# Variational Point Encoding Deformation for Dental Modeling

Johan Ziruo Ye [1 2] Thomas Ørkild [1] Peter Lempel Søndergaard [1] Søren Hauberg [2]

## Abstract

We introduce *VF-Net*, a probabilistic extension of *FoldingNet*, for learning representations of point cloud data. VF-Net overcomes the limitations of existing models by incorporating a 1-to-1 mapping between input and output points. By eliminating the need for Chamfer distance optimization, this approach enables the development of a fully probabilistic model. We demonstrate that VF-Net outperforms other models in dental reconstruction tasks, including shape completion and tooth wear simulation. The learned latent representations exhibit robustness and enable meaningful interpolation between dental scans.

## 1. Introduction

Recent advances has lead to large scale adoption of intraoral dental scanners. Our research is motivated by the need to analyze, search, and organize large collections of such dental scans. These 3-dimensional dental mesh models are used for surgical planning, tooth crown generation, tooth wear estimation, etc. The sensitivity of the such tasks necessitates robustness to noisy data and feedback on model uncertainty to the responsible dentist. Treating these meshes as point clouds enables us to efficiently represent the shape and topology of patients' teeth using a sparse set of points, leading to improved computational efficiency. However, consequently any modeling of point clouds must be invariant to any reordering and variability in cardinality that may be present. The foundation of our paper is a new dataset, the *FDI 16 Tooth Dataset*, which provides a large collection of dental scans. Our primary objective is to learn useful and reliable representations of this data. However, in our paper, we also highlight other crucial tasks such as shape completion of the sides of the tooth unable to be scanned by the intraoral scanner, as well as shape completion of areas previous obstructed by braces or other orthodontic devices.

**Point cloud representation learning for teeth.** As most teeth both move and degrade continuously with time, it is reasonable to seek a continuous vectorial representation when organizing extensive collections of dental scans. This motivates the development of autoencoder-style representation learning models for point cloud data (Rumelhart et al., 1986; Yang et al., 2018). An obvious candidate model is *FoldingNet* (Yang et al., 2018), which reconstructs the original point cloud by deforming points from a 2D plane, as it shares topology with the FDI 16 dataset. FoldingNet and other encoder-decoder models for point clouds reconstruct input point clouds using a learned vectorial representation by minimizing reconstruction error. Since there are no correspondences between points in two given clouds, permutation invariant metrics are used, with the *Chamfer distance* (Barrow et al., 1977) being popular,

$$\begin{aligned}\mathrm{CD}(\mathbf{X},\mathbf{Y}) = \frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}}\min_{\mathbf{y}\in\mathbf{Y}}\|\mathbf{x}-\mathbf{y}\|_2 \\ + \frac{1}{|\mathbf{Y}|}\sum_{\mathbf{y}\in\mathbf{Y}}\min_{\mathbf{x}\in\mathbf{X}}\|\mathbf{y}-\mathbf{x}\|_2\end{aligned} \tag{1}$$

for point clouds $\mathbf{X}$ and $\mathbf{Y}$. Although this metric solves the invariance problem, it poses a new one: *The Chamfer distance* (1) *does not lead to a likelihood, preventing its use in probabilistic models.* For instance, when used in the Gaussian distribution, the function $\mathbf{X} \mapsto {}^1\!/_c \exp(-\mathrm{CD}^2(\mathbf{X},\mu))$ cannot be normalized to have unit integral due to the explicit minimization in Eq. 1. We need robustness to noise and general quantification of uncertainty, and consider the lack of an explicit likelihood detrimental.

**In this paper**, we propose a new architecture that allows us to sidestep the use of Chamfer distances, which, in turn, allow for straight-forward constructions of models akin to *variational autoencoders (VAEs)* (Kingma & Welling, 2014; Rezende et al., 2014). We call the resulting model the *Variational FoldingNet* (Sec. 3), as it bridges FoldingNet (Yang et al., 2018) and VAEs. A key aspect of our model is that it avoids the usage of Chamfer distances, and instead relies on a more appropriate encoder. Moreover, we contribute a new dataset of dental scans (Sec. 2) of the first maxillary molar tooth on the right side of the upper jaw[1] — one of the most common teeth to receive dental treatment/restoration. Using

[1]3Shape A/S [2]Technical University of Denmark. Correspondence to: Johan Ziruo Ye <Johan.Ye@3shape.dk>.

*Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] Referred to as *FDI 16* according to the ISO 3950 notation.

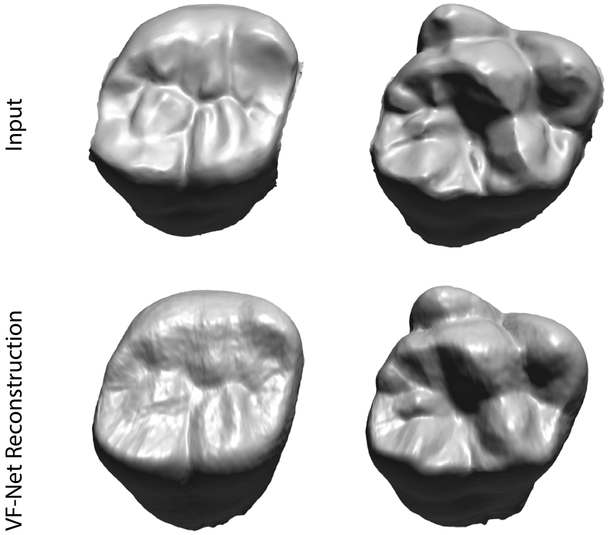

Figure 1: Top row: two teeth from the FDI 16 dataset. Bottom row: reconstructions from VF-Net.

this dataset, we explore keys tasks, such as shape completion in cases where neighboring teeth obstruct the view or shape completion impeded by orthodontic treatment. We also showcase the potential for future tasks in representation learning and style transfer. Finally, we demonstrate that for dental scans, our model performs superior or competitively on various standardized generative modeling tasks when compared to current state-of-the-art models (Sec. 4).

