Geometric Consistency for View Conditioned Diffusion Models

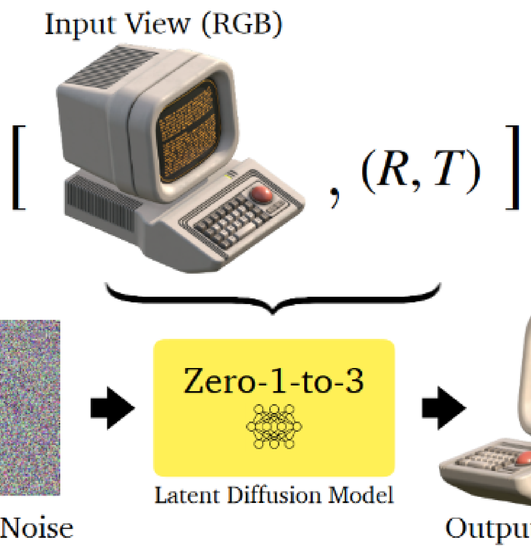
Josef Bengtson
Computer Vision Group
Supervisors: Fredrik Kahl, David Nilsson

CHALMERS
UNIVERSITY OF TECHNOLOGY

Single Image Novel View Synthesis using Diffusion Models

- Single image NVS requires generative ability
- Utilize trained diffusion model
  - **Finetune** on changing camera viewpoint [1].
- No explicit 3D representation ⇒
  No guarantee of geometric consistency

Our Goal: Improve geometric consistency of generated views

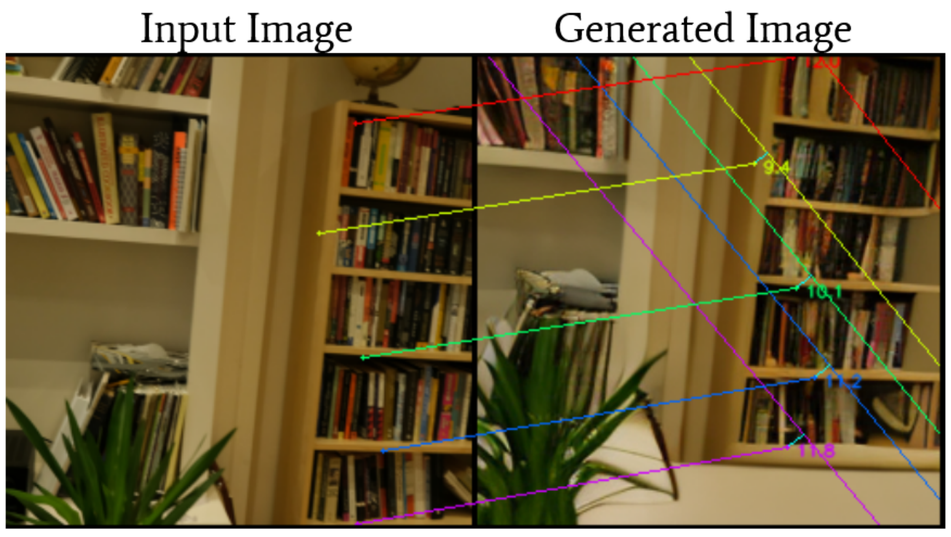


Scene-level novel view synthesis (NVS)

- Trained on large-scale real world images [2]
- Condition on warped images
- Generated sequences have significant geometric inconsistencies

Geometric Consistency Metric

Epipolar Geometry: Matching points should lie on corresponding epipolar lines



1. Perform feature matching
2. Compute epipolar lines using known cameras P_{input} and P_{target}
3. Compute distance between matching points and corresponding epipolar line

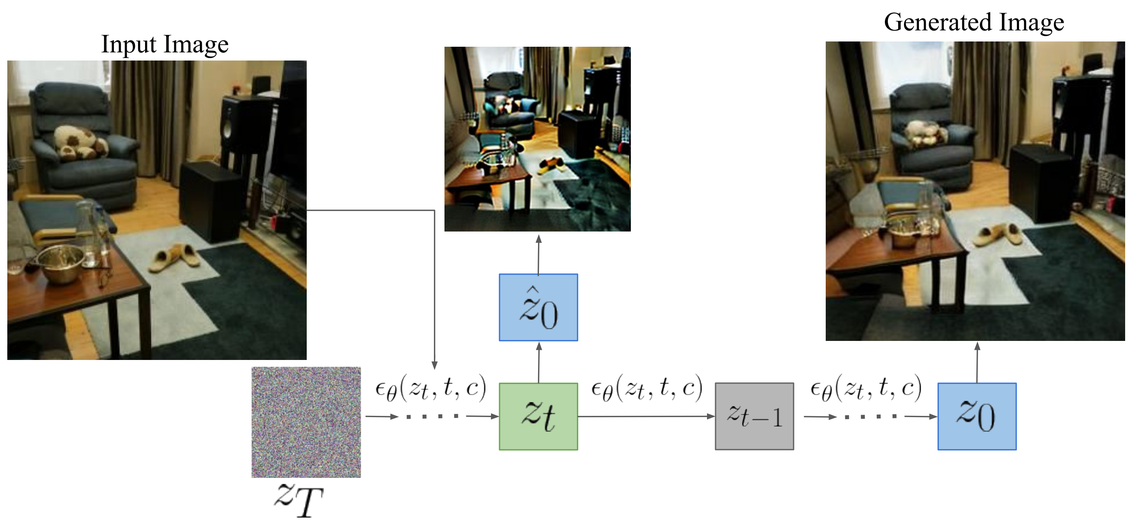
Evaluating Consistency

Pose accuracy of generated images

1. Generate a sequence of images
2. Estimate camera poses
3. Compute rotation and translation distances compared to target pose

Guiding Diffusion Process

Improve consistency **without additional training**



Optimize initial noise z_T so that generated image is more geometrically consistent.
Universal Guidance [3] Guide each diffusion step based on making $\hat{z}_0$ more consistent

$$\hat{z}_0 = \frac{z_t - (\sqrt{1-\alpha_t})\epsilon_\theta(z_t, t, c)}{\sqrt{\alpha_t}}, \tag{1}$$

$$\hat{\epsilon}_\theta(z_t, t) = \epsilon_\theta(z_t, t, c) + s(t) \cdot \nabla z_t \ell_{\text{epipolar}}(c, \hat{z}_0) \tag{2}$$

Main Contributions

- Explicitly optimize for **geometric consistency**
- Can be combined with existing methods **without additional training**

References

[1] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *ICCV 2023*.

[2] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale, 2024.

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models, 2023.

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM