EscherNet: A Generative Model for Scalable View Synthesis

Xin Kong[1], Shikun Liu[1], Xiaoyang Lyu[2], Marwan Taher[1], Xiaojuan Qi[2], Andrew J. Davison[1]

[1] IMPERIAL, dyson Robotics Lab
[2] 香港大學 THE UNIVERSITY OF HONG KONG

CVPR JUNE 17-21, 2024 SEATTLE, WA

kxhit.github.io/EscherNet

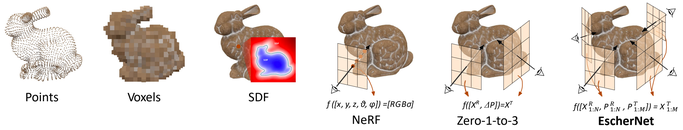
Motivation

Goal: Learn generative 3D priors from scalable multi-view training

Input: Any number of 2D views with poses

Output: Any number of 3D consistent novel views

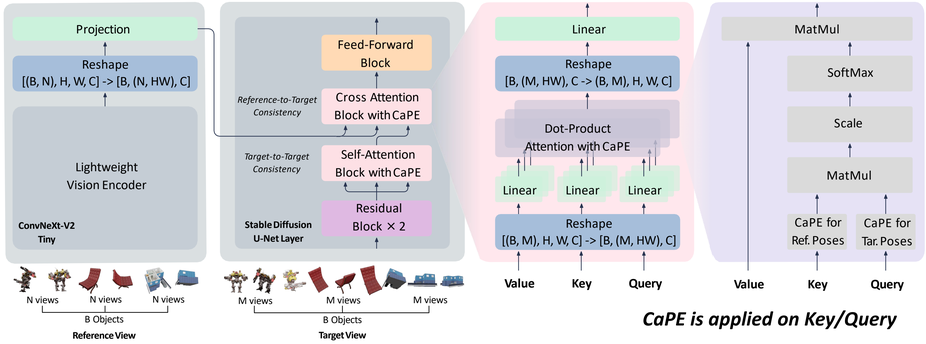
- ✓ ***Generative****: conditional generative model (more given, more constrained)*
- ✓ ***Consistency****: reference-to-target and target-to-target view consistencies*
- ✓ ***Scalability****: not scene-specific (NeRF); large-scale 2D posed images data*
- ✓ ***Generalisation****: trained with fixed number views (3-to-3), infer any-to-any*





Multiview Diffusion Model

- Inheriting 2D diffusion model (SD 1.5) priors



CaPE is applied on Key/Query

Camera Positional Encoding (CaPE)

- 3D Novel View Synthesis as Seq Image-to-Image Generation

Language $\langle \mathrm{RoPE}(v_1, P_1), \mathrm{RoPE}(v_2, P_2) \rangle = \langle \mathrm{RoPE}(v_1, P_1 - P_2), \mathrm{RoPE}(v_2, 0) \rangle$

Token: *The quick brown fox jumps over the lazy dog.*

Position: 0 1 2 3 4 5 6 7 8

3D Vision $\langle \mathrm{CaPE}(v_1, P_1), \mathrm{CaPE}(v_2, P_2) \rangle = \langle \mathrm{CaPE}(v_1, P_2^{-1} P_1), \mathrm{CaPE}(v_2, I) \rangle$

Token:

Position:

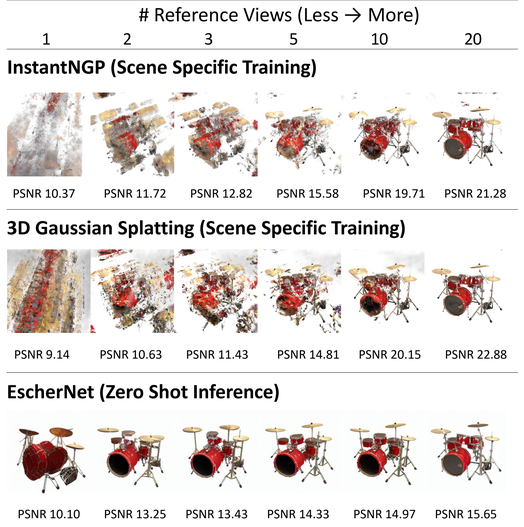
		Type		Length	Origin
Language		Discrete	Linear	Unbounded	0 (sentence beginning)
3DV	Rotation	Continuous	Cyclic	Bounded	N/A (coordinate system doesn't matter)
	Translation		Linear	Unbounded	

- Each view's **absolute position** rotates its token features
- Only **relative position** is obtained in dot-product self attention

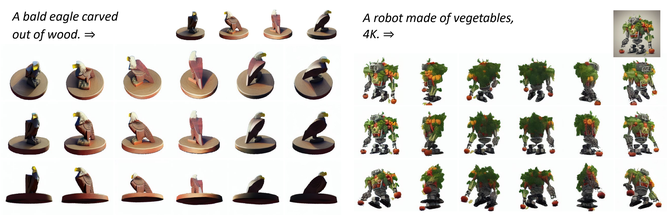


Results

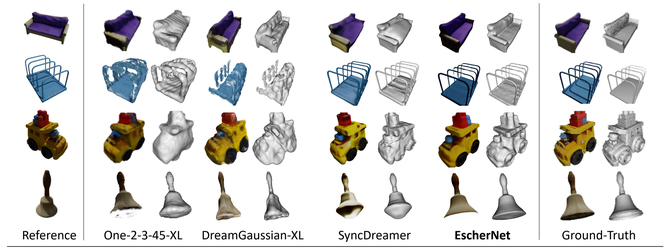
Novel View Synthesis (NeRF)



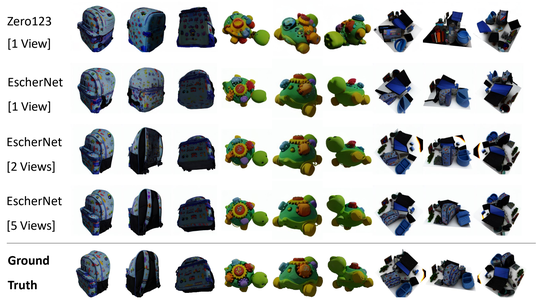
Novel View Synthesis (Text-to-3D)



3D Reconstruction



Novel View Synthesis (GSO/RTMV)



Novel View Synthesis (Real-world 6DoF)

