Pacific Graphics (2025)
S.-S. Lin, T. Schneider, and E. Zhang (Editors)

Poster

Self-Supervised Neural Global Illumination for Stereo-Rendering

Ziyang Zhang[1] and Edgar Simo-Serra[1]

[1]Waseda University, Japan

Abstract

We propose a novel neural global illumination baking method for real-time stereoscopic rendering, with applications to virtual reality. Naively, applying neural global illumination to stereoscopic rendering requires running the model per eye, which doubles the computational cost making it infeasible for real-time virtual reality applications. Training a stereoscopic model from scratch is also impractical, as it will require additional path tracing ground truth for both eyes. We overcome these limitations by first training a common neural global illumination baking model using a single eye dataset. We then use self-supervised learning to train a second stereoscopic model using the first model as a teacher model, where we also transfer the weights of the first model to the second model to accelerate the training process. Furthermore, our spatial coherence loss encourages consistency between the rendering for two eyes. Experiments show our method achieves the same quality as the original single-eye model with minimal overhead, enabling real-time performance in virtual reality.

CCS Concepts

• ***Computing methodologies*** → ***Rendering;*** *Neural networks;* • ***Human-centered computing*** → ***Virtual reality;***

1. Introduction

Virtual reality (VR) applications require stereo rendering for two eyes to perceive depth, which can significantly increase the computational cost of the rendering process compared to traditional tasks. Although numerous works have been proposed to reduce overhead, such as multi-view rendering with a single GPU pass [HZP07], a recent rendering technique, neural global illumination (GI) baking, has not been studied in the context of stereo rendering. Neural GI baking is a technique that utilizes neural networks to approximate the GI effects learned from path tracing images, which has been shown to be effective in achieving real-time performance with high-quality effects [DPD22, ZSS24].

Naively, applying neural GI to stereo rendering requires running the model per eye. This essentially doubles the computational cost, making it infeasible for real-time VR applications. It is also impractical to train a stereoscopic model from scratch, as it requires generating expensive path tracing ground truth for both eyes.

In this work, we propose a novel neural GI baking method for real-time stereo rendering. We take a trained neural GI model as a teacher model, and perform self-supervised learning to train a second stereoscopic model. The stereoscopic model is initialized with part of the weights from the teacher model in order to accelerate the training process. During the training, we also apply a spatial coherence loss to the training loss function to improve the rendering coherence between the two eyes. Our contributions are summarized as follows:

- A training method can convert arbitrary single-eye neural rendering models to stereoscopic models without requiring additional path tracing ground truth for the second eye.
- A transfer learning method that utilizes the weights of a single-eye model to accelerate the training of a stereoscopic model.
- A spatial coherence loss that encourages consistency between the rendering of two eyes.

2. Method

Our method overview is shown in Fig. 1. As a common approach, we use U-Net [RFB15] as the architecture of the single-eye model. The model takes G-buffer features and simple approximated illumination from a real-time rasterization renderer as input and outputs the path-tracing quality GI image. The G-buffer features include position, normal, albedo, and depth. The approximate illumination is by direct illumination with light maps.

After fully training the single-eye model, we used it as a teacher model to generate pseudo ground truth for the stereoscopic dataset, with interpupillary distances varying from 55 to 70 mm. The stereoscopic model also uses U-Net as the architecture with two output heads at the end of the decoder. It uses the same inputs as the single-eye model (one set per eye). The weights of the single-eye model are copied to the stereoscopic model, except for the first layer and the output head of the right eye.

During the training, we also apply an additional spatial coherence loss to the reconstruction loss:

$$\mathcal{L}_{\text{coherence}} = \frac{1}{N}\sum_{i=1}^{N} \left\| \hat{I}_{\text{left}}(x_i) - \mathcal{W}(\hat{I}_{\text{right}}, m)(x_i) \right\|_2^2 \qquad (1)$$

