Shape Translation of Dental Point Clouds

Stefan Engelmann Jensen[1] and Johan Ziruo Ye[1,2]

[1] Technical University of Denmark, Kgs. Lyngby 2800, Denmark
[2] 3Shape, Niels Juels Gade 13, 1059 Copenhagen, Denmark
Johan.Ye@3Shape.com

Abstract. Unpaired shape-to-shape translation remains a largely unexplored area, particularly in three-dimensional contexts. This paper explores its potential in dentistry, focusing on the translation of point cloud representations of teeth between young and old patients. We propose a novel approach that combines the latent overcomplete GAN framework with dual diffusion implicit bridges (DDIB) to enhance shape translations improving the applicability of these models in dental contexts. DDIB, a diffusion-based approach leveraging optimal transport properties, demonstrates significant improvements in generating more diverse and cycle-consistent samples that better resemble the target distribution. While these advancements show promise, further research is necessary to develop an autoencoder that balances high reconstruction accuracy with effective shape translation, addressing the unique challenges of dental morphology. Our findings establish a foundation for future research and applications in dentistry, potentially enabling personalized treatments and proactive interventions for various dental conditions.

Keywords: Shape Translation · Generative Modeling · Point Clouds

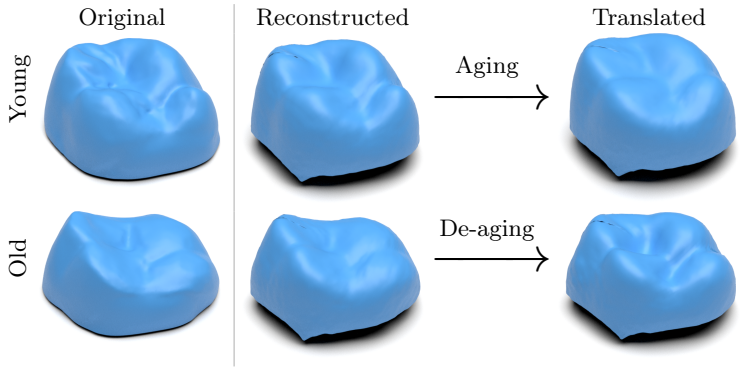

Fig. 1: Shape-to-shape translation between young and old teeth using our approach. The original young and old tooth (*left*) are reconstructed (*middle*), and their latent codes are translated to age/de-age the tooth (*right*), highlighting our method's effectiveness in capturing detailed morphological changes over time.

1 Introduction

Shape transformation and translation across domains is a rapidly growing field in computer vision. In the 2D domain, image translation has enabled the transformation of an image's visual style while preserving its content, influencing fields such as art, design, and media. Recently, methods like CycleGAN [22] have emerged to translate images between domains without explicit paired inputs.

Our research extends these concepts to 3D shape translation, with a focus on geometric transformation between young and old teeth, see Fig. 1. Previous works mostly address color or texture changes in point clouds [2,8,21], however, our approach targets shape translation by capturing the morphological changes that occur over time in teeth. We follow the approach by Yin *et al.* [20], and translate shapes by translating latent codes in an autoencoder. We aim to develop a model that can accurately translate these subtle and complex changes between young and old dental structures by improving their autoencoder and using dual diffusion implicit bridges (DDIBs) [15] as the translation model. We expect aging-related changes to be visible in the tooth curvature and overall morphology, primarily as a result of wear and grind across time. Understanding how age-related changes manifest in dental shapes can provide invaluable insights for forensic science, orthodontics, and prosthodontics. Additionally, this research lays the groundwork for more complex 3D translations, such as the removal or addition of brackets or aligners, and crown or bridge design.

Our Contribution. We present a step forward in bridging the gap between 2D image translation and 3D shape translation, showcasing the potential of advanced neural techniques in shaping the future of 3D modeling and transformation. To the best of our knowledge, we are the first to perform unpaired shape-to-shape translation using dental point clouds and the first to translate between shapes with a diffusion-based approach.

