Depth Estimation from Monocular Images Using Geometric Foundation Models

Rareș Ambruș, PhD

Senior Manager

Large Behavior Models

TOYOTA RESEARCH INSTITUTE

Outline

Introduction

Mono-Depth

MultiView-Depth

Conclusion

Mission

TRI envisions a future where Toyota products, enabled by TRI technology, dramatically improve quality of life for individuals and society.

TOYOTA RESEARCH INSTITUTE

https://www.youtube.com/watch?v=kKz2mtYdQAA

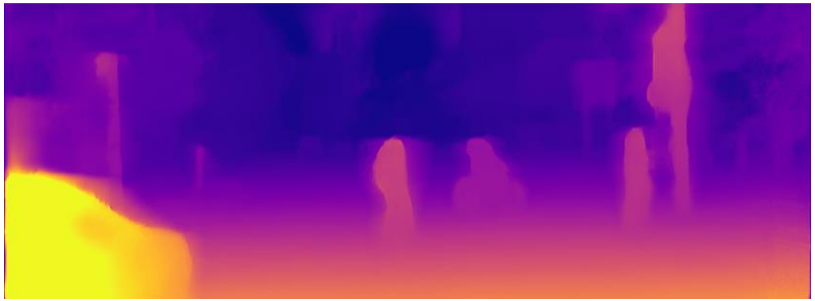
TOYOTA RESEARCH INSTITUTE

Learning Robust 3D Perception from Cameras

EMBEDDED VISION SUMMIT

Single RGB Image

Predicted Depth Image



MonoDepth Network

TOYOTA RESEARCH INSTITUTE

Outline

Introduction

Mono-Depth

MultiView-Depth

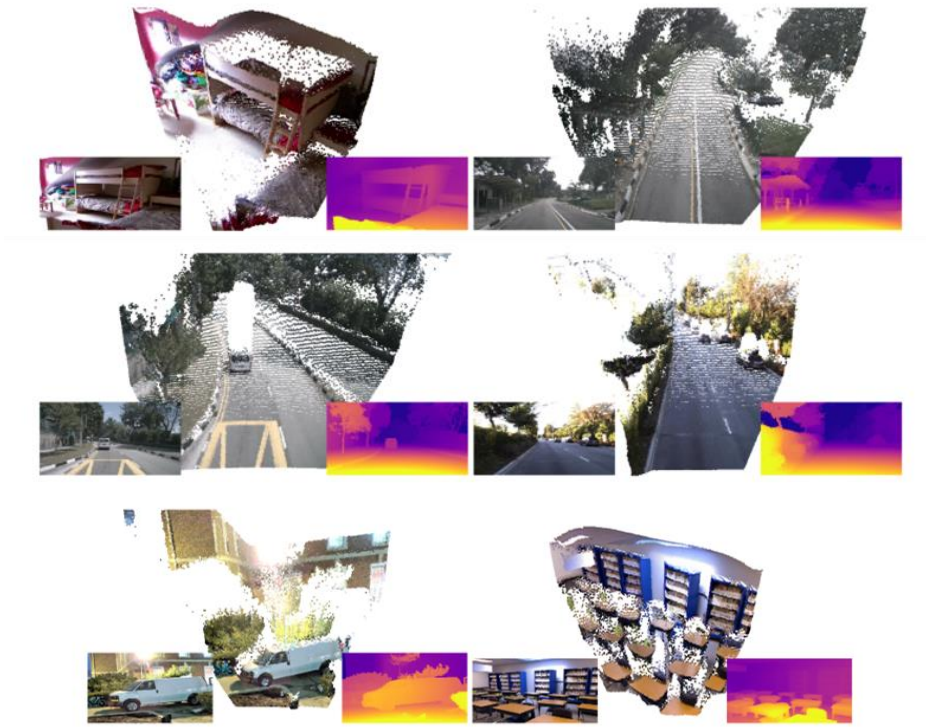
Conclusion

Monocular Depth Estimation

Zero-shot on any domain without fine-tuning (appearance gap)

Scale-aware (metric) predictions on any camera (geometric gap)

