PaReNeRF: Toward Fast Large-scale Dynamic NeRF with Patch-based Reference - Supplementary Material

Xiao Tang*[1], Min Yang[1], Penghui Sun[1], Hui Li[1], Yuchao Dai[2], Feng Zhu[1], Hojae Lee[1]
[1]Samsung R&D Institute China Xi'an (SRCX)
[2]Northwestern Polytechnical University
{xiao1.tang, min16.yang, penghui.sun, hui01.li}@samsung.com
daiyuchao@nwpu.edu.cn, {f15.zhu, hojae72.lee}@samsung.com

1. Details of reference decoder

In Section 3.2.3 of our main text, we propose a novel reference decoder to merge prior information into the decoding and upsampling network, thus improving rendering quality. In Table 1, we show the architecture of the reference decoder designed based on [1].

The network architecture of our CARN feature encoder is shown in Fig. 1. The CARN feature encoder has global cascading connections represented as the blue arrows in Fig. 1 (a). The outputs of intermediary cascading blocks are cascaded into the higher layers, and finally converge on a basic block (see Fig. 1 (b)). Each cascading block hosts local cascading connections themselves, shown in Fig. 1 (d), and such local cascading is almost identical to a global one, except that the unit blocks are residual blocks (see Fig. 1 (c)).

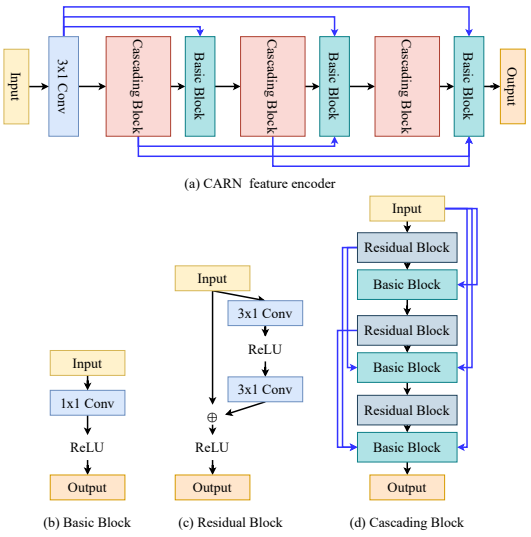

Figure 1. Network structure of CARN feature encoder

*Corresponding author.

Input(dimension)	Module	Layers	Output(dimension)
$\frac{W}{3} \times \frac{H}{3} \times 3$	1st CARN feature encoder	3×1 Conv, 64	$\frac{W}{3} \times \frac{H}{3} \times 64$
$\frac{W}{3} \times \frac{H}{3} \times 64$		Cascading Block	$\frac{W}{3} \times \frac{H}{3} \times 64$
$\frac{W}{3} \times \frac{H}{3} \times 128$		Basic Block	$\frac{W}{3} \times \frac{H}{3} \times 64$
$\frac{W}{3} \times \frac{H}{3} \times 64$		Cascading Block	$\frac{W}{3} \times \frac{H}{3} \times 64$
$\frac{W}{3} \times \frac{H}{3} \times 192$		Basic Block	$\frac{W}{3} \times \frac{H}{3} \times 64$
$\frac{W}{3} \times \frac{H}{3} \times 64$		Cascading Block	$\frac{W}{3} \times \frac{H}{3} \times 64$
$\frac{W}{3} \times \frac{H}{3} \times 256$		Basic Block	$\frac{W}{3} \times \frac{H}{3} \times 64$
$\frac{W}{3} \times \frac{H}{3} \times 64$	Upsamling	3×1 Conv, 64	$\frac{W}{3} \times \frac{H}{3} \times 64$
$\frac{W}{3} \times \frac{H}{3} \times 64$		PixelShuffle, factor 3	$W \times H \times 64$
$W \times H \times 3$	Feature encoder	3×1 Conv, 64	$W \times H \times 64$
$W \times H \times 128$	2nd CARN feature encoder	3×1 Conv, 64	$W \times H \times 64$
$W \times H \times 64$		Cascading Block	$W \times H \times 64$
$W \times H \times 128$		Basic Block	$W \times H \times 64$
$W \times H \times 64$		Cascading Block	$W \times H \times 64$
$W \times H \times 192$		Basic Block	$W \times H \times 64$
$W \times H \times 64$		Cascading Block	$W \times H \times 64$
$W \times H \times 256$		Basic Block	$W \times H \times 64$
$W \times H \times 3$	Feature decoder	3×1 Conv, 3	$W \times H \times 3$

Table 1. Architecture of the reference decoder. "Conv" is 2D convolution, and detailed structure of cascading block and basic block is shown in Fig. 1.

