MSI-NeRF: Linking Omni-Depth with View Synthesis through Multi-Sphere Image aided Generalizable Neural Radiance Field

Supplementary Material

Dongyu Yan Guanyu Huang Fengyu Quan Haoyao Chen

Harbin Institute of Technology (ShenZhen)

1. Additional Implementation Details

Since MatryODShka [1] has different input and output format to our method, to achieve fair comparison, we re-implement MatryODShka and match its data format with ours. MatryODShka uses two omnidirectional stereo (ODS) images as input and sphere sweep the images as a sphere sweep volume (SSV). We find that four fisheye images can provide larger FoV than the original two ODS images, so we directly use fisheye images to construct the SSV. For the network MatryODShka used to process SSV, we use the same 3D CNN structure as our MSI decoder by just changing the input and output channel number. We modify the output of the network to generate a blending weight using soft max layer. Alpha blending is then conducted and the final color and depth images can be rendered. Compared with OmniMVS [5], our method additionally requires neural rendering to achieve image and depth synthesis, which takes more memory usage. To this end, we designed a shallower and narrower network structure as the backbone of our method. The OmniMVS used for comparison is also reimplemented using the same backbone for fairness.

For camera calibration in the real-world experiments, we use the Scaramuzza's camera model [4] for intrinsics and Kalibr [2] for extrinsics. The reason for the unsatisfactory results in real-world scenes is that although the selected camera claim to have a 220° FoV, pixels that exceed 180° only account for a small portion of the total pixels. This leads to a serious loss of wide-angle information, which affects both calibration and image synthesis.

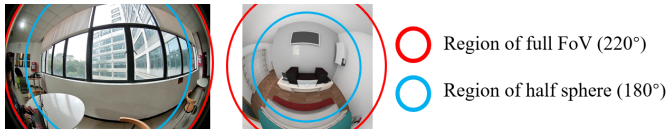

Figure 1. FoV comparison between real-world and virtual camera.

2. Network Architecture

We basically follow the CNN structure proposed in [5] and modify the output format to connect it with our NeRF [3] MLP. The CNN network channel is declined to save GPU memory for the additional MLP and neural rendering. The input of the network is four fisheye images. We use the residual convolution blocks for the image feature extraction, and the dilated convolution for the larger receptive field. The output feature map size is half of the input image. We use a shallow MLP that takes coordinate, direction, projected color and interpolated feature vector as input. The output color and occupancy are then rendered into color and depth map.

3. Mechanisms

Mechanism of Sphere Sweeping: During 3D MSI reconstruction process, our method creates a cost volume by warping the feature maps using intrinsics and extrinsics. Therefore, the network are trained to match the features rather than remembering the camera parameters, and the camera arrangement is not limited to the configuration in the paper. The best configuration is the one with the most camera overlap.

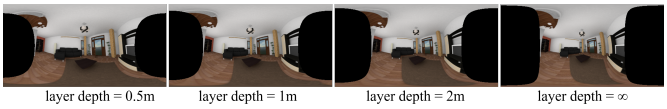

Figure 2. Visualization of projected MSI layer of different depth.

Mechanism of Unsupervised Training: In Fig. 7 in the original paper, we show the results of our method with only color loss. We can see that the depth information can be learned (despite noisy and blurry) from the intrinsic parallax between cameras. Although the input and target become

identical, the strong inductive bias in the MSI and spatial neural rendering process prevent the network from overfitting. Our depth loss can provide additional prior guidance for faster convergence and higher quality.

4. Comparisons

Compare to MatryODShka: Different from our direct raw fisheye input, MatryODShka uses an ODS format as input, which is still synthesized from fisheye or pinhole cameras. During such pre-processing, the parallax information within the original multi-view is eliminated.

Compare to Optimization-based 360 NeRF Methods: After detailed survey, we can't find suitable 360 NeRF method with the same input format as ours. To compare with the optimization-based NeRF method, we implement NeRF-360-NGP which borrows the concept from Mip-NeRF-360 and Instant-NGP. The results can be seen in Fig. 5 of the original paper. Although EgoNeRF and SOMSI also perform 360 NeRF reconstruction, they take pre-processed panoramic images as input, which is different from our approach (raw fiseye images as input). Such synthesized panoramas have lost the parallax information, which requires camera movement for 3D reconstruction.

