Supplementary Materials: Neural Rays for Occlusion-aware Image-based Rendering

YUAN LIU, The University of Hong Kong, China
SIDA PENG, Zhejiang University, China
LINGJIE LIU, Max Planck Institute for Informatics, Germany
QIANQIAN WANG, Cornell University, U.S.A
PENG WANG, The University of Hong Kong, China
CHRISTIAN THEOBALT, Max Planck Institute for Informatics, Germany
XIAOWEI ZHOU, Zhejiang University, China
WENPING WANG, Texas A&M University, U.S.A

CCS Concepts: • **Computing methodologies** → **Image-based rendering**.

Additional Key Words and Phrases: novel view synthesis, neural scene representation, image-based rendering

ACM Reference Format:
Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. 2021. Supplementary Materials: Neural Rays for Occlusion-aware Image-based Rendering. *ACM Trans. Graph.* 1, 1 (July 2021), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 NETWORK ARCHITECTURE

Initialization network. The initialization network takes the estimated depth map and the image as input and outputs the initialized $\boldsymbol{F}^{Init}$ for this view. The architecture is shown in Fig. 1. Except for the estimated depth map and the image, we project the points at the depth of current view to other neighboring view and compute the mean and variance of color differences and depth differences, called *consistency feats* in the figure, as the input to the initialization network. All convolution layers use 3×3 kernel and stride 1. "Conv-ReLU(d_{in},d_{out})" means the input channel number is d_{in} and output channel number is d_{out}. The "Residual-IN(d)" block contains layers of "InstanceNorm-ReLU-Conv(d,d)-InstanceNorm-ReLU-Conv(d,d)" and there is a residual connection between the input and the output of this block. "*3" means the block is repeated 3 times.

Convolutional neural network M. The CNN M is also illustrated in Fig. 1, which concatenates the image with the F^{Init} and processes them by a convolution layer and two Residual-IN blocks. The size of $\boldsymbol{F}^{Init}$ is actually 1/4 of the input image size, which is 200×200 for the NeRF synthetic dataset and 200×150 for the DTU dataset and the LLFF dataset. The input image of CNN M is downsampled to match the size of $\boldsymbol{F}^{Init}$.

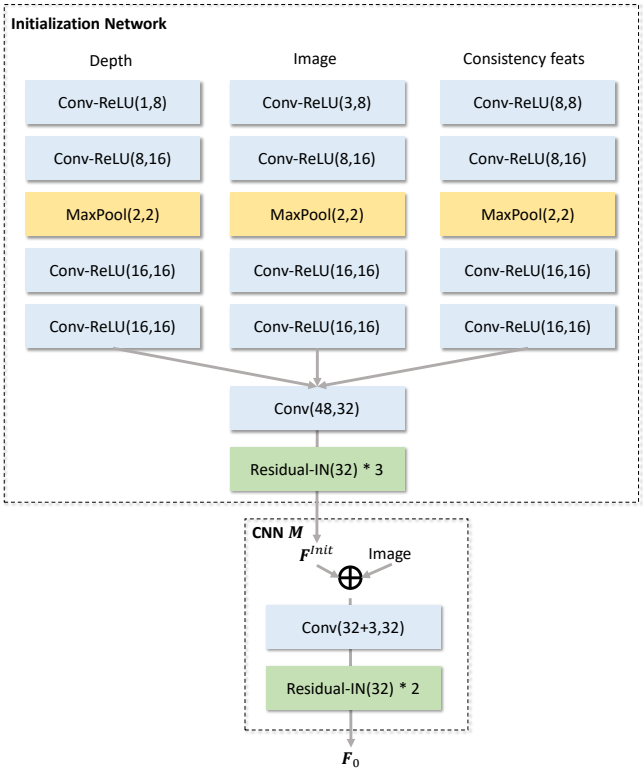

Fig. 1. Architecture of the initialization network and the convolutional neural network M.

Distribution decoder ϕ. The architecture is illustrated in Fig. 2. Note we apply SoftPlus before the mean and variance decoding because both of them are positive. The mean is also positive because the ray can only hit a surface in front of the camera. As for the probability α, we apply a sigmoid before it to scale it to $(0, 1)$.

