Supplementary Material: Learning Geometry-Aware Joint Latent Space for Simultaneous Multimodal Shape Generation

Artem Komarichev, Jing Hua, Zichun Zhong

Department of Computer Science, Wayne State University, Detroit, Michigan 48202, USA

In this Supplementary Material, we provide some additional description and discussion about this work, such as the architecture of the autoencoder on image in Sec. 1, training details in Sec. 2, quantitative evaluation metrics in Sec. 3, and more experimental results and analysis in Sec. 4.

1. The Architecture of the Autoencoder on Images

Fig. 1 shows the detailed architecture of the image autoencoder.

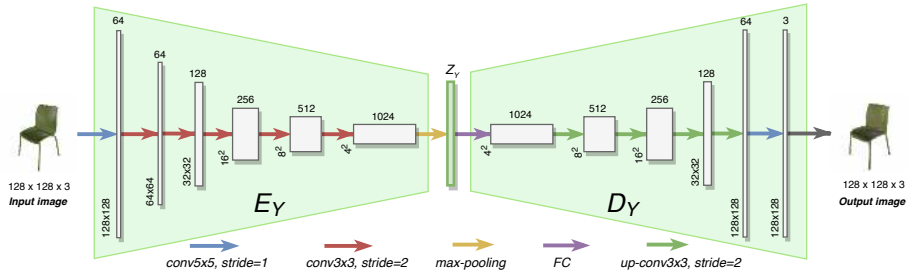


Figure 1: The architecture of the autoencoder on images. The encoder $E_{\boldsymbol{Y}}$ maps a given image into the image latent space. $Z_{\boldsymbol{Y}}$ is a 1024-dim feature vector that encodes the given image. The decoder $D_{\boldsymbol{Y}}$ reconstructs the feature vector from the latent space back to the 2D color image space.

2. Training Details

In this section, we introduce the training details for each part of our entire framework. First, we train our autoencoders on point cloud and images, respectively. In our experiments, we train our autoencoder on point clouds for 200 epochs. We use Adam optimizer ($\beta_1 = 0.9$) with an initial learning rate of 0.01 and a batch size of 32. The learning rate is reduced with decay rate of 0.7 every 20 epochs. The training time on point clouds of *chair* class takes less than 4 hours. In our experiments, we train our autoencoder on images for 400 epochs. We use Adam optimizer ($\beta_1 = 0.9$) with an initial learning rate of 0.0005 and a batch size of 32. We keep the same learning rate for first 200 epochs and linearly decay the learning rate to 0 for the next 200 epochs. The training time on images of *chair* class takes around 4 hours.

As for training the mixer, we use Adam optimizer ($\beta_1 = 0.9$) with initial learning rate of 0.0002 and batch size of 128. We keep the same learning rate for first 300 epochs and linearly decay it to 0 for another 300 epochs. The training time of our mixer on *chair* class takes approximately around 10 minutes. We train our mixer on different categories for 600 epochs.

As for training the joint generative model, we train our joint generative model with WGAN-GP [1] objective. We use Adam optimizer ($\beta_1 = 0.5$) with an initial learning rate of 0.0005 and a batch size of 50. The learning rate is halved every 300 epochs until it reaches 0.0001. The generator takes the input of a noise vector of dimension 256 with zero mean and 0.2 standard deviation, and generates joint latent vector $\hat{Z}_{XY}$. We train each joint generative model for 2000 epochs. The training time of our joint generative model takes around 10 minutes.

The architecture of the binary classification network for cross-modality similarity evaluation is described in Sec. 2.4 in the main paper. For training this network, we use the binary classification loss. We train this network on a set of the shape and image latent vectors calculated by pre-trained geometry-aware autoencoder (GAE)'s shape encoder and image encoder, respectively. We label

each such pair as "1". For each given batch, we generate all possible unmatched pairs that are labeled as "0"s. We train this binary network on *chair* class from *ShapeNet Core* dataset for 200 epochs. We use Adam optimizer ($\beta_1 = 0.9$) with an initial learning rate of 0.0001 and a batch size of 256. We keep the same learning rate for first 100 epochs and linearly decay it to 0 for another 100 epochs. After that, we use this pre-trained network to evaluate the CMSS metric for each method from Tab. 2. in the main paper and Tab. 2 below.

3. Quantitative Evaluation Metrics

Shape Reconstruction Metrics. Let $\mathcal{P}$ be the ground truth point cloud and $\mathcal{P}'$ be a reconstructed point cloud. We assume that the number of points in $\mathcal{P}$ matches the number of points in $\mathcal{P}'$. First, we evaluate F-score [2] at a given threshold δ by calculating the harmonic mean of precision and recall between $\mathcal{P}$ and $\mathcal{P}'$. The threshold value that we choose for our evaluation is 0.0002. We calculate two F-score values F1 and F2 with the distance δ and distance 2δ, respectively. We also report the Chamfer Distance (CD) and Earth Mover's Distance (EMD) as additional reconstruction metrics.

