMD, UHD and MMD are multiplied by $\times 10^2$, $\times 10^2$ and $\times 10^3$ respectively.

ple three different reconstructions for each and keep the one with the lowest CD for evaluation. Table 2 and Figure 6 show quantitative and qualitative results respectively. This study supports the superiority of our approach in producing high quality reconstructions across different metrics. We attribute this to the high quality representation enabled by RIVQ-VAE as highlighted in the reconstruction experiments above, and whose benefits extend to downstream tasks. Importantly, we demonstrate that our canonicalized latent space formulation can be efficiently leveraged for learning auto-regressive models, and that our decoder can accurately rotate and assemble diverse local sequences into plausible global geometries.

4.3. Shape completion

We evaluate our shape completion scheme on partial observations that are synthetically generated by cropping random regions from test shapes using two different ambiguity levels: low ambiguity where we crop 25 to 50% of the shape surface, and high ambiguity where we crop 50 to 75% of the shape surface.

We compare our generations against two baseline point-cloud completion methods: PoinTr [48] and ShapeFormer [46]. The former produces a single completion while the latter can produce multiple plausible shapes. Similarly to ours, both approaches use a single model to handle shape completion scenarios across different categories. We use ShapeFormer [46] model released by authors with R=16 trained on a different train/test split since we obtained lower performance when re-training their latent transformer model. We sample three completions per input partial shape for ShapeFormer [46] and our approach. For evaluation, we use the TMD (Total Mutual Difference) and UHD (Unidirectional Hausdorff Distance) metrics from [43] to respectively measure the diversity of the generated shapes and their faithfulness toward input, as well as the Minimal Matching Distance (MMD) with respect to train shapes [1] to capture the plausibility and quality of completions.

The quantitative and qualitative results are shown in Table 3 and Figure 7 respectively. For low ambiguity experiments, our method achieves better diversity (higher TMD) while ensuring better geometric quality and plausibility (lower MMD). Note that ShapeFormer [46] yields closer shape completion results to partial input (lower UHD). We attribute this to the lower diversity of the generated content that inherently leads to less geometric shift. For high ambi-

