# Sketch-guided Cage-based 3D Gaussian Splatting Deformation

Supplementary Material

## 8. Supplementary Material

### 8.1. Harmonic deformation map

A mapping $f : \Omega \to R^d$ is a harmonic mapping if and only if there exist two $C^2$ mappings $a$ and $b$ such that

$$f(p) = \oint_{q\in S} a(q)(\nabla G(q,p)\cdot \hat{n}(q)dA - \oint_{q\in S} b(q)G(q,p)dA, \tag{9}$$

where $G(q,p)$ is the fundamental solution of the Laplace equation and $\hat{n}(q)$ is the unit normal direction to the surface $S = \partial\Omega$ at the point $q$.

Considering the discrete setting in $d = 3$, suppose a $3D$ region $\Omega$ bounded by a triangular mesh $S = (V, F)$, $V$ is the vertices of $S$ and $F$ are faces. The discrete harmonic deformation map of the points $p \in \Omega$ can be expressed as[2]

$$f_{a,b}(p) = \sum_{t\in F} \int_{q\in t} a(q)\hat{\phi}(q,p)dA - \sum_{t\in F} \int_{q\in t} b_t\hat{\psi}(q,p)dA. \tag{10}$$

where $\hat{\phi}(q,p) = G(q,p)$, $\hat{\psi}(q,p) = G(q,p) \cdot \hat{n}(q)$, $a(q)$ is the piecewise linear interpolation of the values $a_i$, $a_j$,$a_k$ on the vertices of the triangle $t = (i,j,k) \in F$, $b_t$ is piecewise constant map defined by values at the faces $t$. This equation was considered by different papers[2, 22, 35], we used the analytics solution from [2, 35], that has a geometric interpretation and easier computation of the gradients.

By having the analytical solutions of the integrals, the discrete harmonic deformation map of the points $p \in \Omega$ can be expressed as a linear interpolation of the cage's vertices and the face normals respectively:

$$f_{a,b}(p) = \sum_{v\in V} a_v\phi(p) + \sum_{t\in F} b_t\psi(p), \tag{11}$$

where $a_v$ is the Cartesian coordinate of vertex $v$, $b_t$ is the normal of face $t$, $\phi(p)$ and $\psi(p)$ are two sets of scalar functions that the analytical expression was shown in [2, 35].

### 8.2. 3D-aware Score Distillation Sampling

Given a reference image $I_{ref}$ of an object $O$ in viewpoint $(elev, azim)$ ($elev$ is the elevation angle and $azim$ is the azimuth angle), a relative viewpoint transformation $(\Delta elev, \Delta azim)$, and a text embedding $z$, the image of the object $O$ seen in viewpoint $(elev + \Delta elev, azim + \Delta azim)$ can be given by denoising model $\epsilon_\phi$ with parameters $\phi$. Given an arbitrary differentiable parametric function $g_\theta$ that renders images, the 3D-aware SDS guidance that using 3D-aware diffusion model as the backbone is:

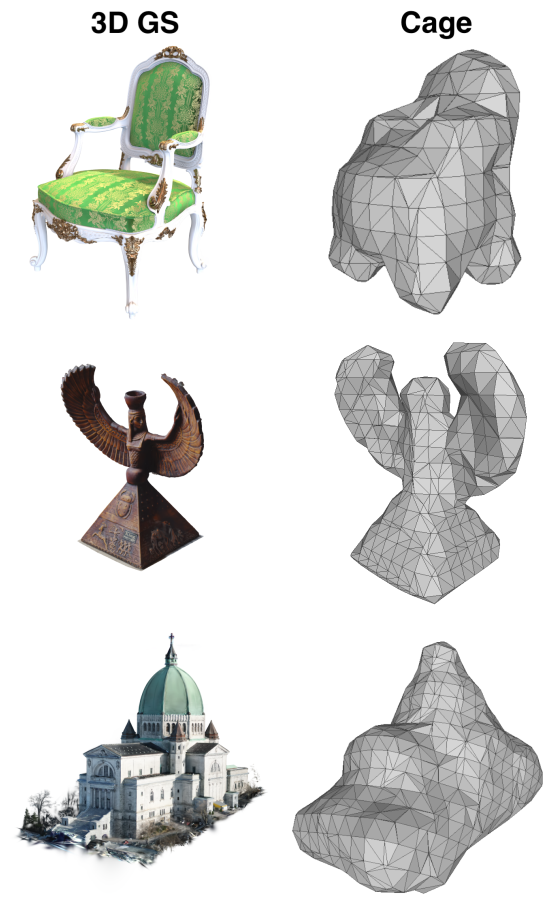


Figure 10. Example of 3D GS and their corresponding cage.

$$\nabla_\theta \mathcal{L}_{SDS}(I_{ref}, (\Delta elev, \Delta azim), x, z, \epsilon, t) = \\ \epsilon_\phi(x_t, I_{ref}, (\Delta elev, \Delta azim), z, t) - \epsilon)\frac{\partial x_t}{\partial \theta}, \tag{12}$$

where the $x$ is the rendering of the object in transformed viewpoint $(elev + \Delta elev, azim + \Delta azim)$, $t$ is a randomly sampled time step $t \sim \mathcal{U}(0,1)$ drawn from the uniform distribution, and noise $\epsilon \sim \mathcal{N}(0, I)$ following a normal distribution, and $x_t$ is the $x$ added the noise at time step $t$.

### 8.3. Example of Cages

We showed in figure 10 some examples of the cages used in our method.


Figure 11. Ablation of directly optimizing 3D GS.

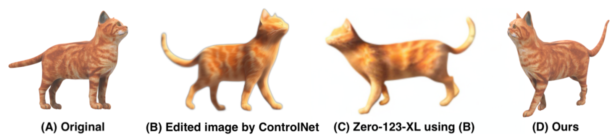


Figure 12. Comparison of using SVR and our method. (B): An edited rendering of (A). (C): A novel-view rendering of (B) using Zero-123-XL. (D): Deformed 3D GS rendering of our method from the same view direction of (C)

## 8.4. System's robustness to imprecise or rough sketches

The ControlNet has a guidance strength parameter ranging from 0 to 1 that decides how much it has to follow the guidance's direction. We set 1 to achieve the most precise control in all our experiments. However, we can set a lower guidance strength when the sketch is rough to mitigate some errors.

## 8.5. Ablations

**Finetune the 3D GS directly.**

As shown in Figure 11 (A), since directly optimizing the 3D GS lacks geometry regularization, the geometry can't be moved efficiently and the local structure was severely compromised.

**Using Single-View reconstruction(SVR).**

We show an example in Figure 12 of directly using SVR to edit 3D GS. 1) Editing rendering by sketch using ControlNet can drastically change the texture, as shown in Figure 12 (B). 2) Zero-123-XL can produce geometry inconsistency in novel-view synthesis, as shown in Figure 12 (C). In contrast, our method can preserve the identity better and have good geometry consistency as shown in Figure 12 (D).

**Numbers of views for 3D-aware SDS.**

We use a pseudo-random viewpoint sampling to prevent severe overlap, with 4 views covering the front, left, back, and right. Experiments with more views (e.g. 8) showed no significant difference but increased memory usage, as shown in figure 13.

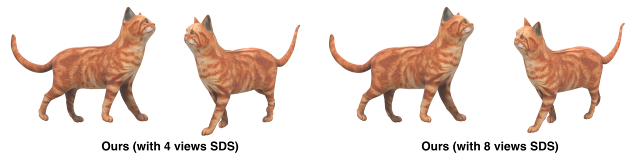


Figure 13. Ablation of numbers of views for 3D-aware SDS.


Figure 14. Failure case and mitigation: In side view A, the legs of the dinosaur are overlapped, which makes failure when we deform the two legs separately. We can rotate to sketch view B to achieve the deformation.

