EUROGRAPHICS 2024 / A. Bermano and E. Kalogerakis
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 2

Supplementary of Perceptual Quality Assessment of NeRF and Neural View Synthesis Methods for Front-Facing Views

H. Liang[1] T. Wu[1] P. Hanji[1] F. Banterle[2] H. Gao[1] R. Mantiuk[1] and C. Oztireli[1]

[1]University of Cambridge, UK
[2]ISTI-CNR, Italy

1. Supplementary Video

We have included reference video clips of the scenes in both of our new datasets. However, we were unable to include the synthesized clips generated by the NVS methods due to file size restrictions on the supplementary materials. We will make them available on the project web page upon acceptance of the paper.

2. More Details of Experimental Setup

Pairwise Comparison Previous works [POMZ*20,HME*22] have shown that the pairwise comparison protocol has higher sensitivity than direct rating. Thus, it provides a meaningful scale and, when combined with active sampling, is more efficient than direct rating (eg. double stimulus). We follow the pairwise comparison procedure adopted in these works.

Specifically, in each trial of the experiment, a participant was shown a pair of videos side-by-side on the same display and was instructed to pick the video of higher quality. Participants could also press the space bar to view the reference video of the displayed scene (except for videos from the LLFF dataset as reference videos were not available). The reference videos were included as one of the compared conditions. To encourage participants to directly compare the presented videos, we displayed a black-frame for half a second when switching videos. This prevents participants from picking up on minute, localized differences and instead forces them to assess overall visual quality. Each participant was shown 300∼400 video pairs. A pairwise comparison experiment can be easily extended to include new NVS methods - ASAP will prioritize comparisons with newly added methods.

ASAP Sampling The comparison samples are determined by ASAP [MWPO*21], an active sampling method. This algorithm is based on approximate message passing and expected information gain maximization and was shown to outperform heuristics, such as the Swiss chess system. ASAP maximizes the information gained from each trial, reduces the measurement error under the same number of comparisons, and ensures that all compared conditions are adequately sampled. ASAP also ensures each method is compared at least once in each batch.

Subjective Score Scaling and JOD Calculation The results of pairwise comparison was scaled under Thurstone's case V model into Just-Objectionable-Difference (JODs) using the pwcmp software [POM17]. A difference of 1 JOD unit means that 75% participants preferred one method over another. As an equivalent model to Bradley-Terry used for scaling pairwise comparison data, Thurstone's case V assumes that participants made their selections by assigning a single quality value to each video and approximates this quality by a normally distributed random variable with the same inter- and intra-observer variance. Meanwhile, the ASAP method also relies on Thurstone's model.

3. More Details of Bootstrapping

Our bootstrapping simulates many repetitions of the experiment: each bootstrap instance samples N scenes, whereas each scene additionally samples K observers that have assessed the scene (sampling with replacement, 'N' and 'K' are bound by the total number of scenes and observers respectively). Each bootstrap instance independently scales the JOD values using a subset of data and computes a point estimation of the correlation.

Bootstrapping enables us to recover the distribution of correlation scores, from which we draw conclusions regarding the ranking of metrics. Without this, we have to conclude from the mean correlations averaged across each single scene estimation, which may be incorrect because we observe from Fig. 6 in the main document that the distributions are spread out. In that respect, we are more rigorous than most works comparing quality metrics. With a large number of bootstrapped samples (2000 in our experiments), the mean converges toward the true correlation value, thereby reducing estimation error. It is easy to see that the mean correlations of our bootstrapped distributions (Fig. 6 in the main document) are close to the mean without bootstrapping (Fig. 6-8 in supplementary).

4. More Details of Non-parametric Test

To determine whether the differences between the metrics are statistically significant, we performed a non-parametric test since the distributions were non-normal (after Fisher's transform). We distinguish different metrics at the 95% level by computing the distribution of the difference of bootstrap samples. If the corresponding percentage of difference is larger than zero, one metric is significantly better than the other. Note that the probability of passing the statistical test will not increase with the number of bootstrap samples.

