High-Resolution Synthetic RGB-D Datasets for Monocular Depth Estimation –Supplementary Material–

Aakash Rajpal[1,2,], Noshaba Cheema[2,3,4], Klaus Illgner-Fehns[1],
Philipp Slusallek[2,4], and Sunil Jaiswal[1,*]

[1] K|Lens GmbH, Germany, [2] Saarland Informatics Campus, Germany,
[3] MPI Informatics, Germany, [4] German Research Center for Artificial Intelligence (DFKI), Germany

1. Evaluation Metrics

Recall for evaluating different depth estimation algorithms; we use Abs-Rel, RMSE, and percentage of correct pixel. Below are the equations used for each error metric:-

- Absolute relative error (abs-rel):-

$$\frac{1}{N}\sum_{p=1}^{N}\frac{|y_p-\hat{y}_p|}{\hat{y}_p} \tag{1}$$

- Root mean squared error (rmse):-

$$\sqrt{\frac{1}{N}\sum_{p=1}^{N}\frac{\|y_p-\hat{y}_p\|^2}{\hat{y}_p}} \tag{2}$$

- Accuracy with threshold t: percentage(%) of $\hat{y}_p$, subject to max

$$\left(\frac{\hat{y}_p}{y_p},\frac{y_p}{\hat{y}_p}\right)=\delta<t\,(t\in[1.25]) \tag{3}$$

where y_p and $\hat{y}_p$ are the ground-truth depth and the estimated depth at pixel p respectively; N is the total number of pixels of test images.

2. Proposed HRSD datasets

The proposed HRSD dataset consists of images varying in captured environments and objects (indoor/outdoor scenes, dynamic objects, homogeneous scenes). This enables improved training of monocular depth estimation algorithms leading to a better overall generalization of diverse scenes. We show the statistics of our HRSD dataset in Table 1 which includes details on a number of indoor scenes, outdoor scenes, etc. We generated a total of 100,000 pairs of RGB and ground truth depth. Some more examples can be seen in Figure 1.

HRSD (1920*1080)	Train	Validation	Test
Indoor	29,000	5000	4000
Outdoor	46000	10000	6000

Table 1. Statistics of Proposed HRSD datasets

Scripthook V Recall we used G2D [3] to extract depth maps from Gbuffers during game-play in the rendering pipeline. We also want to credit ScriptHook V as G2D is formulated on Scripthook V [2], an Alexander Blade library that allows access to GTA-V's native functionality. Adopting Scripthook V distinguishes our dataset from Richter et al. [7] to create modifications ("mods") for meddling within the detailed virtual environment of GTA-V. G2D allows users to collect hyper-realistic computer-generated imagery of an urban scene under controlled 6DOF camera poses and varying environmental conditions. Users directly interact with G2D while playing the game; specifically, users can manipulate the conditions of the virtual environment on the fly. The original aim of Scripthook V is to provide a framework to construct modifications ("mods") to the game. Currently, a wide range of fascinating mods is available, e.g., the Invisibility Cloak, which can make the protagonist invisible. The list of native functions supported by Scripthook V can be found on [1].

3. Ablation Studies

Recall that we propose to use an attention-based supervision loss (AL) term in addition to the default L_1 loss in the proposed algorithm. Here, we discuss the effects of choos-

This work was partially funded by the German Ministry for Education and Research (BMBF).
*Corresponding author: sunil.jaiswal@k-lens.de

ing different loss functions for training DPT [6] networks on the proposed HRSD datasets.

Effect of loss function To verify the effectiveness of the additional attention loss term, we trained the model (DPT-B + R) with three different loss terms, and they are as follows: *1)* L1 loss term, *2)* DPT's [6] loss and *3)* L_1+ L_{AL}.

This study is performed on a reduced HRSD dataset of only 30000 images. We used 25000 images for training and 5000 images for the validation set. We report results on validation sets with the metric described in equation (1) - (3), and the results are shown in Table 2. From Table 2, it is obvious that the L_1+ L_{AL} gives a relatively low error and higher accuracy compared to other loss functions, demonstrating attention-loss effectiveness.

Method	AbsRel ↓	RMSE ↓	$\delta > 1.25$↑
L_{DPT} [6]	0.107	0.394	0.872
L_1	0.095	0.324	0.891
L_1+ L_{AL}	**0.074**	**0.288**	**0.921**

Table 2. Evaluation of different loss functions on the proposed HRSD datasets.

