A. Supplementary material

A.1. Details of network architectures

We provide detailed architectures of the generator and the discriminator used in our system in Figure 1. We train individual models for different categories for 20 epochs on one Nvidia Tesla V100 GPU. Training each model takes approximately 6 hours for category chair, 12 hours for airplane, and 24 hours for car. The training batch size is set to 1. We use Adam optimizer with lr=0.0001, beta1=0.9, beta2=0.999.

A.2. The generator and discriminator masks

First, To ensure each empty voxel in the input content shape leads to empty voxels in its corresponding area of the output, we mask out voxels generated outside a predefined valid region. The region is denoted as the generator mask. There are two masking options: the "strict" generator mask, by upsampling the occupied voxels in the content shape to the desired resolution; and the "loose" generator mask, by upsampling the occupied voxels in the content shape after dilating them by 1 voxel. In both cases we use nearest-neighbor upsampling. In our system, we apply the "loose" generator mask to the raw generator output to keep only voxels within the area of the mask. The reason for using the "loose" mask is to allow the generator to have some freedom to accommodate for different styles as well as allow for more topological variations as the dilation may close holes. The generator mask enables the generator to focus its capacity solely on producing plausible voxels within the valid region.

Second, to ensure each occupied voxel in the input content shape leads to creation of fine voxels in its corresponding area of the output, we require that an occupied coarse voxel is also occupied in the downsampled version of the generator output. We achieve this by training the discriminator to penalize lack of voxels. If all real patches used in training have at least one voxel occupied at their center 4^3 areas, then any patches that have empty 4^3 center areas will be considered fake under the view of the discriminator. Therefore, the discriminator will encourage all input patches to have occupied voxels in their center areas. Hence, we can encourage voxels to be generated inside the desired region by *a.* training the discriminator using patches with occupied center areas as real patches, and *b.* training the generator by feeding to the discriminator those local patches that should have their center areas occupied. These two can be done easily by applying binary masks to the discriminator to only keep the signals of the desired patches. For the real patches, given a detailed shape, we can obtain a *discriminator mask* by checking each local patch for whether their center areas are occupied by at least one voxel. For the fake (generated) shape, we obtain its *discriminator mask* by upsampling the content shape via nearest-neighbor. In our experiments, we use discriminator masks with 1/2 of the resolution of the detailed shapes so that the entire model can fit into the GPU memory.

A.3. Style-content hybrids

We show more results of style-content hybrid shapes in Figure 2 3 4 5 6 7 8. Note that we lift the bilateral symmetry assumption for category motorbike, laptop, and plant.

A.4. Latent space

We show a visualization of the style space for airplanes in Figure 9 and cars in Figure 10. The visualization for chairs can be found in the main paper.

A.5. Evaluation metrics

To quantitatively evaluate the quality of the generated shapes, we propose the following metrics.

Strict-IOU and Loose-IOU. (higher better) Ideally, the downsampled version of a generator output should be identical to the input content shape. Therefore, we can use the IOU (Intersection over Union) between the downsampled voxels and the input voxels to evaluate how much the output shape respects the input. We use max-pooling as the downsampling method, and the Strict-IOU is defined as described above. However, since we relaxed the constraints (see Sec 3.1 of the main paper) so that the generator is allowed to generate shapes in a dilated region, we define Loose-IOU as a relaxed version of IOU to ignore the voxels in the dilated portion of the input:

$$\text{Loose-IOU} = \frac{|V_{in} \cap (V_{out} \cap V_{in})|}{|V_{in} \cup (V_{out} \cap V_{in})|} = \frac{|V_{in} \cap V_{out}|}{|V_{in}|}. \quad (1)$$

where V_{in} and V_{out} are input voxels and downsampled output voxels, and $|V|$ counts the number of occupied voxels in V. Note that our generated shape is guaranteed to be within the region of the dilated input due to the generator mask.

LP-IOU and LP-F-score (higher better). If all local patches from an output shape are copied from the given detailed shapes, it is likely that the output shape looks plausible, at least locally. Therefore, we define the Local Plausibility (LP) to be the percentage of local patches in the output shape that are "similar" to at least one local patch in the detailed shapes. Specifically, we define the distance between two patches to be their IOU or F-score. For LP-IOU, we mark the two patches as "similar" if the IOU is above 0.95; for LP-F-score, we mark "similar" if the F-score is above 0.95. The F-score is computed with a distance threshold of 1 (voxel). In our experiments, we sample 12^3 patches

in a voxel model. The patch size is a bit less than the receptive field of our discriminator to reduce computational complexity. In addition, we want to avoid sampling featureless patches that are mostly inside or outside the shape, therefore we only sample surface patches that have at least one occupied voxel and one unoccupied voxel at their center 2^3 areas. We sample 1000 patches in each testing shape, and compare them with all possible patches in the detailed shapes.

Div-IOU and Div-F-score (higher better). For the same input shape, different style codes should produce different outputs respecting the styles. Therefore, we want to have a metric that evaluates the diversity of the outputs with respect to the styles. During the computation of the LP, we obtain N_{ijk}, the number of local patches from input i, upsampled with style j, that are "similar" to at least one patch in detailed shape k. In an ideal case, any input i upsampled with style j only copies patches from detailed shape j, therefore we have $j = \max_k N_{ijk}$. However, since the input shape might introduce style bias (e.g., a local structure that can only be found in a specific detailed shape), we denote N_{ik} to be the mean of N_{ijk} over all possible j, and use it to remove such bias. The diversity is defined as

$$\text{Div} = \mathbb{E}_{i,j}[\mathbb{1}(j = \text{argmax}_k(N_{ijk} - N_{ik}))]. \qquad (2)$$

We obtain Div-IOU and Div-F-score based on the distance metrics for patches.

Cls-score (lower better). If the generated shapes are indistinguishable from real samples, a well-trained classification network will not be able to classify whether a shape is real or fake. We can evaluate the plausibility of the generated shapes by training such a network and inspect the classification score. However, the network may easily overfit if we directly input 3D voxel models, since we have limited amount of real data. Therefore, we opt to use rendered images for this task. We train a ResNet [2] using high-resolution voxels (from which content shapes are downsampled) as real samples, and our generated shapes as fake samples. The samples are rendered to obtain 24 256^2 images from random views. The images are randomly cropped to 10 64^2 small patches and feed into the network for training. We use the mean classification accuracy as the metric for evaluating plausibility, denoted as Cls-score.