Rendering networks. The rendering network is the same as IBRNet [Wang et al. 2021], except that we add some inputs to the

Authors' addresses: Yuan Liu, The University of Hong Kong, China; Sida Peng, Zhejiang University, China; Lingjie Liu, Max Planck Institute for Informatics, Germany; Qianqian Wang, Cornell University, U.S.A; Peng Wang, The University of Hong Kong, China; Christian Theobalt, Max Planck Institute for Informatics, Germany; Xiaowei Zhou, Zhejiang University, China; Wenping Wang, Texas A&M University, U.S.A.

0730-0301/2021/7-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

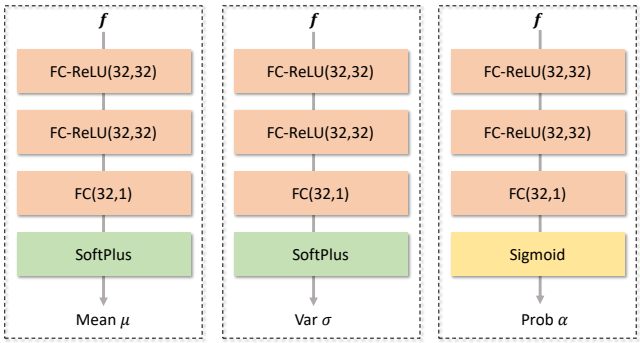

Fig. 2. Architecture of the distribution decoder.

IBRNet. We first apply a fully connected layer on the feature vector $\boldsymbol{f}$, the hitting probabilities and the visibilities to get a feature of dimension 32. This feature is added with the original features from the ResNet34 in IBRNet as the input to the MLP and transformer of IBRNet. For the detail of the subsequent design of MLP and transformer, we recommend the reader to refer to the Figure 2 in IBRNet [Wang et al. 2021].

High resolution model. To produce high resolution rendering, we adopt another smaller network due to the limited GPU memory. The high resolution model applies a ResNet-34 on the concatenation of depth, image and consistency features to produce a F^{Init} at the 1/16 of the original image size. The CNN M applies another ResNet-34 on the image to produce a feature map as the same size with F^{Init}, which is concatenated with the F^{Init} and subsequently processed by two Residual-IN layers. The subsequent distribution decoder and IBRNet structure of the high-resolution version are the same as the low-resolution version.

2 IMPLEMENTATION DETAIL

The initial learning rate is 1e-3 for the scene-specific setting and the finetuning setting and 2e-4 for the generalization setting. The learning rate decays to its half every 20k steps in the scene-specific setting and the finetuning setting while it decays to its half every 50k steps in the generalization setting. The batch size for training is 512.

2.0.1 Normalization of depth maps. For every input view and query view, we have a depth range consisting of near plane depth d_{near} and far plane depth d_{far}. Given a depth, we normalize it by

$$d_{norm} = (-\frac{1}{d} + \frac{1}{d_{near}})/(\frac{1}{d_{near}} - \frac{1}{d_{far}}), \tag{1}$$

where d_{norm} is the normalized version of d. We apply such depth range normalization because the disparity is proportional to $1/d$ and such a normalization will map the near plane depth to 0 and the far plane depth to 1.

2.0.2 Sampling strategy. In the default setting with resolution of 400×400, we do not use the hierarchical sampling [Mildenhall et al. 2020] but directly uniformly sample 128 points along the query ray on the normalized depth in Sec. 2.0.1. When rendering for the high-resolution images of 800×800, we use the hierarchical sampling, in which coarse sampling uniformly samples 64 points while fine sampling has 64 points as done in [Wang et al. 2021].

3 QUALITATIVE RESULTS ON THE DTU DATASET

We provide qualitative results on the DTU dataset in Fig. 3.

4 QUALITATIVE RESULTS ON THE FORWARD-FACING DATASET

We provide qualitative results on the forward-facing LLFF dataset in Fig. 4.

5 HIGH-RESOLUTION RESULTS ON THE NERF SYNTHETIC DATASET

We provide rendering results on the NeRF synthetic dataset with the resolution of 800×800 in Fig. 5.