Shape Generation Metrics. For quantitative comparisons between different point cloud generation methods (including simultaneous multimodal generation and single-modal generation tasks), we use the evaluation metrics in [3] to measure the quality of generated point clouds. Specifically, it suggests using the Jensen-Shannon Divergence (JSD) to evaluate the marginal distribution defined in the 3D Euclidean spaces. Another two metrics are Minimum Matching Distance (MMD) and Coverage (COV) to measure the fidelity and coverage of the generated point clouds and the data distribution, respectively. Based on the point cloud-based CD and EMD losses, it can yield four different metrics: MMD-CD, MMD-EMD, COV-CD, and COV-EMD. Both MMD-CD and MMD-EMD metrics are multiplied by 10^2 and 10, respectively, for better viewing in Tabs. 2 and 3 in the main paper.

Image Generation Metrics. For quantitative comparisons between differ-

ent image generation methods (including simultaneous multimodal generation and single-modal generation tasks), we use the following two common metrics to measure the quality of generated images. The Frechet Inception Distance (FID) [4] is the most common metric that uses feature space extracted with pre-trained inception V3 model [5] for evaluation. Another metric is the Kernel Inception Distance (KID) [6] that uses the same feature extraction model. The FID metric is divided by 10 for better viewing in Tabs. 2 and 3 in the main paper.

Similarity Metric. In our paper, we propose a new CMSS metric (in Sec. 2.4 in the main paper) to evaluate the multimodal correspondence in the latent feature space. This new metric evaluates the similarity correspondence between the generated shape and image in the latent space for simultaneous multimodal generation task.

4. More Experimental Results and Analysis

4.1. Comparison with Single-Modal Generation

We provide the qualitative comparisons of different single-modal point cloud and image generation methods with our method in the joint latent space as follows.

In Figs. 2, 3, we qualitatively compare our generated point clouds in joint latent space with the results generated by the single-modal generation methods, such as r-GAN [3], l-GAN (AE-EMD) [3], l-GAN (AE-CD) [3], TreeGAN [7], and ShapeGF-GAN [8]. As shown in Fig. 2, 3, our method outperforms other single-modal generation methods in the quality of generated point clouds on both *ShapeNet Core chair* and *airplane* classes with respect to high-fidelity geometry and topology of a variety of 3D objects.

In Fig. 4, we qualitatively compare the quality of our generated images with four alternatives, including rendering the reconstructed meshes for image generation by using the generated point clouds, and three state-of-the-art and

well-known generative models, i.e., DCGAN [9], PlatonicGAN [10], and WGAN-GP [1]. As shown in Fig. 4, our approach generates better quality images compared to other alternatives on *ShapeNet Core chair* class. Our results have less artifacts and more realistic textures / colors. More image generation results by our simultaneous multimodal (SMM) generation method are also provided in other related experiments as shown in the following figures.

4.2. Simultaneous Multimodal Generation

Comparison with the State-of-the-Arts. We provide more qualitative evaluation results of our mixer on the joint SMM generation task with other state-of-the-art methods, such as Shape Unicode (img / pc feats) [11], $\mathcal{Z}_{\mathcal{X}}$ concat $\mathcal{Z}_{\mathcal{Y}}$, $\mathcal{Z}_{\mathcal{X}}$ + rendering, and $\mathcal{Z}_{\mathcal{Y}}$ + PSGN [12], as shown in Figs. 5, 6, and 7 on *ShapeNet Core chair*, *airplane*, and *car* classes, respectively.

Joint Latent Space Interpolation. In Figs. 8, 9, and 10, we show more results of the linear interpolation in the proposed joint latent space between the selected left- and right-most shapes and images with large variations in terms of shape geometry and topology, and image texture on different object classes, such as *chair, airplane*, and *car* classes correspondingly.

4.3. Analysis

Ablation Study on Shape Auto-Encoding. The goal of this ablation study is to show the importance of the proposed GAE components. We evaluate two proposed components, i.e., adaptive query ball (in Sec. 2.1.1 in the main paper) and particle-based loss (in Sec. 2.1.2 in the main paper) on the point cloud reconstruction of the *ShapeNet Core chair* class as shown in Tab. 1.

On the one hand, we investigate the importance of the proposed adaptive query ball. In our first ablation experiment, we replace the adaptive query ball with randomly generated query ball sizes. For each centroid, we replace estimated radius with the random radius generated within the bounds, i.e., $R_{rand} = [R_i^{clip} - 0.05, R_i^{clip} + 0.05], i \in \{1, 2, 3\}$. In the second experiment, we replace each estimated radius with a fixed radius R_i^{clip} at different point cloud

Table 1: Ablation experiments on our GAE for the point cloud reconstruction on *ShapeNet Core chair* class. QB – query ball, and PL – particle loss.

	F1 ↑	F2 ↑	CD ↓	EMD ↓
our (random QB)	60.15	82.72	0.651	0.845
our (fixed QB)	67.30	87.56	0.528	0.686
our (mean QB)	68.04	88.07	0.519	0.668
our (local QB, no clip)	68.06	88.01	0.516	0.646
our (w/o PL, fixed QB)	65.74	86.63	0.557	0.722
our (w/o PL)	66.53	87.13	0.542	0.678
our (GAE)	**68.49**	**88.37**	**0.504**	**0.635**

resolution levels. In the third experiment, we replace every local estimated query ball by its mean size per shape. In the fourth experiment, we turn off the radius clipping operation after the radius estimation step as shown in Tab. 1.

