Supplementary Material for Stereo Radiance Fields (SRF): Learning View Synthesis from Sparse Views of Novel Scenes

Julian Chibane[1] Aayush Bansal[2] Verica Lazova[1] Gerard Pons-Moll[1]
[1]University of Tübingen, Max Planck Institute for Informatics, Germany [2]Carnegie Mellon University, USA
{jchibane, vlazova, gpons}@mpi-inf.mpg.de, aayushb@cs.cmu.edu

Abstract

In this supplementary paper we give the implementation details of Stereo Radiance Fields (SRF) in Section 1, define the data split we used in Section 2 and provide a qualitative comparison with COLMAP in Section 3. Code is available for research purposes at our project page https://virtualhumans.mpi-inf.mpg.de/srf/. Please also consider the supplementary video for further results.

1. Implementation Details

1.1. Image encoder

Given an image and a projected point location, we extract RGB color directly in the first stage using bilinear interpolation. We apply a convolutional neuronal network [3] (CNN) with 16 output channels, kernel size of 3, padding 1 and stride 1 followed by batch normalization. From the normalized output we again extract neural features.

The rest of the CNN encoder is a repetition of the following CNN block structure: max-pooling (size 2), two convolutional layers and a batch normalization layer followed by a bilinear feature extraction. We repeat the block six times, always with kernel size of 3, padding 1 and stride 1. The first block has both convolutional layers with 16 output channels, the second 32, third 64, fourth and subsequent 128.

Therefore, the feature encoding $\mathbf{I}_i(\mathbf{p})$, for each view $\mathbf{I}_i$, has dimension $3+16+32+64+128+128+128+128 = 627$.

1.2. Unsupervised Stereo Module

Each possible pair $(\mathbf{I}_i(\mathbf{p}), \mathbf{I}_j(\mathbf{p})) \in \mathbb{R}^{2\times 627}$ with $i, j \in 1, \dots, N, i \neq j$ is stacked into a matrix of dimensions $\mathbb{R}^{(2\cdot S)\times 627\times 1}$, where S is the number of all pairs $S = N^2 - N$. In this form the stereo filters $s_k(\mathbf{I}_i(\mathbf{p}), \mathbf{I}_j(\mathbf{p}))$ can be conveniently represented by convolutional filters $s_k \in \mathbb{R}^{2\times 627\times 1}$ of a CNN. In order to apply the filters once per pair, we use a striding of 2 in the height dimension without any padding. We use $k \in 1, \dots, 128 = K$ filters, by simply defining K as the number of output channel of the CNN. This results in a output matrix $\mathbb{R}^{S\times 1\times K}$ of stereo features.

To aggregate stereo pair information we apply another CNN. The filters aggregate 4 pairs, i.e. each filter has dimensions $\mathbb{R}^{4\times 1\times K}$. We repeat this step once more to create higher order dependencies, in both cases we use a CNN with 128 output channels, padding of 0 and striding of 1. We use max-pooling along height to reduce to a single vector of fixed size (128), independent of the number of views used.

1.3. Decoding

For decoding into density and color we use 4 simple fully connected layers with subsequently 256, 128, 4 output channels, where the last output of 4 channels is interpreted as RGB color (3 channels) and a density value.

1.4. Learning

For learning we use Adam [2] optimizer with betas set to $(0.9, 0.999)$ and a learning rate of $5e-4$. We train all models until validation minimum is reached.

2. Data split

We split the DTU MVS [1] dataset as follows, scans with ID 55, 23, 77, 122, 106, 95, 8, 64, 50, 76 are used for testing, 103, 47, 72, 107, 124 for validation and others for training.

3. Comparison with COLMAP

In Figure 1 we are comparing our fine-tuned results with the classical 3D reconstruction pipeline of COLMAP [5, 4]. We found a trade-off between completeness and precision: COLMAP results are often sharper but are lacking completeness, whereas our 3D reconstruction results are more

Figure 1. **Left and middle**: 3D reconstruction from 10 input images, obtained with COLMAP [5, 4] (left) and our SRF (middle). The COLMAP reconstruction is sharper but lacks complete surface, especially at uniform, textureless regions, see for example the characters of the logo "Shine". 3D reconstruction of SRF can be more prone to noise (see top black spot) but is more complete. **Right**: Sharp novel view synthesised by SRF with full background.

noisy but have better completeness.

Remarkably, Stereo Radiance Fields learned 3D reconstruction as a byproduct of end-to-end training purely from 2D supervision. We also show their main result, a synthesised novel view, in Figure 1 (right). The novel view synthesis combines good aspects of both 3D reconstructions: it shows sharp details and is the most complete, actually rendering the full background.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016. 1

[2] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 1

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[4] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[5] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2