HOLOFUSION: Towards Photo-realistic 3D Generative Modeling Supplementary Material

Animesh Karnewar
UCL
a.karnewar@ucl.ac.uk

Niloy J. Mitra
UCL
n.mitra@ucl.ac.uk

Andrea Vedaldi
Meta AI
vedaldi@meta.com

David Novotny
Meta AI
dnovotny@meta.com

1. Views2Voxel-grid unprojection mechanism

Given a training video s containing frames I_j, we obtain the auxiliary grid $\bar{V} \in \mathbb{R}^{d \times S \times S \times S}$ of auxiliary features $\bar{V}_{:mno} \in [-1, 1]^d$ by using the following procedure. The values $m, n, o \in \mathbb{N}$ index the 3D grid of size $[S \times S \times S]$. We first project the 3D coordinate $\mathbf{p}^{\bar{V}}_{mno}$ of each grid vertex (corner) (m, n, o) to every video frame I_j and sample corresponding 2D image features. The 2D image features f^j_{mno} are obtained using a frozen ResNet-32 encoder [3] $E(I_j)$. We use bilinear interpolation for sampling continuous values and use zero-features for projected points that lie outside the image region on the 3D image plane. Thus, we obtain N_{frames} feature-vectors (corresponding to each frame in the video) for each grid element of the voxel-grid. We accumulate these features using the accumulator MLP $\mathcal{A}_{acc}$. The accumulator $\mathcal{A}_{acc}$ takes as input $[f^j_{mno}; v^j]$, where $[\,;\,]$ denotes concatenation and v^j corresponds to the viewing direction of the camera center of j^{th} frame, and outputs $[\sigma^j_{mno}; f'^j_{mno}]$. Here, the value σ^j corresponds to a weight for j^{th} feature vector and f'^j is an MLP transformed version of the input feature vector. Lastly, we compute the aggregated feature vector (for each of the voxel grid centers) as a weighted sum of the transformed features:

$$\bar{V}_{:mno} = F_{mno} = \sum_j \sigma^j_{mno} f'^j_{mno}. \tag{1}$$

2. Stable-DreamFusion baseline details

Implementation details Since no code is available for either of the two 2D diffusion distillation works, DreamFusion [7] and Magic3D [5], we resort to using the open-source implementation provided by the GitHub user ashawkey titled Stable-DreamFusion [10]. The differences between the implementation and the aforementioned research works are as follows:

1. While DreamFusion uses the Imagen [4] diffusion model, and Magic3D uses a combination of Ediff-I [1] and StableDiffusion [9], the implementation only uses StableDiffusion since Imagen and Ediff-I models are unavilable.
2. Since the used diffusion network of StableDiffusion performs diffusion in the latent-space, the implementation applies the SDS loss in the latent space and back-propagates the SDS gradients through the perceptual encoder.
3. The DIP generator (backbone) is implemented as a vanilla NeRF [6] as opposed to the Mip-NeRF [2] version as used by DreamFusion.

Prompt builder We use a simple procedure to generate the prompts used for distilling the samples from the Stable-DreamFusion [10] baseline. We first start by defining three python lists (aka. Prompt Builder Ingredients), viz. `objects`, `modifiers`, and `colors`. The `objects` list is set to verbose names of the categories from the Co3Dv2 [8] which we used in our experiments. Tab. 1 describes the values for the ingredient lists in full. Then we generate a shuffled list of all prompts by taking a full outer-product of the ingredients and ensuring correct grammar. Some of the generated final prompts are given in the Tab. 2.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 1

[4] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben

Table 1: The values of the builder ingredients `objects`, `modifiers`, and `colors` used for generating the prompts for stable-dreamfusion [10].

Prompt Builder Ingredient	Values used
`objects`	"apple", "water hydrant", "teddy bear", "donut"
`modifiers`	"unreal render of", "zoomed out unreal render of", "wide angle zoomed out unreal render of", "dslr photo of", "zoomed out dslr photo of", "wide angle zoomed out dslr photo of", "plastic", "metallic", "wooden", "furry"
`colors`	"red", "green", "blue", "yellow", "orange", "brown", "pink", "purple", "cyan", "magenta", "sky blue", "baby blue", "crimson", "lime", "teal", "violet", "sea green", "dusk", "gold", "silver"

Table 2: A sampling of the prompts generated after running the prompt-building process.

Sample prompts
"a plastic teddy bear", "a zoomed out unreal render of a water hydrant", "a wide angle zoomed out unreal render of a dusk donut", "a dslr photo of a water hydrant", "a plastic apple", "a sea green apple", "an unreal render of a red apple", "an unreal render of a silver donut", "a baby blue water hydrant", "a metallic donut"

Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[5] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022. 1

[6] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020. 1

[7] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[8] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common Objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *Proc. CVPR*, 2021. 1

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[10] Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion, 2022. https://github.com/ashawkey/stable-dreamfusion. 1, 2