FID-all and FID-style (lower better). Since our method generates shapes for a single category, it is not well suited for evaluation with Inception Score [4]. However, we borrow the idea from Fréchet Inception Distance (FID) [3] and propose a customized FID as follows. We first train a 3D CNN classification network on ShapeNet with 128^3 or 256^3 voxels depending on the input resolution. Afterwards, we use the last hidden layer (512-d) as the activation features for computing FID. We use FID to compare our generated shapes with all high-resolution voxels from which content shapes are downsampled, denoted as FID-all; or with a group of detailed shapes, denoted as FID-style.

Evaluation details For LP and Div, we evaluate on 320 generated shapes (20 contents $\times$ 16 styles) since they are computationally expensive. For other metrics we evaluate on 1600 generated shapes (100 contents $\times$ 16 styles). We evaluate Div and FID-style with the first 16 styles, and LP with all 64 styles.

A.6. Ablation study

We provide all quantitative results for our ablation experiments in this section. The numbers for chairs can be found in Table 1. The numbers for cars can be found in Table 2. The numbers for airplanes can be found in Table 3.

A.7. GUI application

The video is available at https://youtu.be/xIQ0aslpn8g. We obtain the 2D style space via T-SNE embedding. Afterwards, we consider each style as a 2D point and obtain the Delaunay triangulation of the 2D style space. The 8D latent style code for a given 2D point can be computed by finding which triangle it is inside and compute a linear interpolation among the three 8D latent codes of the three vertices via barycentric coordinates.

References

[1] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 4, 5, 6

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017. 2

[4] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 2

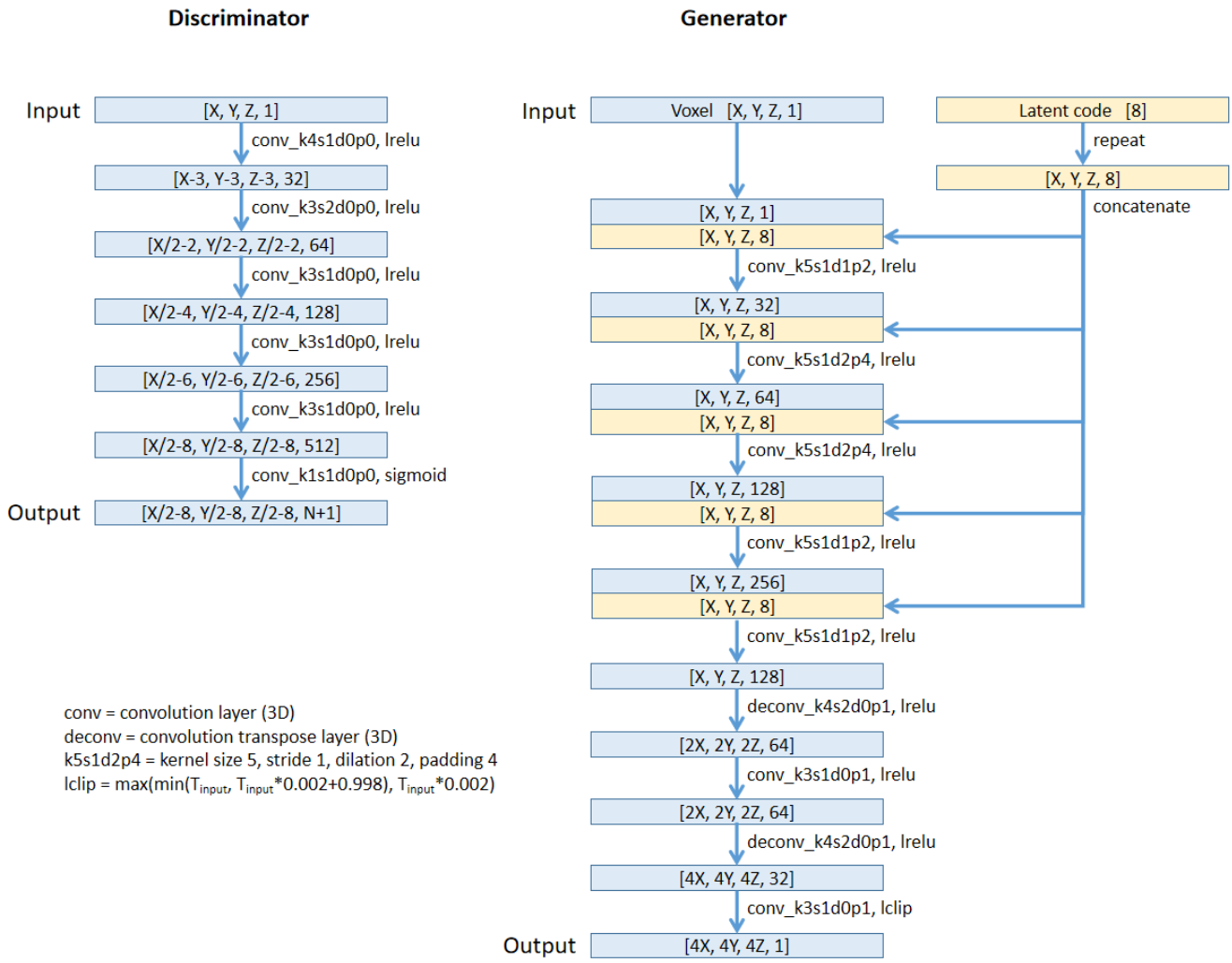

Figure 1: The detailed network architectures. Note that the generator for category chair with 32^3 inputs has smaller receptive fields by replacing all kernel-5 convolution layers with kernel-3 convolution layers.

Figure 2: Results by upsampling coarse chairs with different style codes. We show on the top the detailed shapes that correspond to the input style codes. The input shapes are coarse voxels in the first 6 rows, and downsampled versions of shapes generated by IM-GAN [1] in the last 5 rows. The input resolution is 32^3 and the output resolution is 128^3.