On the other hand, we provide the experiment results on our proposed GAE model in the case that the adaptive query ball is turned on alike the particle loss is turned off. We also include experiments when both the adaptive query ball and the particle loss are turned off which is similar to the method in PointNet++ [13], denoted as "our (w/o PL, fixed QB)" in Tab. 1. Our ablation experiments show the importance of both proposed components on the point cloud reconstruction task. In conclusion, our GAE model benefits the most when both components are turned on as shown in Tab. 1.

Additionally, we include the comparison results for the *flexible* and *deformable* convolutions proposed by [14] as Kernel Point Convolution (KPConv) that learns to adapt kernel points to local geometry. [14] does not provide the architecture for the reconstruction task. We have built a reconstruction model based on the provided classification model by adjusting the last fully-connected layers to the reconstruction task. Also we use the number of the input and the reconstructed points to a fixed number (i.e., 2048 points). The learning rate and the number of epochs are the same as provided in [14]. Our

Table 2: Ablation study of our mixer and joint generative model for simultaneous multimodal generation on *chair* class.

	point cloud					*image*		
Model	JSD ↓	MMD-CD ↓	MMD-EMD ↓	COV-CD ↑	COV-EMD ↑	FID ↓	KID ↓	CMSS ↑
w/ L_{FM} (our)	0.060	**0.128**	**0.522**	**66.94**	**69.30**	**10.53**	**6.53**	**98.67**
w/o L_{FM}	**0.056**	0.130	0.525	65.46	66.42	11.49	7.49	93.43

proposed GAE model with the adaptive query balls significantly outperforms KPConv [14] model with deformable kernels on the point cloud reconstruction task of *ShapeNet Core chair* class. Their results on F1s ↑, F2s ↑, CDs ↓, EMDs ↓ are: 42.94, 66.06, 1.315, 1.102, respectively. To the best of our knowledge, there are at least two possible reasons for such poor performance on KPConv model. First, there is no provided reconstruction model of KPConv in the original paper. Second, KPConv model is working on much denser point clouds (i.e., more than 6K points) as input on the classification model.

Analysis on Mixer for SMM. The goal of the architecture ablation study is to show the importance of the components in the mixer for SMM generation task. The quantitative evaluation of the ablation study on our *mixer* is provided in Tab. 2. Our study shows the importance of the proposed *intermodality feature mapping loss* L_{FM} for the joint generation task. Additionally, we found the optimal aggregation function $f_{aggr}(Z_X^{'}, Z_Y^{'})$ as *sum* and the best joint latent vector dimension as 1024.

4.4. SMM Semantic-Aware Generation

In this experiment, we show that our mixer is capable of sharing additional semantic segmentation information between modalities, when one of the modalities is lacking such segmentation information. Specifically, we train our mixer on the joint latent codes extracted from point clouds without semantics and images with semantics (i.e., the part-level semantic segmentation information is

only available in images as indicated by colors).

Dataset. We evaluate this task on *chair* class from *Shape COSEG* dataset [15] containing 400 models, where each model is annotated with three parts (i.e., back, seat, and legs). This dataset is challenging due to the small set size and variety of model shapes. Initially we translate all models to the origin, and then normalize and orient them, which is the same process as in [3]. From each model, we sample 10,000 uniform points with normals from meshes and their segmentation labels. Additionally, we approximate surface area for each shape point cloud from dataset for particle loss calculation. From each object, we render one view 128×128 semantic image from a fixed viewpoint.

Training Details. We formulate the semantic shape GAE as a per-point classification problem along with the point cloud reconstruction. We extract bijection information from EMD loss and use it to find point-to-point correspondence to calculate the segmentation loss. The overall loss for our 3D semantic shape autoencoder is a weighted sum of two original GAE losses, i.e., reconstruction loss and particle-based loss, as shown in Eq. 2 from the main paper, plus a per-point classification loss. We train this semantic shape GAE on the *chair* class from *Shape COSEG* dataset. Additionally, we train image autoencoder on the semantic images. After that we train our proposed mixer and joint generative model in the same way as other tasks on *ShapeNet Core* dataset. To be more specific, semantic labels are only needed to train our point cloud autoencoder (i.e., GAE). Training our mixer does not require semantic labels. For SMM semantic-aware generation task in the joint latent space, we do not need semantic labels for point clouds.

Fig. 11 shows the new results of the cross-modal shape-image semantic-aware generation, where images can successfully share semantic information with their corresponding point clouds via our mixer and joint generative model. This new task can help us to do multimodal generation and segmentation tasks simultaneously, which enhances the capability of the corresponding single-modality and single-tasking to the next level, where operations can be conducted on correspondingly appropriate modality, e.g., assigning semantic parts is easier through

annotating on 2D images and transferring to 3D shapes, since, on the contrary, annotating 3D shapes is more difficult and less efficient.

References

[1] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of Wasserstein GANs, in: Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.

[2] A. Knapitsch, J. Park, Q.-Y. Zhou, V. Koltun, Tanks and temples: Benchmarking large-scale scene reconstruction, ACM Transactions on Graphics 36 (4) (2017) 1–13.

[3] P. Achlioptas, O. Diamanti, I. Mitliagkas, L. Guibas, Learning representations and generative models for 3D point clouds, in: Proceedings of the International Conference on Machine Learning, 2018, pp. 40–49.