5. Derivation Details

We give a detailed derivation of Equation (6) in the main body. Given the ray origin $\mathbf{r}_o = (o_x, o_y, o_z)^T$ and ray direction $\mathbf{r}_d = (d_x, d_y, d_z)^T$, the equation of the ray can be given as:

$$\mathbf{r}(z) = \mathbf{r}_o + z\mathbf{r}_d, \tag{1}$$

where z is the parameter indicating the distance along the ray, the depth. Since the sphere equation can be writen as $x^2 + y^2 + z^2 = d^2$, we substitute $\mathbf{r}(z)$ into the sphere equation:

$$(r_{o_x} + zr_{d_x})^2 + \left(r_{o_y} + zr_{d_y}\right)^2 + (r_{o_z} + zr_{d_z})^2 = d^2. \tag{2}$$

We can simplify it to:

$$r_{o_x}^2 + r_{o_y}^2 + r_{o_z}^2 + 2z\left(r_{o_x}r_{d_x} + r_{o_y}r_{d_y} + r_{o_z}r_{d_z}\right) + z^2 = d^2. \tag{3}$$

From this, we can tell that the problem is to solve a quadratic equation and the coefficients of the quadratic term, linear term, constant term can be represented as:

$$a = \mathbf{r}_d^T\mathbf{r}_d, \quad b = 2\mathbf{r}_d^T\mathbf{r}_o, \quad c_n = \mathbf{r}_o^T\mathbf{r}_o - 1/\left(d^{-1}\right)^2. \tag{4}$$

Then the original formula can be established.

6. Application Prospects

Being able to synthesize novel views in 6DoF and estimate depth map panoramically mean a lot to VR and robotics.

Traditional VR contents are usually shot by image stitching, which is simply a stream of panorama images. This results in a limited 3DoF camera movement which can cause omitted parallax information, low immersion, and VR sickness. Instead, our method retains the parallax information within the multi-camera system and achieves 6DoF rendering. This can bring a whole new experience to VR content without changing the shooting equipment. Further more, different from the existing panoramic cameras that only support three-degree-of-freedom re-movement and editing, the video captured by this method can be moved and edited in the translation direction. This is of great significance for filming and video making.

On the robotics side, being able to produce omnidirectional depth estimation also helps a lot. It bring perception in autonomous robots to a new level. Compared to LiDAR, it is much cheaper, with wider FoV, and denser points per frame. The novel view synthesis ability also makes driving cars or drones under a third person's view possible, which benefits a lot in teleoperation and navigation.

7. Limitations

Limited by the memory size of the GPU device, our network capacity is restricted and the resolution of the generated color and depth map is set to 512×256. Also, due to the limited computation performance, our method can only run at around three frames per second. These limitations makes our method hard to applied in applications with strong real-time requirements.

At the same time, our method also requires accurate extrinsic and intrinsic calibration. These parameters are used in the warping process that turns fisheye projection into MSI representation, which is essential to the final result. In our real-world experiment, the generated color and depth image both suffer from poorly calibrated cameras, which results in blurry and artifacts. The quality of real-world generation will degrade compared to the results from synthetic data.

Due to the usage of MSI, the free roaming of the camera is limited inside the smallest sphere. This way, the setting of the smallest radius can be a trade of between the nearest collision limit and the virtual camera moving range. A bigger inner sphere can lead to a wider moving range, while may cause objects close to camera being cut off. This trade off needs careful tuning when faced with different scenarios.

The artifacts mainly occurs when the camera has large motion (exceed the minimum MSI sphere of 0.5m) and observes heavily occluded parts. Since the input views have no observations on such regions, it means that the network

needs to solve a generation problem. Our method follows the idea of MVS and aims to learn feature extraction and matching, which lack generation ability of unseen area and generates blurry artifacts. We hope that future research can address this problem by introducing the ability of generative models (e.g. inpainting methods).

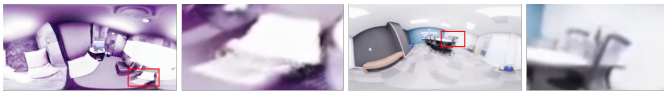

Figure 3. Visualization of selected failure cases.

8. Additional Ablation Studies

We conduct additional ablation studies using regular images and fewer fisheye images settings. The results are shown in the figure below. The ablation study without depth map branch can be reflected in the discussion about unsupervised training.

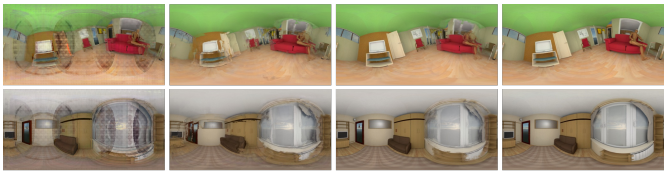

Figure 4. Visualization of more ablation studies.

9. Additional Experiments Result

Here, we show the additional generated depth estimation and novel view synthesis results on the OmniHouse dataset. They are shown in Fig. 5. More visualization results can be seen in the video in the supplementary files.