DOI: 10.2312/pg.20251305

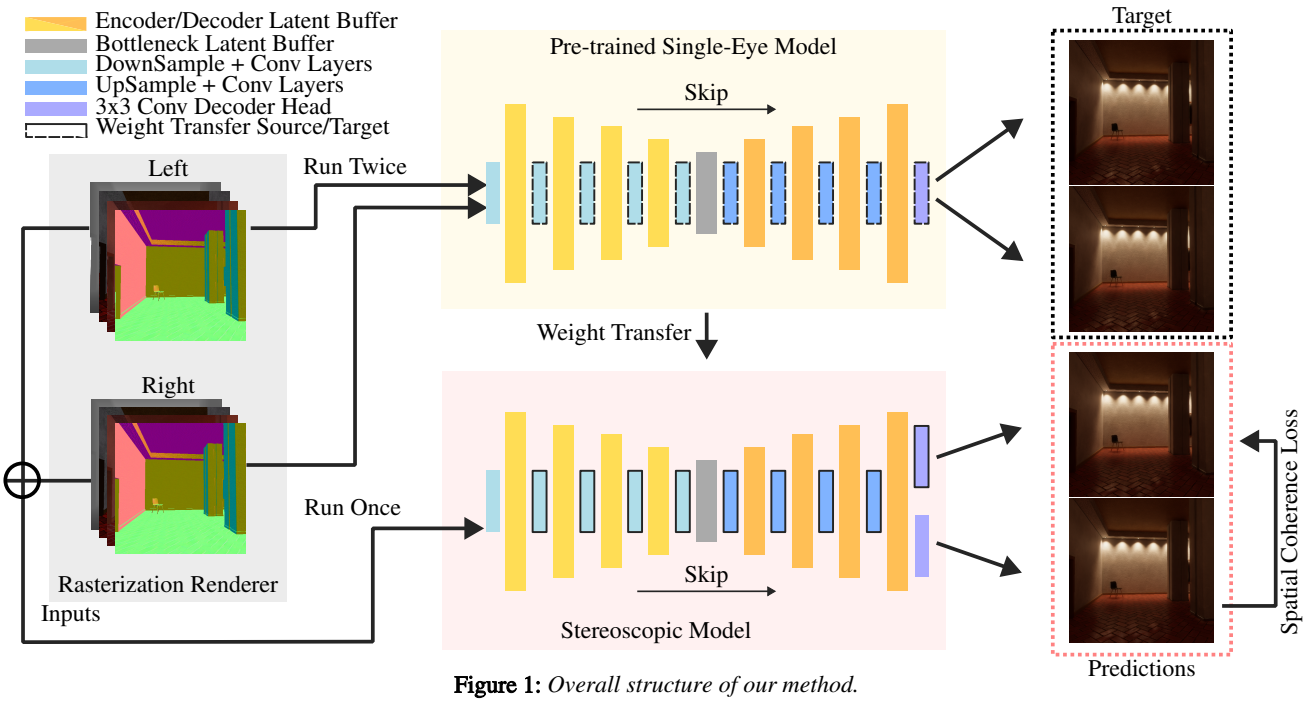

Figure 1: *Overall structure of our method.*

where $\mathcal{W}(\hat{I}_{\text{right}}, m)$ denotes the warping of the right image using motion vectors m.

3. Results and Conclusion

Our preliminary results are shown in Fig. 2. We evaluate our model in a dynamic scene under two challenging lighting conditions: daytime and nighttime. We also tested the effectiveness of the spatial coherence loss. The mean squared error between the two eyes of models trained with the spatial coherence loss term is 0.00538, compared to 0.00804 for models without this term, demonstrating the effectiveness of the spatial coherence loss.

At a resolution of 512×512 on an NVIDIA RTX 2080 Ti, our method achieves an average inference time of just 9.13 ms. By contrast, the original model must be executed twice with a total time of 17.20. Our approach nearly halves the total computation time.

For future work, we consider improving the reconstruction quality from the teacher model and the training speed of the stereoscopic model.

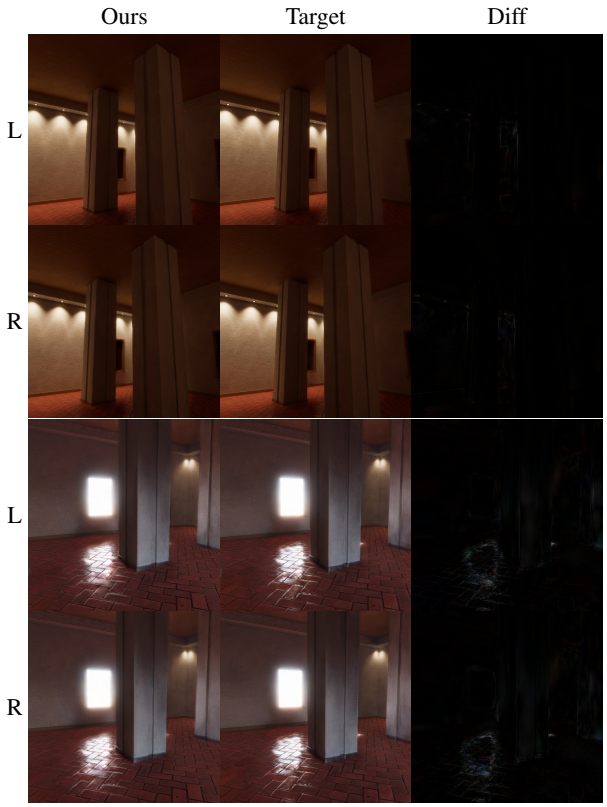

Figure 2: *Results of our method: we show results under two lighting conditions (top: daytime, bottom: nighttime). Our method achieves nearly identical quality to the original model.*

References

[DPD22] DIOLATZIS S., PHILIP J., DRETTAKIS G.: Active exploration for neural global illumination of variable scenes. *ACM Transactions on Graphics (TOG) 41*, 5 (2022), 1–18. 1

[HZP07] HÜBNER T., ZHANG Y., PAJAROLA R.: Single-pass multi-view rendering. *IADIS International Journal on Computer Science and Information Systems 2*, 2 (2007), 122–140. 1

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (2015), Springer, pp. 234–241. 1

[ZSS24] ZHANG Z., SIMO-SERRA E.: Crystalnet: Texture-aware neural refraction baking for global illumination. *Computer Graphics Forum 43*, 7 (2024), e15227. 1