## 2. The FDI 16 Tooth Dataset

We release the FDI 16 dataset, which is a collection of 6,309 irregular anonymous meshes that were collected from intraoral scans by 3Shape[2]. Each tooth in the FDI 16 Tooth dataset is algorithmically segmented and oriented from a scan of an upper jaw. These meshes are from patients undergoing aligner treatment, biasing the data towards younger individuals, who generally have fewer restorations and dental problems. As a result, aligner attachments may be observed in the data. The top row of Fig. 1 displays two such meshes.

The FDI 16 teeth meshes were collected using TRIOS scanners, primarily the TRIOS 3 model. These meshes all share highly similar topologies, so the main differences between them are in their shapes. All teeth have clearly defined boundaries and are consequently open with no representation of the interior object volume. We have made the meshes publicly available. All teeth have been rotated to ensure that the $x$-axis is turned towards the neighboring tooth (FDI 17), while the $y$-axis points in the occlusal direction (direction of the biting surface). Finally, the $z$-axis is given by the cross-product to ensure a right-hand coordinate system. The scale of the data is in millimeters.

## 3. Variational Inference on Point Clouds

**Background: PointNet and FoldingNet.** As stated earlier, point clouds are sets of points with varying size and arbitrary order, and models thereof should unaffected by such changes to the point cloud. Therefore, one of the primary approaches to address such data is to develop neural networks that are invariant to such changes (Qi et al., 2017; Yang et al., 2018). Unfortunately, when it comes to the variational autoencoder, this is not possible with current designs. A variational autoencoder outputs a distribution for each element in which the corresponding input element is evaluated (Kingma & Welling, 2014; Rezende et al., 2014). Current models lack a 1-to-1 connection, resulting in an output permutation that is unlikely to match the input permutation, preventing such aforementioned evaluation. Consequently, any modeling of data with more complicated distributions fails.

FoldingNet becomes invariant to changes in point cloud permutation and cardinality by using a PointNet-like encoder, $e$, which entails utilizing multi-layer perceptrons (MLPs) that operate independently on each point of the point cloud. The folding-based decoder, $\mu : \mathcal{Z} \times \mathbb{R}^2 \to \mathbb{R}^3$, is composed of two MLPs. These are applied to the latent code, $\mathbf{z}$, concatenated with each point in the chosen latent point encoding shape, $\mathcal{G} = \{\mathbf{g}_i\}_{i=1}^{I}$, which in our case is the two-dimensional planar patch $[-1, 1]^2$ (Yang et al., 2018). The folding of the planar patch, $\{\mu(\mathbf{g}_i)\}_{i=1}^{N}$, is determined by the parameter vector $\mathbf{z}$ predicted by the PointNet encoder $e$. Both the encoder $e$ and the decoder $\mu$ are jointly trained to minimize the reconstruction error

$$\mathcal{E} = \sum_{n=1}^{N} \|\mathbf{x}_n - \mathrm{proj}_S(\mathbf{x}_n)\|^2, \tag{2}$$

where $\mathrm{proj}_S : \mathbb{R}^3 \to \mathbb{R}^3$ denotes the projection of a point $\mathbf{x}$ onto the surface spanned by $S = \mu(\mathcal{G})$,

$$\mathrm{proj}_S(\mathbf{x}) = \mu(\hat{\mathbf{g}}) \text{ where } \hat{\mathbf{g}} = \arg\min_{\mathbf{g} \in \mathcal{G}} \|\mathbf{x} - \mu(\mathbf{g})\|^2. \tag{3}$$

This creates a permutation invariant and cardinality invariant autoencoder. FoldingNet approximates this projection during training using Chamfer distances (1).

### 3.1. The Variational FoldingNet

First, we describe the generative process of our proposed *Variational FoldingNet (VF-Net)* and then cover approximate inference and training. Let $p(\mathbf{z})$ be a Normalizing flow prior over the parameters describing the *shape* of an object (Kingma et al., 2017). As FoldingNet, we limit our flat mesh grid $\mathcal{G}$ to be within the planar patch $[-1, 1]^2$. This grid is, as in FoldingNet, subsequently deformed according to $\mathbf{z}$. Let $\mathbf{g} \in \mathcal{G}$ denote a point on this grid, then the corresponding three-dimensional point $\mathbf{x}$ is defined to be distributed as $p(\mathbf{x} \mid \mathbf{z}, \mathbf{g}) = \mathcal{N}(\mathbf{x}|\mu(\mathbf{z}, \mathbf{g}), \sigma^2(\mathbf{z}, \mathbf{g})\mathbf{I})$, where $\mu : \mathcal{Z} \times \mathbb{R}^2 \to \mathbb{R}^3$ and $\sigma^2 : \mathcal{Z} \times \mathbb{R}^2 \to \mathbb{R}_+$ are neural networks.