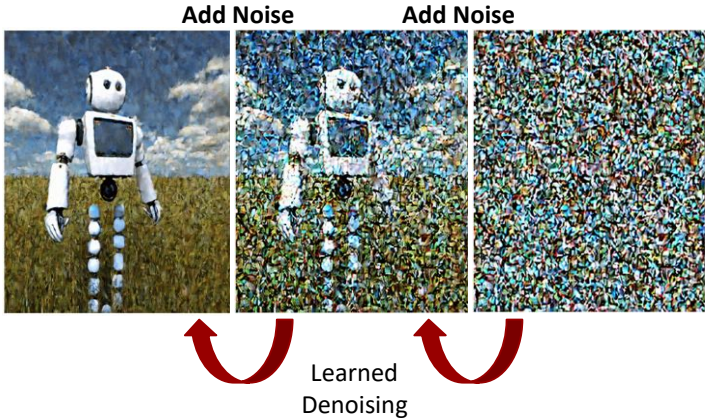
Model uncertainty



Challenges

Large-scale diverse pre-training

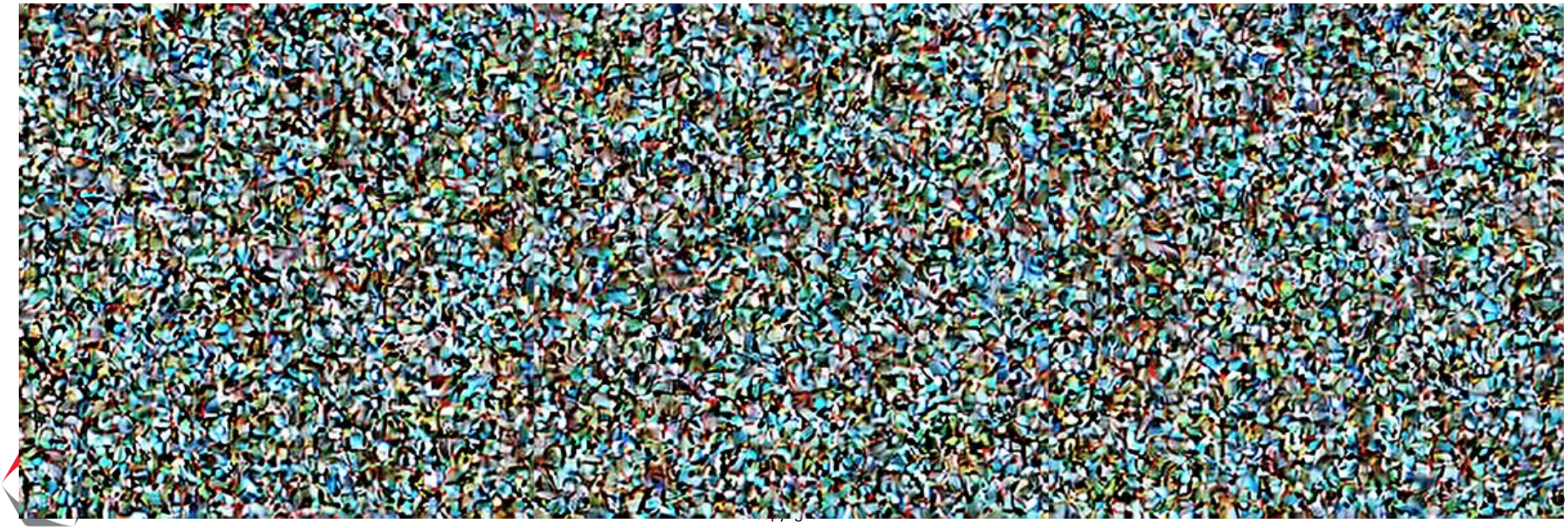
→ **Diffusion models**



Challenges

Prompt

"A photo of a white robot in tall grass staring at clouds in the blue sky"

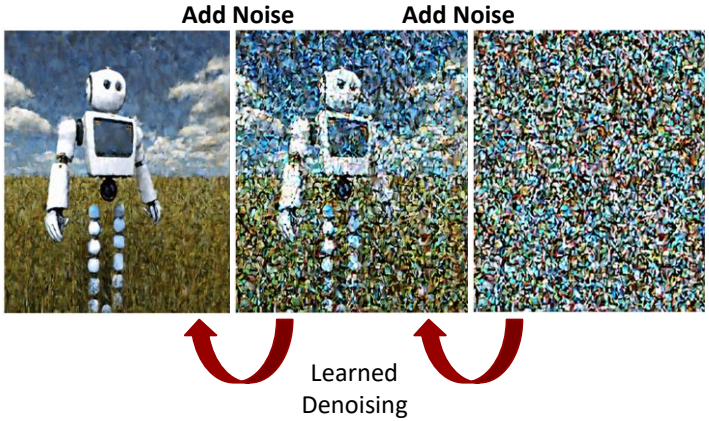


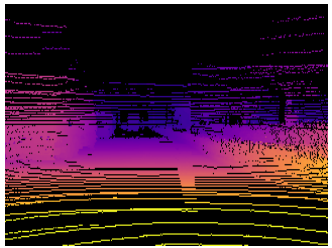
Challenges

Large-scale diverse pre-training
→ **Diffusion models**

Sparse, unstructured training labels
→ **Pixel-level generation**

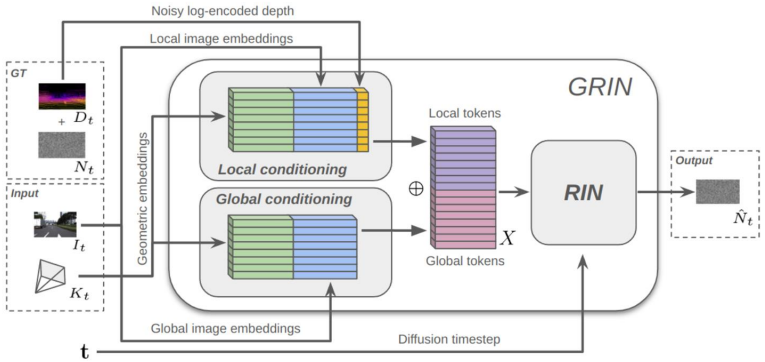
Geometric doman gap
→ **Condition on camera information**





Geometric RIN (GRIN): Efficient Pixel-Level Diffusion with Sparse Labels

Vitor Guizilini, Pavel Tokmakov, Achal Dave, Rares Ambrus, 3DV'25 (Oral)



Architecture: GRIN (Geometric Recursive Interface Networks)

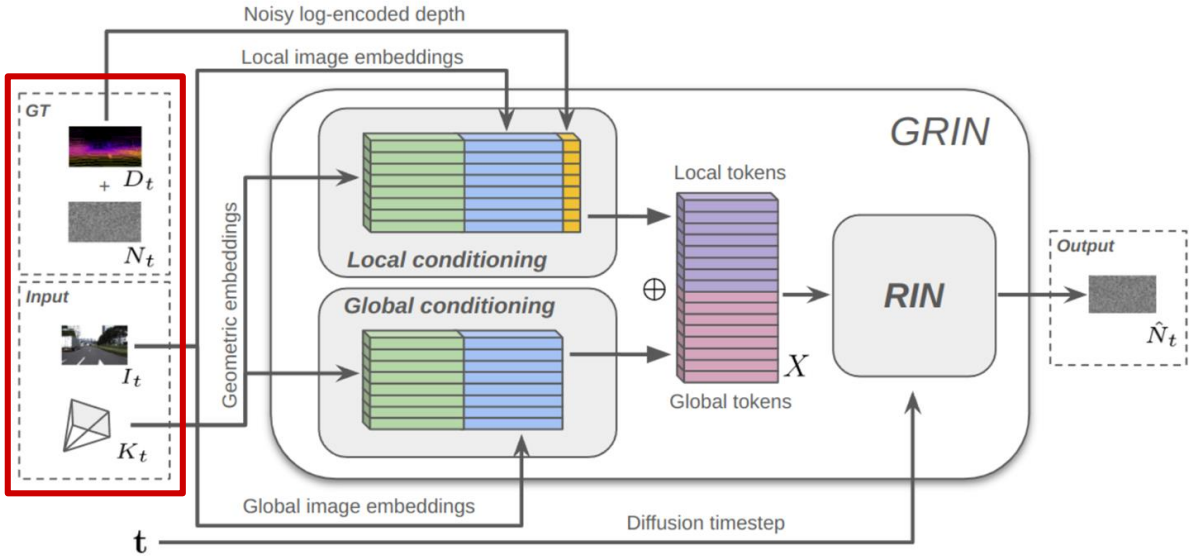
Local conditioning with visual features + 3D geometric embeddings

