Supplementary Material
Doodle Your 3D: From *Abstract* Freehand Sketches to *Precise* 3D Shapes

Hmrishav Bandyopadhyay[1] Subhadeep Koley[1,2] Ayan Das[1] Ayan Kumar Bhunia[1]
Aneeshan Sain[1] Pinaki Nath Chowdhury[1] Tao Xiang[1,2] Yi-Zhe Song[1,2]
[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{h.bandyopadhyay, s.koley, a.das, a.bhunia, a.sain, p.chowdhury, y.song}@surrey.ac.uk

Segmenting Hand-Drawn Sketches We demonstrate the generalisation of sketch-based shape generation to hand-drawn sketches after being trained on synthetic sketches only (Fig. 6). To further explore this generalisation, we include qualitative results from our auxiliary segmentation task on hand-drawn sketches from the AmateurSketch-3D dataset [2] in Fig. S1. We note that despite being trained on synthetic sketches, we can generalise and segment hand-drawn sketches fairly accurately.

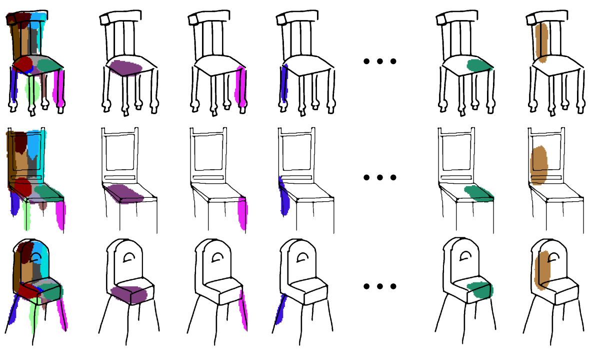


Figure S1. Segmentation results (right) with predicted semantic segmentation map for all 16 parts (left).

Further Details on Shape Decomposition Here, we clarify Sec. 3.2, elaborating on the decomposition of shapes (as meshes) into part-latents Z. To represent a ground truth shape M with decoder $\mathcal{D}$, *(i)* the ground-truth occupancy values for coordinates $X = (x, y, z)$ in and around the shape are recorded along with the coordinates themselves. *(ii)* The decoder $\mathcal{D}$ is trained to decode a randomly initialised part-latent Z to an implicit code as $I = \mathcal{D}(Z)$. *(iii)* Finally I is used in implicit function f_θ to *predict* occupancy values for known coordinates X, as $O_I = f_\theta(I, X)$. During pre-training, $\mathcal{D}, Z$, and f_θ are optimized together with binary cross entropy loss ($\mathcal{L}_{\text{BCE}}$) against the recorded (ground truth) occupancy values.

The decomposition of shape M occurs through its representation as part-latent $Z \in \mathbb{R}^{m \times d}$ using the decoder $\mathcal{D}$. Ideally, after disentanglement each latent code ω_i in part-latent $Z = \{\omega_i\}_{i=1}^{m}$, sufficiently and independently represents individual components of shape M, thus successfully breaking down M into m parts represented as $\{\omega_i\}_{i=1}^{m}$. This disentanglement of part-latents is necessary for independent representation of shape parts. However, $\mathcal{L}_{BCE}$ is not enough for this disentanglement, as it only encourages the final output shape to match shape M, thus ignoring *part-level* correspondence.

To optimize for disentanglement, part-latent Z is projected to part structural representation $Z \rightarrow Z_p$ and part volumetric descriptor $Z \rightarrow Z_g$. Particularly important for this representation, each part's volumetric descriptor is a parametric 3D Gaussian that captures the probability of a 3D coordinate X belonging to that part. This establishes a relationship between coordinates in the 3D space and part-latent Z, thereby representing the volume of each part in 3D. For decomposition and disentanglement respectively, this relation of part and 3D coordinates is pivotal for *(i)* dissipating 3D Gaussians Z_g over the entire shape volume (M) and *(ii)* specifically disentangling overlapped, or closely-placed parts (as information is commonly entangled here [1]), by computing distance between part-Gaussians in 3D.