5. Failure Cases of VMAF and FVVDP

While VMAF and FVVDP generally agree with human perceptual assessments, they are not perfect and it is beneficial to examine the instances where these metrics deviate from subjective preferences. To do so, we investigate per-scene correlations between FVVDP/VMAF scores and bootstrapped perceptual values (illustrated in Fig. 6 of the main document).

For the Lab dataset, despite surpassing PSNR in overall performance, VMAF/FVVDP fares poorly for specific scenes. For example, although the VMAF scores of DVGO and GNT-S are comparable for the `Glossy animals` scene, their subjective scores are quite different. We observe a similar phenomenon for the FVVDP metric. Meanwhile, for the Fieldwork dataset, although VMAF/FVVDP effectively evaluate perceived quality across most scenes, their limitations become apparent when assessing the `Whale` scene. For this scene, NeRF shows lower VMAF and FVVDP scores than IBRNet-S and LFNR; however, subjective scores exhibit minimal disparity.

For the failure cases on VMAF, we speculate that it is because the performance of VMAF is highly dependent on the quality and diversity of the dataset used to train VMAF model. If our testing scenes differ significantly from the training data, VMAF may yield inaccurate results. As for FVVDP, it lacks specific calibration and validation for distortions inherent in novel view synthesis, potentially introducing noise and hindering accurate quality quantification.Fine-tuning VMAF and calibrating FVVDP are beyond the scope of this work. Despite these limitations, the overall performance of VMAF/FVVDP remains superior to the other metrics. Thus, they can be good candidates when evaluating subjective quality in video assessments for NVS.

6. Per-scene Subjective Quality

Due to space limitations, the main document contains subjective scores averaged across all scenes in each dataset (Figure 4 in the main document). Figures 1, 2 and 3 show the subjective results individually for each scene. These results show large variations across the scenes, but they also exhibit common trends:

- The generalizable methods GNT and IBRNet perform poorly on all scenes in our new Lab and Fieldwork datasets (worse than NeRF), but much better on the public LLFF dataset. Per-scene fine-tuning (-S suffix) improves the predictions of both methods.
- Similarly DVGO performs poorly on our new datasets, but much better on the LLFF dataset.
- LFNR has rather uneven performance — it is one of the best methods for some scenes (Lab/CD-occlusions (I/E), Lab/Glossy animals (I), Fieldwork/Naiad statue) but it fails in the others.
- MipNeRF was one the most robust methods, performing typically better or on par with NeRF. In some of the scenes, it matched the quality of the reference (Lab/Glass, Fieldwork/Leopards, Fieldwork/Giraffe, Fieldwork/Naiad statue, Fieldwork/Vespa).
- Plenoxel performed well in most scenes in LLFF dataset (except Room) but was generally worse than NeRF when tested on the Lab dataset. Its performance varied from scenes to scene in the Fieldwork dataset, with a few fail cases (Dinosaur and Whale) but also better-than-NeRF performance (Leopards, Naiad statue, Vespa).

7. Metric Performance: PLCC and RMSE

Apart from Spearman Rank Order correlations (SROCC), we also compute the bootstrapped distributions of Pearson Linear Correlation Coefficient (PLCC) and Root Mean Squared Error between the image metrics score and perceptual quality score on each dataset. The results are shown in Figures 4 and 5. With a few exceptions, the trends shown in those plots are similar to those shown for SROCC in Figure 6 of the main paper. The difference worth noting is that while the correlations (PLCC and SROCC) are much higher for the Fieldwork than for the Lab dataset (indicating good metric performance), the opposite trend is shown by the RMSE. The RMSE values are on average smaller for the Lab dataset, suggesting higher metric accuracy. It must be noted, however, that the range of subjective scores is much larger for the Fieldwork dataset (refer to the scatter plots in Figure 6 and Figure 7). The difference in the RMSE numbers is most likely due to very different magnitudes of distortions in each dataset. If the goal of a metric is to differentiate between NVS methods, the correlation coefficients are better indicators of metric performance.