Global conditioning with dense features to preserve scene-level information

Input Embeddings

Input: RGB image + camera intrinsics

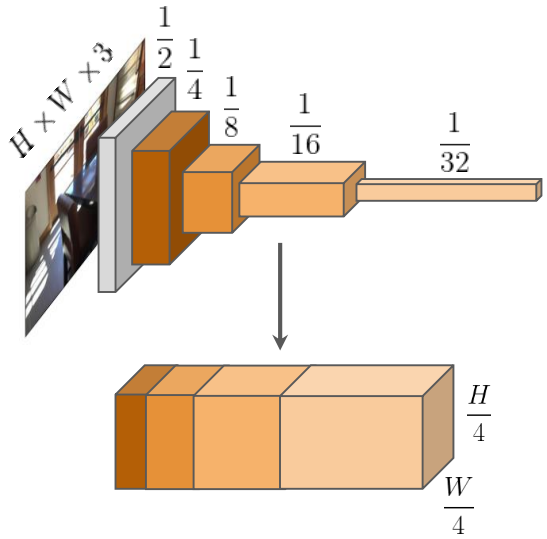
Ground-truth: Sparse depth maps



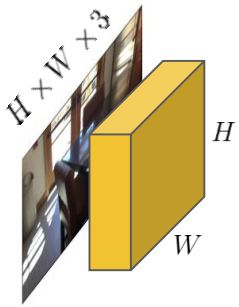
Input Embeddings

Image embeddings

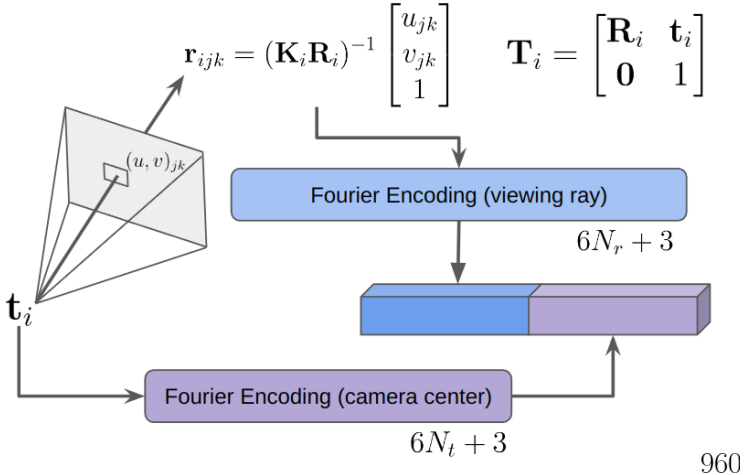
Global



Local



Geometric embeddings

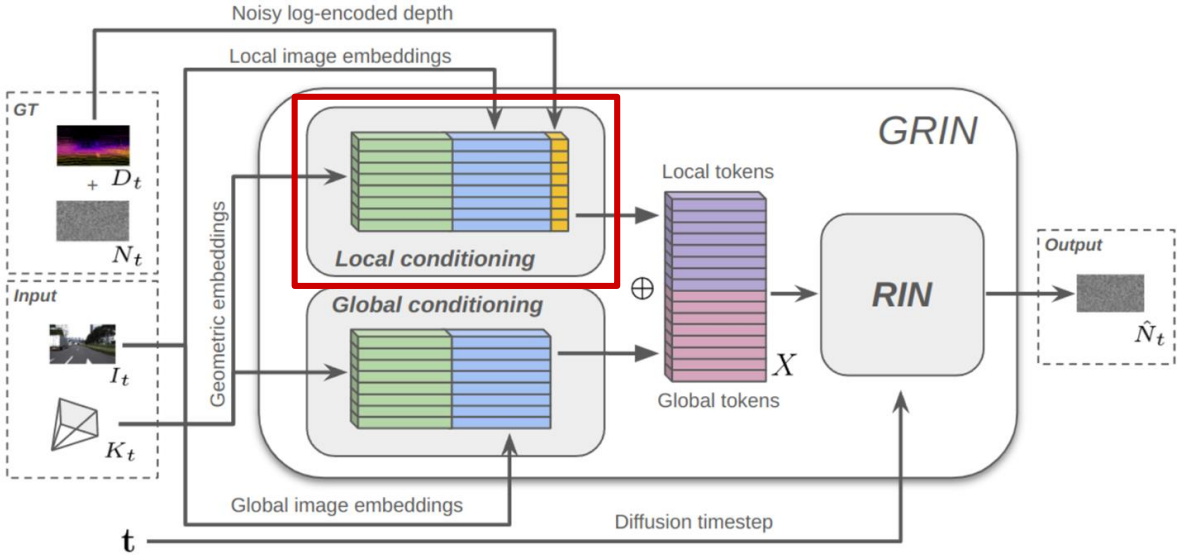


960

Local conditioning

Local conditioning: Image + geometric embeddings

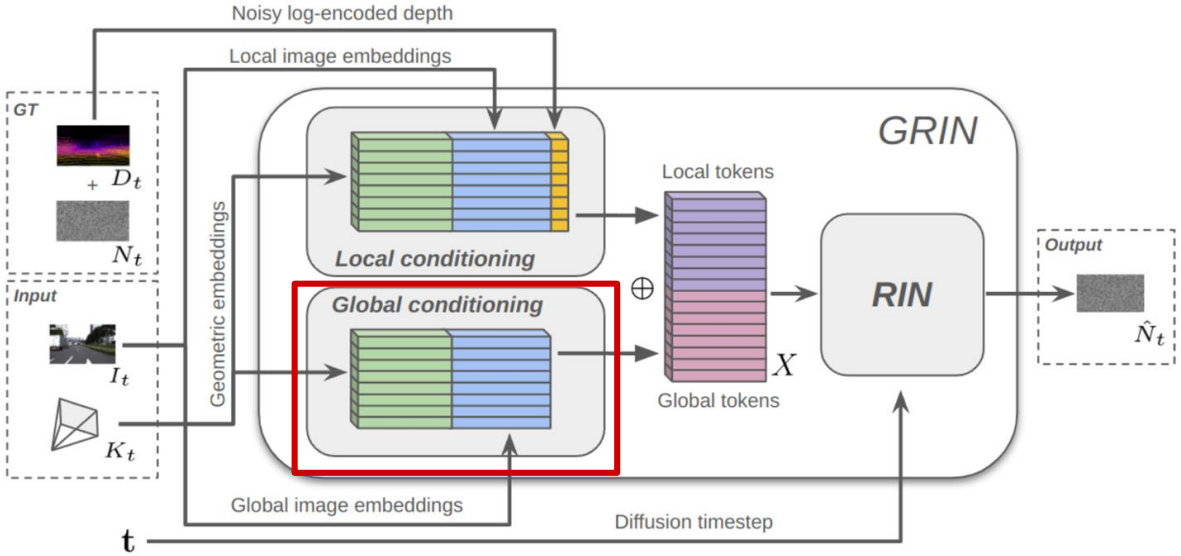
Sparse information from **valid** pixels (log-encoded depth)



