3D-aware Image Generation using 2D Diffusion Models

Jianfeng Xiang[1,2] Jiaolong Yang[2] Binbin Huang[3] Xin Tong[2]

[1]Tsinghua University [2]Microsoft Research Asia [3]ShanghaiTech University

ICCV23 PARIS

Overview

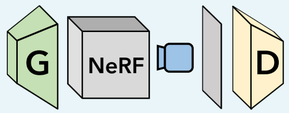
Background

Learning to generate 3D contents has become an increasingly prominent task due to its numerous applications.

Current Limitations

Predominant method rely on GANs for generative modeling.

- **Limited generative power**
- **Hard to scale up**

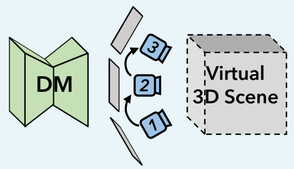


NeRF-GAN paradigm

Our Solution

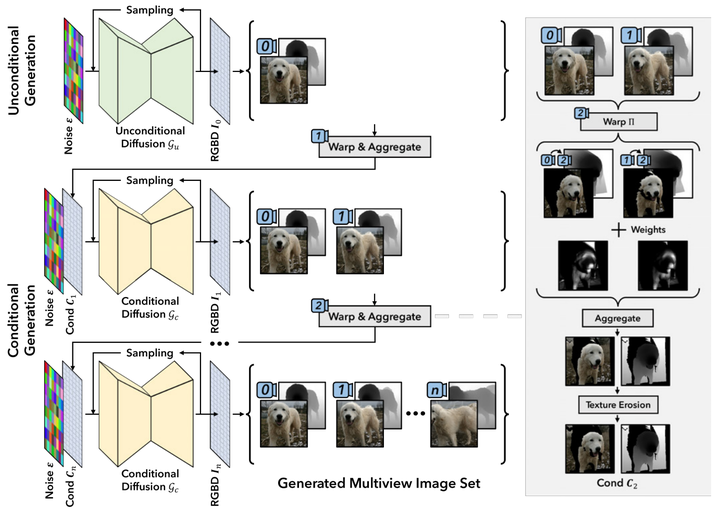
2D diffusion-based paradigm with sequential unconditional-conditional multi-view image generation formulation.

- **Superior generative power**
- **Scaling up to ImageNet**
- **Only using still images** with **depth** from depth estimators to construct the training data



IVID paradigm

We undertake 3D-aware generation on ImageNet. We show the capability of our method for large-angle generation from unaligned data (up to 360 degrees).



Our diffusion-based 3D-aware image generation trained on ImageNet

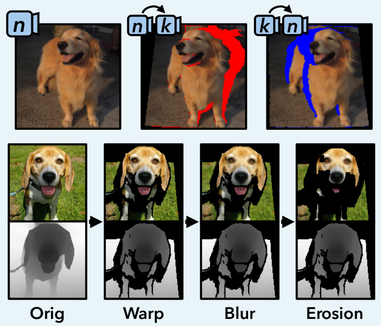
Methodology

Problem Formulation

- The distribution of 3D assets is equivalent to the joint distribution of its multi-view images.
- Factorize it into the production of **an unconditional distribution and a series of conditional distributions**.
- Approximate the conditional distributions by changing conditions to **warped images with holes** with **depth** information.

$$q_u(\mathbf{x}) = q_i(\Gamma(\mathbf{x}, \pi_0)) \cdot q_i(\Gamma(\mathbf{x}, \pi_1)|\Gamma(\mathbf{x}, \pi_0)) \cdot \cdots \cdot q_i(\Gamma(\mathbf{x}, \pi_N)|\Gamma(\mathbf{x}, \pi_0), \cdots, \Gamma(\mathbf{x}, \pi_{N-1}))$$

$$q_u(\mathbf{x}) \approx q_i(\Gamma(\mathbf{x}, \pi_0)) \cdot q_i(\Gamma(\mathbf{x}, \pi_1)|\Pi(\Gamma(\mathbf{x}, \pi_0), \pi_1)) \cdot \cdots \cdot q_i(\Gamma(\mathbf{x}, \pi_N)|\Pi(\Gamma(\mathbf{x}, \pi_0), \pi_N), \cdots)$$



Data Preparation

We construct training data using depth-based image warping. By **forward-backward warping** the images with estimated depth, we can construct training data **without actual multi-view pairs**.

Training

Two **DMs** are trained to fit the unconditional and conditional distributions, respectively.

Blur and texture erosion augmentations are applied to the warped images to boost performance.

Inference

By **sampling views iteratively** following the unconditional-conditional image generation process, the final 3D assets are obtained.

Aggregated conditioning is applied to ensure the novel view is conditioned on all previous sampled views.

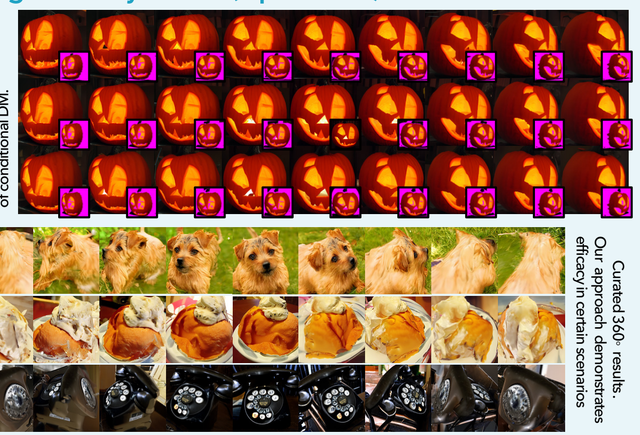
Fast free-view synthesis

Once a set of views uniformly covering the desired viewing range are generated, **fusion-based free-view synthesis** can be applied to interpolate viewpoint in **real-time**.

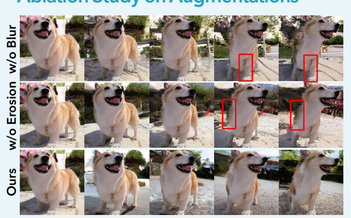
Results

Comparison

Large View Synthesis (Up to 360°)

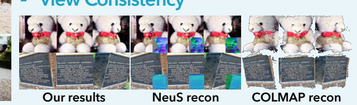


Ablation Study on Augmentations



Geometry Visualization

View Consistency



Project page:

https://jeffreyxiang.github.io/ivid/