REFERENCES

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. 2021. IBRNet: Learning Multi-View Image-Based Rendering. *arXiv preprint arXiv:2102.13090* (2021).

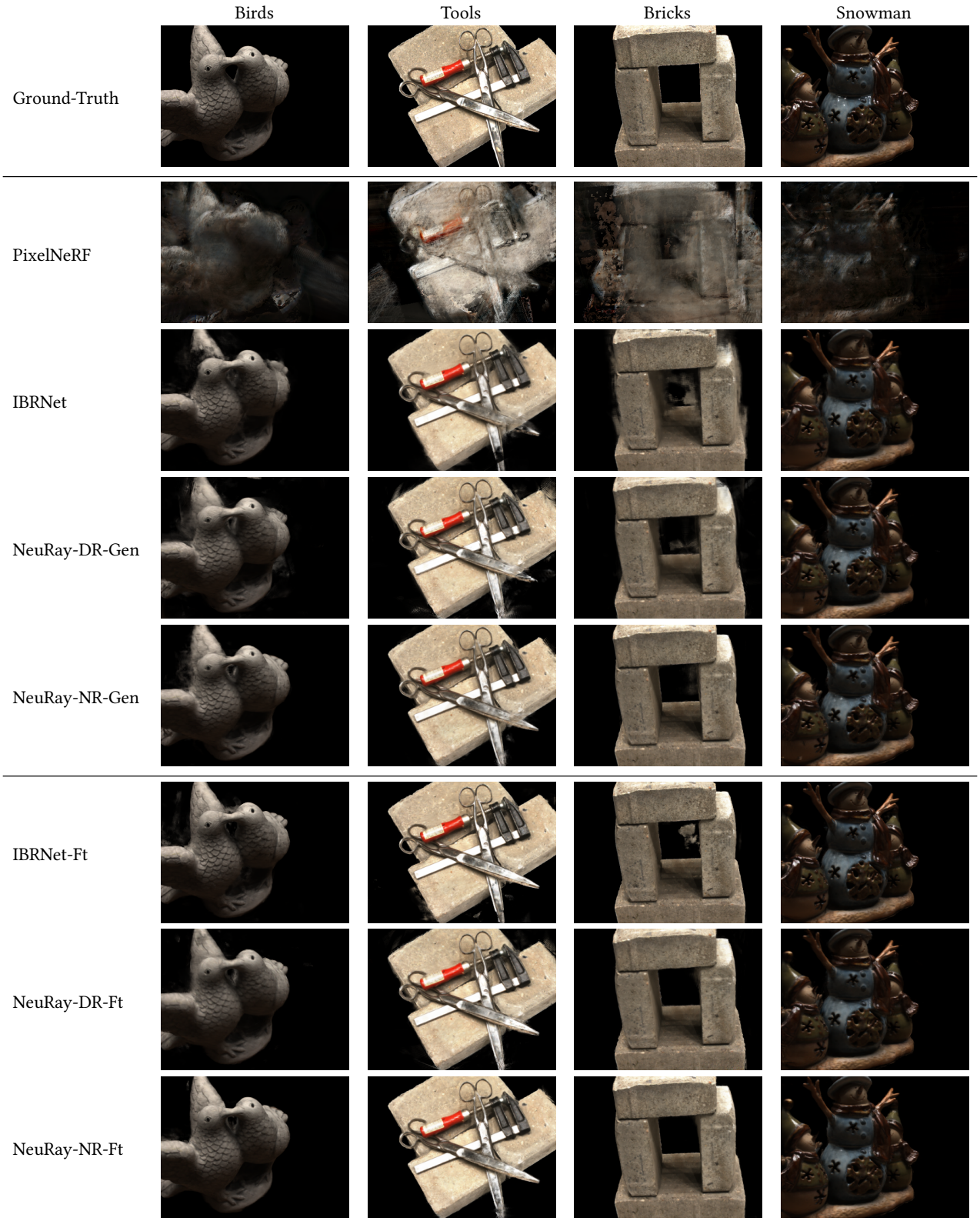

Fig. 3. Qualitative results on the DTU dataset.

