Depth Field Networks for Generalizable Multi-view Scene Representation

3DMV @ CVPR23

Vitor Guizilini   Igor Vasiljevic*   Jiading Fang*   Rareş Ambruş   Greg Shakhnarovich   Matthew Walter   Adrien Gaidon*

TOYOTA RESEARCH INSTITUTE

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

Motivation

Traditional video depth estimation needs explicit geometry:

- ✔ cost volumes, epipolar constraints, bundle adjustment
- ✗ generally not real-time and compute intensive
- ✗ overfits to train set, generalizes poorly

Recent Transformer architectures [2] learn implicity:

- ✔ Attention-based implicit geometry for stereo depth estimation
- ✗ Doesn't match cost-volume-based method accuracy
- ✗ Requires large amounts of diverse data

Solution - Depth Field Networks (DeFiNe):

- ✔ Geometry is learned **implicitly** conditioned on pose video input
- ✔ Geometry-preserving 3D aug. increase **viewpoint diversity**
- ✔ Depth maps can be generated from **arbitrary viewpoints**
- ✔ Achieves a new **state-of-the-art** on Scannet stereo benchmark
- ✔ State-of-the-art by a large margin on 7scenes **zero-shot**

Augmentations

Inductive Biases for Video Depth Estimation

Image Embeddings: Pretrained CNN image features per frame

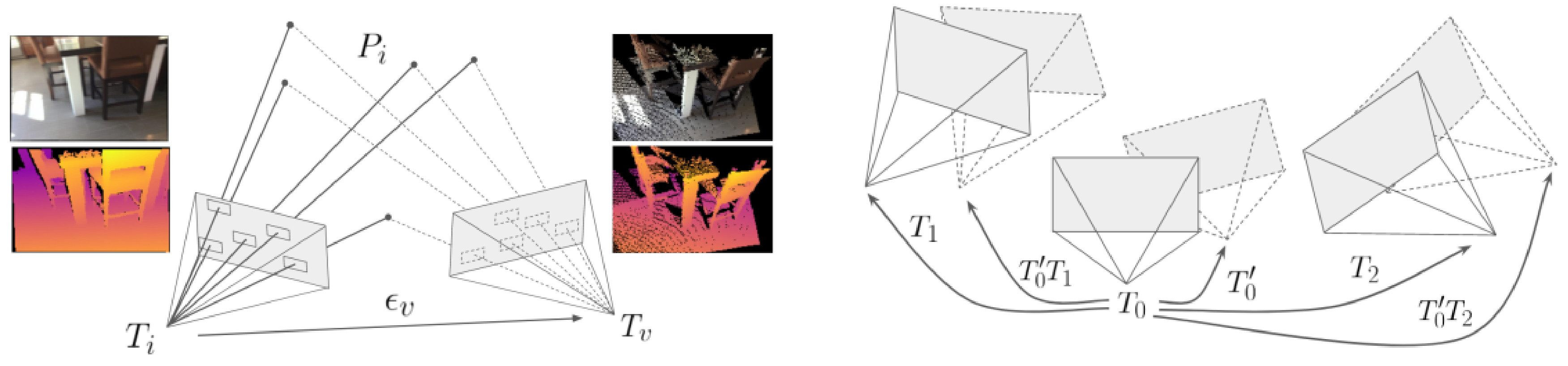
Camera Embeddings: Pose embeddings providing inductive bias for multi-camera relationships between frames

Geometric-Preserving 3D Augmentations

Virtual Camera Projection: Generate virtual RGB-D views for increasing viewpoint diversity at train time, improving generalization