8. Metric Prediction Scatter Plots

Metric predictions for individual scenes are compared with subjective scores in scatter plots in Figures 6, 7 and 8. When metric predictions are accurate, the scatter plot forms a possibly tight curve. The scatter plots for Lab dataset in Figure 6 show the difficulty of the task on this dataset — objective and subjective measures of quality are not well correlated for any of the tested metrics. They correlate even worse on LLFF dataset 8, which demonstrates that testing on sparse views in current evaluation protocol is insufficient to assess the subjective quality of synthesized videos. The scatter plots, however, form much stronger relations for the Fieldwork dataset in Figure 7.

9. Training details

DVGO We follow the training setup as in [SSC22] and set expected numbers of voxels to be $M(c) = 100^3$ and $M(f) = 160^3$ in coarse and fine stages. The points sampling step sizes are set to half of the voxel sizes, i.e., $\delta(c) = 0.5 \cdot s(c)$ and $\delta(f) = 0.5 \cdot s(f)$. The shallow MLP layer comprises two hidden layers with 128 channels. The Adam optimizer [KB14] is employed with a batch size of 8192 rays to optimize the coarse and fine scene representations for 10k and 20k iterations. The base learning rates are 0.1 for all voxel grids and 10^{-3} for the shallow MLP. The exponential learning rate decay is applied.

NeRF We follow the pytorch implementation of NeRF [YC20].

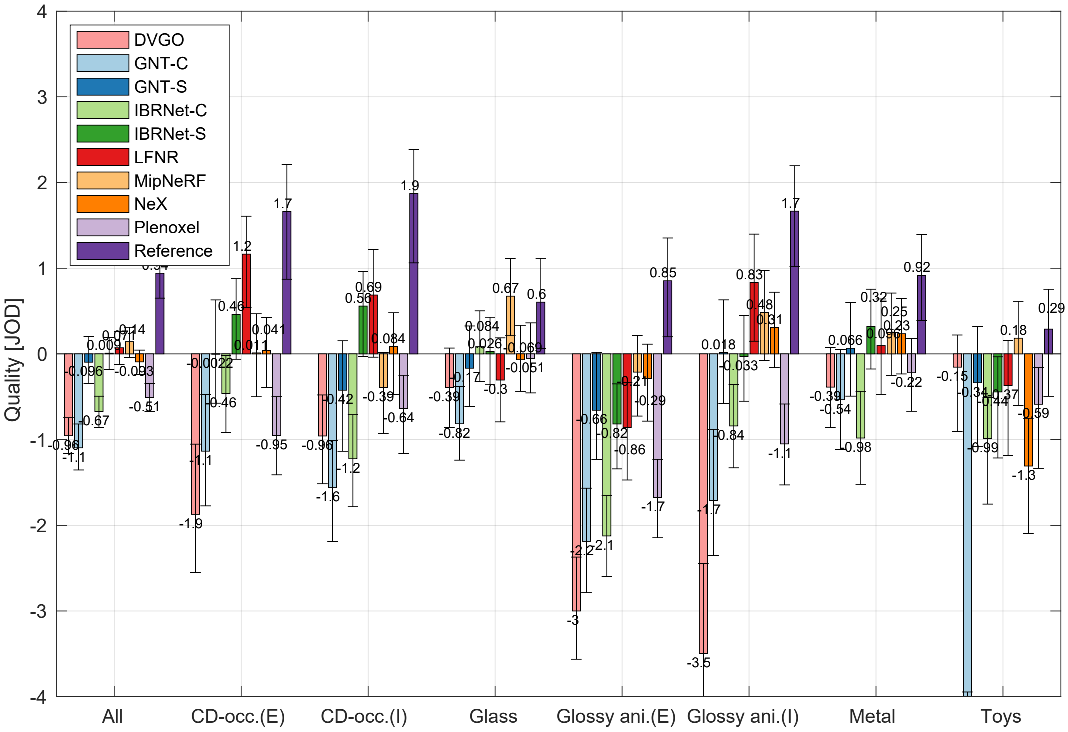

Figure 1: *Perceptual preference of different NeRF methods on the Lab dataset. The notation is the same as in Figure 4 in the main paper. The scenes with (I) in the label used the novel view selection that required only interpolation of the views, while the scenes with (E) required the views to be extrapolated.*

We use a batch size of 1024 rays, each sampled at $N_c = 64$ coordinates in the coarse volume and $N_f = 128$ additional coordinates in the fine volume. We use the Adam optimizer with a base learning rate at 5×10^{-4} and optimize for 2×10^5 iterations.