Global conditioning

Global conditioning: Image + geometric embeddings

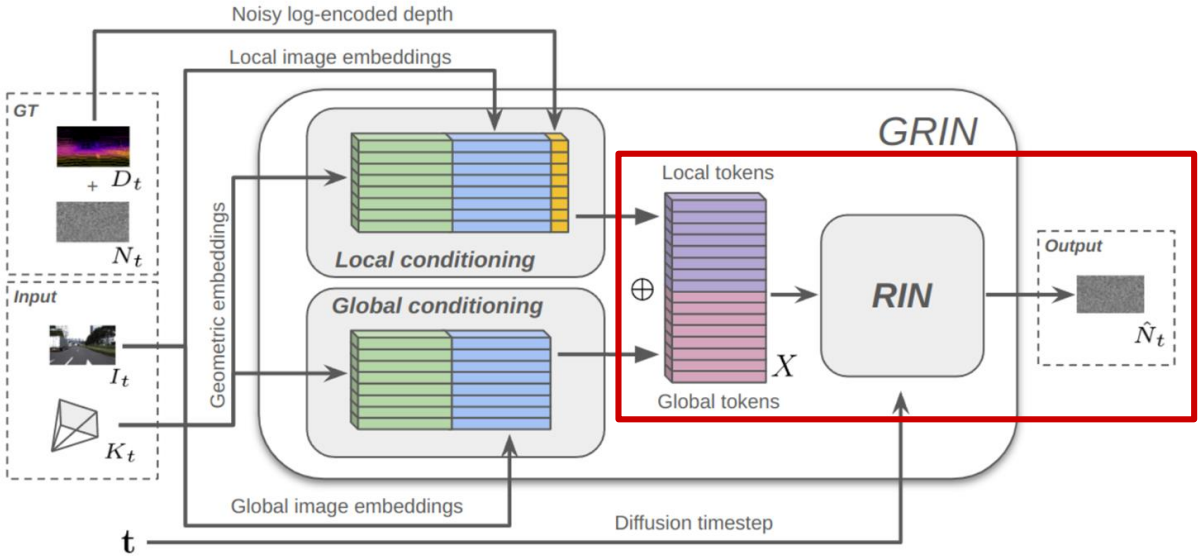
Dense information from the **entire** image



The GRIN Architecture

Local and **global** tokens are concatenated

RIN denoising to generate **local** predictions



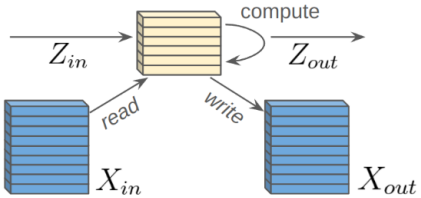
Efficient Diffusion

Recurrent Interface Networks (RIN*)

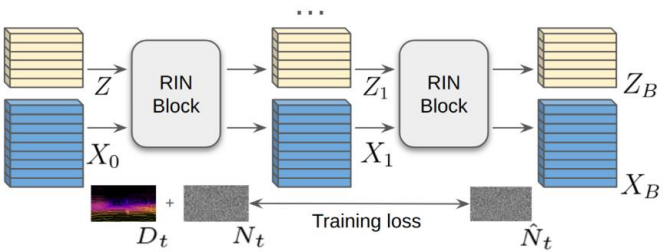
Read: Input tokens are projected (cross-attention) onto a fixed-dimensional latent space

Compute: Self-attention is performed in this latent space

Write: The processed latent space is written back (cross-attention) into the input tokens



(a) RIN block.



(b) RIN model.

**Jabri et al*. Scalable Adaptive Computation for Iterative Generation, ICML 2023

TOYOTA RESEARCH INSTITUTE

Experimental Results

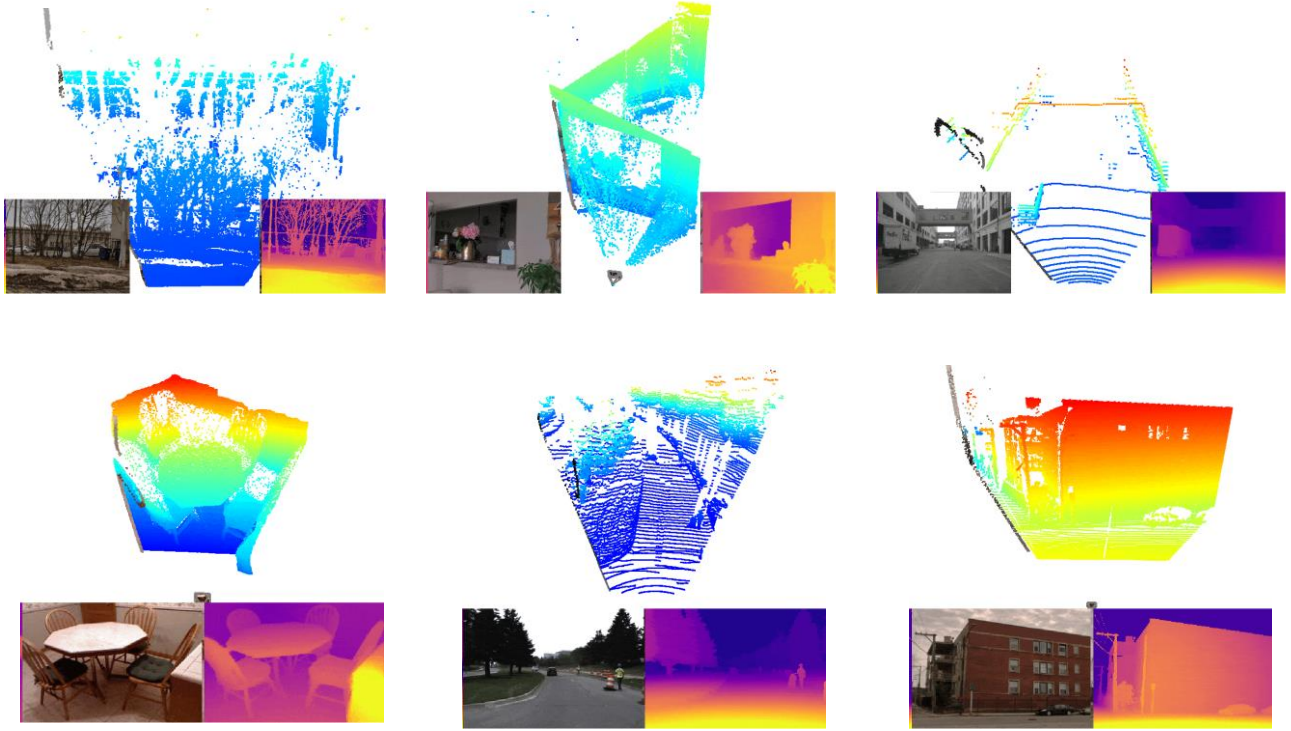
State-of-the-Art zero-shot metric depth estimation