[2] https://www.3shape.com/

In this model, new samples can be generated by first sampling $\mathbf{z}$ and then mapping the grid points through $\mu$ and $\sigma$,

$$\mathbf{x} = \mu(\mathbf{z}, \mathbf{g}) + \sigma(\mathbf{z}, \mathbf{g}) \cdot \epsilon, \qquad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

This defines the likelihood $p(\mathbf{x}) = \int p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) \mathrm{d}\mathbf{z}$ which gives a training objective. Unfortunately, the integral is intractable, and approximations are necessary. Following conventional variational inference (Kingma & Welling, 2014; Rezende et al., 2014), a lower bound on $p(\mathbf{x})$ is

$$\mathcal{L}(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[ \log p(\mathbf{x} \mid \mathbf{z}) \right] - \mathrm{KL}\left( q(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z}) \right) \tag{5}$$

where $q(\mathbf{z} \mid \mathbf{x})$ is any approximation to $p(\mathbf{z} \mid \mathbf{x})$. To evaluate Eq. 5, we first introduce a projection $\mathrm{proj}_{\hat{\mathcal{G}}}(\mathbf{x}) : \mathbb{R}^3 \to \mathcal{G}$ modeled with a neural network. We optimize Eq. 3, where the introduction of $\mathrm{proj}_{\hat{\mathcal{G}}}$ means that $\hat{\mathcal{G}}$ is no longer independent and can now be optimized together with $\mu$. $\hat{\mathcal{G}}$ are our latent point encodings, which give a 1-to-1 mapping throughout the network, allowing for evaluation of the ELBO (5), see suppl. Fig. S1. We use a multivariate normal distribution with isotropic variance as the reconstruction term in the ELBO. This would not be possible with the Chamfer distance as it does not have a distributional counterpart as the associated normalization constant does not exist. The result is a novel method of evaluation for 3D reconstruction networks, which is both probabilistic and avoids the computationally expensive Chamfer distance. Supplementary Fig. S2 demonstrate that our approach can effectively replace Chamfer distances.

## 4. Experimental results

**Limitations.** As a baseline, we sanity-checked VF-Net by reconstructing the airplanes from ShapeNet (Chang et al., 2015). On this dataset, it is clear from the supp. Table S1 that VF-Net and FoldingNet perform great in terms of reconstruction (Kim et al., 2021; Luo & Hu, 2021), but VF-Net poorly samples new meshes. This is due to information on the shape being stored in the latent point encodings, see suppl. Fig. S3, which potentially could be alleviated with a flow or diffusion prior, similar to LION (Zeng et al., 2022). As this is not our focus, we have not pursued such.

**FDI 16 Tooth Data.** We evaluated the reconstruction on FDI 16 using Chamfer distances and earth mover's distance (Rubner et al., 2000), see Table 1. On the FDI 16 data, VF-Net outperforms its peers both when measured using both metrics above. Notably, Point-Voxel Diffusion (PVD) (Zhou et al., 2021) performs poorly in reconstruction as it cannot return the same tooth when embedded. Instead, it returns a randomly sampled tooth. Despite having the lowest reconstruction error, VF-Net's remain overly smooth and lack the desired level of detail, see supp. Fig. S4. A common behavior observed in variational autoencoders (Kingma & Welling, 2014; Vahdat & Kautz, 2021; Tolstikhin et al., 2019).

Table 1: Chamfer distances (CD) and earth mover's distances (EMD) are multiplied by 100. Lower values indicate better reconstruction performance. Bracket sim and Gap sim are untrained extrapolation performances of the models.

| Method | FDI 16 Tooth | | Bracket | Gap |
|---|---|---|---|---|
| | **CD** | **EMD** | **CD** | **CD** |
| DPC | 8.46 | 41.38 | 11.38 | 14.40 |
| SetVAE | 19.86 | 57.52 | 12.28 | 14.16 |
| PVD | 119.62 | 835.82 | 18.10 | 20.32 |
| LION | 5.44 | 22.29 | — | — |
| FoldingNet | 5.25 | 33.59 | 95.38 | 143.56 |
| VF-Net | **1.20** | **6.24** | **5.79** | **4.81** |

**Variance Estimation for Point Clouds.** The relative predicted variance has been visualized in Fig. 2, where red indicates a higher variance and green indicates a lower variance. Interestingly, the network assigns higher variance to the fifth cusp and to aligner attachments. Given that both features are only observed in a subset of individuals, it is natural that these areas would exhibit higher levels of uncertainty. The occlusal surface consistently exhibits a moderate amount of variance.

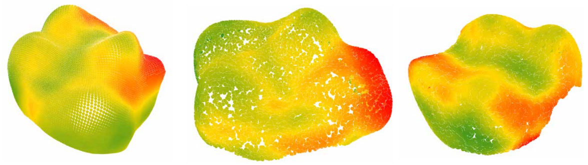

Figure 2: Visualization of VF-Net's predicted relative variance effectively highlights areas of high and low variance, denoted by the colors red and green respectively.