Figure 15. The reconstructed mesh using [10] for our UAV dataset.

## 8.6. Example of failure reconstructed base-mesh for mesh-binding method

The main paper states that the mesh-binding 3D GS method's [7] result significantly depends on the extracted mesh, which can fail in scenes with complex geometry and transparent components. As shown in figure 15, we show a reconstructed mesh of our UAV dataset using the reconstruction method [10] adopted in [7], which produced lots of holes and floating components.

## 8.7. Visualization of silhouettes' deformation.

We showed an example of how the input sketches deform the original silhouettes in figure 16.

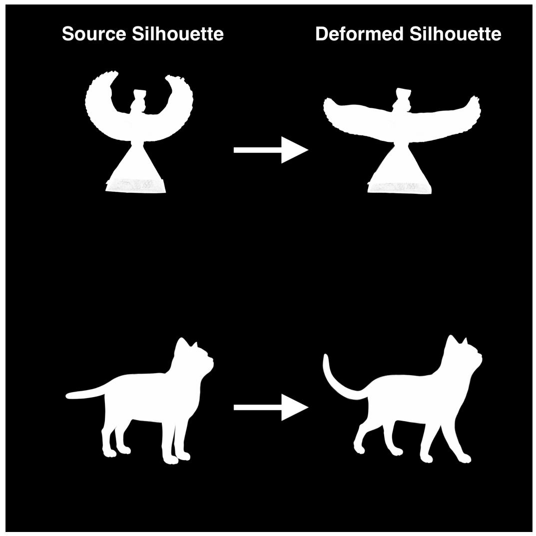


Figure 16. The visualization of the silhouettes' deformation.

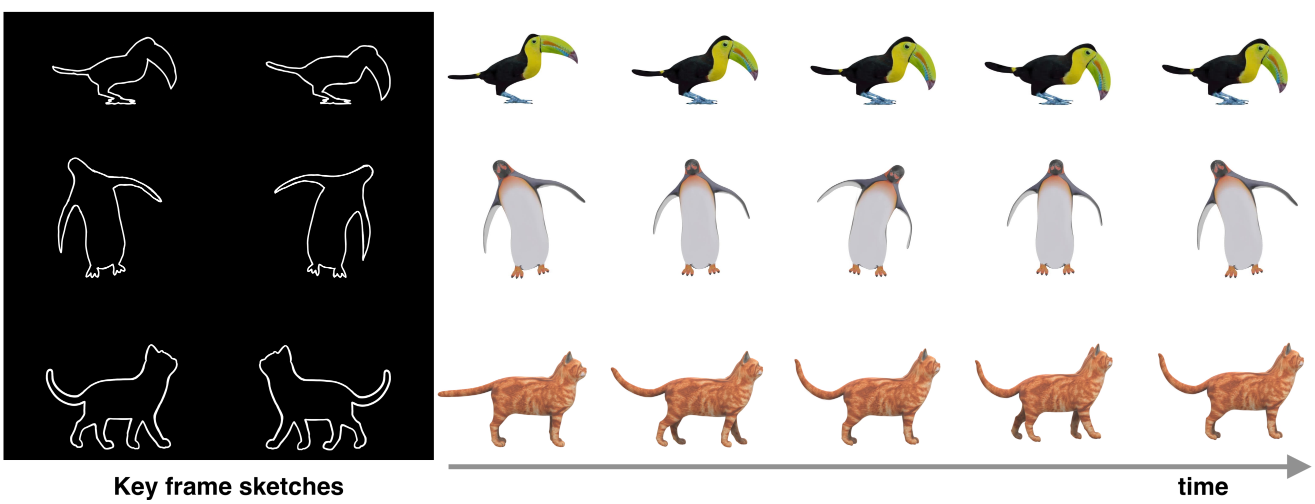


Figure 17. Animate static 3D GS by applying our method to key frames. Users can draw a few sketches for the key frames and apply our method to each key frame. The animation was generated by interpolating the deformed cages between key frames.