[4] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 6626–6637.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[6] M. Bińkowski, D. Sutherland, M. Arbel, A. Gretton, Demystifying mmd gans, arXiv preprint arXiv:1801.01401.

[7] D. Shu, S. Park, J. Kwon, 3D point cloud generative adversarial network based on tree structured graph convolutions, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3859–3868.

[8] R. Cai, G. Yang, H. Averbuch-Elor, Z. Hao, S. Belongie, N. Snavely, B. Hariharan, Learning gradient fields for shape generation, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 364–381.

[9] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434.

[10] P. Henzler, N. Mitra, T. Ritschel, Escaping Plato's cave: 3D shape from adversarial rendering, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9984–9993.

[11] S. Muralikrishnan, V. Kim, M. Fisher, S. Chaudhuri, Shape Unicode: A unified shape representation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3790–3799.

[12] H. Fan, H. Su, L. Guibas, A point set generation network for 3D object reconstruction from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 605–613.

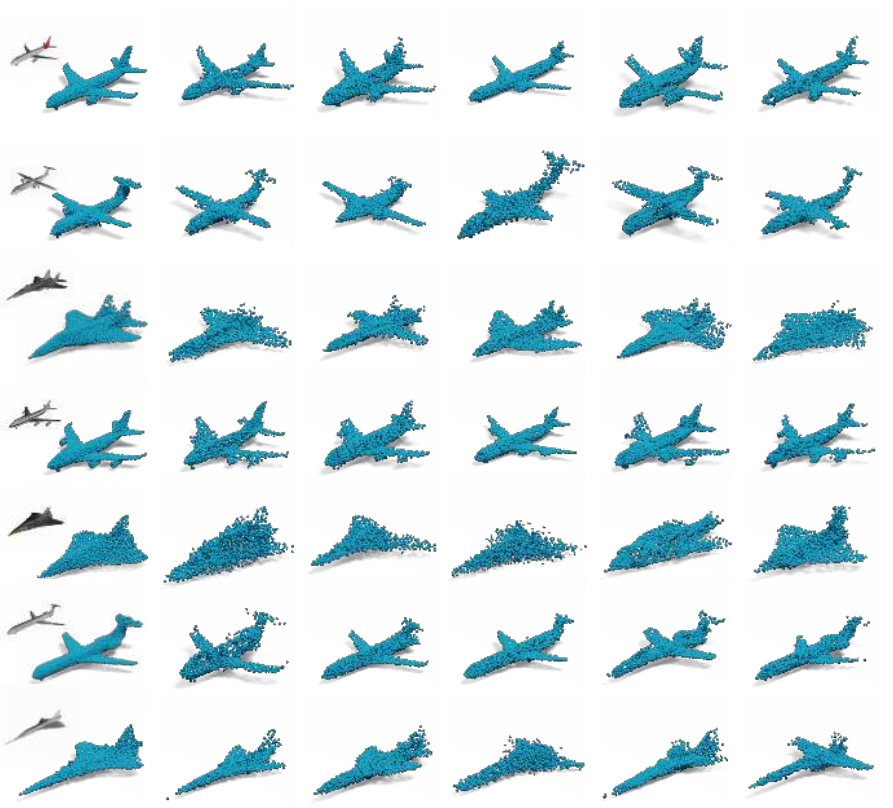
[13] C. Qi, L. Yi, H. Su, L. Guibas, PointNet++: deep hierarchical feature learning on point sets in a metric space, in: Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 5099–5108.

[14] H. Thomas, C. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, L. Guibas, KPConv: Flexible and deformable convolution for point clouds, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6411–6420.

[15] Y. Wang, S. Asafi, O. Van Kaick, H. Zhang, D. Cohen-Or, B. Chen, Active co-analysis of a set of shapes, ACM Transactions on Graphics 31 (6) (2012) 1–10.