**Model Sampling Performances.** We evaluate our generative performance using metrics proposed by Yang et al. (2019), including the Minimum Matching Distance (MMD), Coverage (COV), and 1-nearest neighbor accuracy (1-NNA). MMD measures the average distance to the nearest neighbor point cloud, while COV quantifies the fraction of point clouds in the ground truth test set that are considered nearest neighbors to the generated samples. The 1-NNA metric utilizes a 1-nearest neighbor classifier to determine if a sample is generated or from the ground truth dataset, with a 50% accuracy threshold indicating the data are indistinguishable.

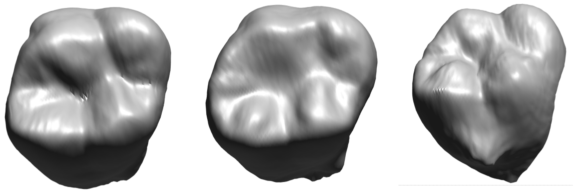

Figure 3: Samples produced by VF-Net. The right-most tooth has been generated with an aligner attachments

Table 2: Sampling performances on FDI 16 Tooth data.

| METHOD | MMD(↓) | | COV(%↑) | | 1-NNA(%↓) | |
|---|---|---|---|---|---|---|
| | CD | EMD | CD | EMD | CD | EMD |
| DPC | 14.79 | 2.40 | 0.068 | 0.14 | 100 | 100 |
| SETVAE | 0.40 | 0.67 | 10.31 | 9.35 | 98.50 | 98.46 |
| PVD | **0.21** | **0.40** | 43.48 | **44.16** | **61.87** | **60.79** |
| LION | 0.22 | 0.44 | **43.89** | 43.83 | 68.67 | 65.19 |
| VF-NET | **0.21** | 0.52 | 41.37 | 31.33 | 67.65 | 73.69 |

A few examples of generated FDI 16 teeth can be seen in figure 3, and generated teeth across all teeth can be found in supplementary figure S5. The sampling performances of the models can be found in Table 2. VF-Net demonstrates superior performance compared to DPC and SetVAE. However, there it is still slightly behind in sampling performance when compared to PVD and LION. (Zhou et al., 2021; Zeng et al., 2022). As our primary task is reconstruction and extrapolation, our performances reflect equally. Due to the shape of the latent point encoding and how it may vary, see Fig. S3, sampling a grid shape each time may perform poorly according to the metrics. Thus, we trained a minor network equating to one fold of VF-Net to output a point encoding from the latent representation. We emphasize that this is completely unnecessary for sampling in general. However, sampling naïvely may lead to systematic sharp edges/corners in the sampled point clouds, which would be detrimental to performance when measured using established metrics due to only the model outputs having such artifacts.

**Simulated Shape Completion.** In dental reconstruction, inferring the obstructed side of a tooth and reconstructing the tooth surface beneath braces' brackets are key challenges. However, such paired data is exceedingly rare. Therefore, it is valuable to develop a model capable of extrapolating these surfaces without explicit training on such data.

As we lack a ground truth paired meshes, we instead simulate the extrapolation on the test set. This is done by sampling a point outside the tooth and deleting the 200 nearest neighbors to that point. An example of the synthetic holes is shown in supp. Fig. S6. The point encodings of the point cloud with a hole in the side and the one without are highly similar. As such, extrapolation can be done by sampling in the point encoding space. For this evaluation, we sample a higher number of points in the latent encoding and calculate the distance from the deleted points to their nearest neighbor in the completed point cloud.

The performance averaged across the test set can be found in Table 1. Bracket and gap denoted simulated braces bracket removal and gap between teeth removal. Here, VF-Net outperforms its peers. Due to its reliance on continuous grid deformation, FoldingNet performs subpar when an area is missing. On the other hand, we were unable to perform shape completion using LION. We attempted to use the latent points from the original tooth, however, as it contained information about the shape and rendered a fair comparison infeasible. Finally, we compared our results to PVD when trained to complete shapes. PVD achieved an reconstruction chamfer distance of 0.97 and 0.74 for simulated bracket holes and gap holes, respectively. Naturally, a trained model for shape completion tasks outperforms its peers attempting untrained shape completion. During shape completion without explicit training, PVD's performance was not as impressive.

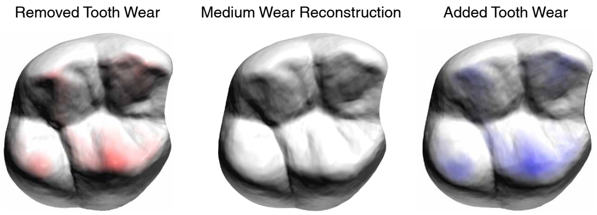

Figure 4: The effect of moving in the direction of tooth wear in the latent space. *Left*: Highlighted in red are areas that have grown. *Middle*: The original reconstruction. *Right*: Areas depicted in blue are lower than the original mesh.

**Representation Learning.** We compared our latent representations to those of FoldingNet, as it is the comparison model with the most interpretable latent variables. To explore the latent space, we manipulated the latent representations by adding and removing tooth wear in this space, see Fig. 4. To determine the direction of tooth wear, we calculated the average directional change in latent representations when encoding 10 teeth with synthetically induced wear, see suppl. Fig. S7. We observe behavior highly similar to the teeth with sculpted tooth wear.