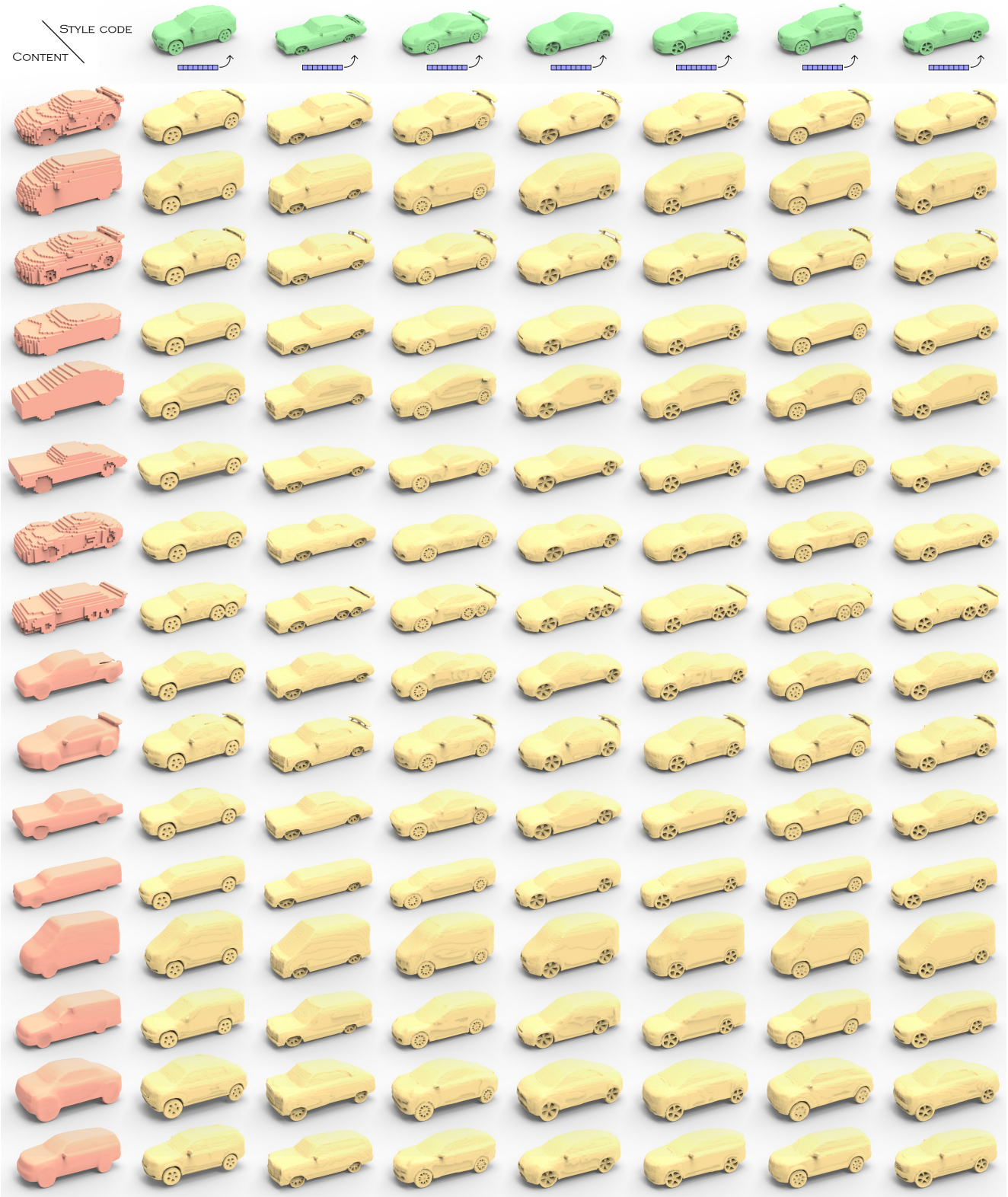

Figure 3: Results by upsampling coarse cars with different style codes. We show on the top the detailed shapes that correspond to the input style codes. The input shapes are coarse voxels in the first 8 rows, and downsampled versions of shapes generated by IM-GAN [1] in the last 8 rows. The input resolution is 64^3 and the output resolution is 256^3.

Figure 4: Results by upsampling coarse airplanes with different style codes. We show on the top the detailed shapes that correspond to the input style codes. The input shapes are coarse voxels in the first 6 rows, and downsampled versions of shapes generated by IM-GAN [1] in the last 7 rows. The input resolution is 64^3 and the output resolution is 256^3.

