Cross-Ray Neural Radiance Fields in the Wild via Monocular Depth Guidance

Group 29
110550093 蔡師睿 110550037 劉又輔 110550110 林書愷

1 Motivation & Goal

1.1 Motivation

Neural Radiance Fields[1] (NeRF) has been a revolutionary approach for novel-view synthesis. To achieve the best performance, NeRF requires input from static scene images. However, in practice we typically have only unconstrained real-world images collected. These images often contains two types of dynamic changes, first is dynamic changes in appearance due to different capturing time and camera settings. The second type involves transient objects such as humans and cars, leading to occlusion and ghosting artifacts. We further observed that transient occluders are often found in front of buildings, suggesting that transient object problem might be a task of particular interest. To address this issue, we hypothesize that depth estimation can be used to guide NeRF in its early stages.

1.2 Goal

Simply put, input images are randomly taken from internet which may include those with transient occluders. Our goal is to enhance the performance of NeRF within effectively reconstruct from unconstrained image collections.

2 Related Works

2.1 Assumption Study

With our further study, we discovered that depth map estimation indeed can improve segmentation in certain ways, as discussed in the paper: Three Ways to Improve Semantic Segmentation with Self-Supervised Depth Estimation[2]. The paper explores three methods by which self-supervised depth estimation (SDE) can enhance semantic segmentation in both semi-supervised and fully-supervised settings. One of the trials involves SDE feature representation can be transferred to semantic

segmentation, by means of SDE pretraining and joint learning of segmentation and depth.

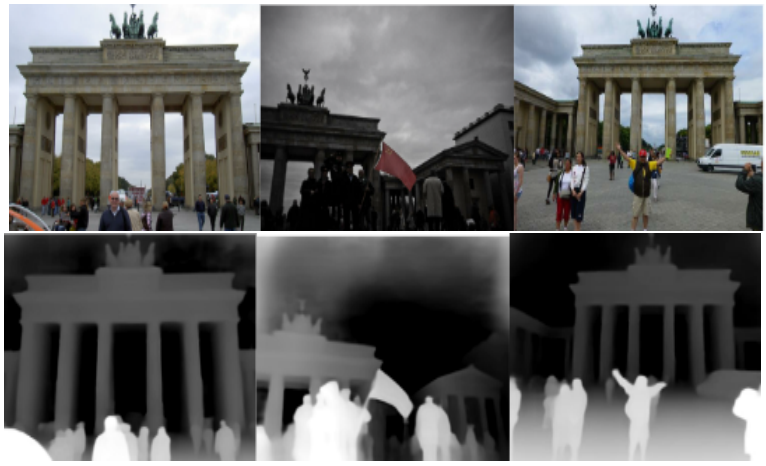

Figure 1: Visualization of unconstrained images and their depth map

These are some examples of our input images and their depth maps. As you can see, the transient objects in front of the gate block the gate, and in the depth map, it clearly show that the transient and target objects are already well separated at this stage.

2.2 Cross-Ray Neural Radiance Fields for Novel-view Synthesis from Unconstrained Image Collections[3]

CR-NeRF is the state-of-the-art method in the series of NeRF in the Wild[4] works. Instead of relying on single-ray information, the interaction of multiple rays used in CR-NeRF more closely mimics the functioning of human vision and is better equipped to handle varying appearances. Another innovative aspect introduced by CR-NeRF is a novel segmentation technique for processing transient objects. This allows CR-NeRF to generate visibility maps that effectively blocking transient objects.