Table 3: The percentage of teeth with the expected increase in classification prediction when pushing the latent representation towards or away from the tooth wear direction. L, M, and H denote light, medium, and heavy wear respectively.

| METHOD | L | M | H |
|---|---|---|---|
| FOLDINGNET | 94.95% | 91.77% | 97.8% |
| VF-NET (OURS) | **96.70%** | **95.71%** | **98.68%** |

To quantify the performance, we trained a small PointNet model (Qi et al., 2017) on a proprietary dataset of 1400 teeth that were annotated with light/medium/heavy tooth wear. Subsequently, we evaluated the latent representations by determining if a movement in the latent space along the tooth wear direction led to the expected change in the PointNet's predicted classification. In Table 3, each class denotes the base class prior to adding/removing tooth wear. To light and heavy we attempted to add and remove tooth wear respectively, while medium tooth wear teeth were evaluated both when adding/removing wear. The findings presented in Table 3 indicate VF-Net's latent representations show greater robustness.

## 5. Conclusion

We propose VF-Net, a fully probabilistic point cloud model closely resembling variational autoencoders. The novel contribution lies in the latent point encodings, which replaces the Chamfer distance and enables working with probability densities. Our experiments, including the reconstruction and extrapolation on FDI 16 Tooth data, showcase the effectiveness of VF-Net. Furthermore, VF-Net demonstrates robust representations for interpolation and modification of reconstructions.

## Acknowledgements

This work was funded by the Innovation Fund Denmark (1044-00172B) and the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). It also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research, innovation programme (757360). SH was supported in part by a research grant (42062) from VILLUM FONDEN.

# References

Barrow, H. G., Tenenbaum, J. M., Bolles, R. C., and Wolf, H. C. Parametric Correspondence and Chamfer Matching: Two New Techniques for Image Matching. August 1977. URL https://openreview.net/forum?id=rkb6wXfdWB.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. ShapeNet: An Information-Rich 3D Model Repository, December 2015. URL http://arxiv.org/abs/1512.03012. arXiv:1512.03012 [cs].

Kim, J., Yoo, J., Lee, J., and Hong, S. SetVAE: Learning Hierarchical Composition for Generative Modeling of Set-Structured Data, March 2021. URL http://arxiv.org/abs/2103.15619. arXiv:2103.15619 [cs].

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, May 2014. URL http://arxiv.org/abs/1312.6114. arXiv: 1312.6114.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improving Variational Inference with Inverse Autoregressive Flow, January 2017. URL http://arxiv.org/abs/1606.04934. arXiv:1606.04934 [cs, stat].

Luo, S. and Hu, W. Diffusion Probabilistic Models for 3D Point Cloud Generation, June 2021. URL http://arxiv.org/abs/2103.01458. arXiv:2103.01458 [cs].

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *arXiv:1612.00593 [cs]*, April 2017. URL http://arxiv.org/abs/1612.00593. arXiv: 1612.00593.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv:1401.4082 [cs, stat]*, May 2014. URL http://arxiv.org/abs/1401.4082. arXiv: 1401.4082.

Rubner, Y., Tomasi, C., and Guibas, L. J. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000. ISSN 1573-1405. doi: 10.1023/A:1026543900054. URL https://doi.org/10.1023/A:1026543900054.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. ISSN 1476-4687. doi: 10.1038/323533a0. URL https://www.nature.com/articles/323533a0. Number: 6088 Publisher: Nature Publishing Group.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein Auto-Encoders, December 2019. URL http://arxiv.org/abs/1711.01558. arXiv:1711.01558 [cs, stat].

Vahdat, A. and Kautz, J. NVAE: A Deep Hierarchical Variational Autoencoder, January 2021. URL http://arxiv.org/abs/2007.03898. arXiv:2007.03898 [cs, stat].

Yang, G., Huang, X., Hao, Z., Liu, M.-Y., Belongie, S., and Hariharan, B. PointFlow: 3D Point Cloud Generation with Continuous Normalizing Flows, September 2019. URL http://arxiv.org/abs/1906.12320. arXiv:1906.12320 [cs].

Yang, Y., Feng, C., Shen, Y., and Tian, D. FoldingNet: Point Cloud Auto-encoder via Deep Grid Deformation. *arXiv:1712.07262 [cs]*, April 2018. URL http://arxiv.org/abs/1712.07262. arXiv: 1712.07262.

Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., and Kreis, K. LION: Latent Point Diffusion Models for 3D Shape Generation, October 2022. URL http://arxiv.org/abs/2210.06978. arXiv:2210.06978 [cs, stat].

Zhou, L., Du, Y., and Wu, J. 3D Shape Generation and Completion through Point-Voxel Diffusion, August 2021. URL http://arxiv.org/abs/2104.03670. arXiv:2104.03670 [cs].