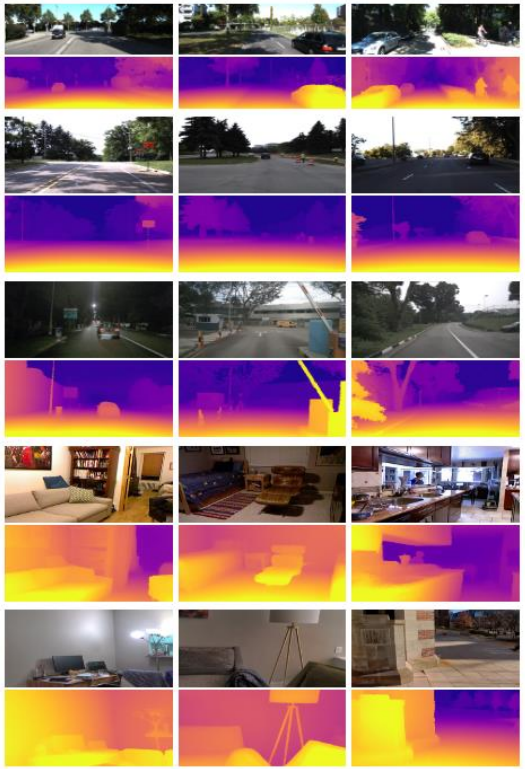
Trained from scratch, on a combination of real-world and synthetic datasets

	AbsRel↓	RMSE↓	$\delta < 1.25$↑	AbsRel↓	RMSE↓	$\delta < 1.25$↑	AbsRel↓	RMSE↓	$\delta < 1.25$↑	AbsRel↓	RMSE↓	$\delta < 1.25$↑
	KITTI [17]			*DDAD* [23]			*nuScenes* [4]			*VKITTI2* [3]		
AdaBins* [1]	0.058	2.360	0.964	0.147	7.550	0.766	0.445	10.658	0.471	0.133	6.248	0.803
NeWCRFs* [78]	0.052	2.129	0.974	0.119	6.183	0.874	0.400	12.139	0.512	0.117	5.691	0.829
ZeroDepth [27]	*0.064*	*2.987*	*0.958*	0.100	6.318	0.889	0.157	7.612	0.822	*0.099*	*4.209*	*0.905*
ZoeDepth† [2]	N/A	N/A	N/A	0.138	7.225	0.824	*0.198*	*8.245*	*0.809*	0.105	5.095	0.850
DMD [55]	N/A	N/A	N/A	0.108	5.365	0.907	N/A	N/A	N/A	0.092	4.387	0.890
Metric3D [76]	0.058	2.770	0.964	N/A	N/A	N/A	0.147	7.889	—	*0.089*	*4.201*	*0.904*
UniDepth [49]	0.047	**2.000**	0.980	*0.097*	*5.399*	*0.919*	*0.143*	*7.425*	*0.839*	*0.078*	*3.850*	*0.923*
GRIN	**0.046**	2.251	**0.983**	**0.093**	**5.307**	**0.922**	**0.138**	**7.217**	**0.857**	**0.074**	**3.501**	**0.937**
	NYUv2 [46]			*SunRGBD* [61]			*DIODE (indoor)* [66]			*DIODE (outdoor)* [66]		
AdaBins* [1]	0.103	0.364	0.903	0.159	0.476	0.771	0.443	1.963	0.174	0.865	10.350	0.158
NeWCRFs* [78]	0.095	0.334	0.922	0.151	0.424	0.798	0.404	1.867	0.187	0.854	9.228	0.176
ZeroDepth [27]	0.100	0.380	0.901	*0.121*	*0.347*	*0.864*	*0.309*	*1.779*	*0.377*	*0.714*	*7.880*	*0.219*
ZoeDepth† [2]	N/A	N/A	N/A	0.123	0.356	0.856	0.331	1.598	0.386	0.757	7.569	0.208
DMD [55]	N/A	N/A	N/A	0.109	0.306	0.914	0.291	1.292	0.380	0.553	8.943	0.187
Metric3D [76]	0.094	0.337	0.926	*0.104*	*0.319*	*0.919*	0.268	1.429	—	0.414	6.934	—
UniDepth [49]	0.063	0.232	**0.984**	*0.106*	*0.316*	*0.918*	*0.237*	*1.329*	*0.408*	*0.401*	*6.491*	*0.278*
GRIN	**0.058**	**0.209**	0.980	**0.098**	**0.301**	**0.927**	**0.221**	**1.128**	**0.439**	**0.393**	**6.011**	**0.303**