2. Qualitative experiment results on VKITTI2 dataset

In Fig. 2 and Fig. 3, we show additional qualitative results on the same Virtual KITTI-2 (VKITTI2) [2] subsequences as in prior work [4–6]. We demonstrate PaReNeRF's capabilities on both the image reconstruction and novel view synthesis tasks. All the presented results show that, PaReNeRF achieves best results across all splits. As the number of training views is reduced, SUDS fails to represent the scene details and produces noise and ghosting artifacts. While our method generates higher-quality renderings.

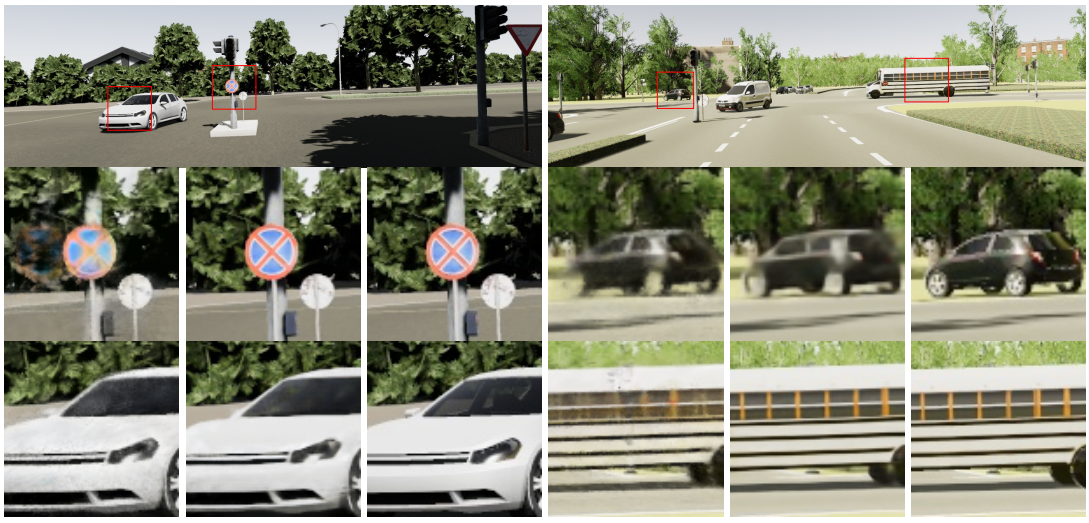

Figure 2. Image reconstruction on VKITTI2 dataset. The first row is a sample of two scenes from the VKITTI2 dataset. The following rows are some local details of the reconstruction results of the SUDS algorithm, our algorithm and the ground truth.

