

# GRIN: Zero-Shot Metric Depth with Pixel-Level Diffusion

## Supplementary Material

In this supplementary material we report additional results, visualizations, and implementation details that could not be included in the main paper due to space limitations. We begin by providing evidence in Section 1 that GRIN can be minimally modified to process multiple images, achieving state of the art results in generalized stereo depth estimation as well. Afterwards, in Section 2 we show additional visualizations on different evaluation benchmarks, and in Section 3 we qualitatively ablate the effects of global conditioning. We then ablate the effect of large-scale, image-text pre-training on GRIN depth estimation performance in Section 4, provide additional architecture details in Section 5, and finally in Section 6 we discuss potential limitations of our proposed architecture.

### 1. Generalized Stereo Depth Estimation

Although our main focus is zero-shot metric *monocular* depth estimation, here we explore how GRIN can be minimally modified to accommodate *multi-view* tasks. This is achieved by globally conditioning the diffusion process on information from multiple images, with geometric embeddings calculated relative to a shared frame of reference. This extension requires known relative camera extrinsics  $\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$ . This information is used to augment the geometric embeddings (Section 4.2 of the main paper) such that  $\mathbf{r}_{jk} = (\mathbf{KR})^{-1} [u_{jk}, v_{jk}, 1]^T$  and  $\mathbf{t}_{jk} = [x, y, z]^T$ . This is aligned with a recent trend of *implicit multi-view geometry* [4, 5, 24], in which explicit constraints such as epipolar projections and cost volumes [7, 23] are eschewed in favor of input-level inductive biases that enable the implicit learning of useful multi-view correlations. Local conditioning remains the same, thus allowing the generation of pixel-level predictions from specific (or multiple) viewpoints.

Note that, because in our monocular experiments we used a canonical camera transformation  $\mathbf{T} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}$  to produce geometric embeddings, the same model can be directly repurposed for the multi-view setting. Thus, we fine-tuned our pre-trained monocular GRIN model on a combination of the stereo split of **ScanNet** [2], with 94212 training pairs; and **DeMoN** [20], with 166285 training pairs collected from three distinct datasets: **SUN3D** [22], **RGBD-SLAM** [19], and **Scenes11** [20]. Evaluation was performed in-domain, on the test splits from the same datasets following [10], with results in Table 1. As we can see, GRIN also achieves state of the art performance in stereo depth estimation, outperforming methods that rely on explicit multi-view geometry as well as recent implicit multi-view geometry methods.




	Method	Abs.Rel.↓	RMSE↓	
ScanNet	DPSNet [7]	0.126	0.315	
	NAS [10]	0.107	0.281	
	IIB [24]	0.116	0.281	
	DeFiNe [3]	0.093	0.246	
	<b>GRIN</b>	<b>0.088</b>	<b>0.224</b>	
SUN3D	DeepMVS [6]	0.282	0.944	
	DPSNet [7]	0.147	0.449	
	NAS [10]	0.127	0.378	
	IIB [24]	0.099	0.293	
	<b>GRIN</b>	<b>0.097</b>	<b>0.274</b>	
RGBD	DeepMVS [6]	0.294	0.868	
	DPSNet [7]	0.151	0.695	
	NAS [10]	0.131	0.619	
	IIB [24]	0.095	0.550	
	<b>GRIN</b>	<b>0.092</b>	<b>0.512</b>	

Table 1. **Stereo depth estimation results.** The same GRIN model was used in all evaluations, with quantitative results on the left and qualitative examples on the right.

### 2. Additional Qualitative Results

In Figure 1 we show additional GRIN qualitative results on different indoor and outdoor images from our evaluation benchmarks. We used the same model from our quantitative evaluation (Table 1, main paper) to produce these results. Due to the generative properties of GRIN, we can obtain multiple depth predictions from the same input image, and use those to (i) improve accuracy by calculating the *median* of all samples, as shown in the middle rows; and (ii) produce an uncertainty map by calculating the *standard deviation* of all samples, as shown in the bottom rows. From these results we can see that the calculated uncertainty maps follow our expectations, i.e., longer ranges are less accurate, as well as object boundaries and sharp discontinuities. Interestingly, these uncertainty maps also accurately detect failure cases of our model, such as the mirror on the bottom of the second column, due to the higher variance between predictions. Similarly, in Figure 2 we show reconstructed pointclouds generated from GRIN predicted depth maps, unprojected to 3D via the camera intrinsics.