TOYOTA RESEARCH INSTITUTE

Experimental Results

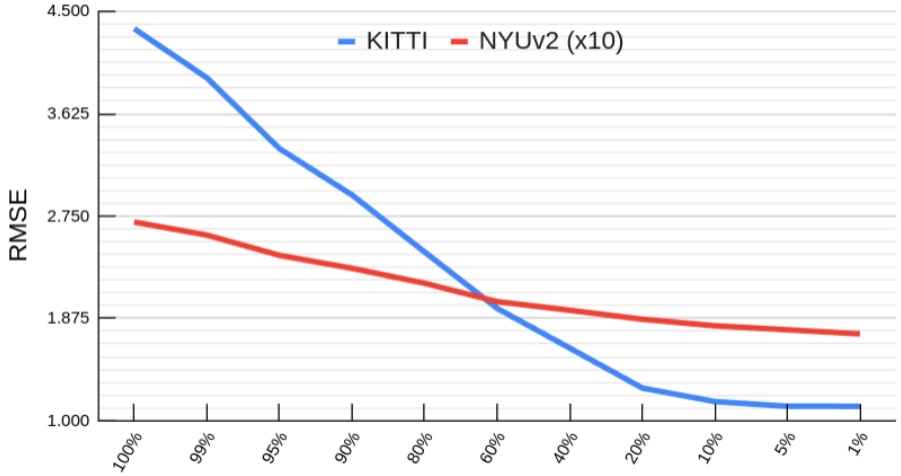


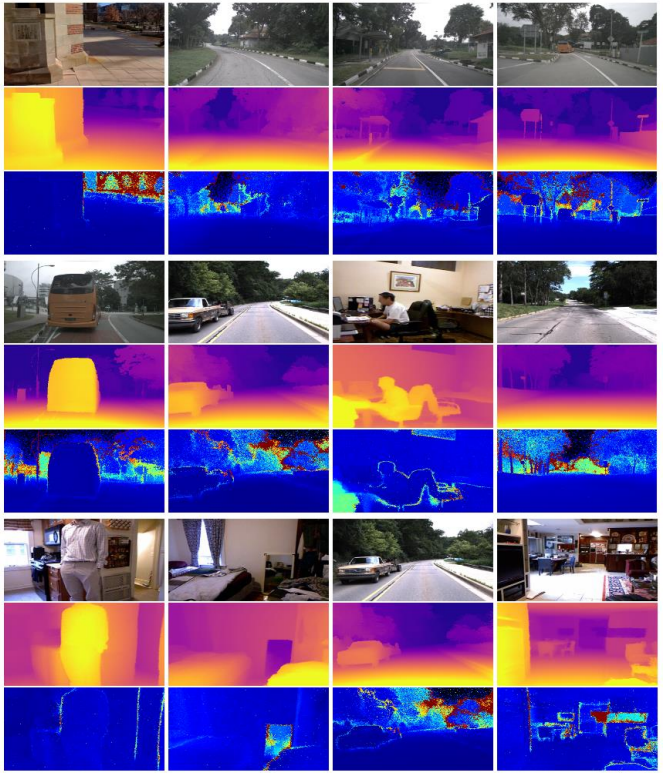


Uncertainty Estimation

Standard deviation from multiple samples

Improvements by **filtering out** inaccurate pixels





Outline

Introduction

Mono-Depth

MultiView-Depth

Conclusion

GRIN Limitations

Single input (monocular)

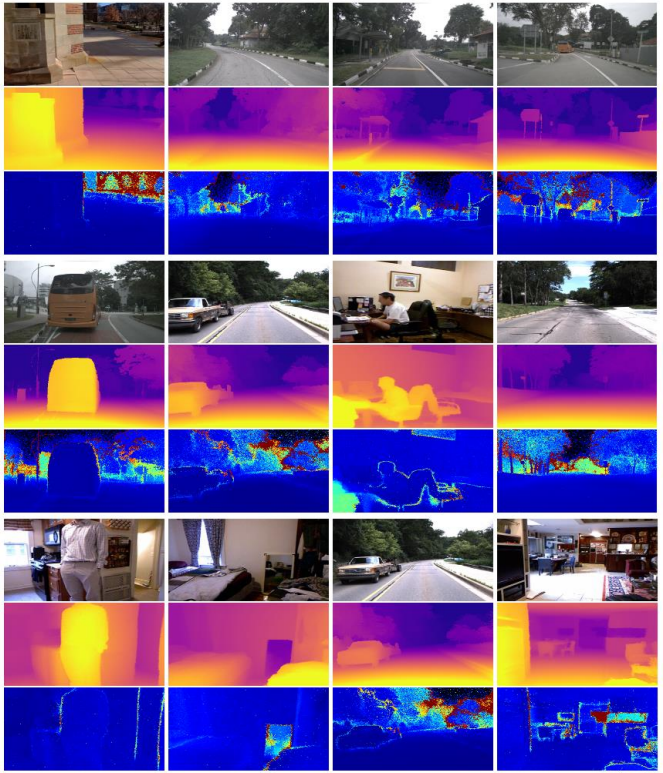
→ **Multiple conditioning cameras**

Single task (depth estimation)