Fig. 4. Qualitative results on the forward-facing dataset.

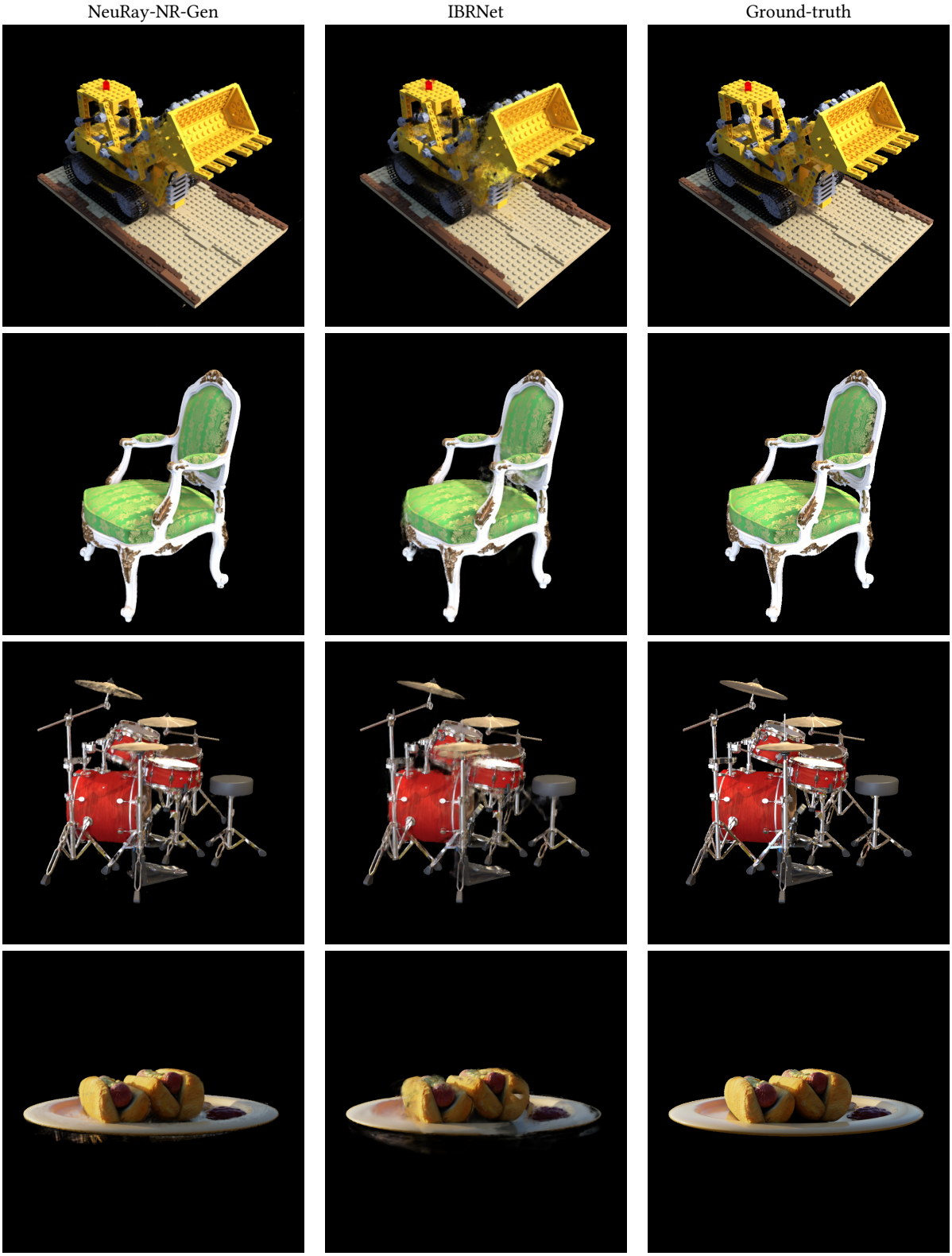

Fig. 5. High resolution rendering results in the generalization setting.