# A. Supplementary Material

## A.1. Model Architecture

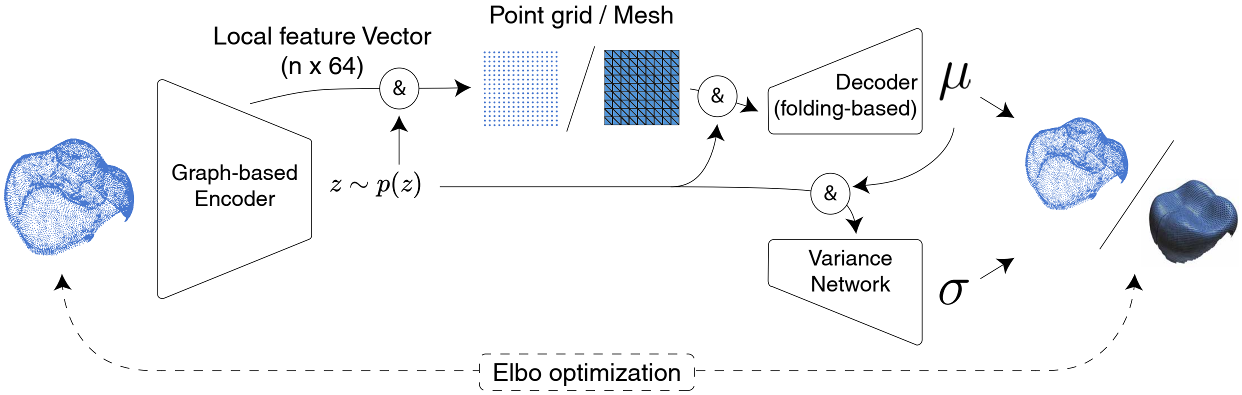

Figure S1: The architecture of VF-Net closely resembles that of FoldingNet, with minor modifications. Notably, a novel addition in VF-Net is the inclusion of a latent point encoding. This encoding allows for a 1:1 mapping throughout the network while maintaining invariance with respect to permutation and cardinality. Furthermore, a variance prediction network has been incorporated to estimate the variance at each output point. Note that ”&” denotes concatenation.

## A.2. Chamfer vs Euclidean

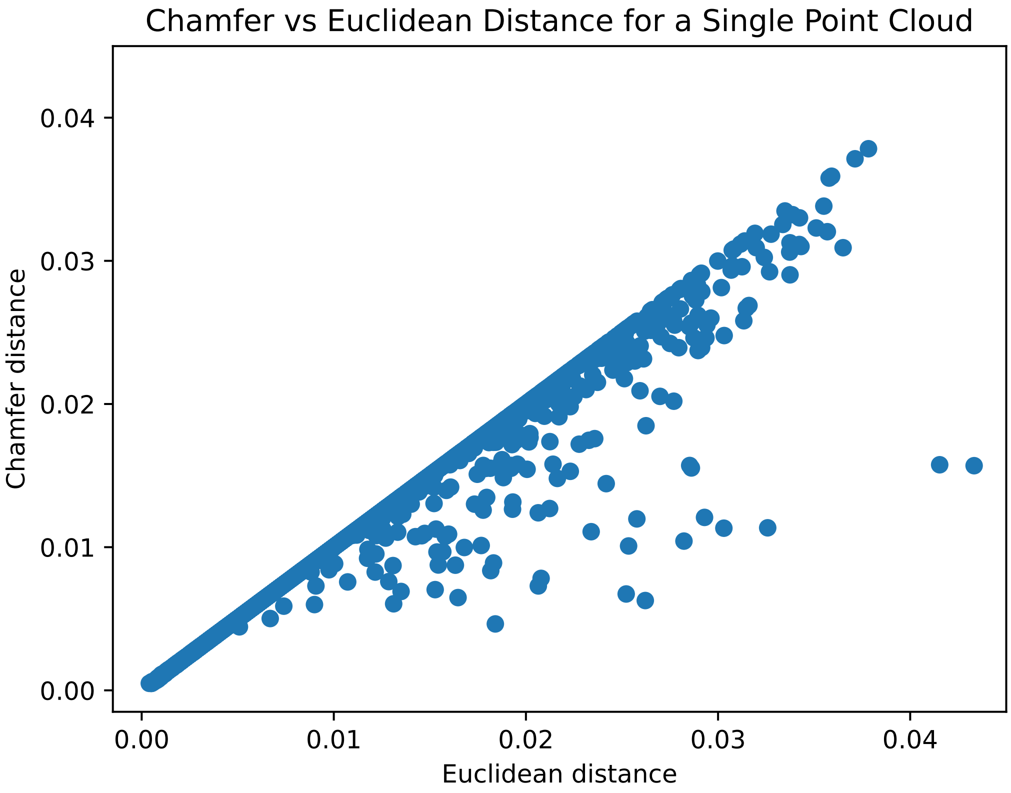

Figure S2: In the above plot, we observe that the euclidean distance acts as an upper bound for the chamfer distance. In the majority of cases, when optimized using VF-Net, these distances are identical. This empirical observation provides support for our claim that the Chamfer distance can be effectively substituted with an appropriate encoder choice.