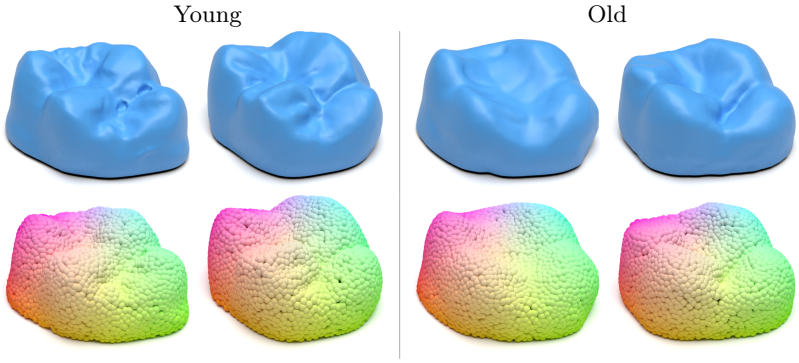

Fig. 2: Sample data from the *young* and *old* classes. Each tooth is shown as a mesh with the corresponding point cloud below. Point clouds were downsampled to 2,048 points.

2 Shape-to-Shape Translation of Young and Old Teeth

The dataset used in this study features algorithmically segmented point clouds of individual first upper right molar teeth, specifically FDI 16 (ISO 3950 notation). The individual teeth were preprocessed following the description by Ye *et al*. [18]. Each jaw was annotated with the patient's age, ranging from 14 to 90 years. The dataset was divided into two classes: *young* (ages below 26, with 7,399 samples) and *old* (ages above 62, with 7,380 samples). The age distribution across these classes is illustrated in Fig. A.1, and the specific data splits can be found in Table A.1. To focus on tooth morphology and minimize the impact of gingival recession, we standardized the border of the dental point clouds. This involved closing the original mesh and cutting the shape in the XZ plane. We excluded the bottom of the teeth from the analysis, as translation of this algorithmically generated portion of the tooth is not of interest. Examples of the meshes and point clouds from each class are illustrated in Fig. 2.

3 Expanding on LOGAN's Approach

Improving the Autoencoder. We follow latent overcomplete generative adversarial network's (LOGAN) [20] proposed approach of using an autoencoder to facilitate the shape-to-shape translation. Designing an autoencoder for point clouds is particularly challenging due to the need to process data that is both permutation invariant and cardinality invariant. This means that the autoencoder must maintain these invariances throughout the encoding and decoding process. LOGAN addresses this by using a PointNet++ [9] encoder, which processes each point individually. Furthermore, the network consistently outputs 2,048 points, regardless of the input's cardinality. We train this autoencoder to reconstruct samples from both domains. Outside the standard reconstruction term, Yin *et al*. also introduce an additional loss term. This term ensures that the output of each set abstraction layer from the PointNet++ [9] encoder should be sufficient to reconstruct the input sample on its own. This encourages global information to be saved at the four set abstraction levels. As we will later see, this leads to robustness with respect to numerical errors from translation, due to repeated information, but it also severely hampers reconstruction performance. We substitute this autoencoder for Variational FoldingNet (VF-Net) [18]. VF-Net extends FoldingNet [17] to gain a one-to-one correspondence throughout the network. This is achieved by substituting the static grid FoldingNet deforms for a learned 2D projection. This allows the model to model the point distribution of its input. We chose VF-Net for its reconstruction precision on dental point clouds.

A Diffusion Approach to Translation. After training the autoencoder to reconstruct both input domains, LOGAN employs a CycleGAN [22] for translation between the latent codes of each domain. Despite not having direct sample correspondences in each domain, it is crucial to enforce a relationship between the young dental samples and the old ones. To address this issue, CycleGAN

introduced the cycle consistency term, which measures similarity after a sample has been mapped back and forth between domains ($\circlearrowleft$). This will encourage the model to save information of the original sample in the translated one. Originally, this was enforced using an L1-term, which encourages the model to have pixel-to-pixel correspondence in source and target domain. While this is highly beneficial in introducing a relationship between the two samples, it also prevents larger content changes outside simple pixel-to-pixel color modifications. In terms of dental point clouds, we expect the sides of the teeth to remain rather constant and the cusps and fissures of the occlusal surface to remain similar.

We propose using DDIB [15] as the unpaired translator network in the latent space. DDIBs exploit that score-based generative models (SGMs) are implicit optimal transport models since they can be considered a special case of the Schrödinger Bridge Problem (SBP) [15]. To train DDIB for unpaired translation between the latent codes of the two domains, we employ a separate diffusion model for each domain. Sampling is performed according to the denoising diffusion implicit models [13] approach, where the forward process uses the source domain model to generate an intermediate uniquely identifiable Gaussian-distributed encoding, followed by the reverse process with the target domain model to obtain the target domain latent code, effectively solving the probability flow ordinary differential equation (ODE) of an SGM [14]. This fully deterministic approach ensures unique and reversible mappings between the latent domains, achieving exact cycle consistency, only up to discretization errors introduced by the ODE solver. This approach is particularly suited for our task of unpaired shape-to-shape translation between young and old dental structures, as it ensures that the significant morphological differences between the two age groups are captured and accurately transformed, preserving the inherent geometric characteristics of each domain while adhering to the principles of optimal transport.