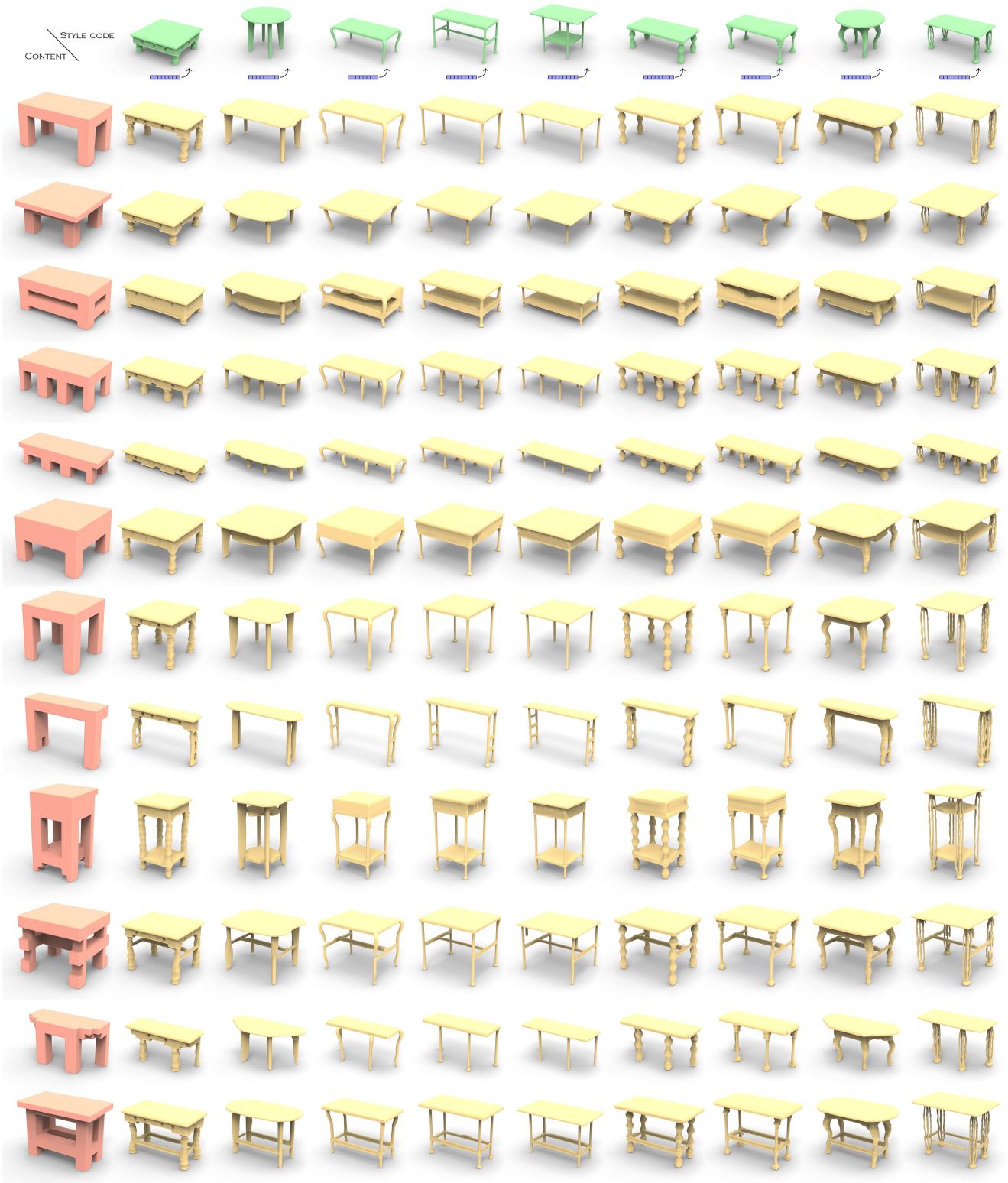

Figure 5: Results by upsampling coarse tables with different style codes. We show on the top the detailed shapes that correspond to the input style codes. The input shapes are coarse voxels. The input resolution is 16^3 and the output resolution is 128^3.

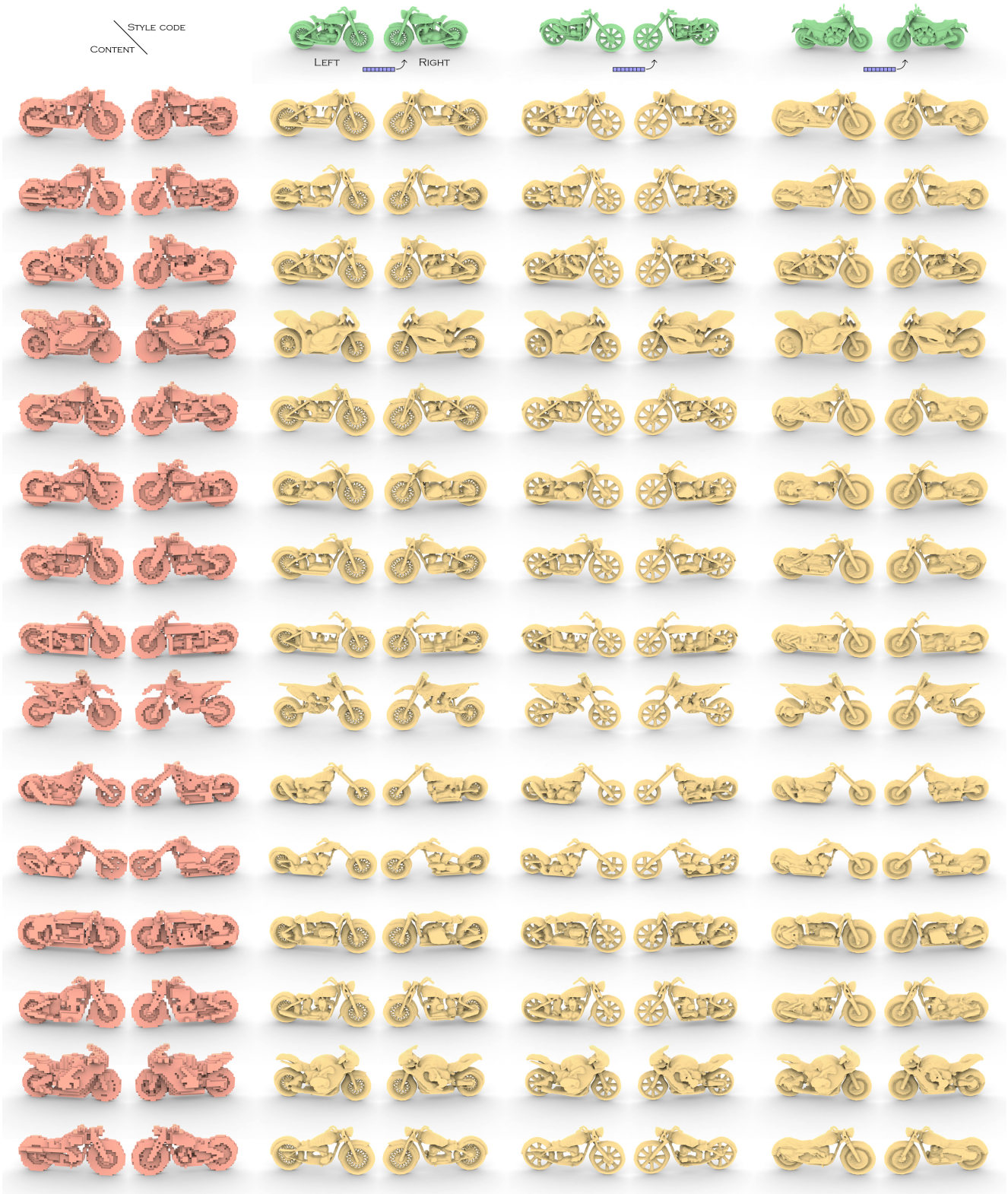

Figure 6: Results by upsampling coarse motorbikes with different style codes. Note that we lift the bilateral symmetry assumption for this category. We show on the top the detailed shapes that correspond to the input style codes. The input shapes are coarse voxels. The input resolution is 64^3 and the output resolution is 256^3.

