Google Brain
Skoltech

RePAST:
Relative Pose Attention Scene Representation Transformer

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Aleksandr Safin
aleksandr.safin@skoltech.ru

Daniel Duckworth
duckworthd@google.com

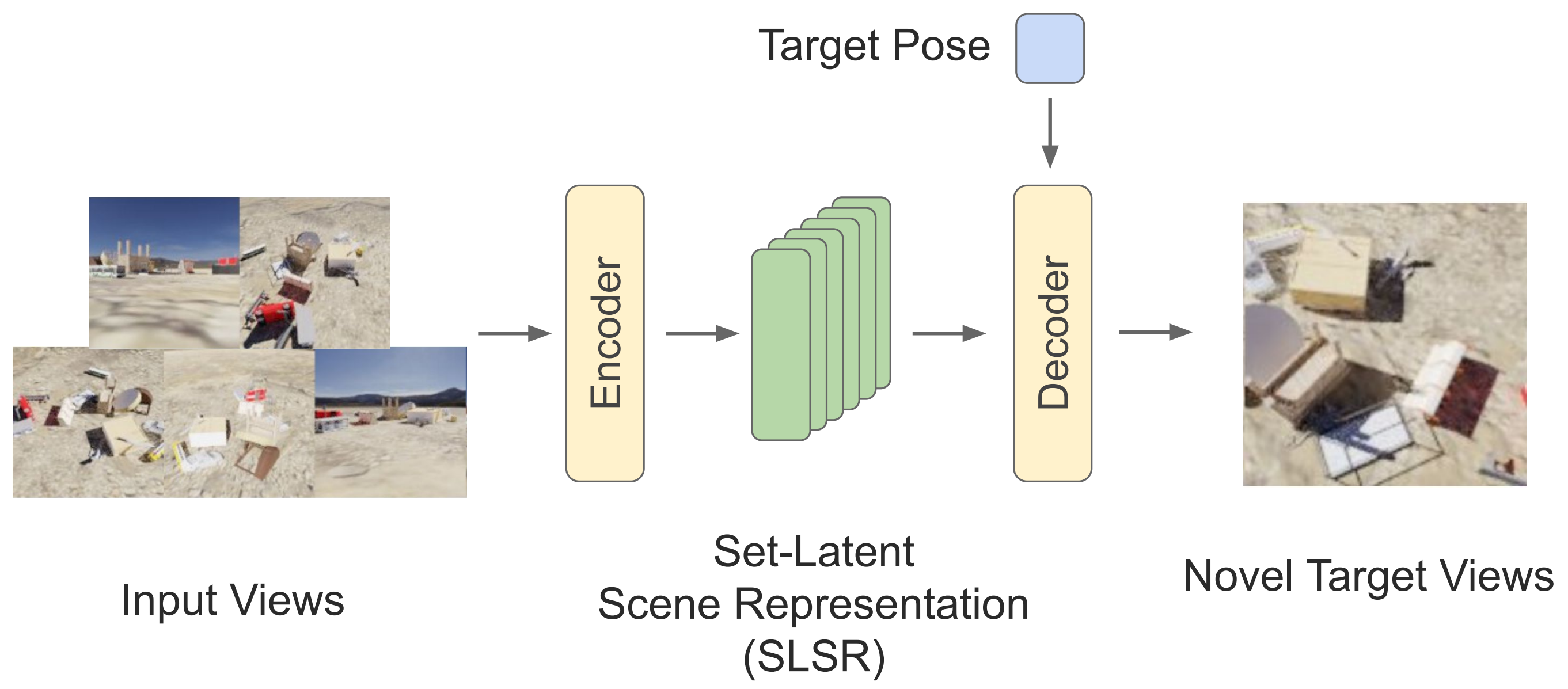
Mehdi S. M. Sajjadi
repast@msajjadi.com

Introduction

Scene Representation Transformer (SRT)

Transformer-based method for NVS.

