

MV2UV: Generating High-quality UV Texture Maps with Multiview Prompts

Supplementary Material

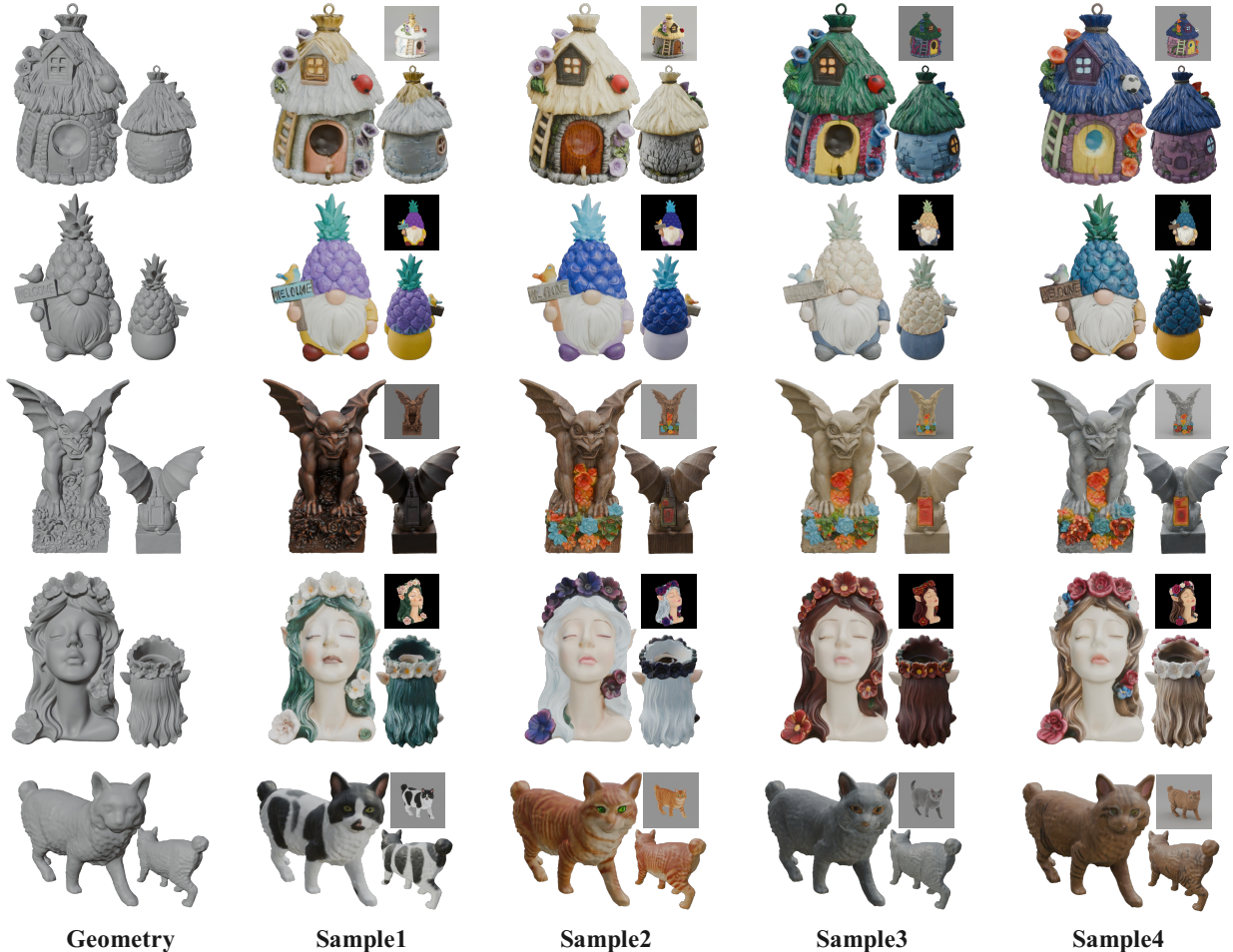


Figure 10. MV2UV can generate diverse, high-quality textures from different prompts, enabling flexible and controllable texture synthesis.