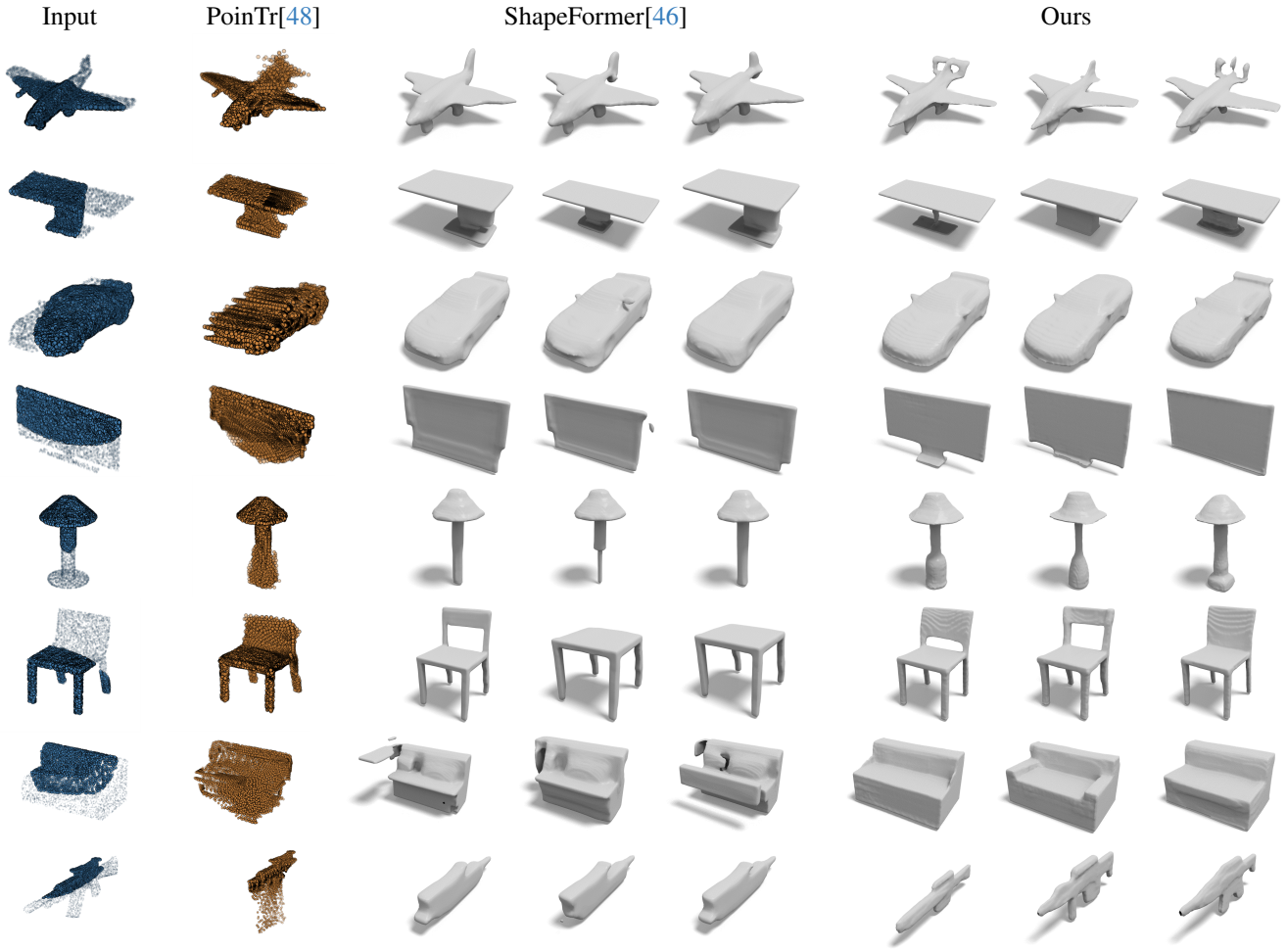

Figure 7. Qualitative results for shape completion. Given an input point cloud with a cropped surface (in light blue), we visualize shape completions using different methods. Rows 1 to 4 correspond to low ambiguity input. Rows 4 to 8 correspond to high ambiguity input. Our approach yields diverse generations, while providing plausible shapes.

guity, ShapeFormer [46] achieves the best TMD. This high completion diversity is however achieved with lower MMD, or equivalently, with lower plausibility with respect to the target category than our approach. The 6th and 7th rows from Figure 7 illustrates examples where ShapeFormer [46] completions diversity is achieved at the expense of the faithfulness toward target category, which is reflected by the MMD. Our approach, however, provides diverse completions while better preserving the target shape structure.

5. Conclusion

We have presented RIVQ-VAE a generative model that encodes shape geometry into a local, discrete and rotation- and translation-invariant representations sampled from a learnable codebook. Our key technical novelty is the handling of this canonicalized formulation to provide high quality local geometries while ensuring global consistency. We achieved this by a careful architecture design that disentangle local rotation estimation from surface prediction. When combined with Transformer [39] to learn to sample plausible shape code sequences from the learned codebook, our approach allows multiple generative tasks including shape completion and single-view reconstruction.

A key limitation of RIVQ-VAE is that capturing local canonical similarities is, similarly to rotation-invariant point cloud processing methods, closely tied to the quality of the input patches. In the future, we plan to combine RIVQ-VAE with a symmetry regularization that is expected to guide the learning process towards optimal embeddings. Besides, we believe that introducing a global consistency regularization loss may further enhance surface smoothness.

Acknowledgements We thank the anonymous reviewers for their valuable feedback and suggestions. We also thank Kawtar Zaher for her contribution to the experiments section. Parts of this work were supported by the ERC Starting Grant No. 758800 (EXPROTEA) and the ANR AI Chair AIGRETTE.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J Guibas. Learning representations and generative models for 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2017. 2, 7

[2] Jan Bechtold, Maxim Tatarchenko, Volker Fischer, and Thomas Brox. 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 (CVPR)*, pages 15880–15889, 2021. 1, 2

[3] Rohan Chabra, Jan E. Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local SDF priors for detailed 3d reconstruction. *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[4] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. Technical Report 1512.03012, arXiv preprint, 2015. 5

[5] Chao Chen, Guanbin Li, Ruijia Xu, Tianshui Chen, Meng Wang, and Liang Lin. Clusternet: Deep hierarchical cluster network with rigorously rotation-invariant representation for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[6] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 6

[7] Zhang Chen, Yinda Zhang, Kyle Genova, Thomas Funkhouse, Sean Fanello, Sofien Bouaziz, Christian Haene, Ruofei Du, Cem Keskin, and Danhang Tang. Multiresolution Deep Implicit Functions for 3D Shape Representation. In *2021 IEEE/CVF International Conference on Computer Vision*. IEEE, 2021. 2