### A.3. ShapeNet Reconstruction Performances

Table S1: Both Chamfer distances (CD) and earth mover's distances (EMD) are multiplied by 1000, and for both, lower values indicate better reconstruction performance.

| Method | ShapeNet Airplanes | |
|---|---|---|
| | **CD** | **EMD** |
| DPC | 0.18 | 47.82 |
| SetVAE | 0.14 | 30.60 |
| PVD | 3.12 | 90.45 |
| LION | 0.061 | 10.19 |
| FoldingNet | 0.079 | 31.47 |
| VF-Net (ours) | **0.031** | **7.31** |

### A.4. Latent Point Encodings

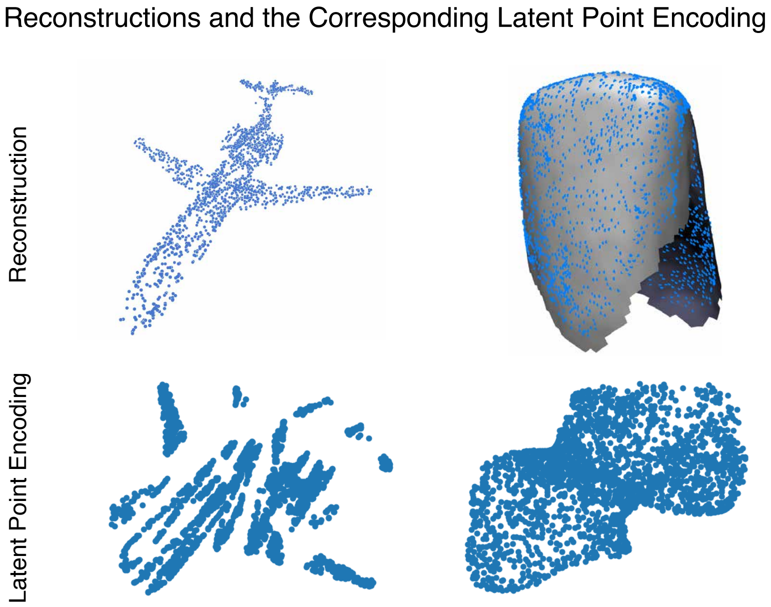

Figure S3: *Left*: While the airplane from ShapeNet is accurately reconstructed, it poses a challenge in terms of sampling due to its non-continuous distribution in the latent point encoding. *Right:* An incisor and its corresponding point encodings. Notably, the encoded points correctly reflects the missing sides of the tooth. Sampling and decoding from this region of the latent point encodings enables extrapolation in 3D space.

## A.5. VF-Net reconstructions

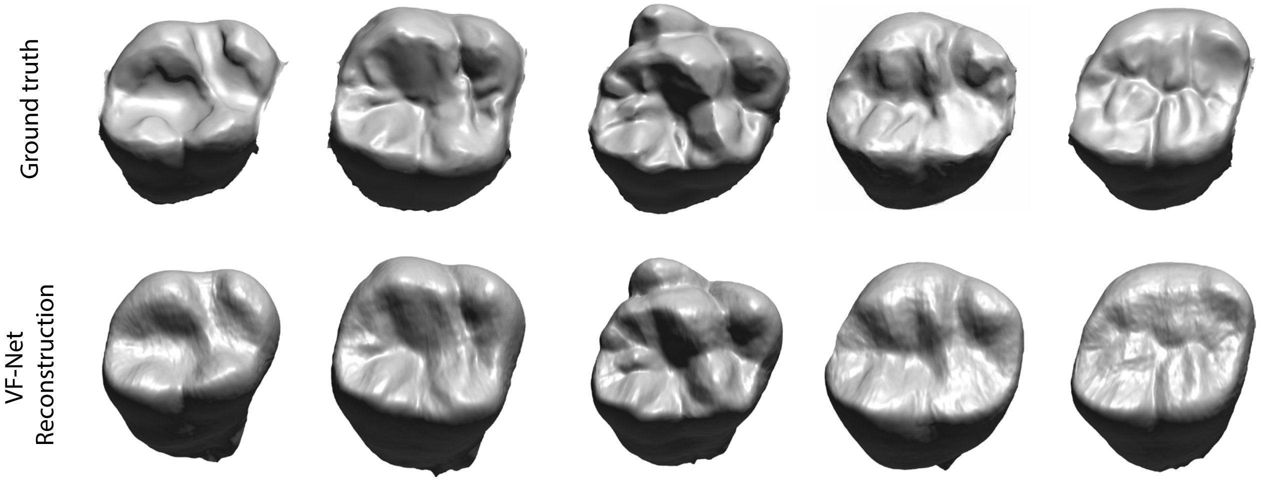

Figure S4: Examples of reconstructions of meshes from the FDI 16 dataset.