### 3. Effects of Global Conditioning

In Figure 3 we ablate the effects of global conditioning by incrementally removing a percentage of global vectors during inference. As we can see, quality degrades as we decrease the amount of global information available to condition the diffusion process, and this degradation takes the form of less-defined boundaries and overall loss of fine-grained details. Interestingly, removing 50% of global vectors does not affect results significantly, and it is still possi-

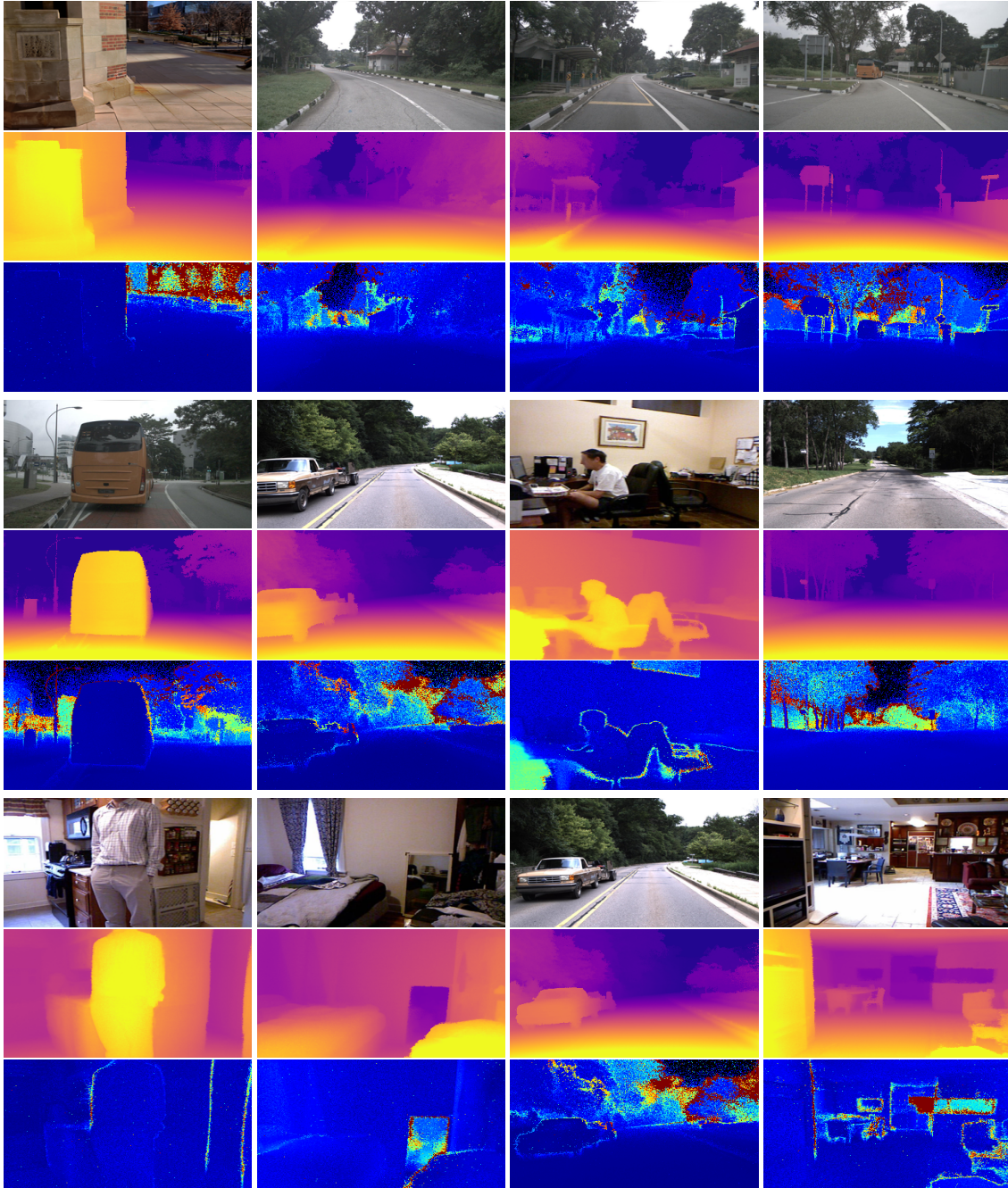


Figure 1. **Zero-shot GRIN qualitative results**, including input image (top), predicted depth map (middle), and uncertainty map (bottom).