- Input images are encoded into Set-Latent Scene Representation (SLSR) through a self-attention transformer
- Novel views are rendered through decoder transformer that cross-attends from target pose into SLSR



Limitations

SRT uses first input view as the ***reference view*** as a global coordinate frame, i.e. all camera poses are transformed w.r.t. ***reference view***

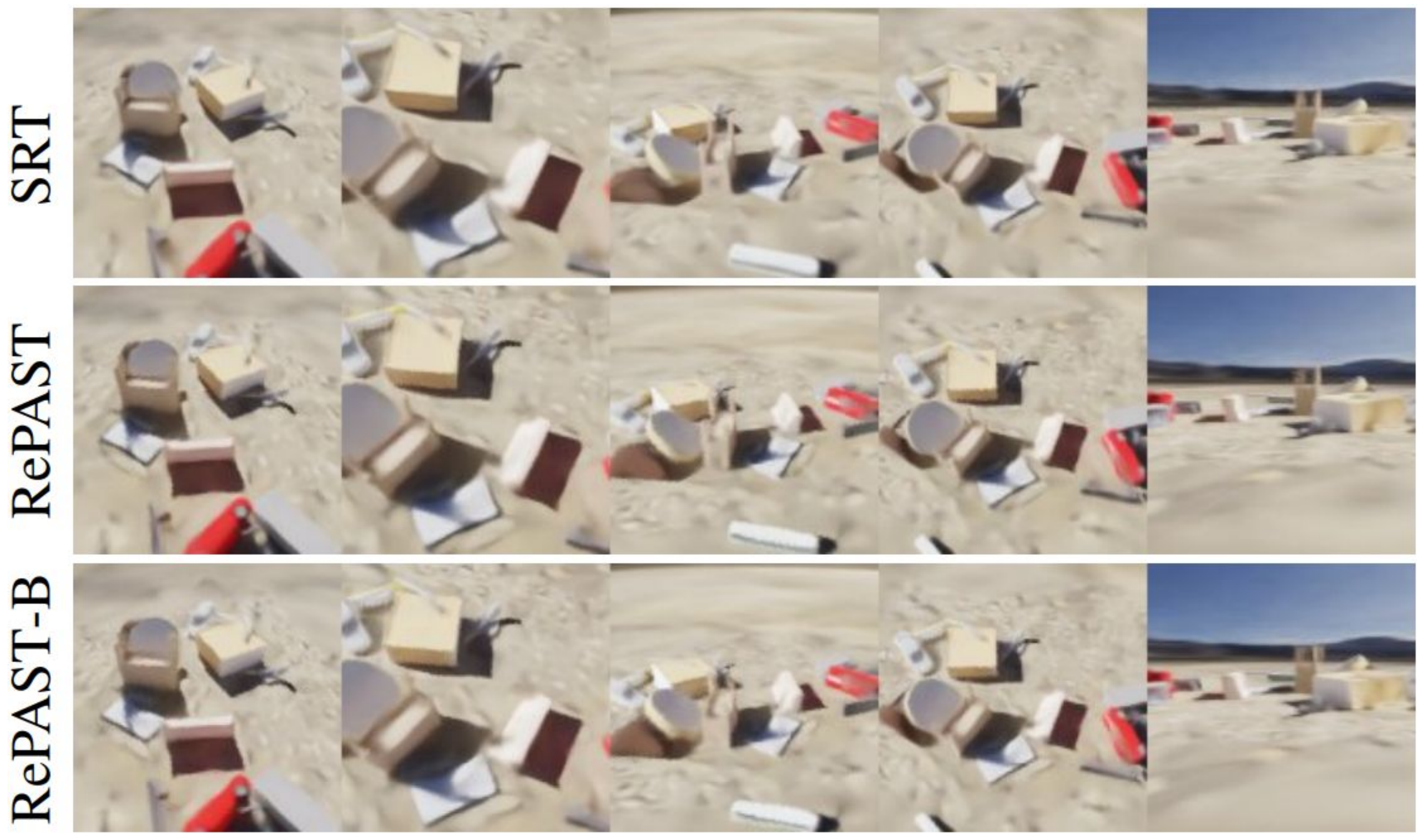
=> SRT's SLSR is not symmetric to the (arbitrary) choice of the reference view

Benefits of Invariants

- No flickering when just order of input frames is changed
- More effective design
- Opens use of SRT on large-scale scenes

Results

- RePAST is by design invariant to the global camera poses transformation
- PSNR for RePAST is slightly higher than for SRT despite the introduction of the new invariance