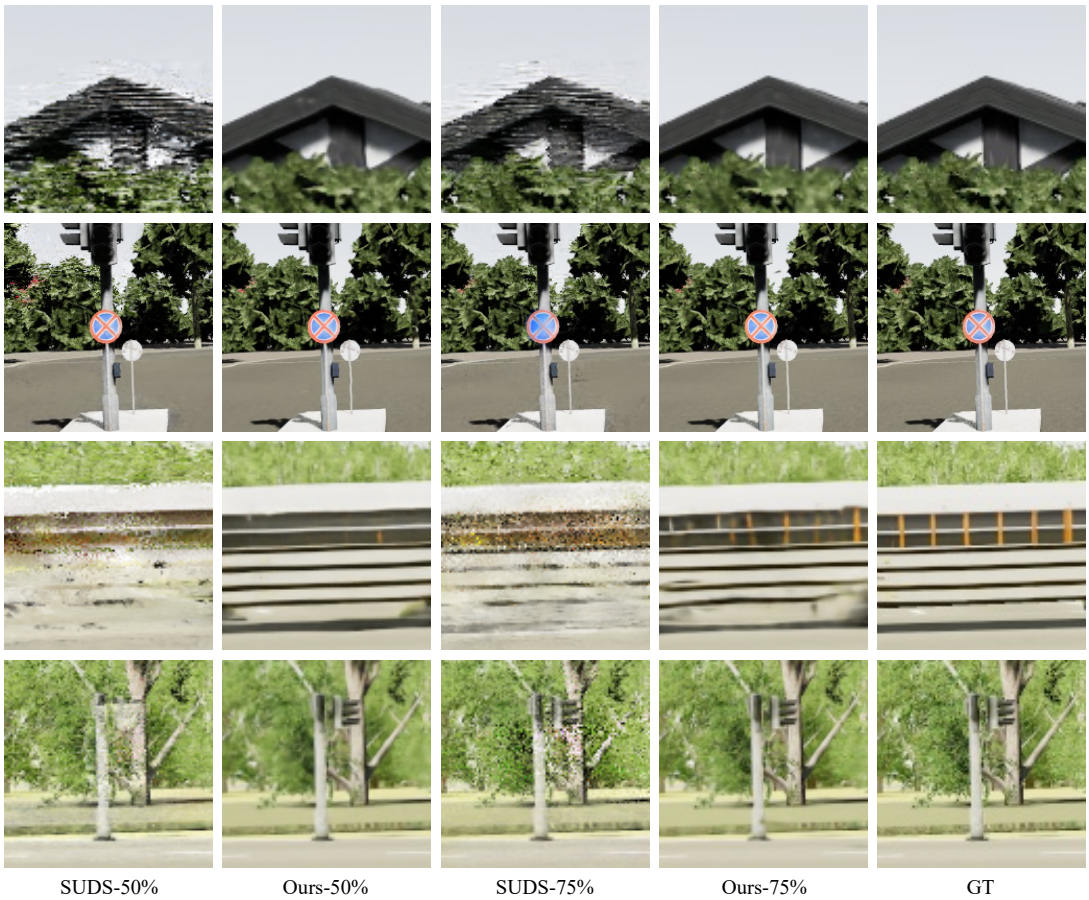

Figure 3. Novel view synthesis on VKITTI2 dataset. We show some local details of the novel view synthesis results trained with different proportion of VKITTI2 subsequence.

3. Ablation Details

In the first experiment, we did not apply any optimization strategy and trained the SUDS model with the same KITTI [3] subsequences and the same experimental setup described in [6]. Then based on this baseline, we analyzed the effects of applying patch sampling, encoder-decoder structure and reference decoder in sequence.

Effect of patch sampling. Compared with the baseline, in this experiment, we only change random ray sampling to patch sampling, and each batch includes 16 patches with a size of 16×16.

Effect of encoder-decoder structure. In this experiment, we sample patches of rays and volume render a 2D feature map on a lower-resolution (414×125), and then leverage a 2D CNN to generate a high-resolution (1242×375) RGB image. To further analyze the effect of introducing reference information later, we set the 2D CNN to the CARN super-resolution network as shown in Fig. 4.

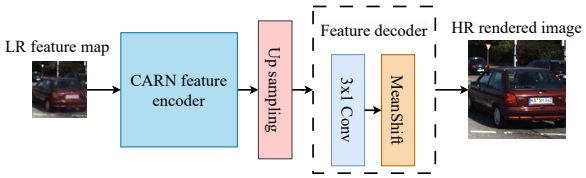

Figure 4. CARN super-resolution network

To optimize this system, we use the losses described in [6] to jointly optimize the radiance fields and the CNN network. We only change $C(r)$ in the L2 photometric loss $\mathcal{L}_c(r) = \left\| C(r) - \hat{C}(r) \right\|^2$ from the rendered RGB image output by the SUDS radiance field to the reconstructed image output by the CNN network.

Additionally, we tested the impact of the number of iterations on training time and image quality.

Effect of reference decoder. This model is our PaReNeRF. Detailed design and experimental setup can be found in Section 3 and 4 of our main text. We also evaluate the performance under different iterations. Our PaReNeRF with full components achieved the best results.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 252–268, 2018. 1

[2] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 1

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 3

[4] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. 1

[5] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021.

[6] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12375–12385, 2023. 1, 3