## A.6. Samples Across all Teeth

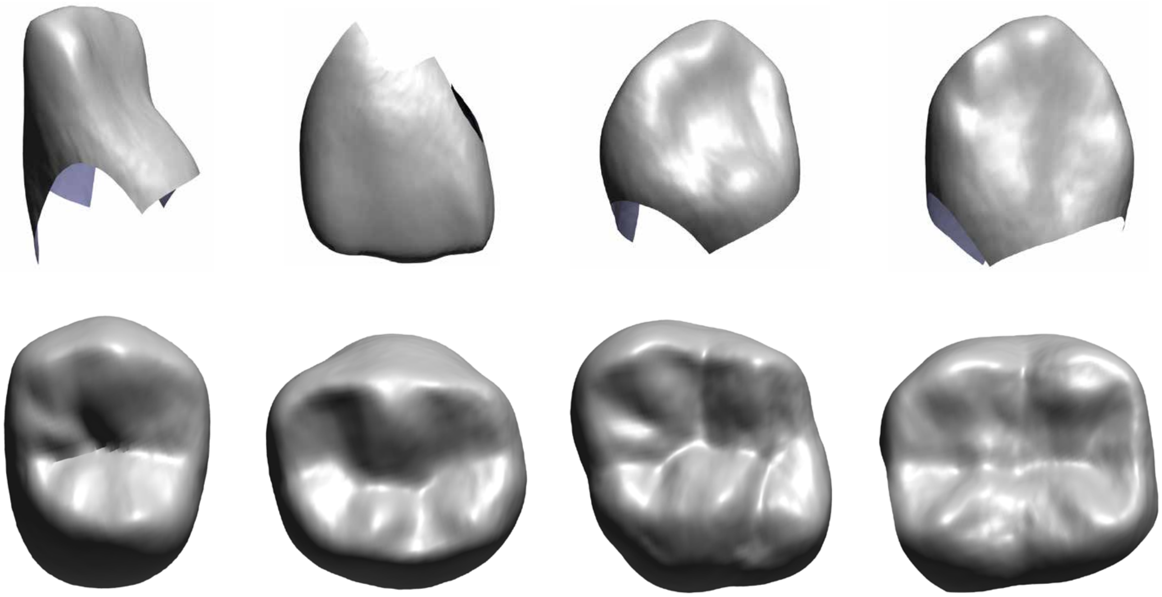

Figure S5: A showcase of meshes sampled by VF-Net. All four major modalities are covered: Incisor, canine, pre-molar, and molar, the four major types of teeth.

### A.7. Hole Completion

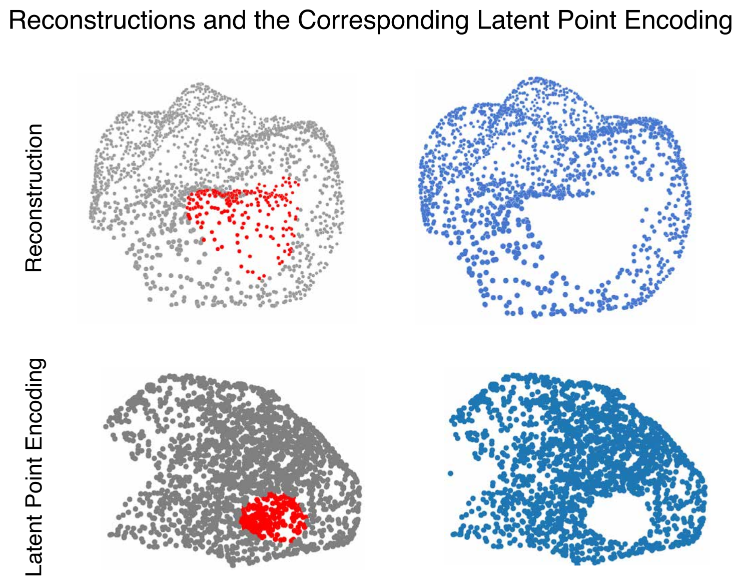

Figure S6: *Left*: To illustrate the hole reconstruction problem, we present an example where the red points are removed from the point cloud. *Right*: The latent point encodings remain highly similar to the shape of the encoded points prior to the deletion of points. Sampling the missing area becomes a straightforward task by sampling within the corresponding empty region of the latent point encoding.

### A.8. Synthetic Toothwear Teeth

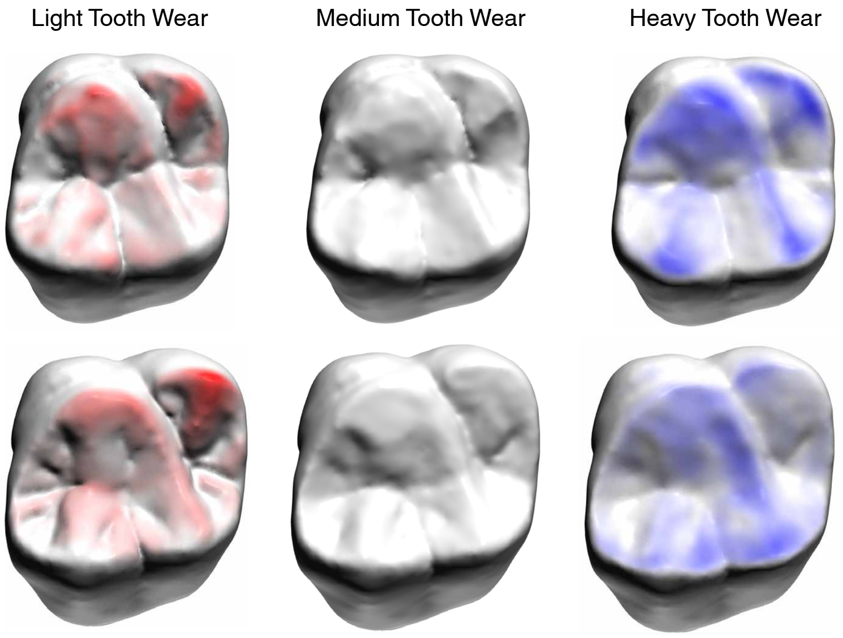

Figure S7: Two of the ten manually sculpted teeth to simulate tooth wear. *Left*: Highlighted in red are areas that have higher values compared to the original reconstruction. *Middle*: The original reconstruction. *Right*: Areas depicted in blue are lower than the original mesh.