GNT For the cross-scene generalizable GNT model (denoted as GNT-C), we use the pre-trained model released by [WCC*22]. For finetuned version of GNT model on each scene (denoted as GNT-S), we finetune the cross-scene model with Adam optimizer with base learning rates for the feature extraction network and GNT as 10^{-3} and 5×10^{-4} respectively, which decay exponentially over training steps. For all our experiments, we train for 50,000 steps with 4096 rays sampled in each iteration.

IBRNet For the cross-scene generaliable IBRNet model (denoted as IBRNet-C), we use the pre-trained model from [WWG*21]. During fine-tuning stage for IBRNet-S, we optimize both 2D feature extractor and IBRNet itself with Adam optimizer using base learning rates of 5×10^{-4} and 2×10^{-4}).

LFNR The architecture of transformer is the same as the ones recently introduced for vision related tasks [DBK*20]. Each transformer has 8 blocks and the internal feature size is 256. In each training step, we randomly choose a target image and sample a batch of random rays from it. The batch size is 128. We train for 250 000 iterations with the Adam optimizer and a linear learning rate decay schedule with 5000 warm-up steps.

MipNeRF We follow the training procedure specified by [BMV*22]: 1 million iterations of Adam with a batch size of 4096 and a learning rate that is annealed logarithmically from $5 \cdot 10{-}4$ to $5 \cdot 10^{-6}$.

NeX As in [WPYS21], we use an MPI with 192 layers with $M = 12$ consecutive planes sharing one set of texture coefficients. We sample and render 8,000 pixels in the training view for photometric loss computation. The network is trained for 4,000 epochs using Adam optimizer with a learning rate of 0.01 for base color and 0.001 for both networks and a decay factor of 0.1 every 1,333 epochs.

Plenoxel The implementation of Plenoxel is based on a custom PyTorch CUDA [NVF20] extension library to achieve fast differentiable volume rendering. We use a batch size of 5000 rays and optimize with RMSProp [TH12]. For optimization of density, we use the same delayed exponential learning rate schedule as MipNeRF [BMV*22], where the exponential is scaled by a learning rate of 30 and decays to 0.05 at step 250000, with an initial delay period of 15000 steps. For SH we adopts a pure exponential decay learning rate schedule, with an initial learning rate of 0.01 that decays to 5×10^{-6} at step 250000.

References

[BMV*22] Barron J. T., Mildenhall B., Verbin D., Srinivasan P. P., Hedman P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24,*