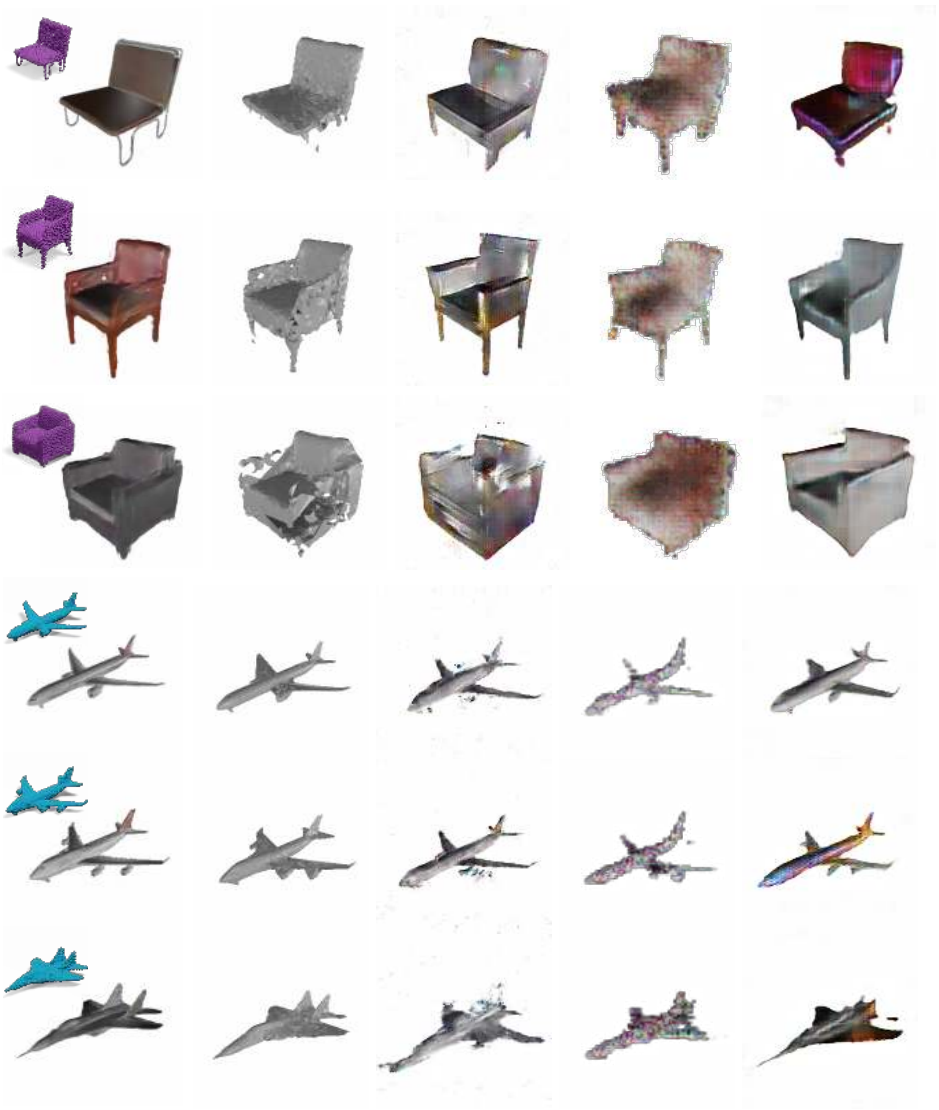
(a) **Our (joint)** (b) r-GAN [3] (c) l-GAN (CD) [3] (d) l-GAN (EMD) [3] (e) TreeGAN [7] (f) ShapeGF-GAN [8]

Figure 2: Qualitative comparison of different single-modal shape (point cloud) generation methods on *chair* class. Our joint SMM (shape + image) generation results are also provided correspondingly.



(a) **Our (joint)** (b) r-GAN [3] (c) l-GAN (CD) [3] (d) l-GAN (EMD) [3] (e) TreeGAN [7] (f) ShapeGF-GAN [8]

Figure 3: Qualitative comparison of different single-modal shape (point cloud) generation methods on *airplane* class. Our joint SMM (shape + image) generation results are also provided correspondingly.



(a) **Our (joint)** (b) Rendering (c) DCGAN (d) PGAN (e) WGAN-GP

Figure 4: Qualitative comparison of different image generation methods on *chair* and *airplane* classes for Rendering, DCGAN [9], PlatonicGAN [10] (PGAN), and WGAN-GP [1]. Our joint SMM (image + point cloud) generation results are also provided correspondingly.