ble to observe details in the predicted depth map with as few as 25%. We attribute this robustness to our dropout strategy (Section 4.4, main paper), that promotes robustness to sparse global conditioning. However, as shown in our ablation study (Table 2, main paper), the introduction of global conditioning significantly improves results, relative to the baseline of using only local conditioning on sparse data.

#### 4. Image-Text Pre-Training

We now ablate the effect of large-scale pre-training and report the results in Table 2. To this end, we follow [13] and pre-train GRIN on 400M text-image pairs [17] for the task of text-to-image diffusion. We set the image resolution to  $256 \times 256$  and use a VQ encoder [26] with a down-sampling factor of 4. Captions are encoded with CLIP [12] (the ViT-



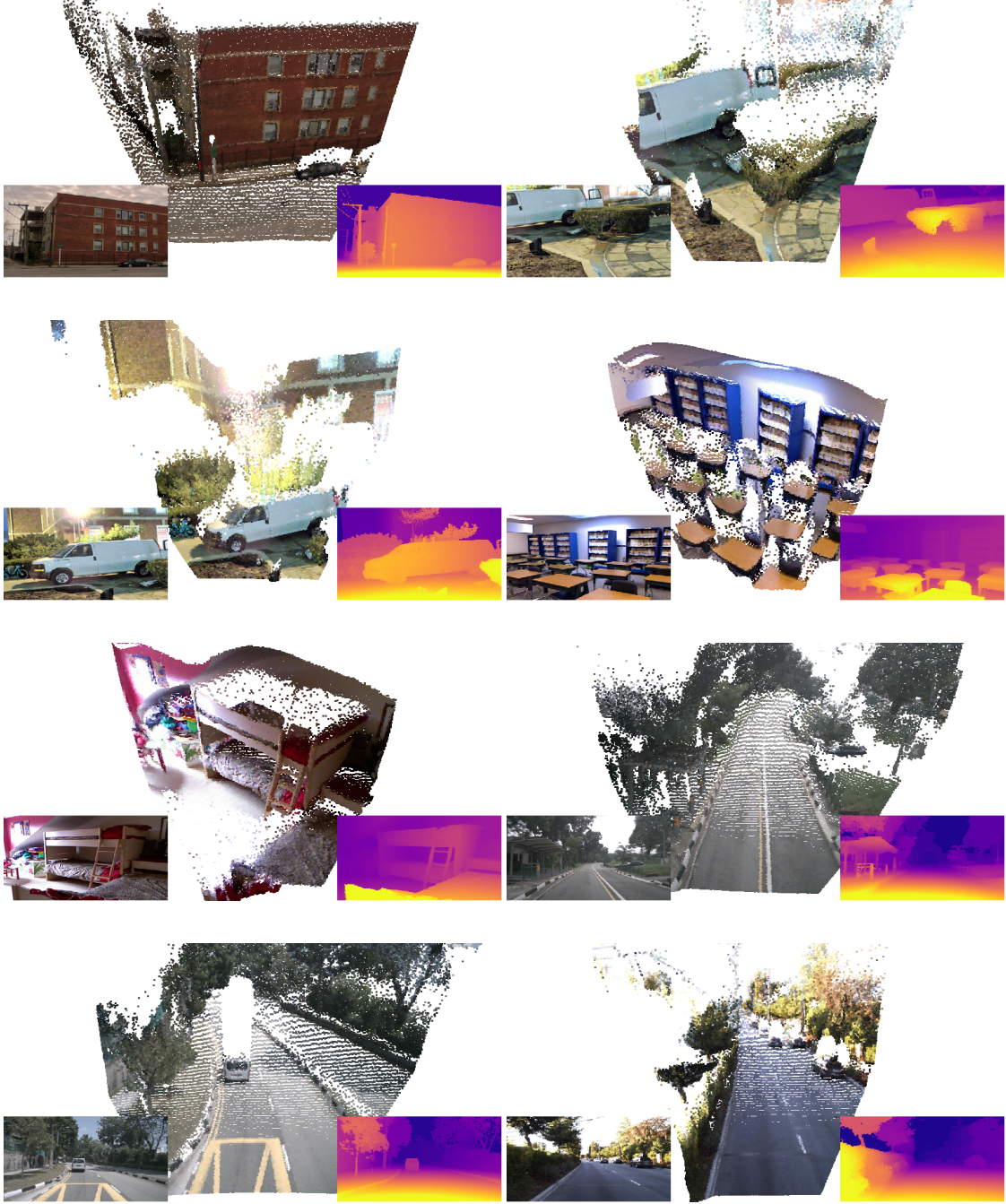


Figure 2. **Zero-shot reconstructed pointclouds**, obtained by unprojecting RGB pixels onto 3D space using GRIN depth predictions and camera intrinsics.

L/14 variant) and conditioning is performed by concatenating the caption encoding to the RIN latent tokens. We train the model for 1 million iterations with a batch size of 1024 using the LION optimizer [1] and a learning rate of  $1.5e-5$  and weight decay of  $2.0e-1$ . Following [8], we use a warm up schedule and a cosine learning rate decay. The resulting