4 Experiments and Results

We evaluate the effectiveness of our shape-to-shape translation model for dental point clouds using generative metrics, classification-based accuracy scores, and cycle consistency to ensure robust domain transfer while preserving age-invariant features from the source point cloud.

Reconstruction Quality. Since LOGAN's approach translates shapes by manipulating the autoencoder's latent codes and subsequently decoding, performance is tightly linked to the autoencoder's reconstruction quality. Thus, we first focused on improving the autoencoder. Replacing LOGAN's autoencoder with VF-Net resulted in over a 7-fold improvement in Chamfer distance (CD) and earth mover's distance (EMD). VF-Net achieves an average CD of 0.57 (EMD: 3.85) versus LOGAN's CD of 6.50 (EMD: 28.93), translating to a deviation per point of 0.053 mm vs. 0.18 mm. For a class-specific breakdown, refer to Table A.2, and see the visual results in Fig. 3.

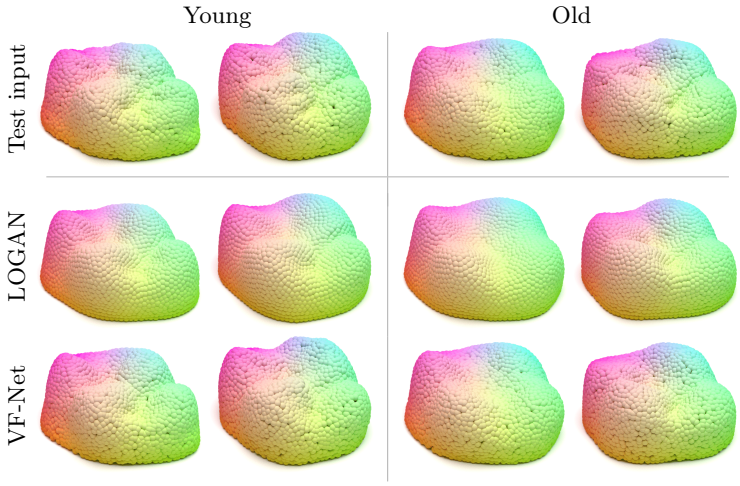

Fig. 3: Autoencoder reconstructions. First row: *original test input*. Second row: *LOGAN autoencoder reconstructions*. Third row: *VF-Net autoencoder reconstructions*.

Shape-to-Shape Translation. Due to the lack of long-term intra-oral scanner data for direct aging analysis of teeth, we evaluate the translated samples using established 3D generation metrics [16]. The metrics used include minimum matching distance (MMD), which measures the average distance to the nearest neighbor point cloud; coverage (COV), which measures the fraction of point clouds in the ground truth test set considered the nearest neighbor for each generated sample; and 1-nearest neighbor accuracy (1-NNA), which uses a 1-NN classifier to determine whether a sample is generated or from the ground truth dataset, with 50% indicating that generated samples are indistinguishable from the test set. Each metric should be interpreted in conjunction with the others for meaningful insights. For perspective, we include a comparison between a random subset of the training set's target class and the test set's target class as an oracle result, providing a performance ceiling since the model cannot be expected to exceed the inherent quality of the training data. The results can be found in Table 1.

Table 1: Translation results. ↑: higher is better, ↓: lower is better. Results are averaged over two translation tasks, with best scores are in bold. The training set is subsampled to match the test set size. Real shape scores worse than some generated shapes are marked in gray. MMD-CD and MMD-EMD scores are scaled by 10^2.

Model		MMD(↓)		COV(%, ↑)		1-NNA(%, ↓)	
Autoencoder	Translator	CD	EMD	CD	EMD	CD	EMD
LOGAN	LOGAN	14.99	33.46	36.07	35.71	70.38	64.20
LOGAN	DDIB	14.03	**32.86**	**44.77**	**45.49**	66.51	**61.28**
VF-Net	LOGAN	13.65	35.17	34.00	33.68	67.00	81.06
VF-Net	DDIB (ours)	**13.28**	35.45	43.13	42.26	**65.24**	85.82
——Training set ——		13.91	33.60	49.82	49.17	50.19	49.11

VF-Net generates samples with a lower MMD than even the oracle results. Moreover, VF-Net+DDIB achieves more accurate shape-to-shape translation results and nearly matches the diversity of LOGAN+DDIB when measured with CD. However, discrepancies arise in performance evaluation using CD and EMD. We attribute this to VF-Net's point encodings, which determine the point distribution in the point cloud. While this flexibility enables accurate reconstructions, it hinders EMD performance since the encodings are not changed during shape translation. Overall, DDIB translation yields significantly more diverse sample sets, consistently outperforming LOGAN's CycleGAN approach.

PointNet++ Classification. To further assess domain transfer, a binary PointNet++ [9] classifier, trained on the same dataset, was used to evaluate the translated shapes. The PointNet++ results, detailed in Table 2, indicate that translating from young to old is easier, as this mainly involves the removal of high-frequency occlusal surface details due to tooth wear. Since both autoencoders struggle to model high-frequency details, removing them is easier. Thus, three out of four models surpass the oracle results. Conversely, translating from old to young, which involves reversing dental indications like tooth wear, is more challenging. The baseline model achieves around 26% accuracy, while LOGAN+DDIB performs just above random guessing, likely due to LOGAN's overly smooth reconstructions that fail to capture high-curvature details. However, VF-Net+DDIB significantly outperforms the others with an accuracy above 84%, demonstrating that precise modifications and accurate decodings are crucial for effective performance in this task. Qualitative results can be found in Fig. 4.

Table 2: PointNet++ classification accuracy of translated shapes, best scores are in bold. The last row shows the test set results for the target class. Real shape scores worse than generated shapes are marked in gray.

Model		ACC(%, ↑)		
Autoencoder	Translator	Young → Old	Old → Young	Average
LOGAN	LOGAN	96.44	26.04	61.24
LOGAN	DDIB	**99.78**	51.27	75.53
VF-Net	LOGAN	79.27	77.89	78.58
VF-Net	DDIB (ours)	93.09	**84.22**	**88.66**
——— Test set ———		92.50	91.42	91.96

Cycle Consistency. Finally, we evaluate cycle consistency to see if distinctive features are preserved during translation. To reduce the autoencoder's influence, we encode the input point cloud, perform the translation only in the latent space, decode, and compare it to its reconstruction. Theoretically, DDIB should achieve exact cycle consistency, however, numerical limitations of ODE solvers like DDIM [13] introduce errors, making perfect cycle consistency challenging in

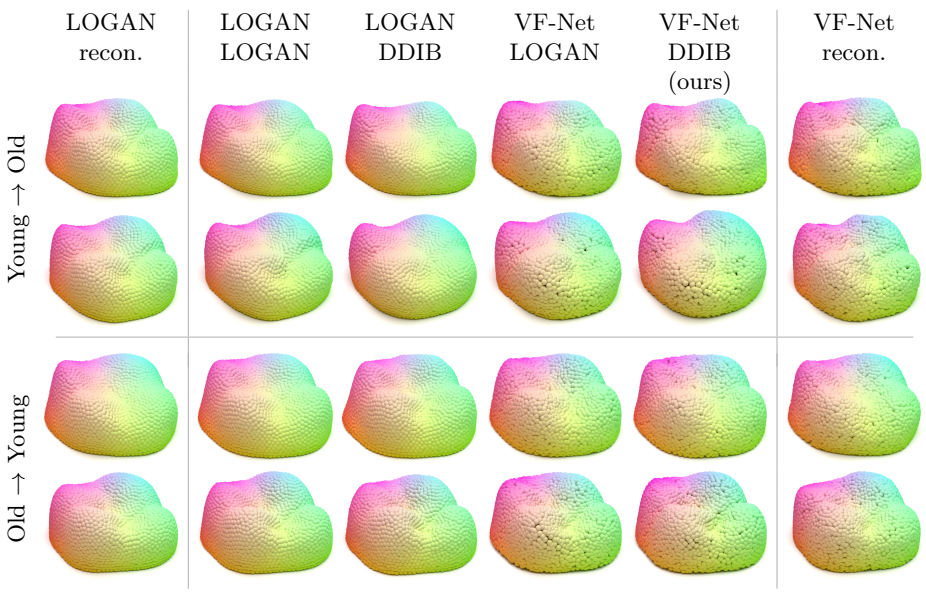

Fig. 4: Shape-to-shape translation examples. The first and second text rows indicate the autoencoder and translator used, respectively. Top two rows: *young* → *old translation*. Remaining rows: *old* → *young translation*.