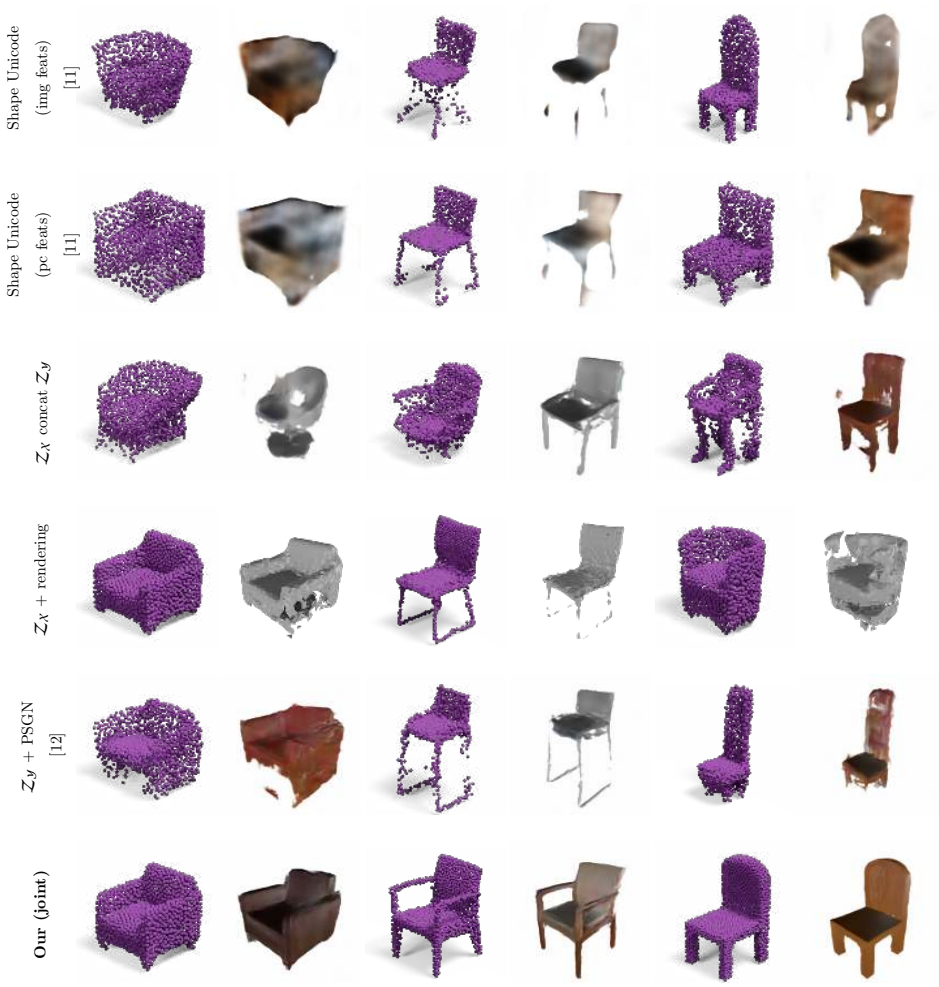


Figure 5: Qualitative comparison of our method with the state-of-the-art methods on multimodal shape and image generation on *chair* class. $\mathcal{Z}_{\mathcal{X}}$ - the shape latent space, $\mathcal{Z}_{\mathcal{Y}}$ - the image latent space. Left: point clouds, right: images.

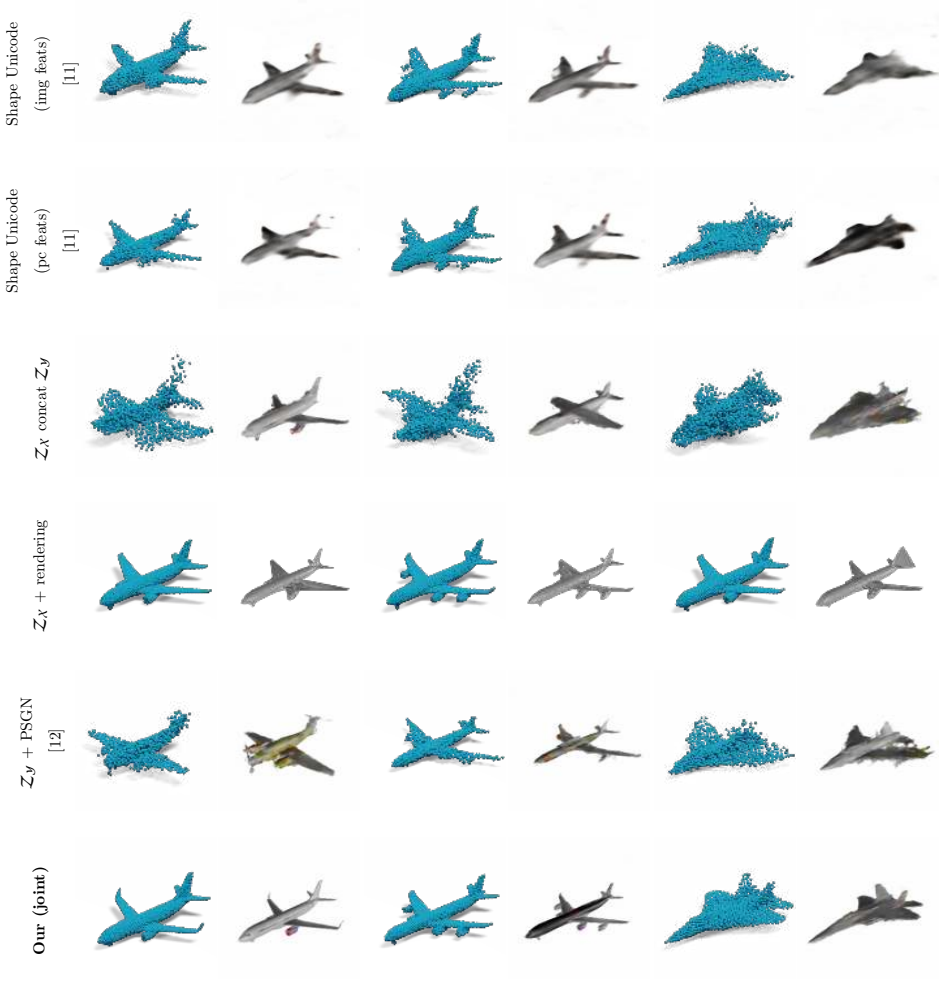


Figure 6: Qualitative comparison of our method with the state-of-the-art methods on multi-modal shape and image generation on *airplane* class. $\mathcal{Z}_\mathcal{X}$ - the shape latent space, $\mathcal{Z}_\mathcal{Y}$ - the image latent space. Left: point clouds, right: images.