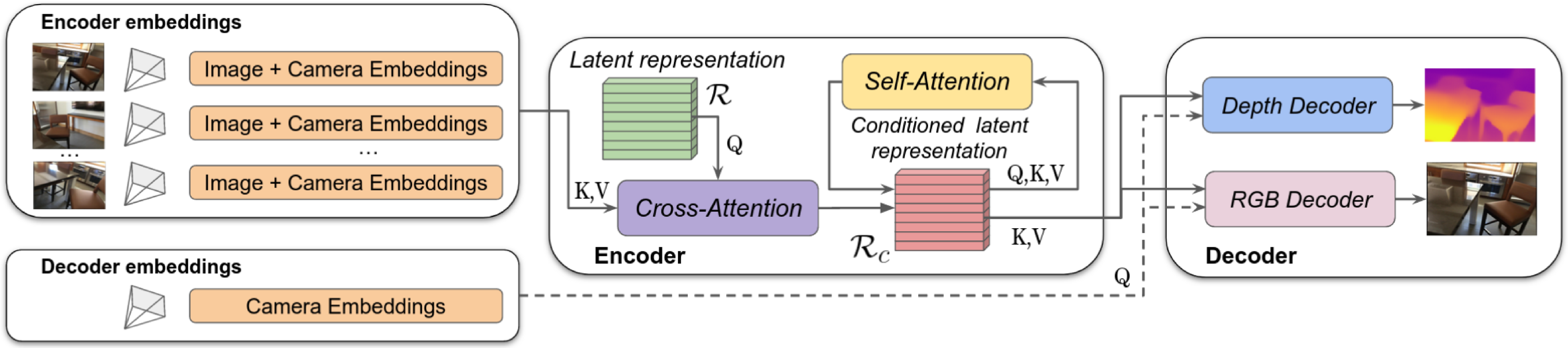
Canonical Jittering: promote translation and rotation equivariance

Canonical Randomization: increase scene diversity



(a) Virtual Camera Projection.   (b) Canonical Jittering.

Pipeline



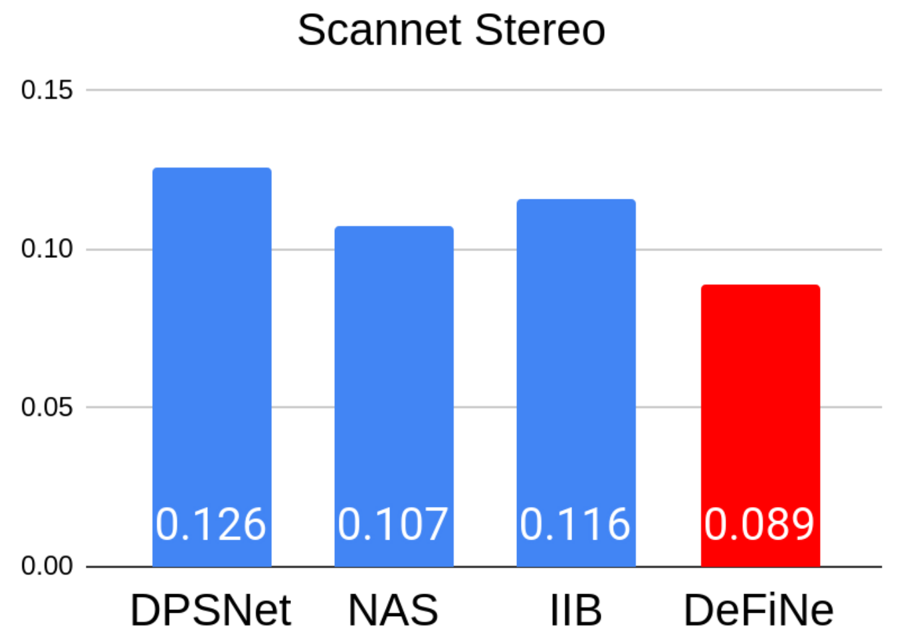
PercieverIO [1] backbone: Arbitrary inputs projected into low-dim. latent, task-specific decoders for arbitrary outputs

DeFiNe Encoder: video frame CNN features and Fourier-encoded pose embeddings (single cross-attention layer projects to latent)