model achieves a comparable Inception Score [14] to that of StableDiffusion [13] on the COCO validation set [11], confirming the effectiveness of our pre-training approach.

Quantitative results obtained when training GRIN initialized from this checkpoint, relative to training GRIN from scratch, are shown in Table 2. Interestingly, starting from a

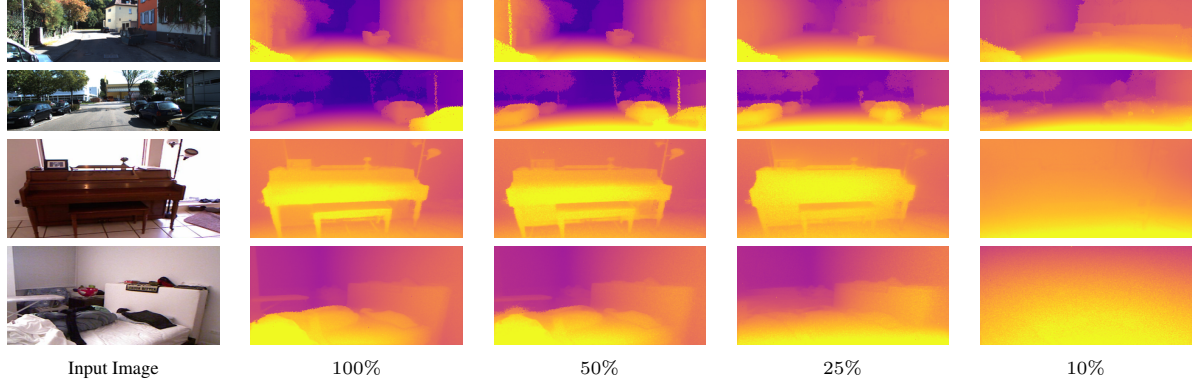


Figure 3. **Degradation in depth estimation performance** when removing global conditioning vectors during inference. The percentage value indicates how many global conditioning vectors are maintained, randomly sampled from the total of  $\frac{HW}{16}$  vectors.

Method	<i>KITTI</i>			<i>NYUv2</i>		
	AbsRel	RMSE $\delta < 1.25$		AbsRel	RMSE $\delta < 1.25$	
<b>GRIN</b> (LAION)	0.047	2.228	0.979	0.058	0.213	0.978
<b>GRIN</b> (scratch)	0.046	2.251	0.983	0.058	0.209	0.980

Table 2. **Effect of large-scale image-text pre-training** on zero-shot GRIN depth estimation results.

pre-trained checkpoint did not improve results significantly. We hypothesize that this is due to the different target tasks (i.e. RGB vs depth generation), as well as the availability of ample target domain data (i.e. depth labels). We leave as future work a more in-depth evaluation on the trade-off between text-image pre-training versus the number of target domain labels.

