NeRF Director: Revisiting View Selection in Neural Volume Rendering

Supplementary Material

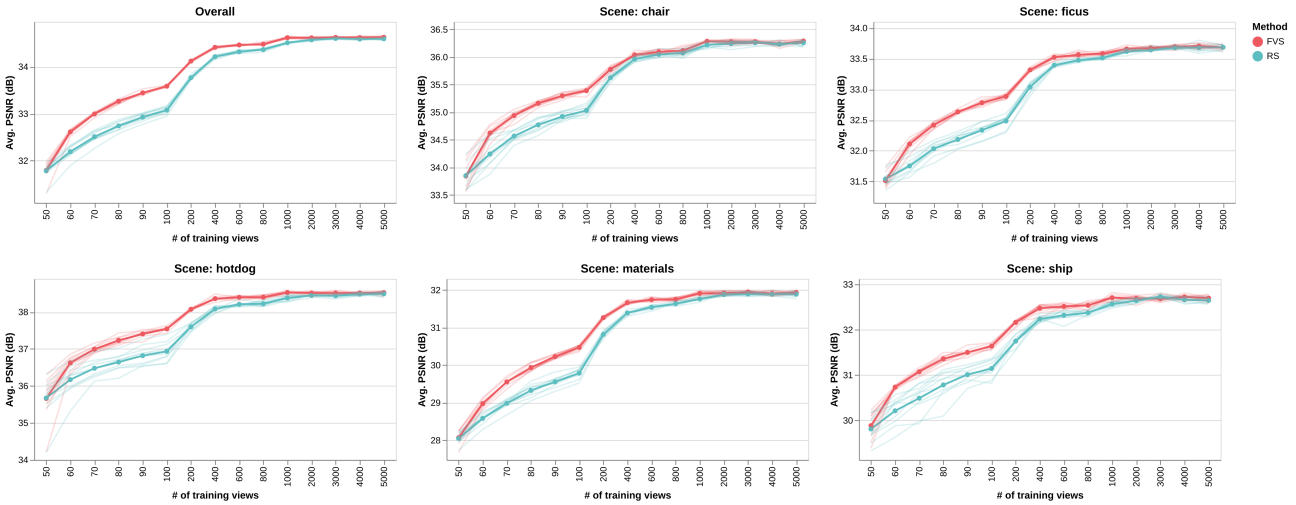


Figure 1. Asymptotic quantitative results of adding more views on NeRF Synthetic datasets in terms of PSNR. The first plot is the overall results across five scenes; the others are scene-specific results. Low-opacity lines present the results for each repetition, while high-opacity lines present the average result across five repetitions.

In this supplementary material, we provide details about the following topics:

- *Additional details on our motivation* in Appendix A;
- *Relaxation in information gain-based sampling* in Appendix B;
- *Additional implementation details* in Appendix C;
- *Additional results and visualization* in Appendix D;

A. Additional Details on Our Motivation

Asymptotic performance of adding more views. To gain a deeper insight into the impact of view selection on novel view synthesis, we train InstantNGP [8] on the NeRF Synthetic dataset with training splits of varied sizes ranging from 50 to 5000 views. Figure 1 demonstrates overall and scene-specific results. It illustrates that with sufficient training time and sampled views, RS achieves the same asymptotic performance as FVS. However, it can be noted that RS samples cameras independently, which may require more views to achieve the same rendering quality.

Sparse 3D-reconstruction runtime analysis. The proposed technique also offers the advantage of reducing the computation required for the initial sparse reconstruction needed to estimate the camera parameters. Before training any NeRF, one has to compute camera intrinsics and extrinsics, by solving a structure-from-motion problem, which may become costly as the number of camera n increases. Traditional approaches rely on four steps: feature extraction ($\mathcal{O}(n)$), feature matching ($\mathcal{O}(n^2)$), SfM ($\mathcal{O}(n^3)$) and bundle adjustment ($\mathcal{O}(P^3)$), preventing its use for a large number of images.[2]

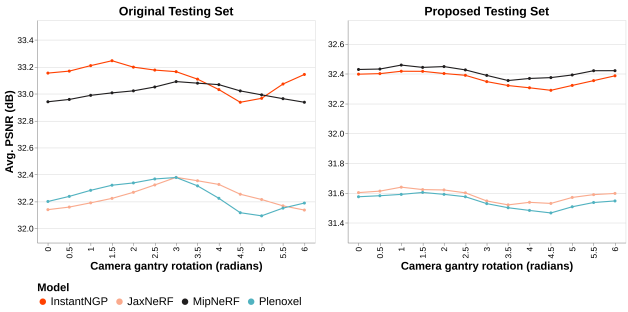


Figure 2. The rendering performance of four distinct NeRF models, in terms of PSNR, under various *z-axis* rotations of the test camera poses. Left: original test set, Right: proposed test set.

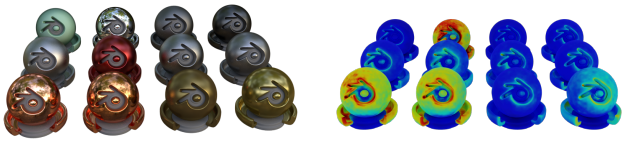
It is worth observing that our current framework and FVS could be amenable to performing view selection before solving SfM, as the presented algorithm does not require high localization accuracy. For instance, one can imagine a scenario where a real-time slam algorithm (inertial + visual odometry) [1, 9, 13] estimates the camera poses. Similarly, with the coarse camera overlap, one could swiftly compute the matrix $\mathcal{A}$ adopted in Equation (5) using recent fast feature matchers [5].

[2] where P denotes the number of camera parameters and 3D points.

Rankings inversion of SOTA methods. As discussed in Section 3, view selection is important in the robust evaluation of different NeRF models. Figure 2 and Table 1 provide detailed quantitative results, in terms of PSNR, on the original and our proposed test set, each with thirteen sets of rotations.

B. Relaxation in Information Gain-based Sampling

Limitations without relaxation. Varying material or geometry complexity may lead to diverse reconstruction outcomes. Figure 3 illustrates an example from the NeRF Synthetic dataset. In this scene, spherical objects in Figure 3a have different material complexity. Specifically, intricate surface parts or highly complex materials may contribute to an increased reconstruction error, as shown in Figure 3b. As a result, deterministic IGS methods selecting the view with the highest error or uncertainty tend to stack new training views on these complex areas. For example in Figure 4, newly sampled training views cluster in the forward face of the *chair* due to its increased texture complexity. This overfitting is counterproductive as the performances are inherently inferior. This can be seen in Figure 5 where deterministic IGS exhibits worse performance than RS.



(a) (b)

Figure 3. Scene *materials* ground truth image (a); Its re-projected PSNR from all rendered images to the mesh (b), where *red* means lower PSNR.

Lloyd relaxation. To alleviate the aforementioned oversampling effect, we propose a modification based on the LLoyd-Max Algorithm [6] inspired by optimal transport and stippling theory [3, 11].

More specifically, given a set of k selected views with camera centers $(\mathbf{c}_1, \ldots, \mathbf{c}_k)$ and m proposed views, we offer a modification of the Lloyd iteration described in Algorithm 3.

C. Implementation Details

Re-split the test set for TanksAndTemples. The test sets in the TanksAndTemples datasets comprise one or two video clips, showcasing parts of the reconstructed scene. Figure 6a visualizes the original test view coverage of *M60* and *Truck*. Notably, a significant portion of the objects are not covered by the original test cameras. As motivated in Section 3, we propose a novel split of the test set

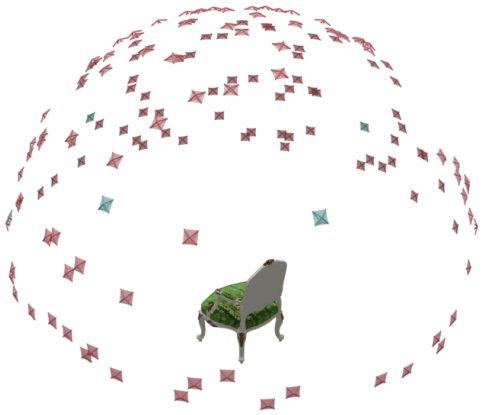


Figure 4. Visualization of the distribution of initial training cameras (in green) and selected cameras (in red) through IGS without relaxation.

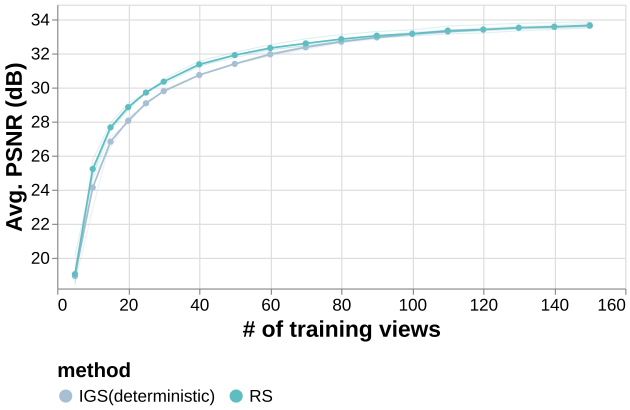


Figure 5. Quantitative comparisons of deterministic IGS and RS on the NeRF Synthetic dataset. IGS without relaxation behaves worse than RS.

Algorithm 3: Lloyd Relaxation

input : $\mu \in \mathcal{U}(\Omega)$, $d = dim(\Omega)$, $\Omega = \mathbb{S}^2$ or $\mathbb{R}^3$
Discrete measure $\frac{1}{N}\sum \delta_{\mathbf{x}_i}$
$v \in \mathbb{R}^{k \times d}$, $p \in \mathbb{R}^{m \times d}$

output: c

$c = \{v, p\} \in \mathbb{R}^{(k+n) \times d}$
for $i \leftarrow 1$ **to** N_{iter} **do**
  $\mathcal{V}^c \leftarrow \texttt{Voronoi}(c)$
  $b^c \leftarrow \texttt{computeBarycenter}(\mathcal{V}^c, \mu)$
  $c \leftarrow \{v, b^c_{k+1\ldots m}\}$
return c

for all four scenes in the TanksAndTemples dataset which aims at providing a more robust evaluation of different view selection methods. We first put together all training and

Table 1. Quantitative results in terms of PSNR of four SOTA NeRF models under various *z-axis* rotations of the test camera poses.

Original test set	**0.0**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**4.0**	**4.5**	**5.0**	**5.5**	**6.0**
InstantNGP [8]	33.15	33.17	33.21	33.25	33.20	33.18	33.16	33.11	33.03	32.94	32.97	33.07	33.14
MipNeRF [2]	32.94	32.96	32.99	33.01	33.02	33.05	33.09	33.08	33.07	33.02	32.99	32.96	32.94
JaxNeRF [7]	32.14	32.16	32.19	32.22	32.27	32.32	32.38	32.35	32.33	32.25	32.21	32.17	32.14
Plenoxels [4]	32.20	32.24	32.28	32.32	32.34	32.37	32.38	32.32	32.22	32.12	32.09	32.15	32.19

Proposed test set	**0.0**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**3.5**	**4.0**	**4.5**	**5.0**	**5.5**	**6.0**
InstantNGP [8]	32.40	32.40	32.42	32.42	32.40	32.39	32.35	32.32	32.31	32.29	32.32	32.35	32.39
MipNeRF [2]	32.43	32.43	32.46	32.44	32.45	32.43	32.39	32.36	32.37	32.37	32.39	32.42	32.42
JaxNeRF [7]	31.60	31.61	31.64	31.62	31.62	31.60	31.55	31.52	31.54	31.53	31.57	31.59	31.60
Plenoxels [4]	31.58	31.58	31.59	31.60	31.59	31.58	31.53	31.50	31.48	31.47	31.51	31.54	31.55

test views for a particular scene. Then, an equal number of test views were selected as in the original test set for each scene using FVS. The distance metric considered during this process encompassed both spatial distance defined in Equation (4) and photogrammetric distance defined in Equation (6). We can observe from Figure 6b that our proposed test set is able to cover the reconstructed scene more uniformly.

ActiveNeRF [10]: ActiveNeRF regards information gain as the reduction of uncertainty. Thus, it selects the candidate view that maximizes the information gained at each selection step. The ActiveNeRF[3] is implemented on the backbone of vanilla NeRF. We re-implemented it using Instant-NGP backbone and coined it Active-InstantNGP. Direct re-implementation of ActiveNeRF on InstantNGP failed to learn the 3D scene due to the reformulation of the NeRF framework as well as the training and rendering process. We highlighted the learning of the radiance field in our adopted loss function. Due to disparities in the rendering performance of Active-InstantNGP, we report the rendering performance of InstantNGP trained on the training views selected with Active-InstantNGP.

Density-aware NeRF Ensembles [12]: We referred to the experiment of the next best view selection in [12] and implemented Density-aware NeRF Ensembles on Instant-NGP within our NeRF Director framework. More specifically for each selection, we trained 5 models on the same training views with different random seed initializations. The training process of each ensemble model comprises 2000 training steps. Then, we computed the uncertainty for all remaining training views using these 5 models as described in [12]. We selected the training view with the highest uncertainty each time.

[3] https://github.com/LeapLabTHU/ActiveNeRF/tree/main

View #	FVS		RS		Speedup
	mean	σ	mean	σ	
50	0.7	$\pm$0.26	2.6	$\pm$0.26	4.03
100	1.4	$\pm$0.43	2.6	$\pm$0.30	1.96
150	1.7	$\pm$0.48	2.7	$\pm$0.32	1.57

Table 2. Averaged training time (in minutes) and standard deviation (σ) comparisons of FVS against RS at the converged quality.

D. Additional Results

Runtime cost analysis on the NeRF Synthetic dataset. We report the runtime cost result in Table 2 measuring the training time of InstantNGP across 5 scenes of NeRF Synthetic and averaged for 10 runs (with different views). For a fixed view budget, our proposed FVS reaches the performances of the traditional RS significantly faster (up to 4$\times$ Speedup).

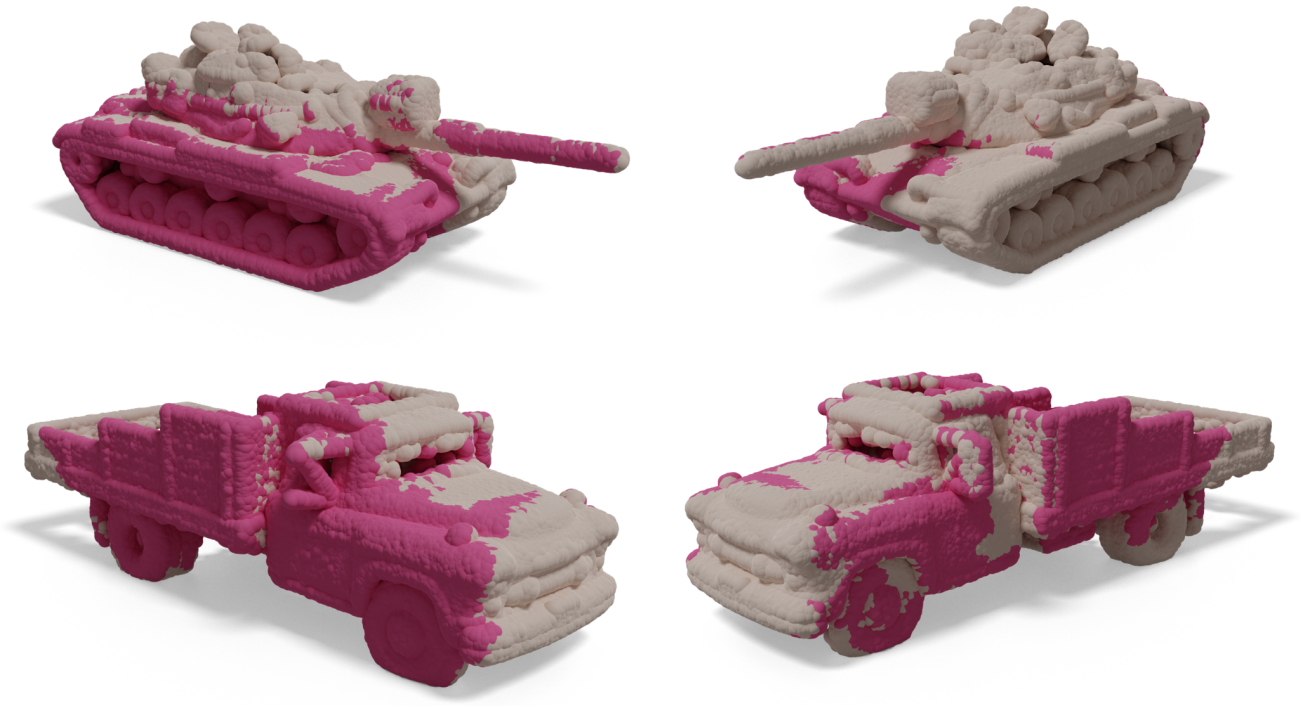
Quantitative results of Plenoxels. We also provide the quantitative results of Plenoxels' [4] asymptotic performance of adding more views, in terms of PSNR and SSIM (Figure 7). We reported the results on 5 scenes of the NeRF Synthetic dataset and 3 scenes (*M60*, *Playground*, and *Truck*) of the TanksAndTemples dataset for 5 repetitions. Similar trends in performance, relative to the number of views, can be observed with this alternative backbone.