practice. Despite this, the use of DDIB with the two autoencoders shows varying degrees of cycle consistency. The LOGAN autoencoder, while less accurate in point cloud reconstruction, exhibits notable robustness to these deviations from the original latent code, resulting in better cycle consistency when combined with DDIB compared to VF-Net+DDIB. This robustness is attributed to LOGAN's overcomplete latent code design, in which features across multiple scales all include global information about the point cloud. This redundancy leads to more robust latent codes, while beneficial for shape translation, they hinder accurate reconstruction. Since minor deviations in the biting surface can frequently lead to patient irritation, this is critical in denal applications. Detailed cycle consistency results are provided in Table A.3, and example cases are illustrated in Fig. A.2.

Limitations. VF-Net, although promising in enhancing reconstructions, relies on learned point encodings derived and predicted from the input shape. This reliance introduces bias in point distribution during the translation process, particularly when substantial geometric changes are needed. This can hamper its shape translation performance, particularly if larger geometric changes are required. This limitation is evident in EMD generation results for VF-Net, which reflect its challenges in achieving optimal shape-to-shape translation. Furthermore, while DDIB is theoretically perfectly cycle-consistent, in practice, numerical errors from the ODE solver can introduce inconsistencies. When using LOGAN's approach, these numerical errors propagate through the autoencoder, leading to significant deviations, as seen in the cycle consistency metrics. This situation leads to a counterintuitive approach to autoencoder design. Typically, in an autoen-

coder, disentanglement and compression efficiency, such as avoiding duplicated information, are desirable. However, in this context, LOGAN's results suggest that retaining repeated global information enhances robustness against translation inaccuracies. This necessity complicates the use of off-the-shelf autoencoders in conjunction with LOGAN's method for effective shape-to-shape translation.

5 Related Work

Unpaired Image Translation. Generative adversarial networks (GANs) have significantly advanced image translation [3–6]. Due to the lack of style separation and paired ground truth data, our available approaches are limited. CycleGAN [22] and DualGAN [19] are prominent models for unpaired image translation, utilizing cycle consistency loss to ensure source-target correspondence. Nevertheless, cycle consistency inherently assumes pixel-to-pixel correspondence, often resulting in changes that are confined to color rather than content. In contrast, point-to-point correspondences are rarely available, necessitating the use of geometric losses such as Chamfer Distance (CD) or Earth Mover's Distance (EMD) to preserve overall shape. While this approach allows for geometric (content) changes, it complicates the preservation of source information in a point cloud. Beyond GANs, diffusion models have shown promise for image translation. Palette [10] employs conditional diffusion models to perform *paired* image translation. BBDM [7] also handles paired image translation but uses Brownian motion in latent space to bridge the paired training data. Since we lack such paired data, these methods are not applicable to our use case. UNIT-DDPM [11] was developed for unpaired image translation using conditional DDPMs that are trained jointly, leveraging a cycle consistency loss at pixel level.

Unpaired Shape-to-Shape Translation. 3DSNet [12] approaches shape translation by disentangling content and style, allowing the combination of a source shape's content with a target shape's style. During inference, unpaired style transfer requires both an input shape and a style shape to generate a new shape that retains one shape's content and another's style. However, this method is not applicable to our work, as all teeth in our dataset have similar shapes. Instead, we focus on learning direct mappings between two unpaired domains. UNIST [1] builds on LOGAN by introducing a novel autoencoder structure using neural implicit representations instead of point clouds. UNIST generates higher-quality shapes than LOGAN while reusing the latter method's translation methodology. However, it involves a discretization step where shapes are converted to binary voxel occupancies, introducing precision loss. This precision loss is particularly problematic for applications involving dental surface scans, where maintaining high accuracy and detail is crucial.

6 Conclusion

In this work, we extended the framework of LOGAN, an unpaired shape-to-shape translation model, to enable bidirectional translation between young and old teeth. While effective overall, LOGAN struggles with tasks requiring detailed, fine-grained alterations, which are crucial for capturing dental morphological changes. Replacing the autoencoder with VF-Net and the translator with DDIB, we observed significant improvements, with our approach outperforming LOGAN across all metrics. Our results demonstrate more diverse, cycle-consistent samples that closely resemble the target distribution.