6. Training Details

We finetune our backbone network based on Stable Diffusion XL (SDXL) with frozen self attention and cross attention modules. To enhance the model’s ability to complete UV textures, we randomly drop each view with a probability of 0.1. We do not employ classifier-free guidance (CFG). Additionally, we shuffle the order of input multiview images as a form of data augmentation. The model is trained for 10 epochs on 80K samples with a learning rate of $5e-5$ and a batch size of 32.

7. Detail of Geometric Position Encoding

Position Embedding. We normalize both the position (x, y, z) and normal (n_x, n_y, n_z) maps to the range $[-1, 1]$

and concatenate them along the channel dimension to form the input 3D coordinates representation. To better capture high-frequency geometric signals, we apply a Fourier-based positional embedding to these coordinates, defined as:

$$\gamma(p) = [\sin(2_0\pi p), \cos(2_0\pi p), \dots, \sin(2_{L-1}\pi p), \cos(2_{L-1}\pi p)]$$

where each dimension p of the input coordinates is projected into a series of sinusoids. The hyperparameter L controls the highest frequency band and thus the fidelity of representable geometric details. We set $L = 10$ in all experiments.

Learnable Position Encoder. We further introduce a learnable position encoder that transforms the initial position embeddings into a multi-scale feature pyramid via a series of

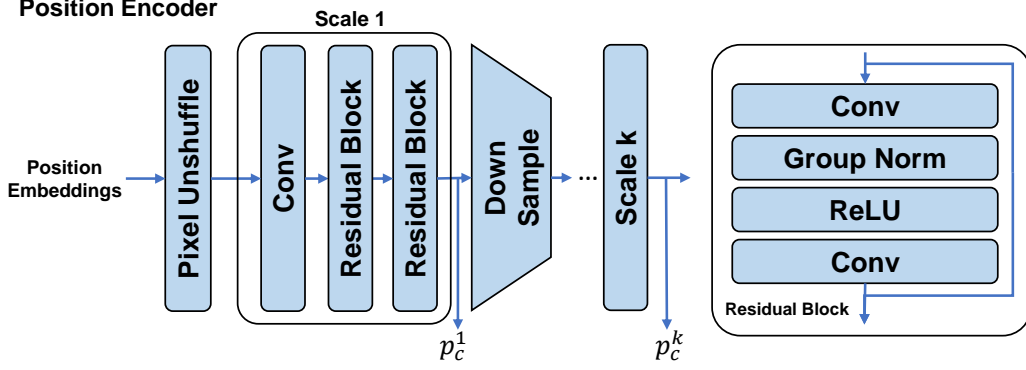


Figure 11. Detail of Learnable Position Encoder.

convolutional residual blocks. As shown in Fig. 11, the input embedding $p_{\text{in}} \in \mathbb{R}^{H \times W \times C}$ is first downsampled via a Pixel Unshuffle layer, rearranging it into a lower-resolution feature map of size $\mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times (C \times 8 \times 8)}$. At each scale, a convolutional layer followed by two residual blocks extracts positional features p_c^k . The final output is a pyramid of features $[p_c^1, \dots, p_c^k]$ whose spatial dimensions align with the corresponding feature maps in the diffusion model’s attention layers. This design enables fine-grained control over the UV texture generation process, facilitating high-precision detail synthesis across multiple spatial scales.

8. Diverse Texturing

Given an untextured 3D mesh, we first render a single view of the model and apply Nano Banana to perform example-based image generation, producing a reference image. This reference image is then fed into MV2UV to generate the corresponding texture map. As shown in Fig. 10, our approach can produce diverse, high-quality textures conditioned on different image prompts, enabling flexible and controllable texture synthesis for 3D shapes.