Zero Shot Novel View Synthesis in the Wild

Tim Tomov[I], Michael Neumayr[I], David Rozenberszki[I]
[I] Technical University of Munich

Code is available

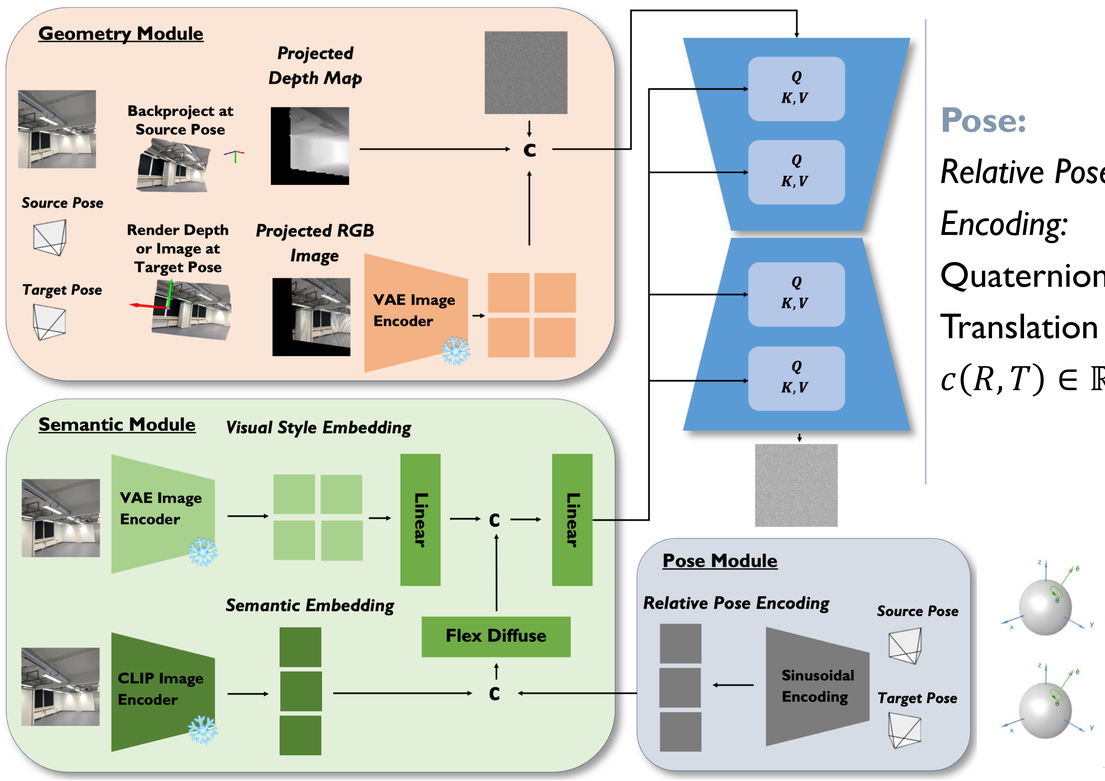
Generalizable Novel View Synthesis

Input: Single posed RGB-D image

Output: RGB image from novel pose

Validate on novel scenes (ScanNet++)

Modified Stable Diffusion



Pose:
Relative Pose Encoding:
Quaternion/ Translation
$c(R,T) \in \mathbb{R}^7$

Geometric conditioning: Provides **local structure**

- *Projected Depth Map:* Enables geometrical understanding of the scene
- *Projected RGB Image:* Enables capturing high frequency details

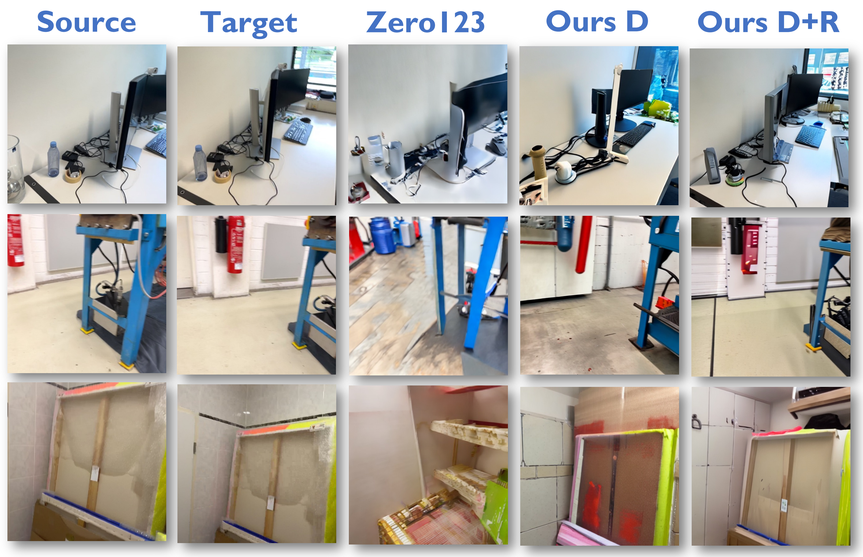
Semantic conditioning: Provides **global features**

- *Semantic Embedding*: Provides semantic context of the scene
- *Visual Style Embedding:* Provides more enriched visual information

Experiments

Conditioning Signal:
Ours D *Only Depth*
Ours D+R *Depth + RGB*

Qualitative Results



Correct geometry and semantics of the scenes are understood

Quantitative Results

Methods	LPIPS ↓	SSIM ↑
Baseline (Zero123)	0.2539	0.1735
Ours (Depth)	0.2012	0.2759
Ours (Depth + RGB)	**0.1624**	**0.4297**

Varying Pose Change (Ours D)

Increasing relative pose change to the right

the left

Ablations

Depth	RGB	Visual Style	LPIPS ↓	SSIM ↑
		✓	0.2632	0.0811
✓		✓	0.2288	0.2036
✓			0.2245	0.2064
	✓	✓	0.1736	0.3984
✓	✓	✓	**0.1624**	**0.4297**

Geometry

Semantic (Ours D)