

# GoLF-NRT: Integrating Global Context and Local Geometry for Few-Shot View Synthesis

## Supplementary Material

In this supplementary material, we provide additional details to supplement the main manuscript, encompassing detailed analysis of per-scene optimization experiments and further visual comparisons with existing works. Alongside these, we have also incorporated a demonstration example in GIF format within the supplementary materials, offering a dynamic illustration of our methodology.

### 1. Supplementary Notes on Experiments

#### 1.1. Visualization Results under a Single Input

Synthesizing novel views from a single input image represents a highly challenging task, as it constitutes the most extreme scenario among few-shot input conditions. The absence of multi-view information makes obtaining accurate depth information significantly more difficult. As shown in Figure 1, inaccuracies in depth estimation can cause misalignments within the synthesized scene, leading to noticeable artifacts. Notably, methods such as GNT [4], which rely heavily on multi-view feature matching, often produce severely distorted rendered images under such conditions. In contrast, our approach introduces global features to enhance the model’s scene comprehension, maintain semantic consistency across objects, and significantly improve the quality and realism of the synthesized views.

#### 1.2. Results on Reflective/Transparent Surfaces

Table 1 summarizes the validation metrics for various methods [3, 4] across five scenarios in the shiny dataset [5]. Our method consistently achieves the best performance across all scenarios, with particularly significant improvements observed in the CD and Lab scenarios. We attribute this to the increased presence of reflective and transparent surfaces in these scenarios, which pose greater challenges for accurately capturing lighting and geometric information. By incorporating global features, our approach enhances the model’s adaptability to such complexities, enabling a more accurate interpretation of variations in lighting and surface textures. As a result, our method generates more realistic and visually coherent view synthesis outputs, even in challenging conditions.

To further validate our approach, we visualized the rendering results for two scenarios, Materials and Ship, from the Blender dataset [2], which are representative scenarios featuring reflective surfaces. As shown in Figure 2, the images rendered by GNT [4] exhibit significant loss of geometric detail, leading to blurred boundaries between ob-

jects. In contrast, our method produces renderings that more accurately capture the overall appearance and structure of the objects, showcasing superior geometric fidelity and visual clarity.

### 2. Per-Scene Optimization Experiments

#### 2.1. Implementation Details.

We conducted fine-tune experiments on each of the eight scenes included in the LLFF dataset: fern, trex, orchids, flowers, etc. To accelerate the training speed, we reduced the number of sampling points from 128 coarse samples and 64 fine samples to 64 coarse samples and 64 fine samples. As neither CaesarNeRF [6] nor EVE-NeRF [3] have open-sourced their experiments related to per-scene optimization experiments, our work is temporarily only compared with GNT [4], our baseline. For each scene, the training process was iterated for 60,000 times.

#### 2.2. Analysis.

The detailed experimental results are presented in Table 2. By comparing GoLF with state-of-the-art methods such as GNT [4] and several others [1, 2, 5], GoLF-NRT has achieved favorable performance in most of the evaluated metrics. In terms of PSNR metrics, we have achieved optimal values in more than half of the scenarios, and at least sub-optimal results for the remaining scenarios. Additionally, on the SSIM metric, we have surpassed the majority of existing methods, while we achieved the best experimental results in terms of the LPIPS metric across every scene. Furthermore, compared to GNT [4], our method has consistently yielded superior results across all scenarios and metrics.

### 3. Visualizations

In this section, We present two different variations, frame-wise results as attached to this document, and the video results in the form of GIF files, which are included in the supplementary material.

#### 3.1. Framewise Results.

We provide additional examples of per-scene optimization on LLFF, specifically comparing our proposed GoLF-NRT method to GNT [4] in terms of both many-shot (i.e., 10 input views) and few-shot (i.e., 1 input views) settings. The results of this comparison are presented in Figure 3 and Figure 4, respectively.

Metric	Method	cd	crest	food	giants	lab
PSNR $\uparrow$	GNT [4]	32.39	21.79	<u>25.68</u>	27.95	<u>25.01</u>
	EVE-NeRF [3]	<u>33.73</u>	<u>23.59</u>	25.61	<u>29.24</u>	24.87
	GoLF-NRT	<b>34.10</b>	<b>23.72</b>	<b>26.06</b>	<b>29.31</b>	<b>25.85</b>
SSIM $\uparrow$	GNT [4]	0.959	0.692	0.853	0.908	0.850
	EVE-NeRF [3]	<u>0.971</u>	<b>0.777</b>	<u>0.856</u>	<u>0.924</u>	<u>0.888</u>
	GoLF-NRT	<b>0.979</b>	<u>0.768</u>	<b>0.864</b>	<b>0.928</b>	<b>0.895</b>
LPIPS $\downarrow$	GNT [4]	0.056	0.237	0.128	0.093	<u>0.144</u>
	EVE-NeRF [3]	<u>0.038</u>	<b>0.169</b>	<u>0.127</u>	<u>0.085</u>	0.146
	GoLF-NRT	<b>0.037</b>	<u>0.189</u>	<b>0.118</b>	<b>0.083</b>	<b>0.136</b>

Table 1. Per-scene quantitative comparison of state-of-the-art view synthesis methods on Shiny with 10 input views. Best results are in bold, second-best are underlined.

Metric	Method	trex	fern	flower	leaves	room	fortress	horns	orchids
PSNR $\uparrow$	LLFF [1]	27.48	<b>28.72</b>	20.72	21.13	24.54	21.79	23.22	18.52
	NeRF [2]	26.80	25.17	27.40	20.92	32.70	31.16	27.45	20.36
	NeX [5]	<u>28.73</u>	25.63	<b>28.90</b>	21.96	32.32	31.67	28.46	20.42
	GNT [4]	28.15	24.31	27.32	<u>22.57</u>	<u>32.96</u>	<u>32.28</u>	<u>29.62</u>	<u>20.67</u>
	GoLF-NRT	<b>28.75</b>	<u>25.65</u>	<u>28.77</u>	<b>23.18</b>	<b>33.09</b>	<b>32.47</b>	<b>29.66</b>	<u>21.27</u>
SSIM $\uparrow$	LLFF [1]	0.857	0.753	0.844	0.697	0.932	0.872	0.840	0.588
	NeRF [2]	0.880	0.792	0.827	0.690	0.948	0.881	0.828	0.641
	NeX [5]	<b>0.953</b>	<b>0.887</b>	<b>0.933</b>	0.832	<b>0.975</b>	<b>0.952</b>	<u>0.937</u>	<u>0.765</u>
	GNT [4]	0.936	0.846	0.893	<u>0.852</u>	0.963	0.934	<u>0.935</u>	<u>0.752</u>
	GoLF-NRT	<u>0.944</u>	<u>0.851</u>	<u>0.915</u>	<b>0.873</b>	<u>0.971</u>	<u>0.936</u>	<b>0.941</b>	<b>0.787</b>
LPIPS $\downarrow$	LLFF [1]	0.222	0.247	0.174	0.216	0.155	0.173	0.193	0.313
	NeRF [2]	0.249	0.280	0.219	0.316	0.178	0.171	0.263	0.321
	NeX [5]	0.193	0.205	0.150	0.173	0.161	0.131	0.173	0.242
	GNT [4]	<u>0.080</u>	<u>0.116</u>	<u>0.092</u>	<u>0.109</u>	<u>0.060</u>	<u>0.061</u>	<u>0.076</u>	<u>0.153</u>
	GoLF-NRT	<b>0.076</b>	<b>0.114</b>	<b>0.068</b>	<b>0.092</b>	<b>0.057</b>	<b>0.058</b>	<b>0.070</b>	<b>0.136</b>

Table 2. Per-scene optimization quantitative comparison of state-of-the-art view synthesis methods on LLFF with 10 input views. Best results are in bold, second-best are underlined.

### 3.2. Video Results.

In addition to the framewise rendering presented, we have also incorporated rendered videos, in the format of GIF files, as part of the supplementary material accompanying this document. While our primary focus has been on achieving generalizable rendering utilizing few-shot reference views for each frame’s reconstruction, for the purpose of video rendering, we showcase examples for two cases, including the rendering results with three reference views for generalizable rendering and per-scene optimization.

In the context of the generalizable setting that utilizes three reference views, we have selectively chosen three scenes from the LLFF dataset characterized by high-frequency pattern variations: "flower", "horns", and "leaves". For these scenes, we conduct a comparative analysis between GoLF-NRT and its baseline method, GNT [4]. Our findings indicate that when the input views are limited yet adequate, GNT tends to produce more inconsistent fragments. Moreover, there are a lot of flickering artifacts that are noticeable in the frame. In contrast, our proposed

GoLF-NRT, due to the incorporation of more global context information, results in smoother rendered videos, particularly noticeable at the boundaries of leaves and other objects, while also yielding a cleaner overall image.

In the context of the per-scene optimization setting, we present an example involving "orchids", comparing GoLF-NRT with GNT [4]. Similarly, GoLF-NRT produces a more consistent rendering.

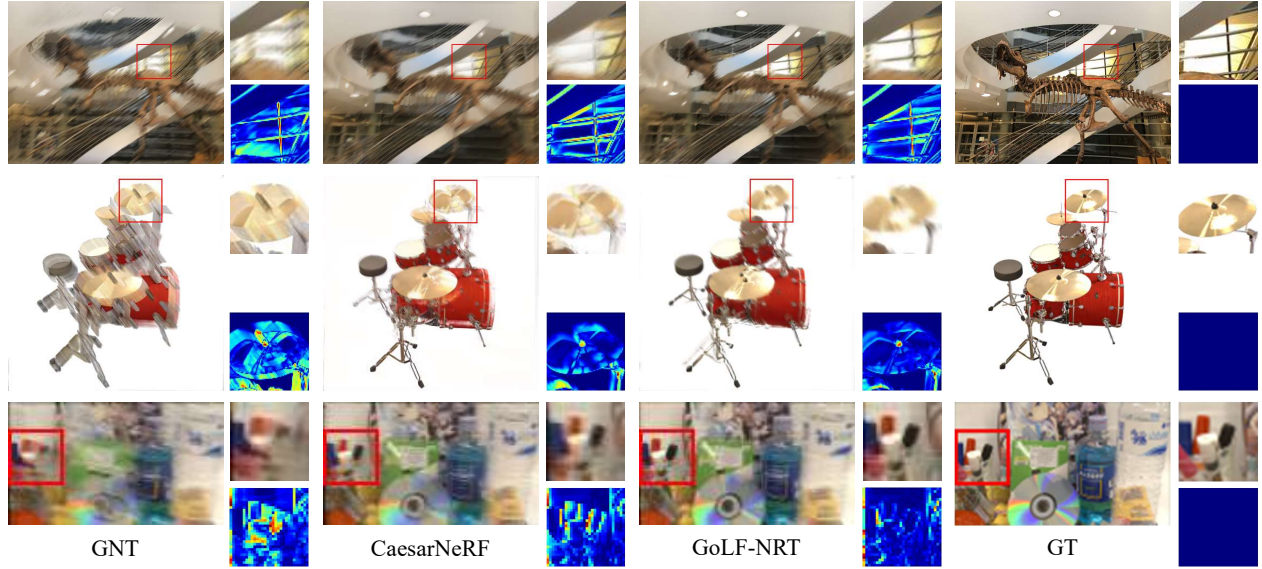


Figure 1. Qualitative comparison of GoLF-NRT with GNT, EVE-NeRF and CaesarNeRF with 1 input views. The first, second, and third rows correspond to the Trex scene from LLFF, the Drums scene from Blender, and the CD scene from Shiny, respectively. Each image triplet includes: the reconstructed image on the left, a zoomed-in view on the upper right, and the error map corresponding to the zoomed-in view on the lower right.

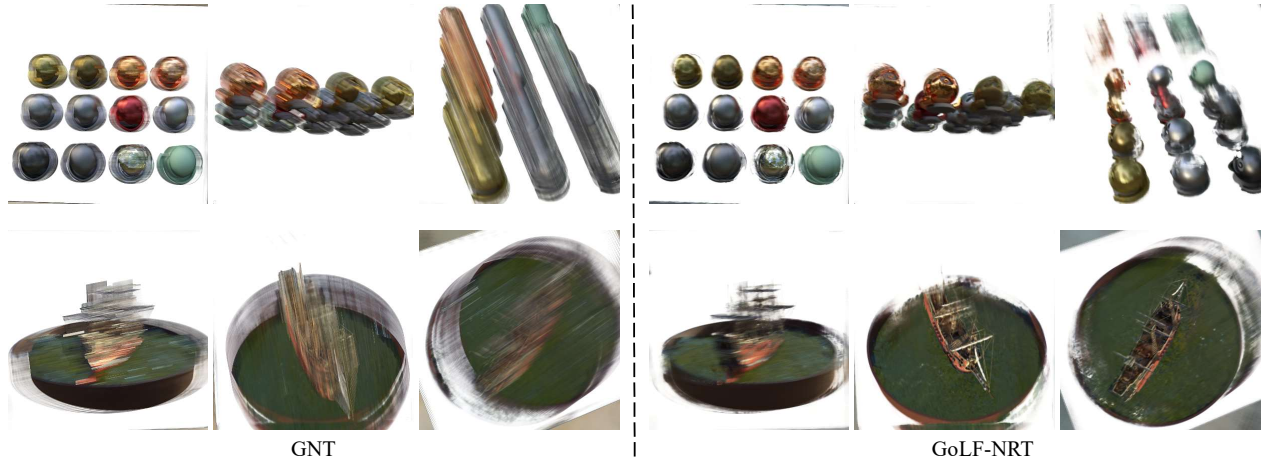


Figure 2. Qualitative comparison between GoLF-NRT and GNT with 1 input views. The first and second rows correspond to the Materials and Ship scene from Blender, respectively.