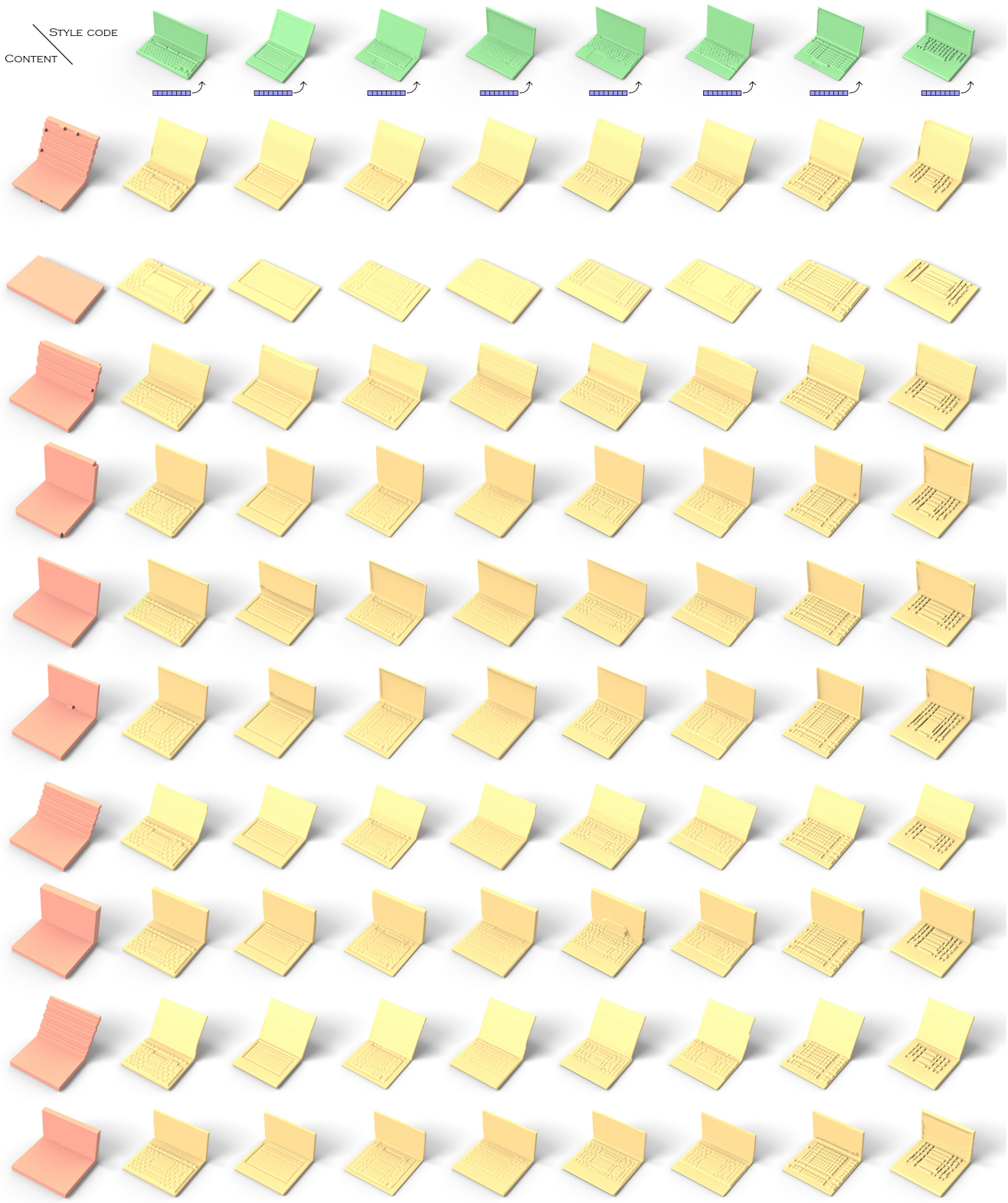

Figure 7: Results by upsampling coarse laptops with different style codes. Note that we lift the bilateral symmetry assumption for this category. We show on the top the detailed shapes that correspond to the input style codes. The input shapes are coarse voxels. The input resolution is 32^3 and the output resolution is 256^3.

Figure 8: Results by upsampling coarse plants with different style codes. Note that we lift the bilateral symmetry assumption for this category. We show on the top the detailed shapes that correspond to the input style codes. The input shapes are coarse voxels. The input resolution is 32^3 and the output resolution is 256^3.