DeFiNe Decoder: camera ray queries decoded to depth and RGB predictions

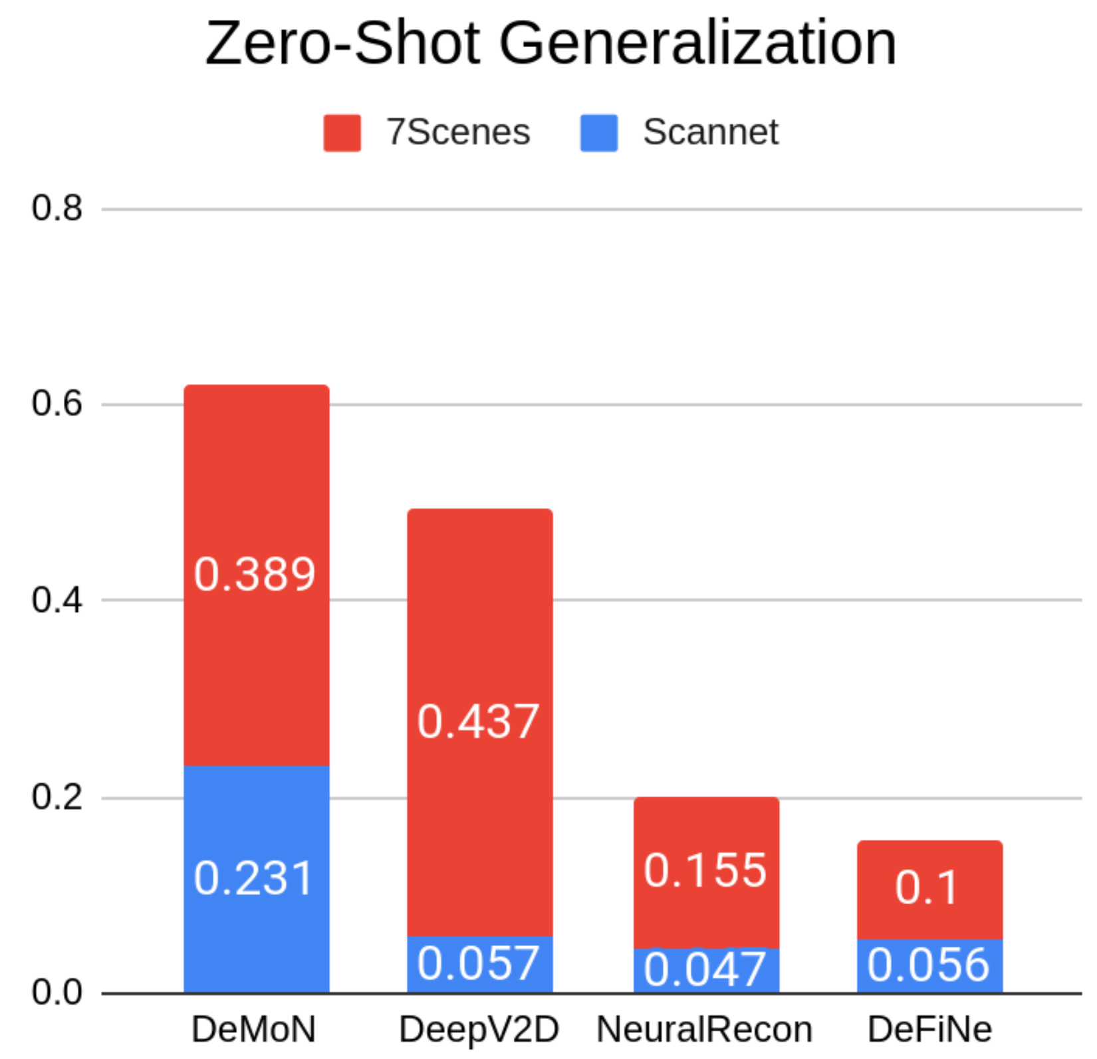
Experimental Results

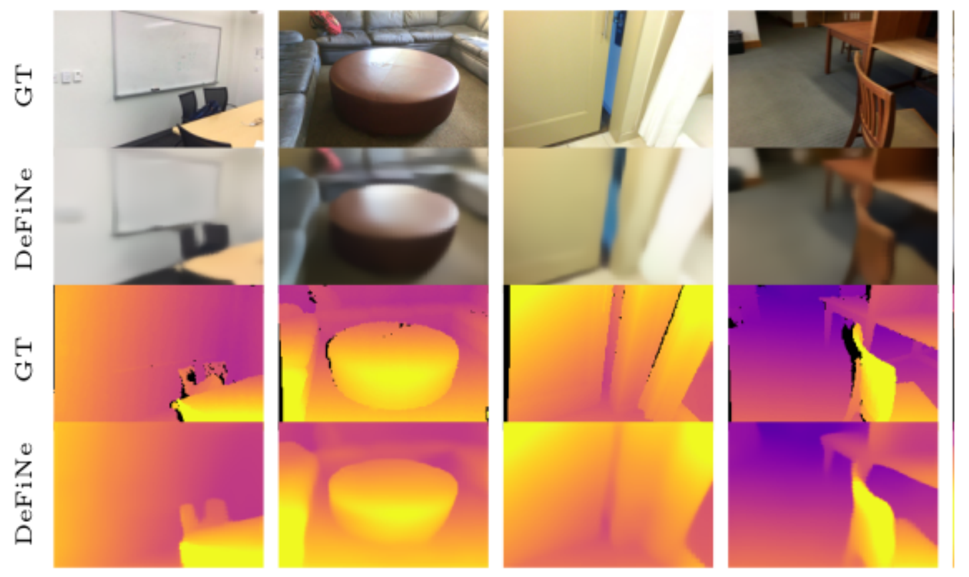
ScanNet-Stereo



✔ beats traditional methods based on expensive cost volumes
✔ added view diversity leads to large (**23%**) improvement over baseline

ScanNet-Video

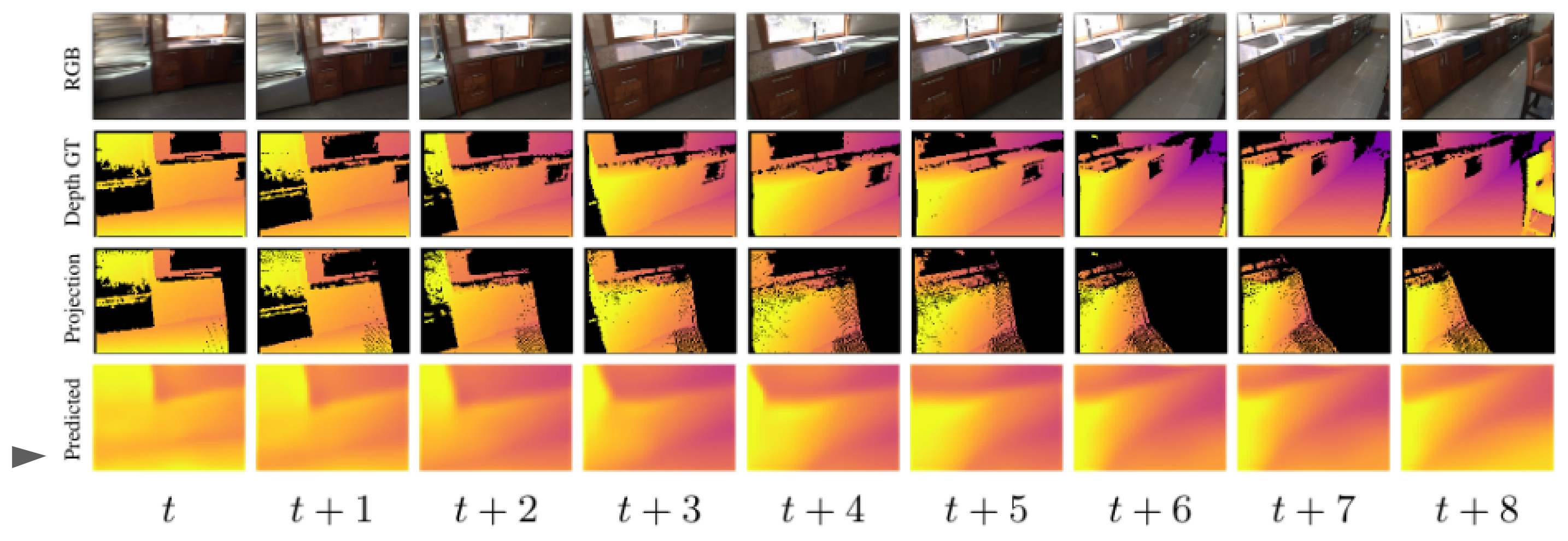




State-of-the-art in **zero-shot transfer** to *7scenes* by a large margin (**35%**)

Depth Extrapolation

DeFiNe allows for depth *synthesis* from unseen viewpoints



References

[1] *Andrew Jaegle et al.* **Perceiver IO: A General Architecture for Structured Inputs and Outputs. ICLR'22**
[2] *Wang Yifan, Carl Doersch, Relja Arandjelovic, Joao Carreira, Andrew Zisserman.* **Input-level Inductive Biases for 3D Reconstruction. CVPR'22**

Code

https://github.com/tri-ml/vidar