Shape Interpolation

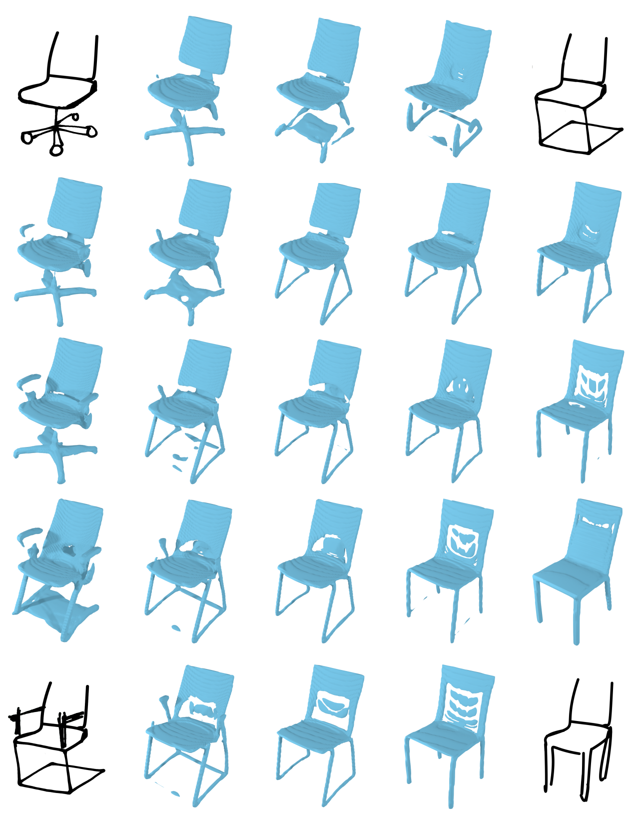
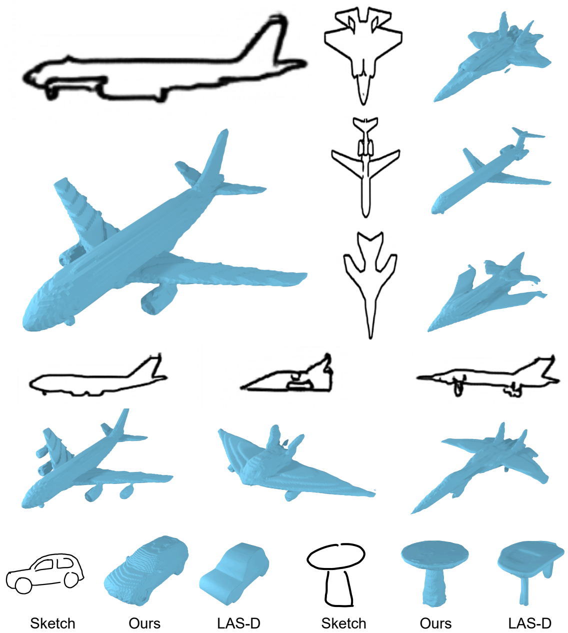


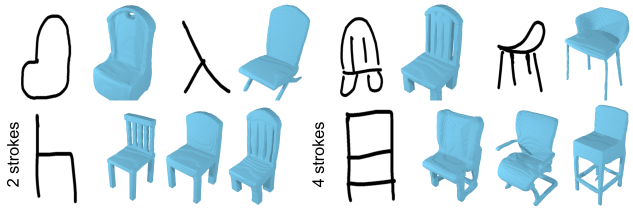
Figure S2. Interpolating shapes among four different sketches (at corners) from left to right and top to bottom.

Generation of Shapes from Other Categories In addition to chairs (Fig. 6), we perform sketch-based 3D generation of *airplanes*, *tables*, *rifles* and *cars*.



Additional Qualitative Results (Chairs)

Model Response to extreme inputs:



User feedback for generated shapes: We build an internal demo using Gradio for shape generation and editing, and ask 30 users to draw 10 sketches each on the demo-canvas and rate *(i)* the generated shapes on score from $1 \rightarrow 5$ (bad $\rightarrow$ excellent) based on how they match their expectation. We then ask the same users to edit their sketches and rate the edited shape based on *(ii)* localisation of edits, and *(iii)* quality of details added, using scores ($1 \rightarrow 5$). Users reported a mean opinion score (MOS) of $4.17/4.00$ for Ours/LAS-D generation quality, $4.91/4.35$ for localisation, and $4.30/3.90$ for quality of edits. We also obtained (iv) a satisfaction score of $4.37/3.55$ for Ours/LAS-D from the same users based on generation speed, shape quality, consistency, and resolution by rating from $1 \rightarrow 5$. None of them were linked to the project to prevent conflicts.

References

[1] Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or. Spaghetti: Editing implicit shapes through part aware generation. *ACM TOG*, 2022. 1

[2] Anran Qi, Yulia Gryaditskaya, Jifei Song, Yongxin Yang, Yonggang Qi, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Toward fine-grained sketch-based 3d shape retrieval. *IEEE TIP*, 2021. 1