→ **Multi-task: depth and image**

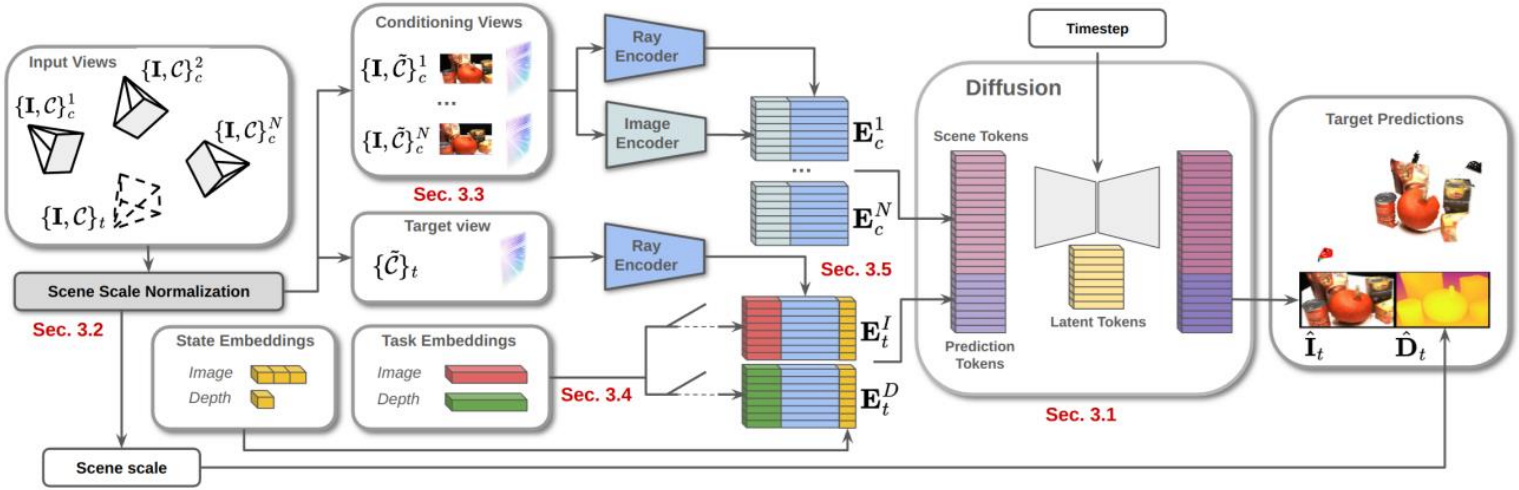
Fixed output viewpoint

→ **Novel view from any viewpoint**



Zero-Shot Novel View and Depth Synthesis with Multi-View Geometric Diffusion

Vitor Guizilini, Zubair Irshad, Dian Chen, Greg Shakhnarovich, Rares Ambrus. CVPR25.



MVGD extends GRIN to **novel view and depth synthesis** from any viewpoint

Efficient network architecture allows **500+ input conditioning cameras**

Scene **scale normalization** for training on a large-scale, heterogeneous dataset

Zero-Shot Novel View and Depth Synthesis with Multi-View Geometric Diffusion

Vitor Guizilini, Zubair Irshad, Dian Chen, Greg Shakhnarovich, Rares Ambrus. CVPR25.

Scene scale normalization allows training on a large-scale, heterogeneous dataset (**60M samples**)

Dataset	Syn.	Dyn.	Met.	I/O	V/M	# Sequences	# Samples
ArgoVerse2 [105]		✓	✓	O	VM	1043	3, 909, 297
BlendedMVG [117]				O	M	502	115, 142
CO3Dv2 [74]				I	V	25, 243	5, 088, 873
DROID [56]			✓	I	M	76, 792	7, 340, 712
LSD [37]		✓	✓	O	VM	17, 647	1, 057, 920
LyftL5 [42]		✓	✓	O	VM	394	347, 508
MVImgNet [122]				I	V	194, 368	5, 768, 120
NuScenes [8]		✓	✓	O	VM	850	204, 894
Taskonomy [123]			✓	O	M	533	4, 584, 462
HM3D [71]			✓	O	M	900	9, 531, 876
Hypersim [75]			✓	O	M	457	74, 619
P. Domain [32, 35]	✓	✓	✓	IO	VM	5, 449	555, 000
RealEstate10k [125]			—	O	V	74, 532	10, 115, 793
RTMV [97]	✓			I	M	1, 909	286, 350
ScanNet [16]			✓	O	V	1, 513	2, 477, 378
TartanAir [102]	✓	✓	✓	O	VM	369	613, 274
VKITTI2 [7]	✓	✓	✓	O	VM	45	38, 268
Waymo [93]	✓		✓	O	VM	1, 000	990, 340
WildRGBD [108]				I	V	23, 049	8, 026, 495
Total						**426, 595**	**61, 126, 321**

TOYOTA RESEARCH INSTITUTE

Zero-Shot Novel View and Depth Synthesis with Multi-View Geometric Diffusion

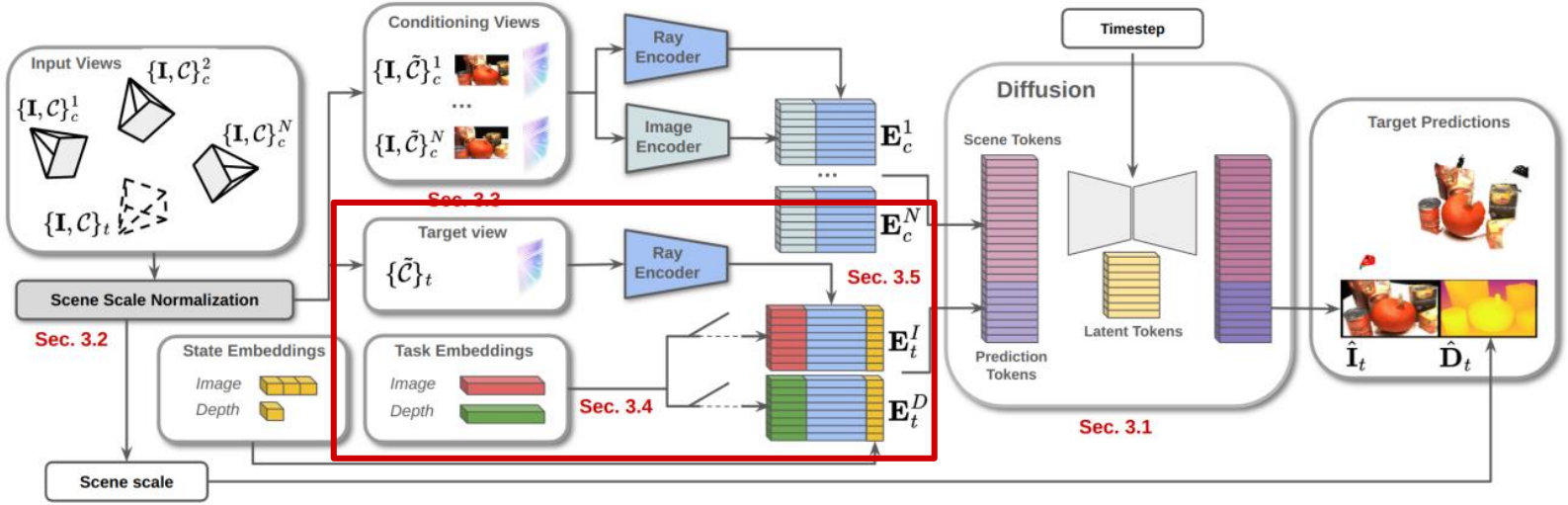
Vitor Guizilini, Zubair Irshad, Dian Chen, Greg Shakhnarovich, Rares Ambrus. CVPR25.