[8] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[9] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 6

[10] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance, 2022. 2

[11] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 1, 2

[12] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so(3) equivariant representations with spherical cnns. 2018. 3

[13] Carlos Esteves, Ameesh Makadia, and Kostas Daniilidis. Spin-weighted spherical cnns. In *Advances in Neural Information Processing Systems*, pages 8614–8625. Curran Associates, Inc., 2020. 3

[14] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6

[16] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. 5

[17] Moritz Ibing, Isaak Lim, and Leif Kobbelt. 3d shape generation with grid-based implicit functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13559–13568, 2021. 1, 2

[18] Chiyu Max Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[19] Chieh-Hsin Lai, Dongmian Zou, and Gilad Lerman. Robust vector quantized-variational autoencoder. *CoRR*, abs/2202.01987, 2022. 2

[20] Leon Lang and Maurice Weiler. A wigner-eckart theorem for group equivariant convolution kernels. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[21] Leon Lang and Maurice Weiler. A Wigner-Eckart Theorem for Group Equivariant Convolution Kernels. In *International Conference on Learning Representations*, 2021. 4

[22] Xianzhi Li, Ruihui Li, Guangyong Chen, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. A rotation-invariant framework for deep point cloud analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 3

[23] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 3

[24] James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Understanding posterior collapse in generative latent variable models, 2019. 1, 2

[25] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. AutoSDF: Shape priors for 3d completion, reconstruction and generation. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7

[26] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas Guibas. Structurenet: Hierarchi-

cal graph networks for 3d shape generation. *ACM Transactions on Graphics (TOG), Siggraph Asia 2019*, 38(6):Article 242, 2019. 2

[27] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[28] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[29] Adrien Poulenard and Leonidas J. Guibas. A functional approach to rotation equivariant non-linearities for tensor field networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13174–13183, 2021. 4

[30] Adrien Poulenard, Marie-Julie Rakotosaona, Yann Ponty, and Maks Ovsjanikov. Effective rotation-invariant point cnn with spherical harmonics kernels. In *2019 International Conference on 3D Vision (3DV)*, pages 47–56. IEEE, 2019. 3

[31] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3, 5, 6

[32] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5099–5108. Curran Associates, Inc., 2017. 3

[33] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 2

[34] Rahul Sajnani, Adrien Poulenard, Jivitesh Jain, Radhika Dua, Leonidas J. Guibas, and Srinath Sridhar. Condor: Self-supervised canonicalization of 3d pose for partial shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[35] Maxim Tatarchenko*, Stephan R. Richter*, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? 2019. 5

[36] Nathaniel Thomas, Tess E. Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. *CoRR*, abs/1802.08219, 2018. 3, 4

[37] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Carsten Stoll, and Christian Theobalt. Patchnets: Patch-based generalizable deep implicit 3d shape representations. *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[38] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017. 1, 2, 4, 5

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1, 2, 3, 4, 5, 8

[40] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019. 3

[41] Yixin Wang, David Blei, and John P Cunningham. Posterior collapse and latent variable non-identifiability. In *Advances in Neural Information Processing Systems*, pages 5443–5455. Curran Associates, Inc., 2021. 1, 2

[42] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 3

[43] Rundi Wu, Xuelin Chen, Yixin Zhuang, and Baoquan Chen. Multimodal shape completion via conditional generative adversarial networks. In *The European Conference on Computer Vision (ECCV)*, 2020. 7

[44] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 7

[45] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 5

[46] Xingguang Yan, Liqiang Lin, Niloy J. Mitra, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 5, 6, 7, 8

[47] Shun Yao, Fei Yang, Yongmei Cheng, and Mikhail G. Mozerov. 3d shapes local geometry codes learning with sdf. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2110–2117, 2021. 2

[48] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *ICCV*, 2021. 7, 8

[49] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. *arXiv preprint arXiv:2205.13914*, 2022. 3

[50] Zhiyuan Zhang, Binh-Son Hua, David W. Rosen, and Sai-Kit Yeung. Rotation invariant convolutions for 3d point clouds deep learning. In *International Conference on 3D Vision (3DV)*, 2019. 3

[51] Zhiyuan Zhang, Binh-Son Hua, Wei Chen, Yibin Tian, and Sai-Kit Yeung. Global context aware convolutions for 3d point cloud understanding. pages 210–219, 2020. 3