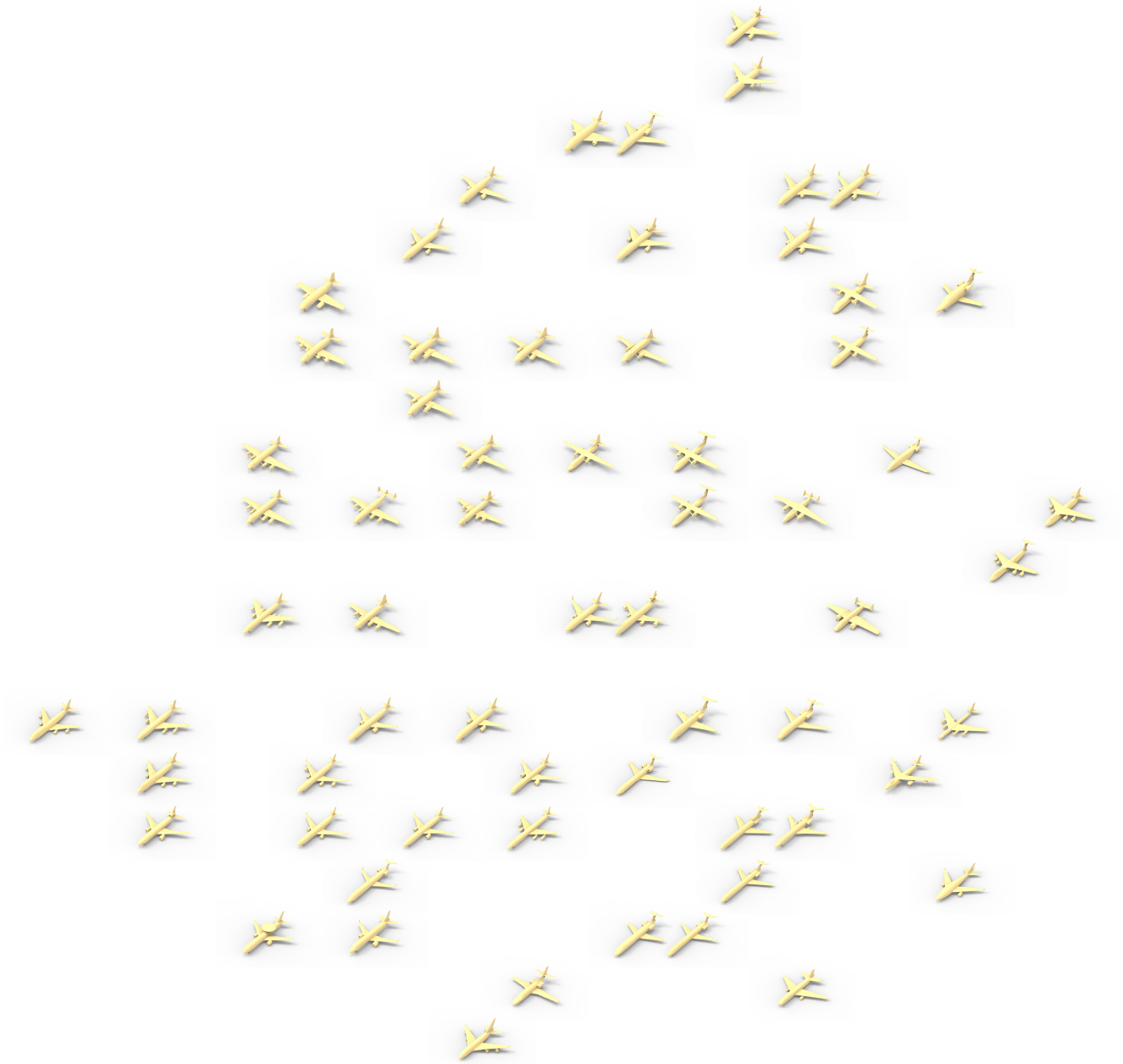

Figure 9: Visualization of 64 latent codes for airplanes via T-SNE embedding. For each latent code, the corresponding style shape is displayed in its location.

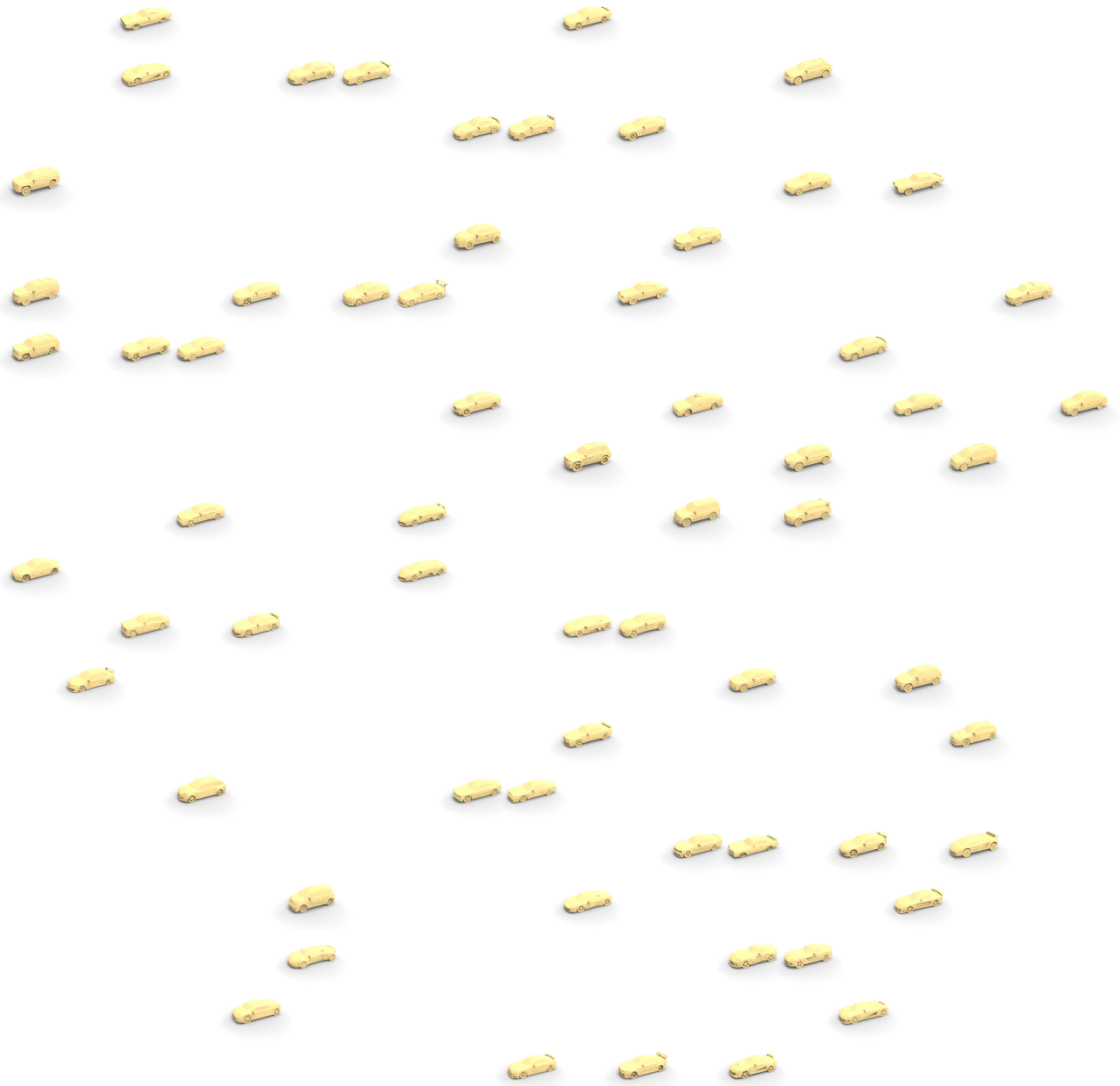

Figure 10: Visualization of 64 latent codes for cars via T-SNE embedding. For each latent code, the corresponding style shape is displayed in its location.

	Strict-IOU ↑	Loose-IOU ↑	LP-IOU ↑	LP-F-score ↑	Div-IOU ↑	Div-F-score ↑	Cls-score ↓	FID-all ↓	FID-style ↓
Recon. only	0.976	0.993	0.260	0.935	0.325	0.188	0.627	53.2	411.7
No Gen. mask	0.655	0.792	0.452	0.973	0.825	0.806	0.672	121.9	379.9
Strict Gen. mask	0.587	0.587	0.344	0.941	0.150	0.100	0.750	305.5	548.2
No Dis. mask	0.145	0.167	N/A	N/A	N/A	N/A	0.843	2408.9	2714.1
Conditional Dis. 1	0.947	0.981	0.259	0.949	0.291	0.194	0.593	51.3	402.7
Conditional Dis. 3	0.928	0.977	0.246	0.963	0.197	0.206	0.603	55.8	418.2
Proposed method*	0.673	0.805	0.432	0.973	0.800	0.816	0.644	113.1	372.5
$\alpha = 0.0$, $N = 16$	0.704	0.840	0.604	0.956	0.147	0.128	0.695	111.2	409.7
$\alpha = 0.2$, $N = 16$	0.583	0.750	0.527	0.971	0.875	0.934	0.667	115.5	371.5
$\alpha = 0.5$, $N = 16$	0.570	0.738	0.506	0.970	0.997	0.972	0.690	114.1	367.1
No L_{GAN}^{global}, $N = 16$	0.558	0.735	0.491	0.963	1.000	0.981	0.692	125.9	390.3
$\alpha = 0.0$, $N = 32$	0.763	0.864	0.551	0.962	0.184	0.156	0.598	131.2	391.7
$\alpha = 0.2$, $N = 32$	0.652	0.812	0.495	0.974	0.838	0.831	0.636	103.6	390.1
$\alpha = 0.5$, $N = 32$	0.598	0.757	0.470	0.974	0.934	0.934	0.662	111.1	380.0
No L_{GAN}^{global}, $N = 32$	0.561	0.728	0.462	0.969	0.997	0.984	0.690	109.1	368.2
$\alpha = 0.0$, $N = 64$	0.798	0.868	0.496	0.983	0.163	0.128	0.589	162.5	405.2
$\alpha = 0.2$, $N = 64$	0.781	0.864	0.423	0.985	0.353	0.334	0.619	109.2	370.3
$\alpha = 0.5$, $N = 64$*	0.673	0.805	0.432	0.973	0.800	0.816	0.644	113.1	372.5
No L_{GAN}^{global}, $N = 64$	0.578	0.741	0.426	0.965	0.950	0.988	0.669	116.3	381.8
$\sigma = 0.0$	0.915	0.952	0.435	0.943	0.153	0.125	0.544	71.9	385.7
$\sigma = 0.5$	0.869	0.919	0.493	0.952	0.172	0.144	0.580	101.2	379.5
$\sigma = 1.0$*	0.673	0.805	0.432	0.973	0.800	0.816	0.644	113.1	372.5
$\sigma = 1.5$	0.592	0.719	0.296	0.985	0.944	0.903	0.667	171.2	413.0
$\sigma = 2.0$	0.565	0.614	0.208	0.982	0.575	0.666	0.711	244.8	482.7
$\beta = 0.0$	0.730	0.815	0.279	0.967	0.178	0.269	0.669	129.9	391.1
$\beta = 5.0$	0.652	0.785	0.448	0.974	0.822	0.775	0.642	135.4	378.7
$\beta = 10.0$*	0.673	0.805	0.432	0.973	0.800	0.816	0.644	113.1	372.5
$\beta = 15.0$	0.677	0.803	0.443	0.974	0.788	0.744	0.660	132.2	391.2
$\beta = 20.0$	0.672	0.794	0.422	0.976	0.797	0.813	0.651	125.0	380.8

Table 1: Quantitative results for our ablation experiments on chairs. "N/A" is due to empty outputs. The models with * are the same model.

