Google DeepMind

RUST: Latent Neural Scene Representations from Unposed Imagery

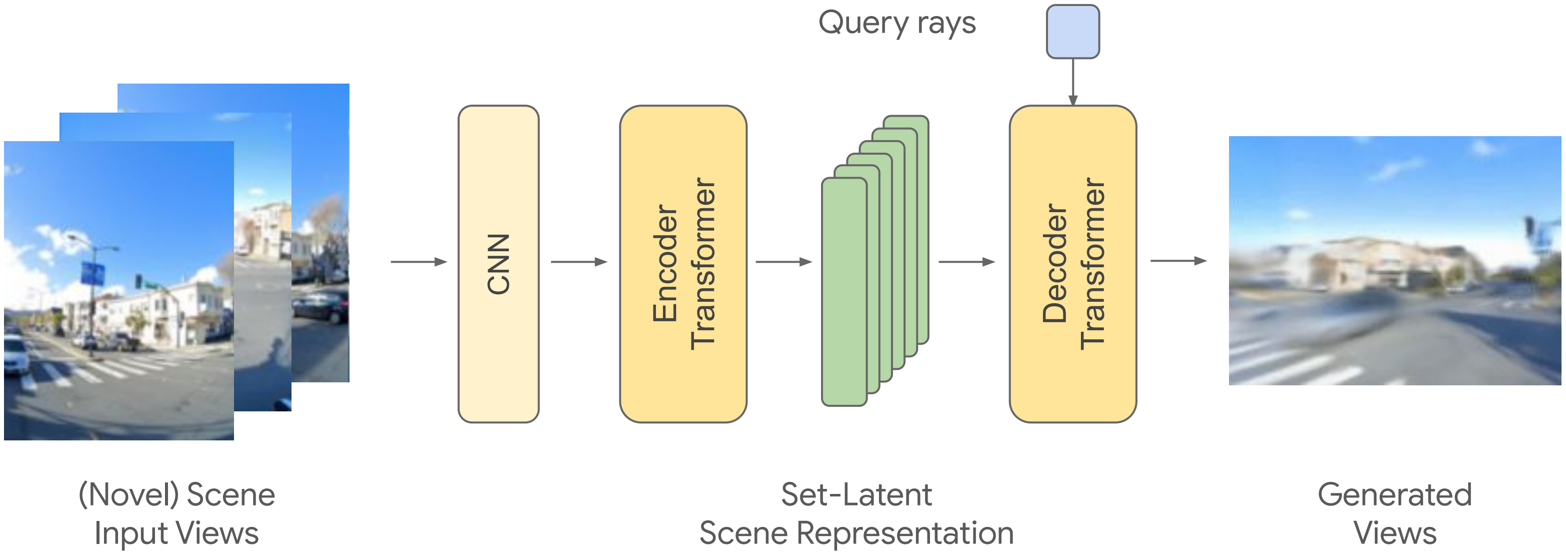
Mehdi S.M. Sajjadi, Aravindh Mahendran, Thomas Kipf, Etienne Pot, Daniel Duckworth, Mario Lučić, Klaus Greff

TAG: THU-AM-078

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Why?

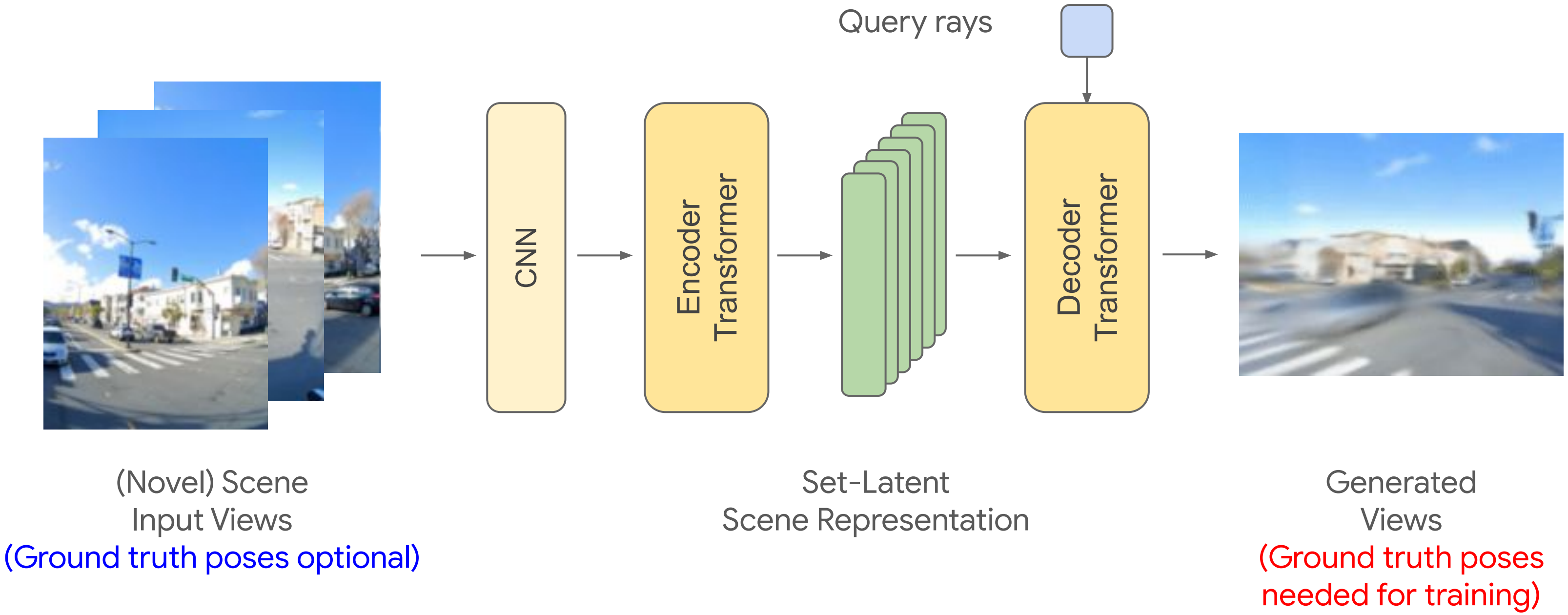
Scene Representation Transformer (**SRT**): Novel view synthesis based 3D latent scene representations are powerful.



Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations, Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, Andrea Tagliasacchi, CVPR 2022.

Why?

… but they need a lot of posed data to train effectively

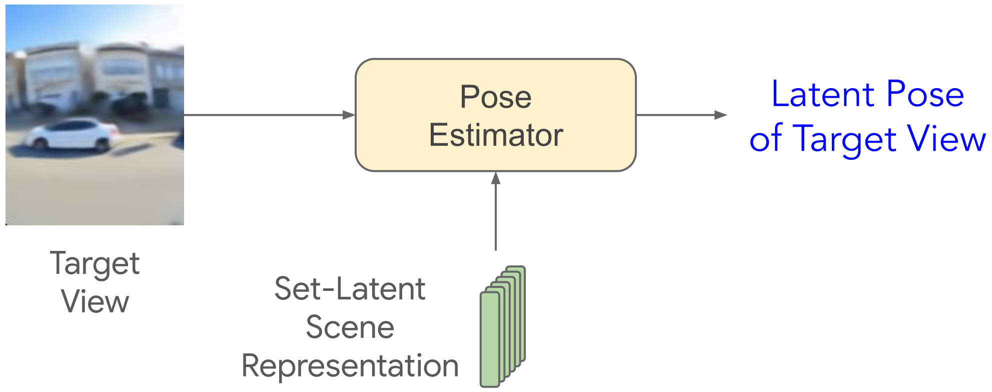


How?

We present **RUST**: **R**eally **U**nposed **S**cene representation **T**ransformer

1. Peaks at the target view.
2. Infers an implicit **8D** latent pose.

Key component:



RUST: Results

We traverse the latent pose space to generate target views and stitch them into a video.

MSN-Hard dataset

Height

Rotation

Distance

End of preview … stay on for more details or check-out our website:

RUST: Results

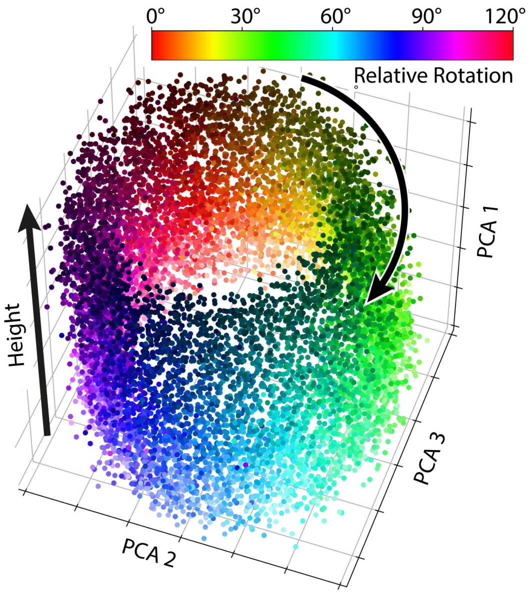
We traverse the latent pose space to generate target views and stitch them into a video.