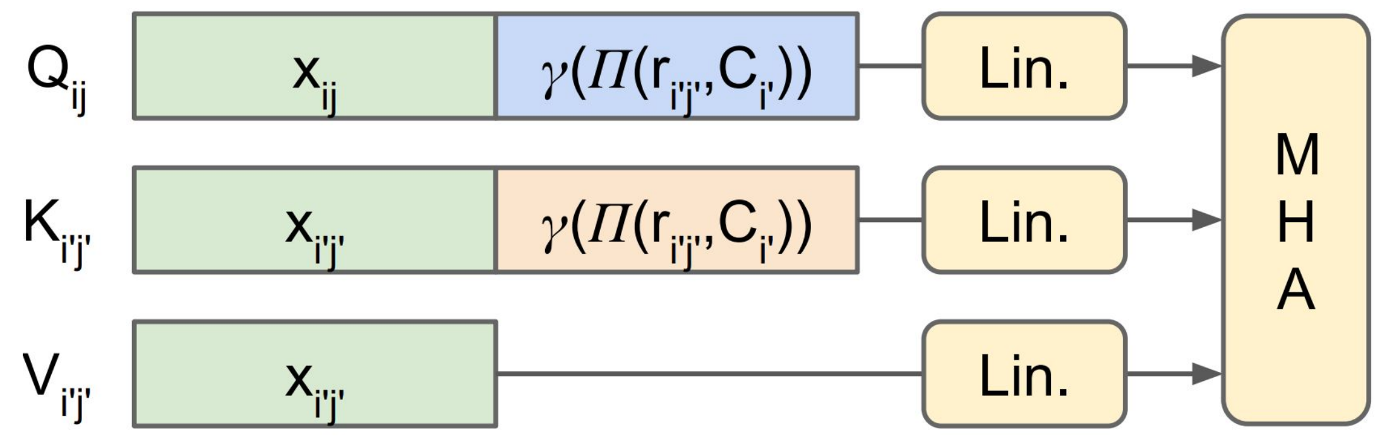
	↑PSNR	↑SSIM	↓LPIPS
SRT [12]	24.61	0.784	0.223
RePAST	24.89	0.794	0.202
RePAST-B	24.71	0.788	0.211

Quantitative results – RePAST modestly improves over SRT across all metrics. Removing the relative camera injection in the Decoder (RePAST-B) leads to slightly lower quality.

Method

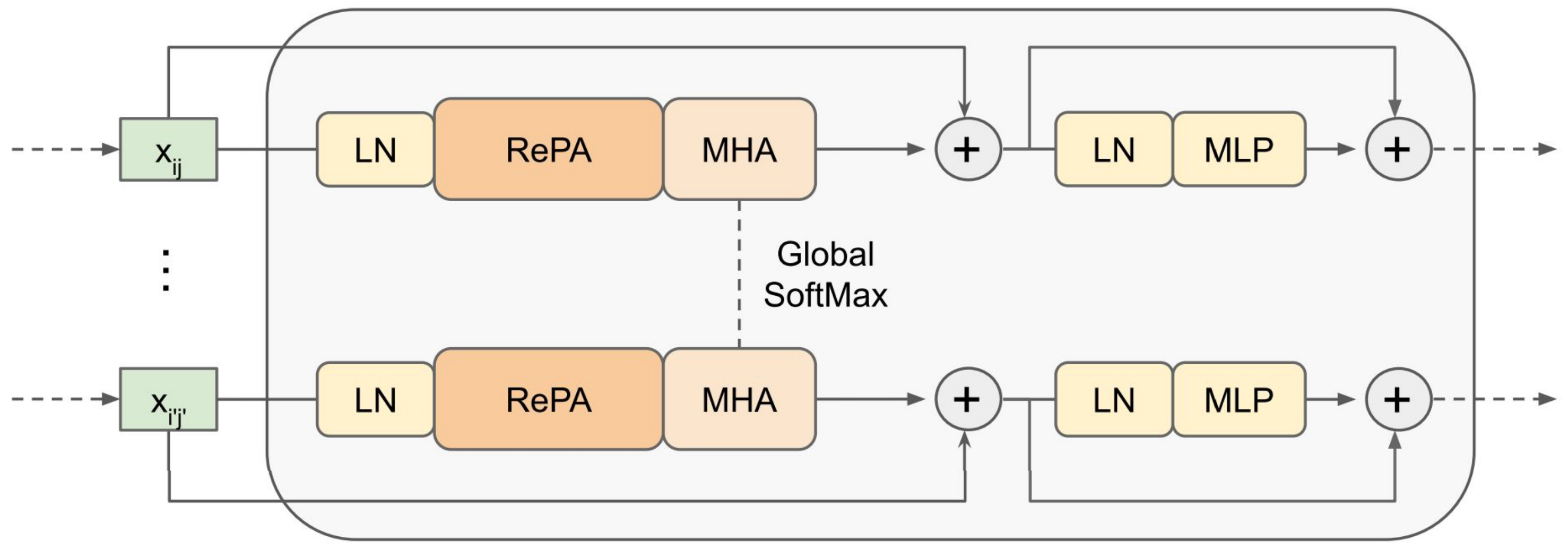
We propose the novel Relative Pose Attention (**RePA**) block.