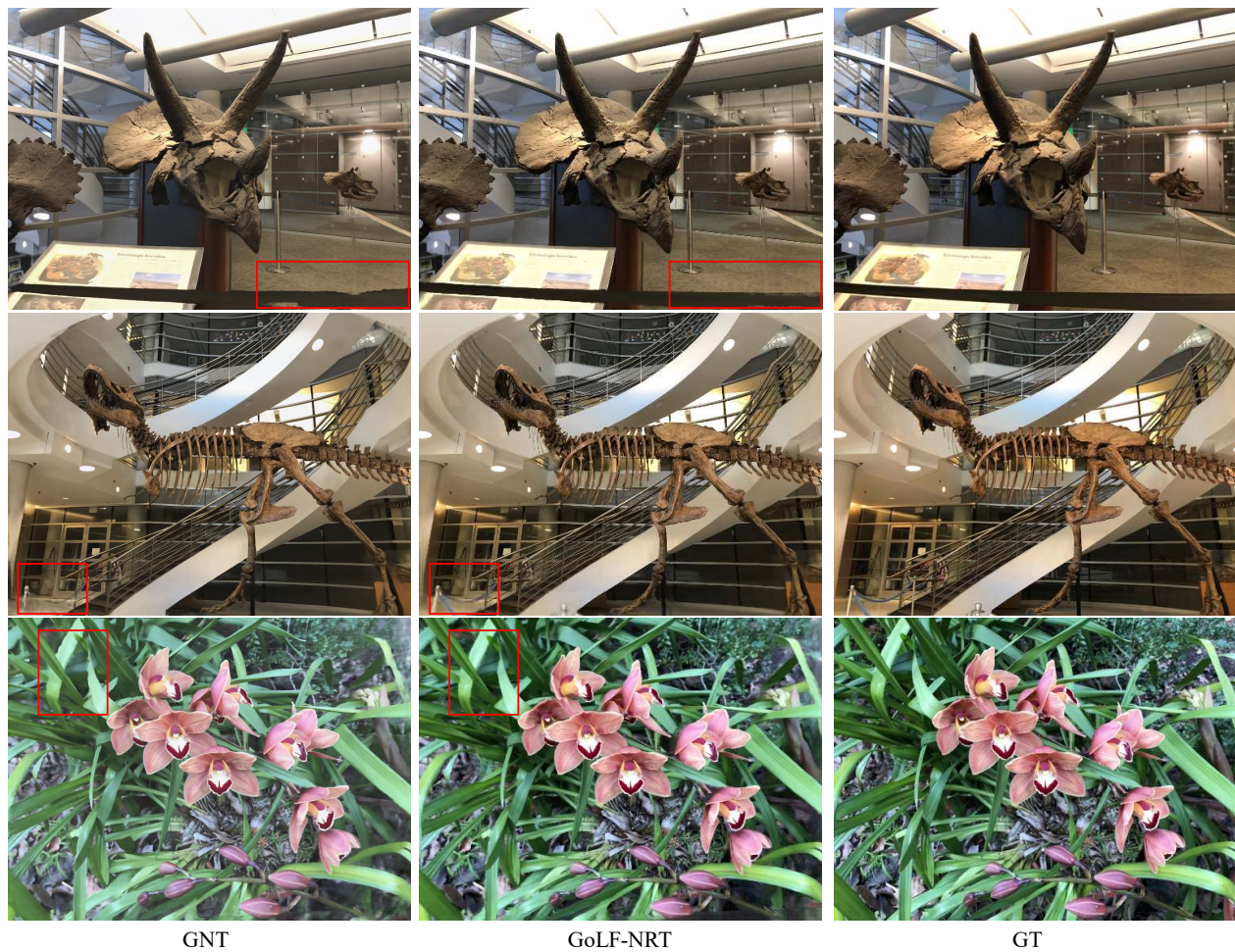


Figure 3. Qualitative comparison between GoLF-NRT and GNT with 10 input views. The first, second, and third rows correspond to the Horns, Trex, and Orchids scene from LLFF, respectively.