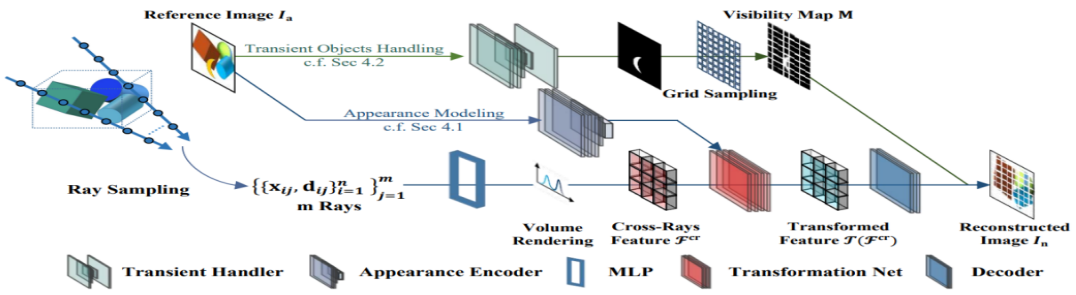

Figure 2: Pipeline of CR-NeRF

CR-NeRF has two main components: the transient handler and the ray handler. For the ray handler, given the position x and direction d of multiple rays, it generates

a cross-ray feature that accumulates multi-view information from the scene. An appearance encoder learns the reference image's appearance features, which are fused using a transform network. The decoder then synthesizes the colors of multiple pixels in the reconstructed image. To eliminate transient objects, the transient handler generates a visibility map and uses a grid sampling strategy to match this map with the rays during training.

2.3 Zoedepth: Zero-shot transfer by combining relative and metric depth[5]

We select Zoedepth for depth estimation due to its excellent generalization and metric scale retention, using MiDas[6] as its training strategy. This is crucial for our diverse dataset, which varies in light sources, camera angles, and image scales. Zoedepth's generalization ensures consistent output across different datasets with minimal fine-tuning.

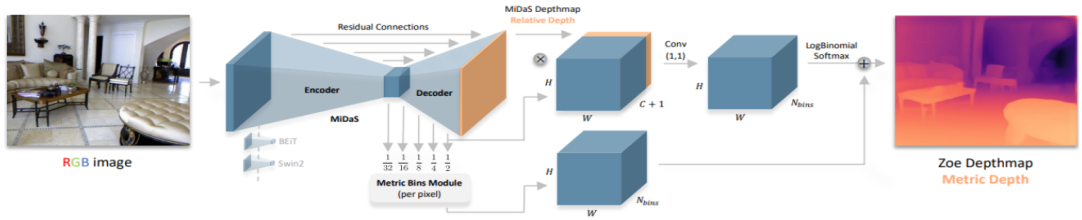

Figure 3: Pipeline of zoedepth

Giving a brief introduction to Zoedepth pipeline. The input is RGB images and is fed into the MiDaS depth estimation framework. The MiDaS framework is basically a encoder-decoder structure with the bottleneck storing depth information per pixel and then succeeding four hierarchy levels of the MiDaS decoder are hooked into the metric bins module. The metric bins module computes the per-pixel depth bin centers that are linearly combined to output the metric depth.

3 Method

In our method, we mainly focus on leveraging the performance of segmentation with the assistance of predicted monocular depth. For the depth estimation model, we have chosen the pretrained Zoedepth model to estimate the depth map for all unconstrained images and the reason is mentioned above. With the estimated depth maps, we then concatenate them to original input image and feed as a new inputs to segmentation network and follow its training pipeline. With our method, preciser visibility maps will be generated and perform better transfer appearances of target.