Qualitative results of InstantNGP. We provide the qualitative results of InstantNGP on both the TanksAndTemples and the NeRF Synthetic dataset, as shown in Figure 8 and Figure 9 respectively. We compared our proposed FVS and IGS(vMF) with the baseline RS and view selection method in [12]. It can be observed that our proposed methods can generate a clearer and sharper appearance.

References

[1] cuVSLAM (CUDA Stereo Visual SLAM) - NVIDIA Docs. 1

[2] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P.



(a)

(b)

Figure 6. Visualization of test view coverage on objects *M60* and *Truck*. Pink areas indicate that the ray-mesh intersection is greater than 10 in those regions. (a): Original test view; (b): Proposed test view.

Srinivasan. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5835–5844. IEEE, 2021. 3

[3] Fernando De Goes, Katherine Breeden, Victor Ostromoukhov, and Mathieu Desbrun. Blue noise through

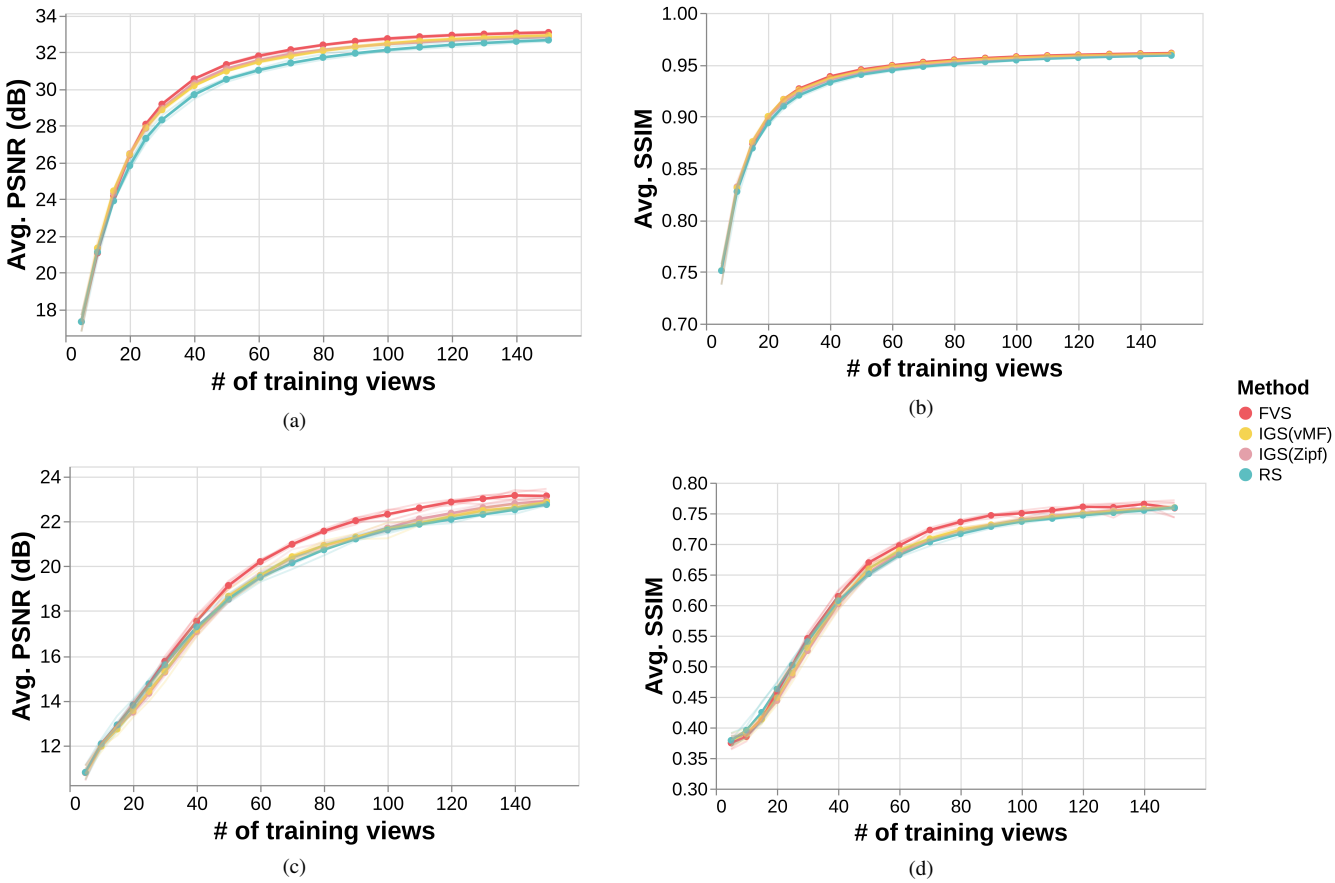


Figure 7. Quantitative comparisons of rendering quality on Plenoxels [4] along with the increase of used training views sampled by different view selection methods. The top row shows the results on the NeRF Synthetic dataset in terms of PSNR (a) and SSIM (b). The bottom row shows the results on the TanksAndTemples dataset in terms of PSNR (c) and SSIM(d). Low-opacity lines present the results for each repetition, while high-opacity lines present the average result across five repetitions.

optimal transport. *ACM Transactions on Graphics (TOG)*, 31(6), 2012. 2

[4] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance Fields without Neural Networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5491–5500. IEEE, 2022. 3, 5

[5] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. 2023. 1

[6] Stuart P. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. 2

[7] 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. In *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing. 3

[8] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):102, 2022. 1, 3

[9] Raul Mur-Artal, J. M.M. Montiel, and Juan D. Tardos. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 1

[10] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. ActiveNeRF: Learning where to See with Uncertainty Estimation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13693 LNCS:230–246, 2022. 3

[11] Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With Applications to Data Science*. Now Publishers Inc, 2019. 2