References

1. Chen, Q., Merz, J., Sanghi, A., Shayani, H., Mahdavi-Amiri, A., Zhang, H.: Unist: unpaired neural implicit shape translation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18614–18622 (2022). https://doi.org/10.48550/arXiv.2112.05381 8
2. Chen, Y., Yuan, Q., Li, Z., Liu, Y., Wang, W., Xie, C., Wen, X., Yu, Q.: Upst-nerf: Universal photorealistic style transfer of neural radiance fields for 3d scene (2022). https://doi.org/10.48550/arXiv.2208.07059 2
3. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation (2018). https://doi.org/10.48550/arXiv.1711.09020 8
4. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks (2018). https://doi.org/10.48550/arXiv.1611.07004 8
5. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019). https://doi.org/10.48550/arXiv.1812.04948 8
6. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan (2020). https://doi.org/10.48550/arXiv.1912.04958 8
7. Li, B., Xue, K., Liu, B., Lai, Y.K.: Bbdm: Image-to-image translation with brownian bridge diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 1952–1961 (2023). https://doi.org/10.48550/arXiv.2205.07680 8
8. Liu, K., Zhan, F., Chen, Y., Zhang, J., Yu, Y., Saddik, A.E., Lu, S., Xing, E.: Stylerf: Zero-shot 3d style transfer of neural radiance fields (2023). https://doi.org/10.48550/arXiv.2303.10598 2
9. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017). https://doi.org/10.48550/arXiv.1706.02413 3, 6
10. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022). https://doi.org/10.48550/arXiv.2111.05826 8
11. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021). https://doi.org/10.48550/arXiv.2104.05358 8

12. Segu, M., Grinvald, M., Siegwart, R., Tombari, F.: 3dsnet: Unsupervised shape-to-shape 3d style transfer. arXiv preprint arXiv:2011.13388 (2020). https://doi.org/10.48550/arXiv.2011.13388 8
13. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020). https://doi.org/10.48550/arXiv.2010.02502 4, 6
14. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020). https://doi.org/10.48550/arXiv.2011.13456 4
15. Su, X., Song, J., Meng, C., Ermon, S.: Dual diffusion implicit bridges for image-to-image translation. arXiv preprint arXiv:2203.08382 (2022). https://doi.org/10.48550/arXiv.2203.08382 2, 4
16. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: PointFlow: 3D Point Cloud Generation with Continuous Normalizing Flows (Sep 2019). https://doi.org/10.48550/arXiv.1906.12320, arXiv:1906.12320 [cs] 5
17. Yang, Y., Feng, C., Shen, Y., Tian, D.: Foldingnet: Point cloud auto-encoder via deep grid deformation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 206–215 (2018). https://doi.org/10.48550/arXiv.1712.07262 3
18. Ye, J.Z., Ørkild, T., Søndergaard, P.L., Hauberg, S.: Variational point encoding deformation for dental modeling. arXiv preprint arXiv:2307.10895 (2023). https://doi.org/10.48550/arXiv.2307.10895 3
19. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE international conference on computer vision. pp. 2849–2857 (2017). https://doi.org/10.48550/arXiv.1704.02510 8
20. Yin, K., Chen, Z., Huang, H., Cohen-Or, D., Zhang, H.: Logan: Unpaired shape transform in latent overcomplete space. ACM Transactions on Graphics (TOG) **38**(6), 1–13 (2019). https://doi.org/10.48550/arXiv.1903.10170 2, 3
21. Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: Lion: Latent point diffusion models for 3d shape generation (2022). https://doi.org/10.48550/arXiv.2210.06978 2
22. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017). https://doi.org/10.48550/arXiv.1703.10593 2, 3, 8