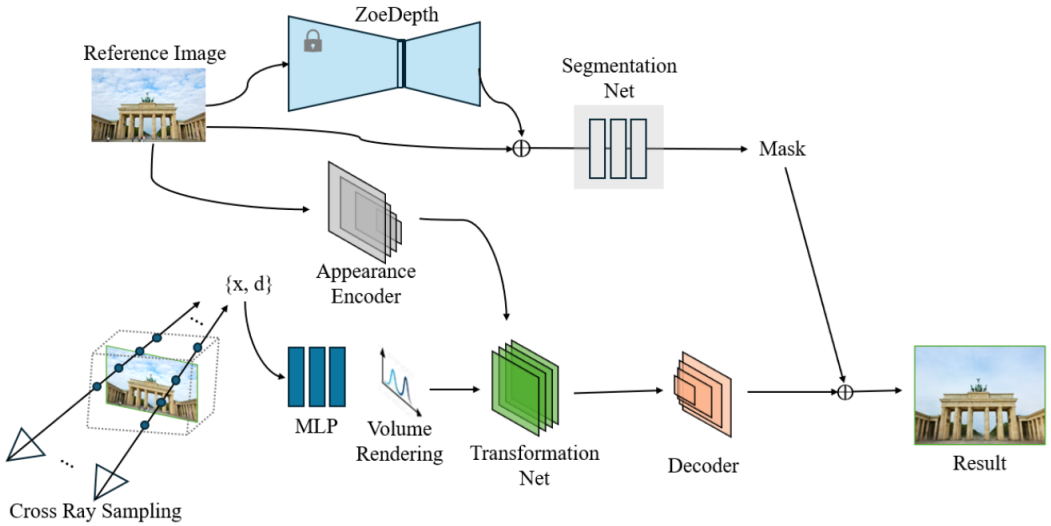

Figure 4: Pipeline of our method

For a fair comparison, we didn't modify the network architecture of the transformation and segmentation models. Moreover, we follow CR-NeRF's default training hyper-parameter settings to determine whether our depth guidance can improve the performance of transient handling.

4 Experiments

4.1 Qualitative Results

Figure 5: Qualitative results between ours and ground-truth

4.2 Quantitative Comparison

	Brandenburg Gate			Sacre Coeur			Trevi Fountain		
	PSNR (↑)	SSIM (↑)	LIPIS (↓)	PSNR (↑)	SSIM (↑)	LIPIS (↓)	PSNR (↑)	SSIM (↑)	LIPIS (↓)
Ha-NeRF[7]	24.58	0.8829	0.0927	20.36	0.7947	0.1317	20.27	0.7270	0.1628
CR-NeRF	26.86	0.9069	0.0733	22.03	0.8369	0.1060	22.02	0.7488	0.1354
Our	27.35	0.8926	0.0964	22.62	0.8346	0.1025	22.13	0.7642	0.1302

Table 1: Quantitative comparison between different methods. The red and the underlined numbers indicate the best and second-best results, respectively.

4.3 Segmentation Mask Visualization

Figure 6: Segmentation mask visualization

4.4 Appearance Modification

Figure 7: Appearance modification in the Trevi Fountain scene

5 Conclusion & Future Works

We propose a simple yet effective method to improve the handling of transient objects. By incorporating monocular depth maps, we enhance the segmentation task, thus boosting CR-NeRF's overall performance. While our method is logically sound, we think that it needs further refinement for robustness, including better merging of depth maps with RGB images and a more suitable segmentation network, because we only directly concatenate the depth maps and input images in this work.

6 Libraries and Open Sources used

- CR-NeRF : https://github.com/YifYang993/CR-NeRF-PyTorch
- ZoeDepth : https://github.com/isl-org/ZoeDepth

7 Contribution

110550093 蔡師睿: paper survey, code, run experiments, report
110550037 劉又輔: proposal and presentation slides, reorganize our code, report
110550110 林書愷: proposal video, code, run experiments, report

The code and data can be found at *https://github.com/Shukkai/CR-NeRF-via-Depth-Guidance*.

References

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[2] L. Hoyer, D. Dai, Y. Chen, A. Koring, S. Saha, and L. Van Gool, "Three ways to improve semantic segmentation with self-supervised depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 130–11 140.

[3] Y. Yang, S. Zhang, Z. Huang, Y. Zhang, and M. Tan, "Cross-ray neural radiance fields for novel-view synthesis from unconstrained image collections," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 901–15 911.

[4] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections," in *CVPR*, 2021.

[5] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.

[6] R. Birkl, D. Wofk, and M. Müller, "Midas v3.1 – a model zoo for robust monocular relative depth estimation," *arXiv preprint arXiv:2307.14460*, 2023.

[7] X. Chen, Q. Zhang, X. Li, *et al.*, "Hallucinated neural radiance fields in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 943–12 952.