4. Repeating visual results

Due to the main paper's space constraint, we repeat all the visual results to give a more detailed visualization.

Our final architecture (DPT-B + R + AL) trained on the HRSD dataset gives fine details while also improving global coherence in challenging areas, for example, large homogeneous regions of trees/grass in the outdoor scene 4 or a cluster of small objects in the background in indoor scenes 3. DPT [6] algorithm on high-resolution images leaves out structures of objects far away in indoor scenes, like the items on the plate 3. On the other hand, multires [5] produce sharper boundaries and perform better on thinner objects such as traffic signs and poles 4 but provide inconsistent depth maps due to depth bleeding in certain areas highlighted by a green rectangle. Compared to both, our method does not exhibit any such artifacts and results in a smoother and more robust depth map closest to ground truth depth on high-resolution images.

References

[1] GTA V KERNEL MODS description. http://www.dev-c.com/natived b/. Accessed: 2010-09-30. 1

[2] Scripthook V description. https://www.gta5-mods.com/. Accessed: 2010-09-30. 1

[3] Anh-Dzung Doan, Abdul Mohsi Jawaid, Thanh-Toan Do, and Tat-Jun Chin. G2d: from gta to data, 2018. 1

[4] Andreas Geiger, P Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: the kitti dataset. *The International Journal of Robotics Research*, 32:1231–1237, 09 2013. 6

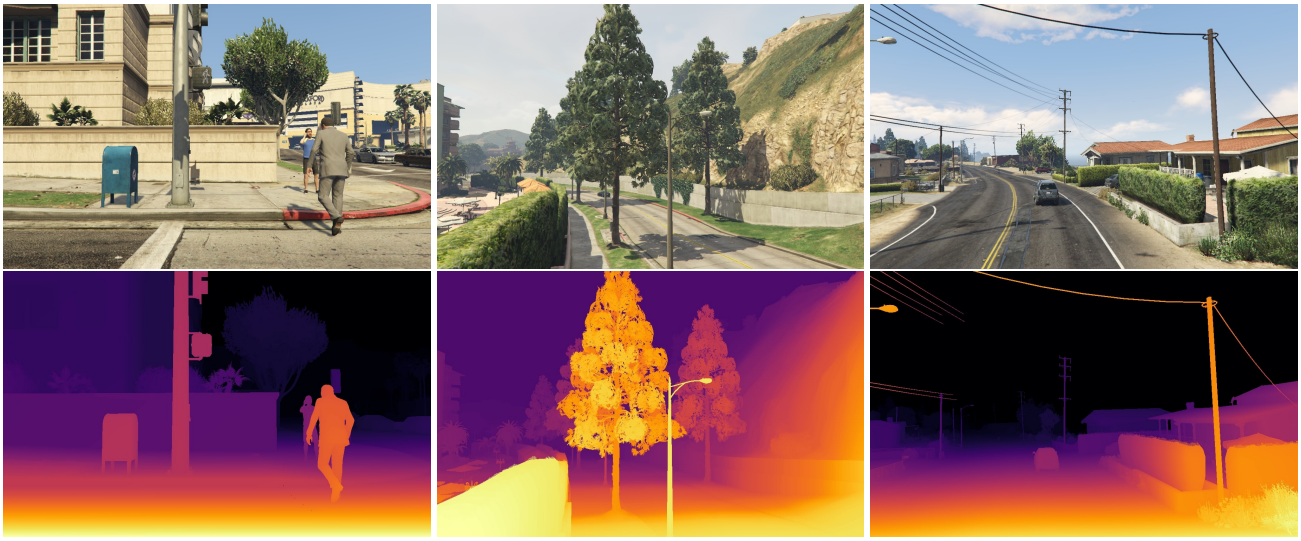
[5] S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9680–9689, 2021. 2, 5, 6