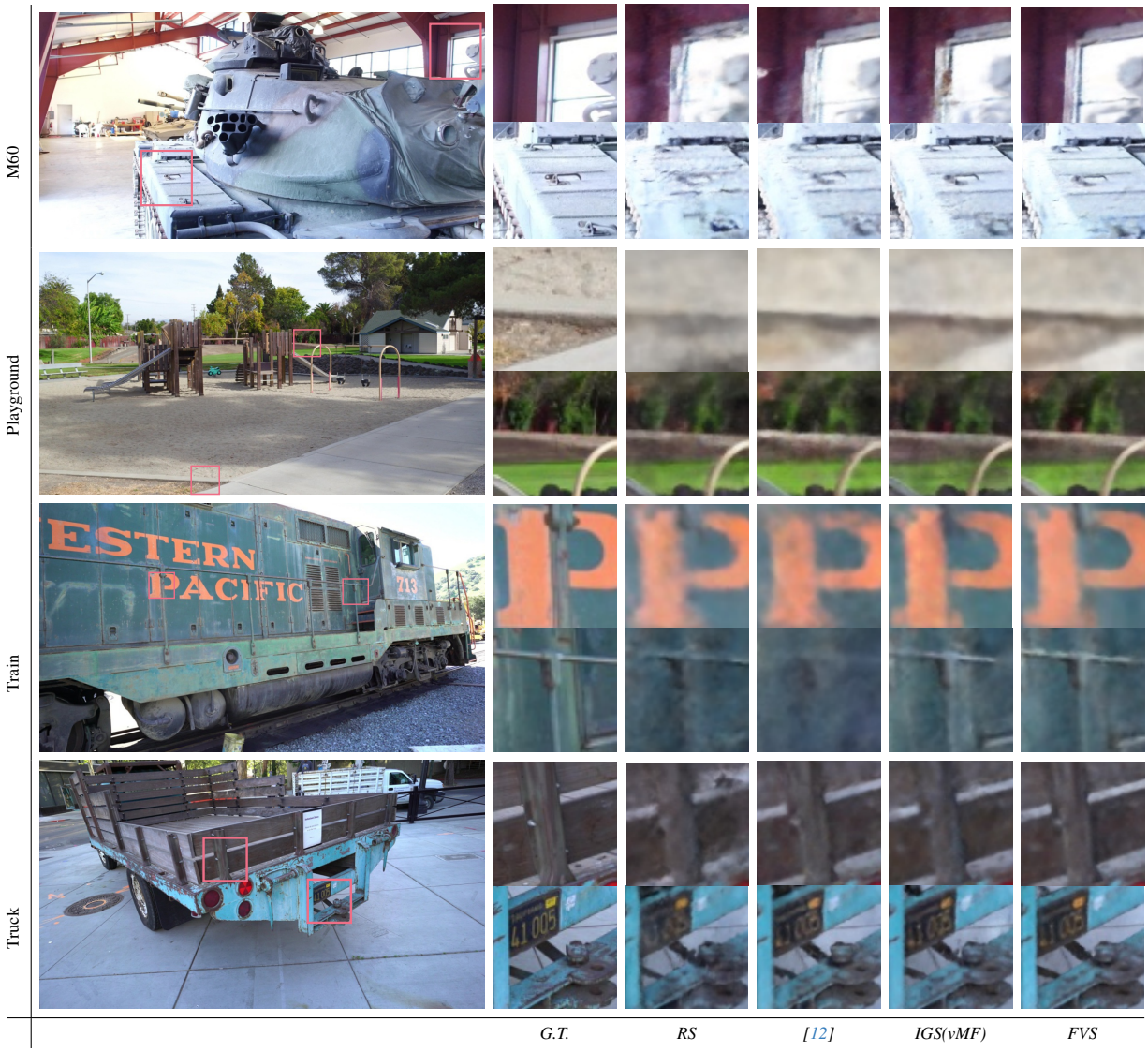


Figure 8. Qualitative comparison results of four view selection methods on the TanksAndTemples dataset with 80 training views.

[12] Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware NeRF Ensembles: Quantifying Predictive Uncertainty in Neural Radiance Fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9370–9376. IEEE, 2023. 3, 6, 7

[13] M. Wudenka, M. G. Muller, N. Demmel, A. Wedler, R. Triebel, D. Cremers, and W. Sturzl. Towards Robust Monocular Visual Odometry for Flying Robots on Planetary Missions. *IEEE International Conference on Intelligent Robots and Systems*, pages 8737–8744, 2021. 1

Figure 9. Qualitative comparison results of four view selection methods on the NeRF Synthetic dataset with 80 training views.