A Appendix

A.1 Data details

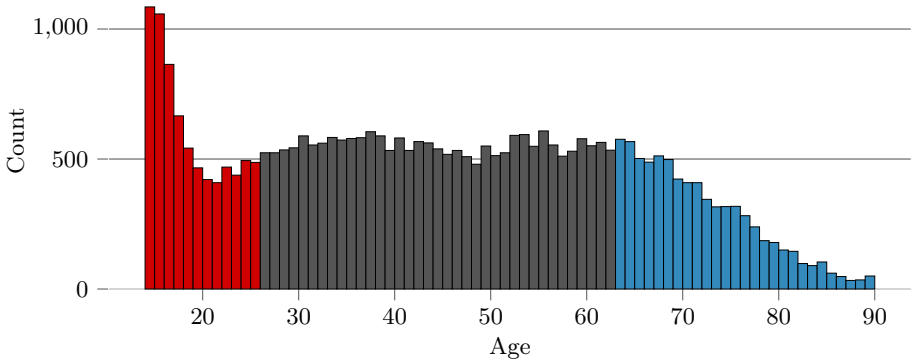

Fig. A.1: Age distribution of data color coded according to young (*red*), and old (*blue*) class. The grey bars in the histogram resemble unused data.

Table A.1: Number of data samples in each split and class

Split	Young	Old
Training	4999	4990
Validation	999	999
Test	1375	1375

A.2 Autoencoder reconstruction results

Table A.2: Autoencoder reconstruction errors in non-normalized space. The best scores are highlighted in bold. CD and EMD are divided by the number of points and multiplied with 10^2.

	Young		Old		Average	
Autoencoder	CD	EMD	CD	EMD	CD	EMD
LOGAN	6.168	28.56	6.831	29.30	6.500	28.93
VF-Net	**0.508**	**3.625**	**0.622**	**4.064**	**0.565**	**3.845**

A.3 Cycle consistency results

Table A.3: Cycle-consistency results. To minimize the influence of the autoencoder, we encode the input point cloud, perform the cycle solely in the latent space, decode, and then compare against its reconstruction. The best scores are highlighted in bold. Chamfer distances (CD) and earth mover's distances (EMD) are multiplied with 10^2.

Model		Young ↻ Old		Old ↻ Young		Average	
Autoencoder	Translator	CD	EMD	CD	EMD	CD	EMD
LOGAN	LOGAN	1.493	7.688	3.413	13.82	2.453	10.75
LOGAN	DDIB	**0.415**	**3.560**	**1.455**	**7.021**	**0.935**	**5.291**
VF-Net	LOGAN	15.16	10.80	5.412	12.20	10.29	11.50
VF-Net	DDIB	1.476	6.470	15.36	23.33	8.418	14.90

A.4 Cycle consistency examples

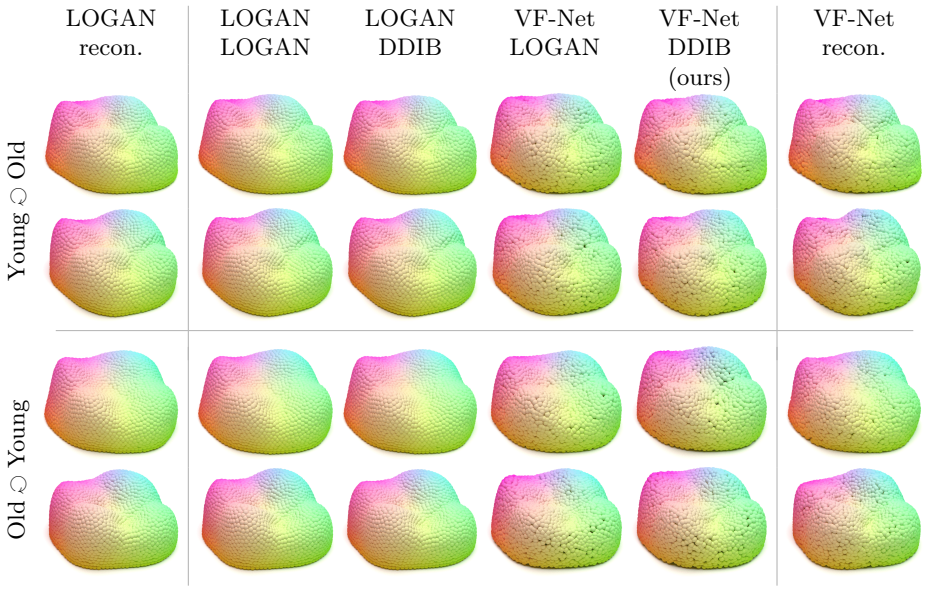

Fig. A.2: Examples of cycled reconstructions. The first and second row of text indicate the autoencoder and translator used respectively. Top four rows: *young* ↻ *old* cycle. Rest: *old* ↻ *young* cycle.