[6] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. 2021. 2, 4, 5, 6

[7] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games, 2016. 1

[8] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5

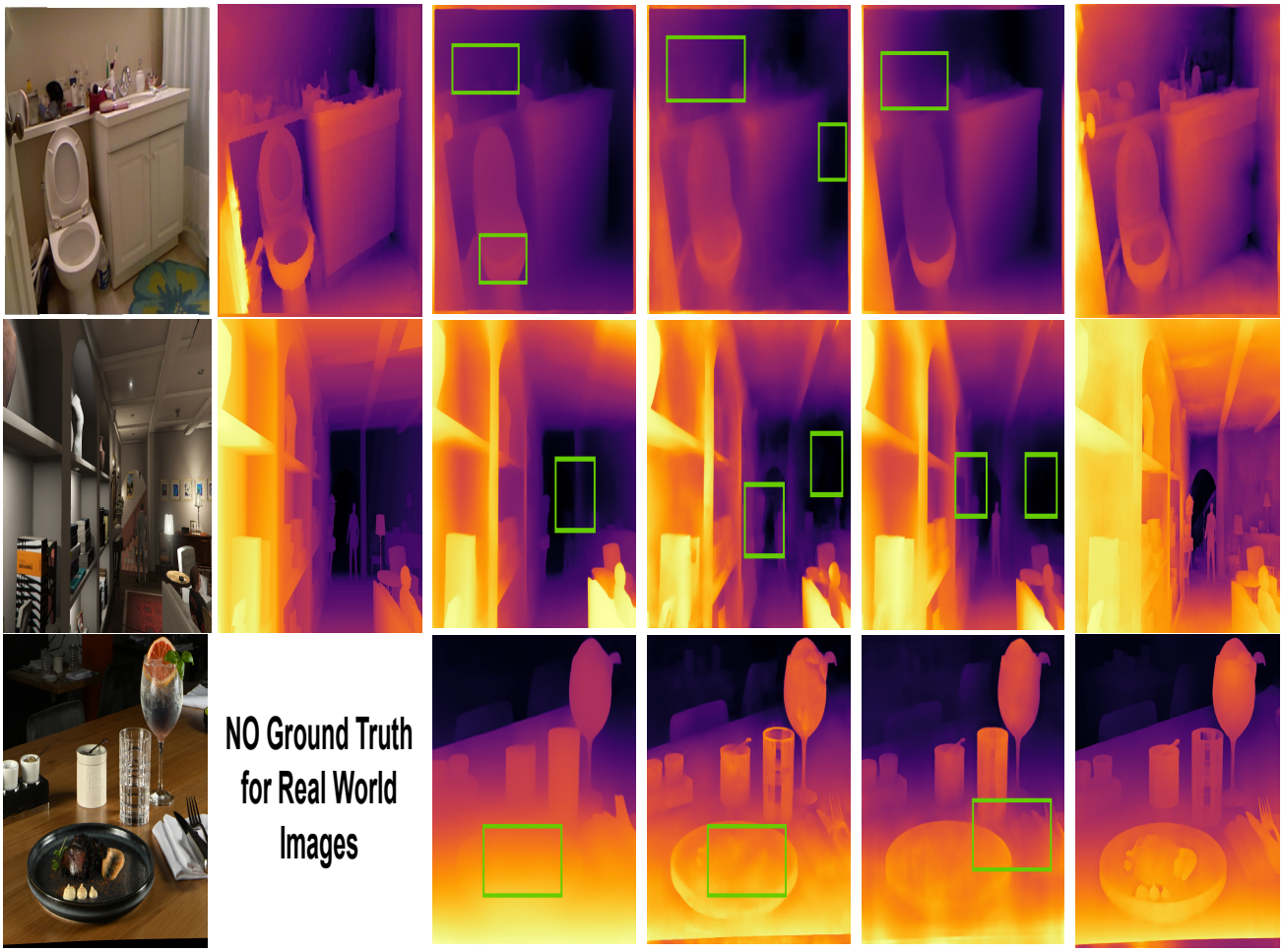
(a) Indoor Scenes



(b) Outdoor Scenes

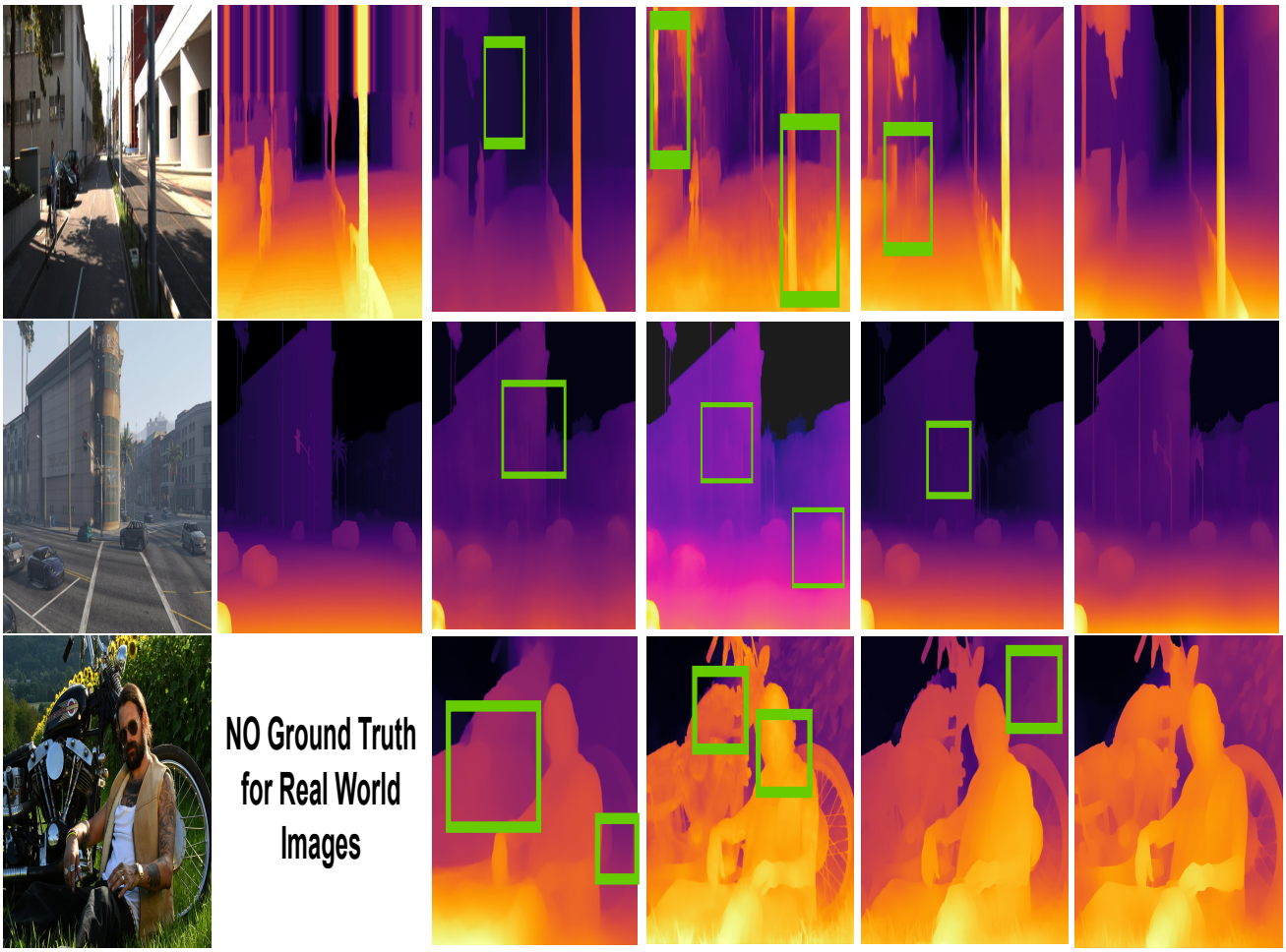
Figure 1. Some scenes from the proposed HRSD dataset.

Figure 2. Improvement over state-of-the-art DPT [6]. Ours DPT-B and Ours DPT-B+R+Al are two variants of DPT [6] trained on the proposed HRSD datasts.



(a) RGB (b) GT-Depth (c) DPT [6] (d) MultiRes [5] (e) ViT-B + R + AL (f) DPT-B + R + AL

Figure 3. Indoor Scenes. 1st Row:- NYU [8]. 2nd Row:- HRSD indoor. 3rd Row:- RealWorld. Our DPT-B + R + AL gives a consistent depth map across all regions and displays sharp structure for overall objects i.e. items on the table in real-world image. Original DPT fails to identify objects in the background as shown by the green rectangles i.e. no structure of human in HRSD indoor. Multires leads to inconsistent depth map highlighted by green rectangles i.e. the toilet seat in NYU image.



(a) RGB (b) GT-Depth (c) DPT [6] (d) MultiRes [5] (e) ViT-B + R + AL (f) DPT-B + R + AL

Figure 4. Outdoor Scenes. 1st Row:- KITTI [4]. 2ndRow:- HRSD outdoor. 3rd Row:- RealWorld. Similar to indoor scenes, our DPT-B + R + AL gives the best performance outputting consistent depth map with precise overall structure i.e. the motorbike in the real-world image. Original DPT, again fails to identify objects in the background as shown by the green rectangle i.e. no structure of background buildings in KITTI image. Multires leads to inconsistent depth map, highlighted by green rectangles i.e. depth around the biker body is fluctuating in real-world image.