Efficient network architecture allows **incremental model upsampling** without restarting from scratch

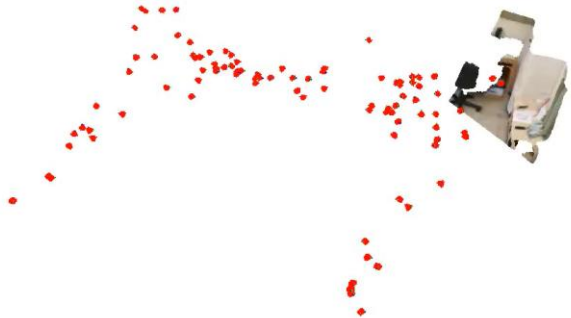
# Latents	# Param.	*RE10K (2-view)*		*CO3Dv2 (3-view)*			
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	Abs.Rel.↓	RMSE↓
256	417.9M	25.89	0.841	18.48	0.567	0.121	5.347
512	418.2M	26.33	0.859	19.56	0.590	0.109	5.048
1024	418.7M	27.73	0.870	20.08	0.622	0.104	4.766
2048	419.7M	28.41	0.891	20.68	0.678	0.101	4.654

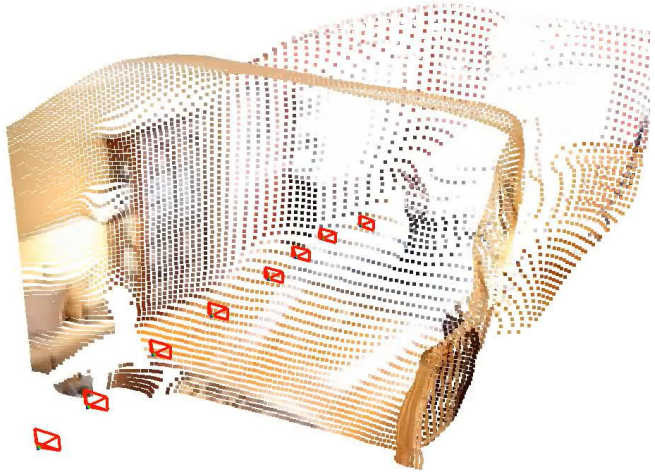
Zero-Shot Novel View and Depth Synthesis with Multi-View Geometric Diffusion

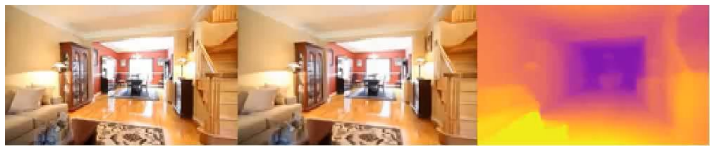
Vitor Guizilini, Zubair Irshad, Dian Chen, Greg Shakhnarovich, Rares Ambrus. CVPR25.



Task and target embeddings allow switching between **novel view and depth synthesis** from any viewpoint







Red - input views; Green - novel viewpoint

Pointclouds created by aggregating depth from all novel viewpoints

Outline

Introduction

Mono-Depth

MultiView-Depth

Conclusion

Conclusion

Geometric RIN (GRIN): Efficient Pixel-Level Diffusion with Sparse Labels → efficient zero-shot metric monocular depth estimation

Zero-Shot Novel View and Depth Synthesis with Multi-View Geometric Diffusion → arbitrary number of conditioning views, large-scale pretraining and incremental upsampling

Thanks to many collaborators!

A Alspach, A Beaulieu, J Bohg, W Burgard, A Bühler, D Chen, H Chiu, A Cramariuc, A Dave, KG Derpanis, Y Du, F Durand, J Fang, Z Fang, K Fragkiadaki, WT Freeman, A Gaidon, A Ganeshan, L Guibas, V Guizilini, AW Harley, N Heppert, X Huang, T Ikeda, MZ Irshad, S Iwase, P Jensfelt, T Kanai, W Kehl, T Kerola, H Kim, Z Kira, K Kitani, T Ko, T Kollar, M Kowal, Y Kudo, N Kuppuswamy, R Lee, KH Lee, J Li, S Lin, K Liu, M Lunayach, S Maeda, H Mei, K Nishiwaki, D Park, S Pillai, A Raventos, D Rempe, G Rosman, T Sadjadpour, G Shakhnarovich, P Sharma, V Sitzmann, J Solomon, C Stearns, X Tan, J Tang, JB Tenenbaum, A Tewari, P Tokmakov, A Valada, A Vallet, I Vasiljevic, M Walter, Y Wang, Y Yang, S Zakharov

TOYOTA RESEARCH INSTITUTE

Rareș Ambruș, PhD
Senior Manager
Large Behavior Models

Questions?

Code & Data: https://github.com/TRI-ML/vidar
Blog posts: https://medium.com/toyotaresearch/
Open Positions: https://tri.global/careers
Twitter: https://twitter.com/ToyotaResearch

TOYOTA RESEARCH INSTITUTE

References

Depth estimation:

- **Geometric RIN (GRIN): Efficient Pixel-Level Diffusion with Sparse Labels**
  Vitor Guizilini, Pavel Tokmakov, Achal Dave, Rares Ambrus. 3DV25.
- **Zero-Shot Novel View and Depth Synthesis with Multi-View Geometric Diffusion**
  Vitor Guizilini, Zubair Irshad, Dian Chen, Greg Shakhnarovich, Rares Ambrus. CVPR25.

Object-centric representations:

- **ReFiNe: Recursive Field Networks for Cross-Modal Multi-Scene Representation**.
  Zakharov, Liu, Gaidon, Ambrus. SIGGRAPH24.
- **ZeroGrasp: Zero-Shot Shape Reconstruction Enabled Robotic Grasping**
  Iwase, Irshad, Liu, Guizilini, Lee, Ikeda, Amma, Nishiwaki, Kitani, Ambrus, Zakharov. CVPR25.
- **OmniShape: Zeroshot Multi-Hypothesis Shape and Pose Estimation in the Real World**
  Liu*, Zakharov*, Chen, Ikeda, Gaidon, Shakhnarovich, Ambrus. ICRA25.

Robot policies: failure detection and statistical analysis:

- **Is your imitation learning policy better than mine? policy comparison with near-optimal stopping**
  Snyder, Hancock, Badithela, Dixon, Miller, Ambrus, Majumdar, Itkina, Nishimura. RSS25.
- **Can We Detect Failures Without Failure Data? Uncertainty-Aware Runtime Failure Detection for Imitation Learning Policies**
  Xu, Nguyen, Dixon, Rodriguez, Miller, Lee, Shah, Ambrus, Nishimura, Itkina. RSS25.