Forward Pitch Roll

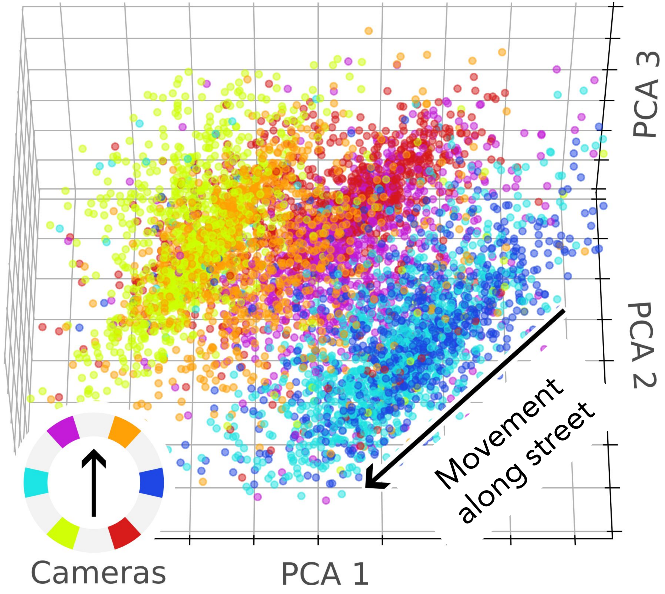
Street View Dataset

RUST: Results

PCA projections of latent pose space reveals ways to control the camera.



MSN-Hard
dataset



Street View
Dataset

RUST: Results

Method	Pose	PSNR
SRT [22]	p_x, p_y	23.31
SRT†	p_x, p_y	24.40
SRT†	$\hat{p}_x, p_y$	23.81
UpSRT†	$\cancel{p_x}, p_y$	23.03
SRT†	$\hat{p}_x, \hat{p}_y$	18.65
UpSRT†	$\cancel{p_x}, \hat{p}_y$	18.64
RUST	$\cancel{p_x}, \cancel{p_y}$	23.49

Ablation	PSNR
Right-half PE	23.88
Stop grad.	23.16
No SLSR	20.83
No self-attn	22.97
3-dim. $\tilde{p}$	20.40
64-dim. $\tilde{p}$	23.40
768-dim. $\tilde{p}$	23.11

Table 1. **Quantitative results on MSN – Left:** Comparison with prior work in various settings: perfect (p_x, p_y), noisy ($\hat{p}_x, \hat{p}_y$) and lack of ($\cancel{p_x}, \cancel{p_y}$) input and target poses. We report SRT both as proposed [22], and with our improved architecture (SRT†, UpSRT†). Despite requiring no poses, RUST matches the performance of SRT and UpSRT† while strongly outperforming all methods when target pose is noisy $\hat{p}_y$. **Right:** Model ablations, see Sec. 4.1.1.

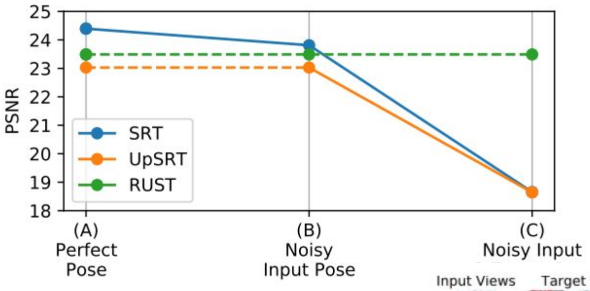


Figure 3. **Robustness to camera noise** – Sajjad
SRT and UpSRT on (A) perfect pose, and (B) n
the more realistic setting (C) where *input & t*
both methods fail as they rely on accurate targ
training. RUST needs no pose, so its performa

Method	# Views	MSE	R^2 (%)	Success (%)
RUST EPE	7	0.08	99.9	[100]
COLMAP	10	0.00	100.0	4.2
COLMAP	80	0.07	99.7	29.5
COLMAP	160	0.38	99.1	58.9
GNeRF	12	29.39	46.7	[100]
GNeRF	150	9.24	83.1	[100]
GNeRF-FG	150	4.05	92.7	[100]

Table 2. **Explicit pose estimation on MSN** – RUST EPE recovers relative camera poses nearly perfectly from the SLSR (5 input views) and the pair of latent target poses. COLMAP [23] requires a much larger number of images, and still has a significantly lower success rate for registration. Similarly, GNeRF [15] requires many views of the scene, and fails to estimate accurate poses even when the background pixels are removed from the data (GNeRF-FG).

Figure 7. **Qualitative results on SV** – Comparison of RUST with prior work using accurate camera pose. RUST outperforms our improved UpSRT variant, while producing similar quality as the fully posed improved SRT model. We further train a dense semantic segmentation decoder on top of the frozen RUST scene representation, showing that it retains semantic information about the scene.

Thank you!

For more results and details please come to our poster or checkout our …

Paper

Website

Paper:
https://openaccess.thecvf.com/content/CVPR2023/papers/Sajjadi_RUST_Latent_Neural_Scene_Representations_From_Unposed_Imagery_CVPR_2023_paper.pd

Website: https://rust-paper.github.io/