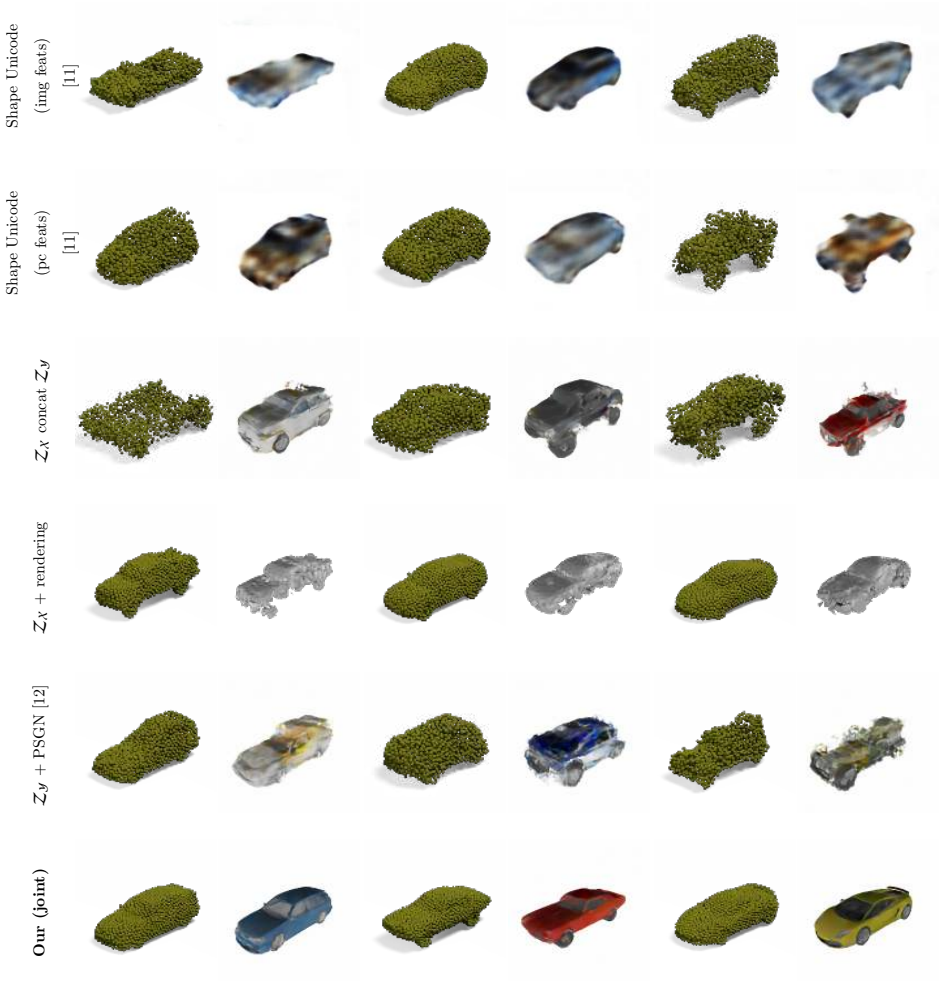


Figure 7: Qualitative comparison of our method with the state-of-the-art methods on multi-modal shape and image generation on *car* class. $\mathcal{Z}_\mathcal{X}$ - the shape latent space, $\mathcal{Z}_\mathcal{Y}$ - the image latent space. Left: point clouds, right: images.

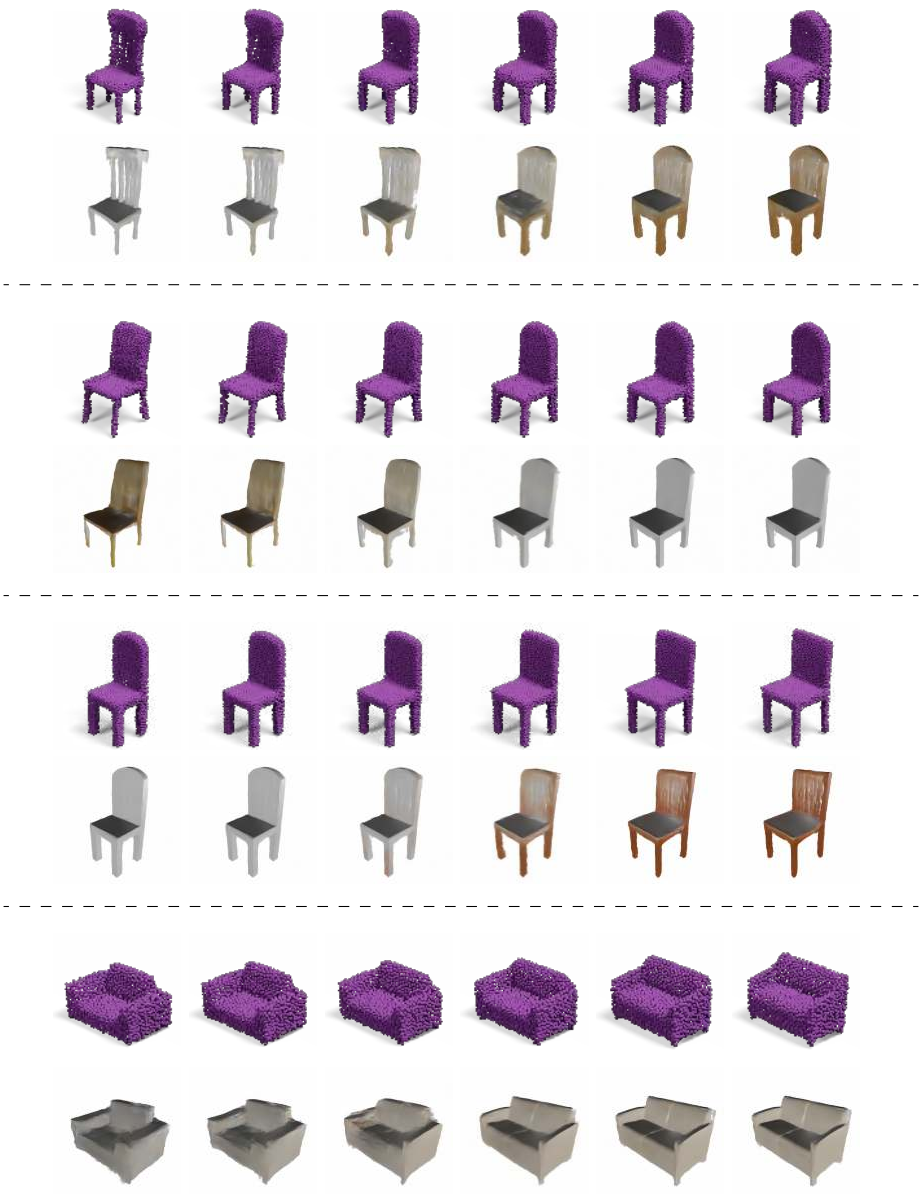


Figure 8: Joint latent space interpolation and generation on *chair* class. Top row: point clouds, bottom row: images.

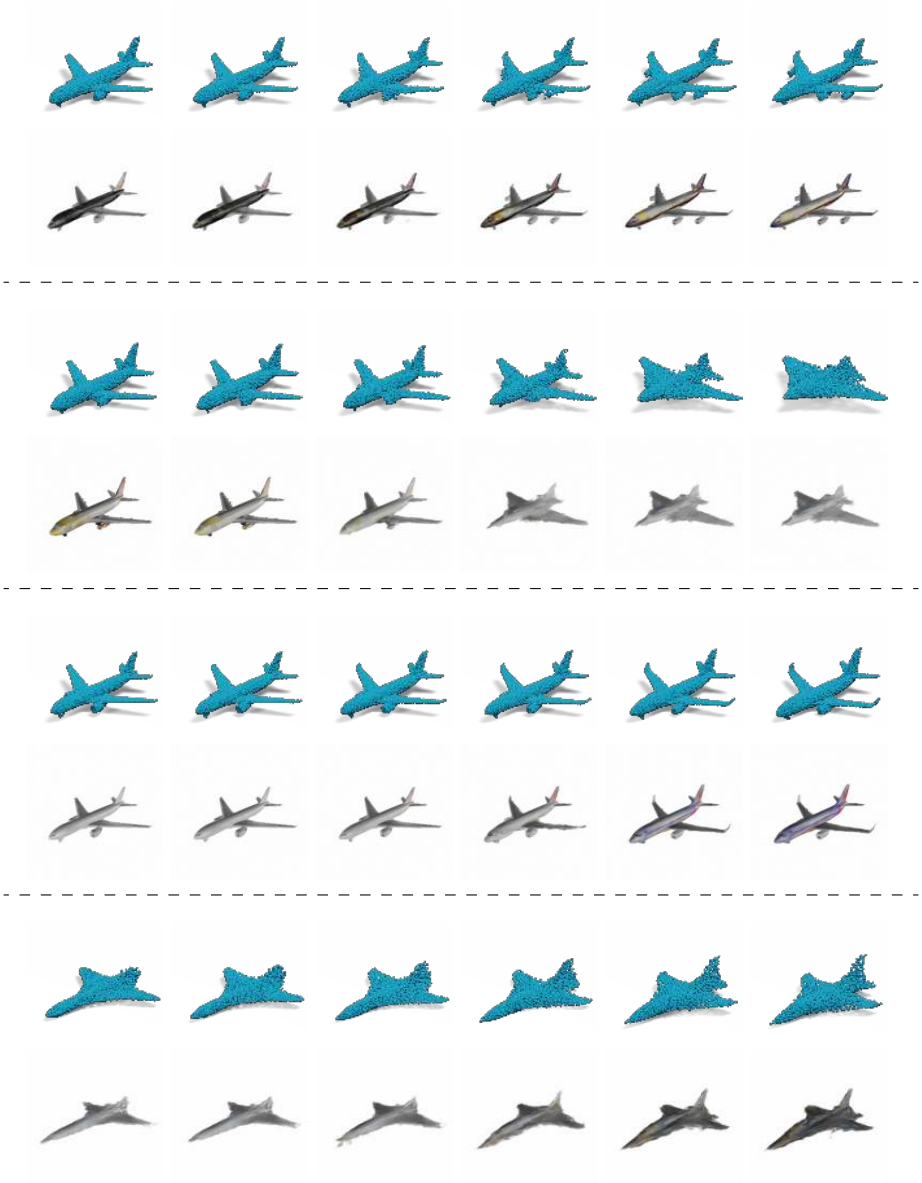


Figure 9: Joint latent space interpolation and generation on *airplane* class. Top row: point clouds, bottom row: images.

Figure 10: Joint latent space interpolation and generation on *car* class. Top row: point clouds, bottom row: images.

Figure 11: Some new SMM semantic-aware generation results with large variations in shape geometry and topology on *chair* class from *Shape COSEG*. Top row: point clouds, bottom row: images.