## 5. Architecture Details

We implemented GRIN with a  $\mathbf{Z} \in \mathbb{R}^{256 \times 1024}$  latent space, 16 read-write heads, 16 latent heads, and a sequence of 4 RIN blocks, each with a depth of 6. Global image embeddings  $\mathbf{F}^{glob}$  used a ResNet18 encoder, resulting in  $\frac{HW}{16}$  960-dimensional vectors that were projected to 512 dimensions using a  $1 \times 1$  convolutional layer. Local image embeddings  $\mathbf{F}^{loc}$  used a  $9 \times 9$  convolutional layer with reflexive padding to generate  $HW$  128-dimensional vectors. Geometric embeddings  $\mathbf{g}_{jk}$  were calculated using 16 bands with a maximum frequency of 2, based on cameras matching the resolution of corresponding image embeddings. Depth estimates were generated between 0.1 and 200 meters, with base 10 for the log-scale parameterization. During training, we subsampled  $L = 1024$  valid pixels as supervision, and  $G = 2048$  global embeddings for conditioning. During inference, following [9] we generate 10 estimates by sampling different noise values and output the median value as our final prediction. In total, our implemented GRIN architecture has 341, 563, 599 parameters. We build upon the

open-source RIN PyTorch implementation from [21]. For additional details, we refer the reader to the supplementary material. Our code and pre-trained models will be made available upon acceptance.

## 6. Limitations

GRIN enables training with unstructured sparse data by operating at the pixel-level, which removes the need for latent autoencoders that require inputs with explicit spatial structure (i.e., 2D image grids). This is possible in large part due to the efficiency inherent to the RIN architecture, with bottleneck latent tokens where self-attention is computed. Although powerful, further work is still required to improve efficiency, especially during inference due to the multiple denoising steps required to produce depth estimates. Recent developments in how to speed up image generation, such as sample efficient denoising [18] and distillation [25], should improve performance significantly. In accordance to [16], we have noticed that log-space depth parameterization improves performance, however there is still some trade-off between accuracy in shorter and longer ranges (Table 2, main paper) that we believe can be addressed with better parameterization and the use of alternative diffusion objectives [15]. Moreover, we noticed some instability during training, both in terms of the optimizer choice (LION [1] performed the best) and learning rate (larger learning rates led to mid-training divergence). We also observed some sensitivity to the training datasets, ensuring a similar ratio of real-world and synthetic datasets, as well as indoor and outdoor datasets, was key to achieving our reported performance.

## References

- [1] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms, 2023. 3, 4

- [2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017. 1
- [3] Vitor Guizilini, Igor Vasiljevic, Jiading Fang, Rare Ambru, Greg Shakhnarovich, Matthew R Walter, and Adrien Gaidon. Depth field networks for generalizable multi-view scene representation. In *European Conference on Computer Vision*, pages 245–262. Springer, 2022. 1
- [4] Vitor Guizilini, Igor Vasiljevic, Jiading Fang, Rares Ambrus, Greg Shakhnarovich, Matthew R Walter, and Adrien Gaidon. Depth field networks for generalizable multi-view scene representation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 245–262. Springer, 2022. 1
- [5] Vitor Guizilini, Igor Vasiljevic, Jiading Fang, Rares Ambrus, Sergey Zakharov, Vincent Sitzmann, and Adrien Gaidon. Delira: Self-supervised depth, light, and radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 1
- [6] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *CVPR*, 2018. 1
- [7] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv:1905.00538*, 2019. 1
- [8] Allan Jabri, David J. Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In *ICML*, 2023. 3
- [9] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation, 2023. 4
- [10] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2199, 2020. 1
- [11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3
- [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2
- [13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3
- [14] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 2016. 3
- [15] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. 4
- [16] Saurabh Saxena, Junhwa Hur, Charles Herrmann, Deqing Sun, and David J. Fleet. Zero-shot metric depth with a field-of-view conditioned diffusion model, 2023. 4
- [17] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2
- [18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4
- [19] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 1
- [20] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 1
- [21] Phil Wang. Recurrent interface network (pytorch), 2022. 4
- [22] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *2013 IEEE International Conference on Computer Vision*, pages 1625–1632, 2013. 1
- [23] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 1
- [24] Wang Yifan, Carl Doersch, Relja Arandjelović, João Carreira, and Andrew Zisserman. Input-level inductive biases for 3D reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2022. 1
- [25] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *CVPR*, 2024. 4
- [26] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *ICLR*, 2021. 2