- RePA enables us to capture dependency in camera space of the key
- RePA is used in both the Encoder and the Decoder
- RePAST is SRT with RePA attention

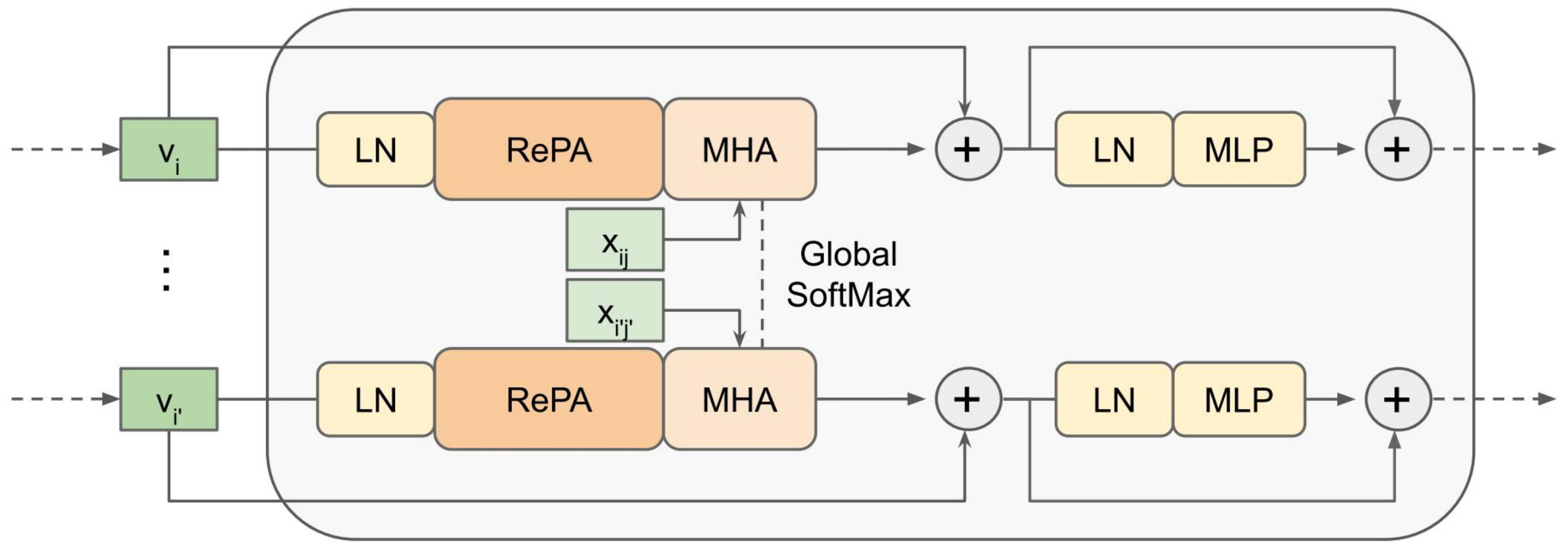


Both query and keys are augmented with their pose relative to the camera belonging to the key token.

RePAST Encoder Layer – Each SLSR token selfattends into all other tokens using RePA. As in SRT's vanilla self-attention, softmax layers are computed globally over all tokens.



RePAST Decoder Layer – The decoder is similar to the encoder layers. Instead of self-attention, the target view queries cross-attend into the SLSR using RePA. The N decoding streams interact through the global softmax, and the results are averaged for a final MLP (not shown here) to produce the target RGB color.



Conclusion

- We propose RePA to make SRT invariant to the global transformations of camera space.
- RePAST is natural extension to SRT, while it is invariant to the order of the input cameras or the arbitrary choice of a particular reference frame.
- RePAST is a step towards applying SRT to large-scale scenes.