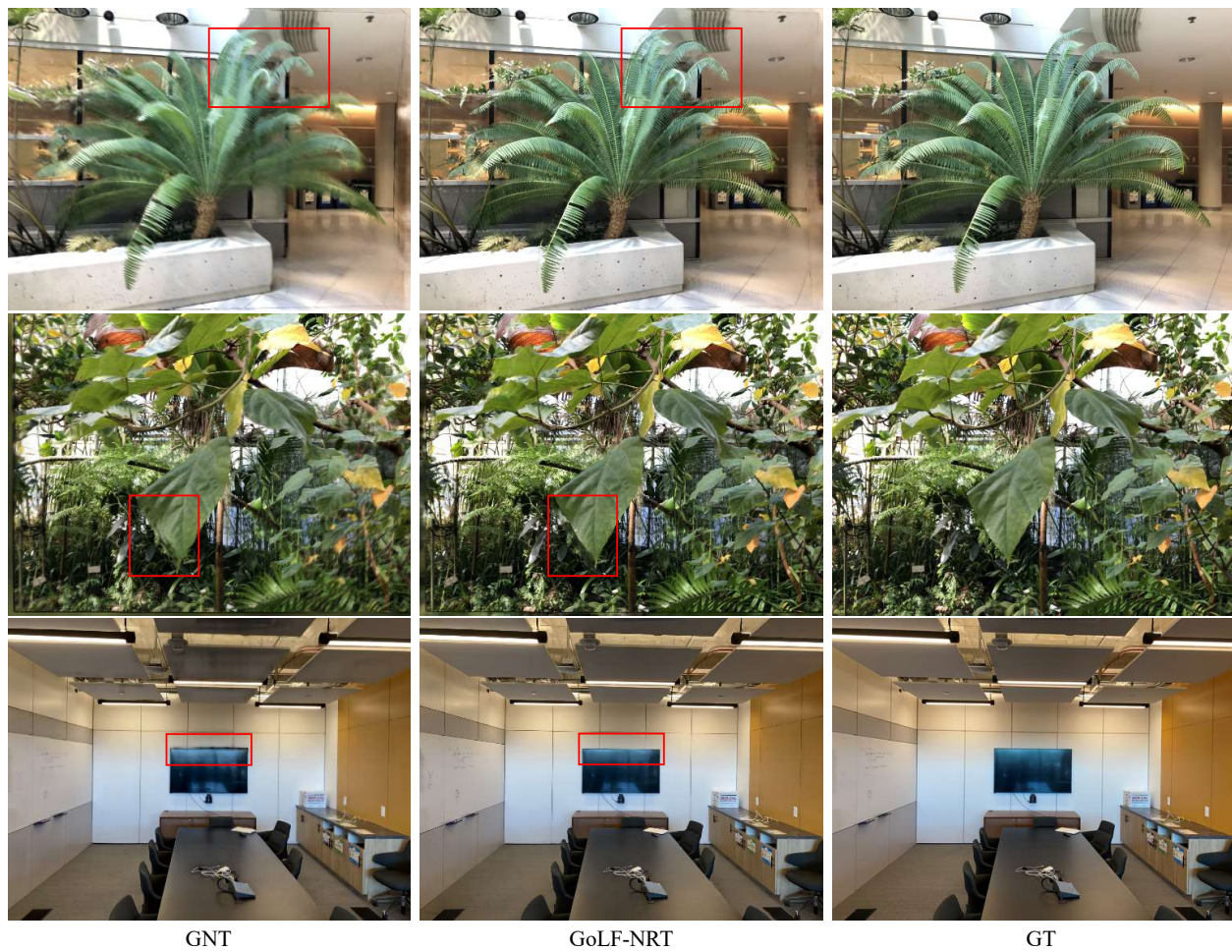


Figure 4. Qualitative comparison between GoLF-NRT and GNT with a single input view. The first, second, and third rows correspond to the Fern, Leaves, and Room scene from LLFF, respectively.

## References

- [1] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019. 1, 2
- [2] 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, 2
- [3] Zhiyuan Min, Yawei Luo, Wei Yang, Yuesong Wang, and Yi Yang. Entangled view-epipolar information aggregation for generalizable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4906–4916, 2024. 1, 2
- [4] Mukund Varma, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all that nerf needs? In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2
- [5] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8534–8543, 2021. 1, 2
- [6] Haidong Zhu, Tianyu Ding, Tianyi Chen, Ilya Zharkov, Ram Nevatia, and Luming Liang. Caesarnerf: Calibrated semantic representation for few-shot generalizable neural rendering. In *European Conference on Computer Vision*, 2024. 1