	Strict-IOU ↑	Loose-IOU ↑	LP-IOU ↑	LP-F-score ↑	Div-IOU ↑	Div-F-score ↑	Cls-score ↓	FID-all ↓	FID-style ↓
Recon. only	0.991	0.998	0.760	0.998	0.172	0.084	0.493	153.4	457.0
No Gen. mask	0.957	0.988	0.741	0.998	0.928	0.825	0.506	72.7	347.2
Strict Gen. mask	0.829	0.829	0.751	0.995	0.159	0.084	0.538	303.2	569.3
No Dis. mask	0.908	0.930	0.722	0.999	0.356	0.359	0.511	81.6	274.4
Conditional Dis. 1	0.924	0.947	0.738	0.999	0.997	0.853	0.501	119.6	427.2
Conditional Dis. 3	0.955	0.988	0.759	0.999	0.956	0.706	0.490	83.1	364.1
Proposed method*	0.953	0.964	0.730	0.998	0.584	0.456	0.494	113.8	401.7
$\alpha = 0.0$, $N = 16$	0.882	0.987	0.832	0.996	0.275	0.238	0.600	1069.9	1478.7
$\alpha = 0.2$, $N = 16$	0.905	0.978	0.766	0.998	1.000	0.934	0.506	79.8	372.3
$\alpha = 0.5$, $N = 16$	0.909	0.975	0.772	0.999	1.000	0.941	0.492	84.5	377.7
No L_{GAN}^{global}, $N = 16$	0.900	0.972	0.764	0.998	1.000	0.947	0.500	79.6	377.2
$\alpha = 0.0$, $N = 32$	0.927	0.987	0.844	0.999	0.134	0.128	0.582	875.2	1251.7
$\alpha = 0.2$, $N = 32$	0.932	0.985	0.753	0.999	1.000	0.831	0.498	86.5	373.2
$\alpha = 0.5$, $N = 32$	0.922	0.979	0.756	0.999	1.000	0.909	0.507	77.2	356.1
No L_{GAN}^{global}, $N = 32$	0.910	0.970	0.745	0.998	1.000	0.928	0.497	68.0	357.1
$\alpha = 0.0$, $N = 64$	0.959	0.987	0.825	0.998	0.091	0.119	0.517	651.9	1019.5
$\alpha = 0.2$, $N = 64$*	0.955	0.988	0.759	0.999	0.956	0.706	0.490	83.1	364.1
$\alpha = 0.5$, $N = 64$	0.942	0.986	0.767	0.999	0.975	0.806	0.500	123.9	414.1
No L_{GAN}^{global}, $N = 64$	0.927	0.976	0.739	0.998	1.000	0.931	0.502	62.6	338.2
$\sigma = 0.0$	0.977	0.994	0.763	0.995	0.119	0.075	0.499	223.3	548.7
$\sigma = 0.5$	0.981	0.996	0.773	0.998	0.084	0.100	0.481	284.4	626.9
$\sigma = 1.0$*	0.955	0.988	0.759	0.999	0.956	0.706	0.490	83.1	364.1
$\sigma = 1.5$	0.938	0.983	0.750	0.999	0.991	0.838	0.490	85.6	363.9
$\sigma = 2.0$	0.953	0.979	0.780	0.999	0.744	0.438	0.505	151.9	448.2
$\beta = 0.0$	0.725	1.000	0.000	0.999	1.000	0.081	0.754	2759.8	3273.4
$\beta = 5.0$	0.946	0.986	0.745	0.999	0.975	0.866	0.490	57.2	320.2
$\beta = 10.0$*	0.955	0.988	0.759	0.999	0.956	0.706	0.490	83.1	364.1
$\beta = 15.0$	0.958	0.989	0.753	0.999	0.894	0.753	0.500	75.8	350.0
$\beta = 20.0$	0.950	0.985	0.750	0.998	0.994	0.878	0.505	64.7	334.8

Table 2: Quantitative results for our ablation experiments on cars. The models with * are the same model.

	Strict-IOU ↑	Loose-IOU ↑	LP-IOU ↑	LP-F-score ↑	Div-IOU ↑	Div-F-score ↑	Cls-score ↓	FID-all ↓	FID-style ↓
Recon. only	0.966	0.980	0.465	0.999	0.166	0.100	0.493	64.8	328.6
No Gen. mask	0.884	0.934	0.477	0.999	0.413	0.259	0.525	66.1	323.7
Strict Gen. mask	0.487	0.487	0.380	0.974	0.069	0.072	0.642	1252.5	1196.9
No Dis. mask	0.508	0.564	0.277	0.998	0.084	0.141	0.539	552.6	859.4
Conditional Dis. 1	0.782	0.855	0.477	0.997	0.347	0.294	0.493	667.5	773.1
Conditional Dis. 3	0.809	0.854	0.443	0.996	0.094	0.119	0.524	717.4	795.9
Proposed method*	0.875	0.947	0.474	0.998	0.516	0.353	0.487	57.3	340.9
$\alpha = 0.0$, $N = 16$	0.843	0.921	0.510	0.999	0.069	0.059	0.516	93.9	331.7
$\alpha = 0.1$, $N = 16$	0.764	0.890	0.487	0.997	0.825	0.659	0.502	100.6	307.7
$\alpha = 0.2$, $N = 16$	0.720	0.845	0.501	0.990	0.994	0.897	0.504	106.3	308.0
No L_{GAN}^{global}, $N = 16$	0.657	0.805	0.516	0.986	1.000	0.947	0.504	132.5	354.9
$\alpha = 0.0$, $N = 32$	0.883	0.946	0.503	1.000	0.059	0.066	0.515	95.2	350.9
$\alpha = 0.1$, $N = 32$	0.835	0.926	0.459	0.999	0.734	0.538	0.503	71.4	329.4
$\alpha = 0.2$, $N = 32$	0.777	0.887	0.481	0.997	0.947	0.788	0.520	79.5	325.6
No L_{GAN}^{global}, $N = 32$	0.675	0.818	0.493	0.989	1.000	0.941	0.493	135.2	377.0
$\alpha = 0.0$, $N = 64$	0.898	0.959	0.499	0.999	0.059	0.056	0.503	80.2	353.6
$\alpha = 0.1$, $N = 64$*	0.875	0.947	0.474	0.998	0.516	0.353	0.487	57.3	340.9
$\alpha = 0.2$, $N = 64$	0.831	0.921	0.463	0.998	0.756	0.600	0.498	67.7	332.0
No L_{GAN}^{global}, $N = 64$	0.707	0.833	0.478	0.991	0.997	0.934	0.489	105.8	354.8
$\sigma = 0.0$	0.802	0.892	0.422	0.988	0.059	0.041	0.495	205.1	363.0
$\sigma = 0.5$	0.883	0.911	0.464	0.997	0.066	0.084	0.508	81.7	327.2
$\sigma = 1.0$*	0.875	0.947	0.474	0.998	0.516	0.353	0.487	57.3	340.9
$\sigma = 1.5$	0.845	0.899	0.451	0.997	0.469	0.409	0.518	175.2	372.7
$\sigma = 2.0$	0.730	0.767	0.534	0.993	0.278	0.384	0.549	847.1	818.3
$\beta = 0.0$	0.384	1.000	0.000	0.940	0.659	0.050	0.811	6342.0	5491.9
$\beta = 5.0$	0.892	0.948	0.460	0.999	0.325	0.188	0.492	73.2	310.0
$\beta = 10.0$*	0.875	0.947	0.474	0.998	0.516	0.353	0.487	57.3	340.9
$\beta = 15.0$	0.895	0.952	0.468	0.999	0.450	0.319	0.500	63.8	354.6
$\beta = 20.0$	0.872	0.946	0.459	0.998	0.531	0.475	0.517	69.2	311.6

Table 3: Quantitative results for our ablation experiments on airplanes. The models with * are the same model.