References

[1] Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. Matryodshka: Real-time 6dof video view synthesis using multi-sphere images. In European Conference on Computer Vision, pages 441–459. Springer, 2020. 1

[2] ETH Zurich Autonomous Systems Lab. Kalibr: Toolbox for multi-camera and imu calibration. https://github.com/ethz-asl/kalibr. 1

[3] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1):99–106, 2021. 1

[4] Davide Scaramuzza, Agostino Martinelli, and Roland Siegwart. A toolbox for easily calibrating omnidirectional cameras. In 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 5695–5701. IEEE, 2006. 1

[5] Changhee Won, Jongbin Ryu, and Jongwoo Lim. Omnimvs: End-to-end learning for omnidirectional stereo matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8987–8996, 2019. 1

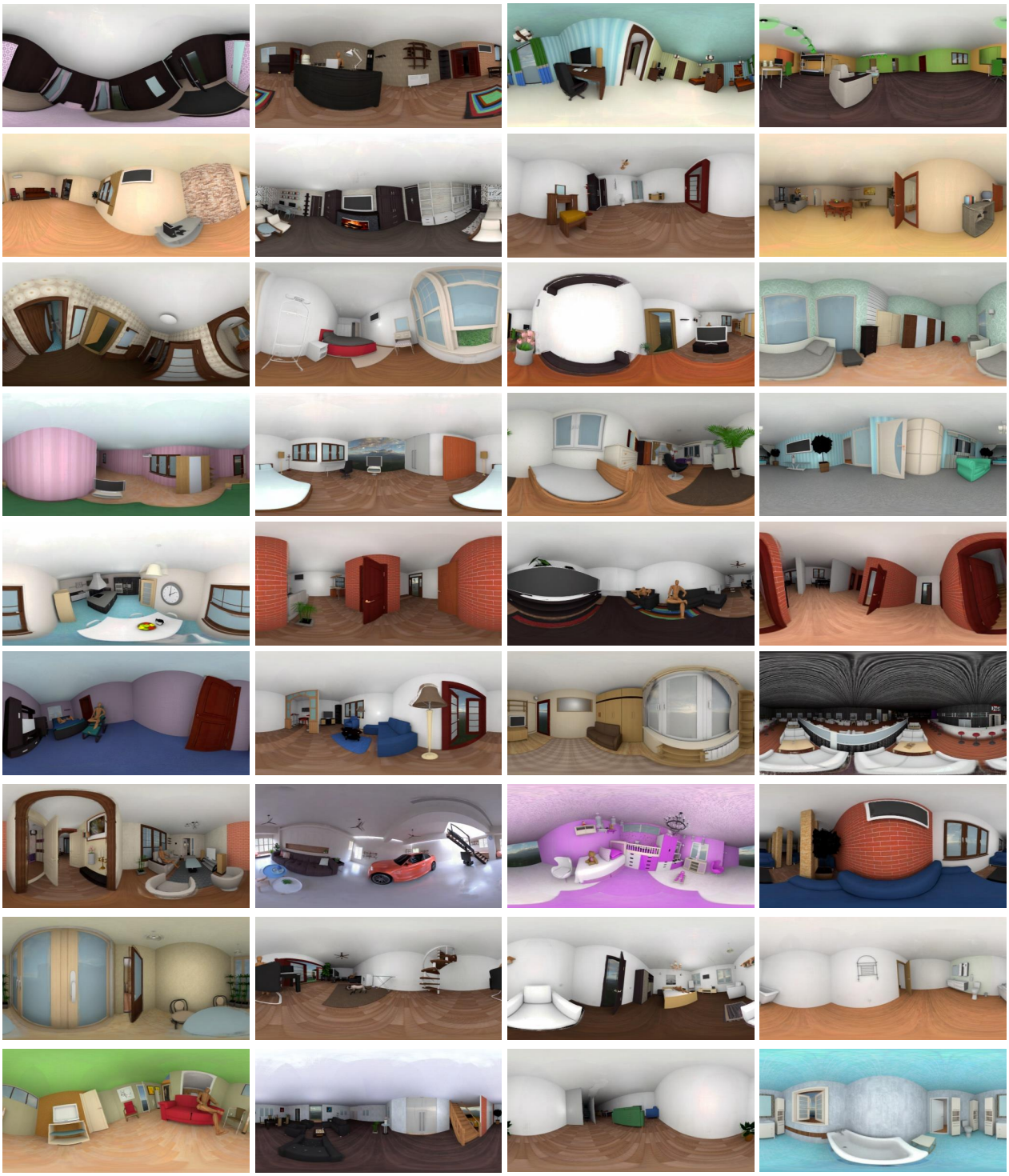

Figure 5. Additional experiment results.