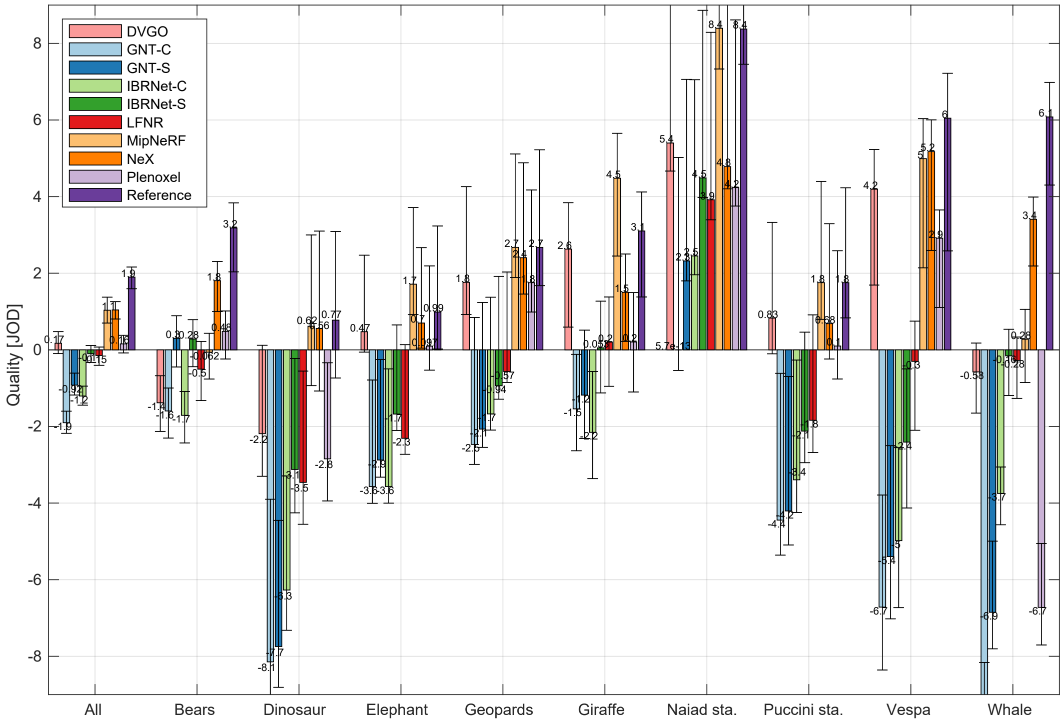

Figure 2: *Perceptual preference of NeRF methods on the Fieldwork dataset.*

2022 (2022), IEEE, pp. 5460–5469. URL: https://doi.org/10.1109/CVPR52688.2022.00539, doi:10.1109/CVPR52688.2022.00539. 3

[DBK*20] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020). 3

[HME*22] Hanji P., Mantiuk R., Eilertsen G., Hajisharif S., Unger J.: Comparison of single image hdr reconstruction methods — the caveats of quality assessment. In *ACM SIGGRAPH 2022 Conference Proceedings* (2022), pp. 1–8. 1

[KB14] Kingma D. P., Ba J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 2

[MWPO*21] Mikhailiuk A., Wilmot C., Perez-Ortiz M., Yue D., Mantiuk R.: Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 IEEE International Conference on Pattern Recognition (ICPR)* (Jan 2021). 1

[NVF20] NVIDIA, Vingelmann P., Fitzek F. H.: Cuda, release: 10.2.89, 2020. URL: https://developer.nvidia.com/cuda-toolkit. 3

[POM17] Perez-Ortiz M., Mantiuk R. K.: A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686* (2017). 1

[POMZ*20] Pérez-Ortiz M., Mikhailiuk A., Zerman E., Hulusic V., Valenzise G., Mantiuk R. K.: From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing 29* (2020), 1139–1151. doi:10.1109/TIP.2019.2936103. 1

[SSC22] Sun C., Sun M., Chen H.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022* (2022), IEEE, pp. 5449–5459. URL: https://doi.org/10.1109/CVPR52688.2022.00538, doi:10.1109/CVPR52688.2022.00538. 2

[TH12] Tieleman T., Hinton G.: Rmsprop: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural Networks for Machine Learning 4*, 2 (2012), 26–31. 3

[WCC*22] Wang P., Chen X., Chen T., Venugopalan S., Wang Z., et al.: Is attention all nerf needs? *arXiv preprint arXiv:2207.13298* (2022). 3

[WPYS21] Wizadwongsa S., Phongthawee P., Yenphraphai J., Suwajanakorn S.: Nex: Real-time view synthesis with neural basis expansion. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021* (2021), Computer Vision Foundation / IEEE, pp. 8534–8543. doi:10.1109/CVPR46437.2021.00843. 3

[WWG*21] Wang Q., Wang Z., Genova K., Srinivasan P. P., Zhou H., Barron J. T., Martin-Brualla R., Snavely N., Funkhouser T. A.: Ibrnet: Learning multi-view image-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021* (2021), Computer Vision Foundation / IEEE, pp. 4690–4699. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Wang_IBRNet_Learning_Multi-View_Image-Based_Rendering_CVPR_2021_paper.html, doi:10.1109/CVPR46437.2021.00466. 3

[YC20] Yen-Chen L.: Nerf-pytorch. https://github.com/yenchenlin/nerf-pytorch/, 2020. 2

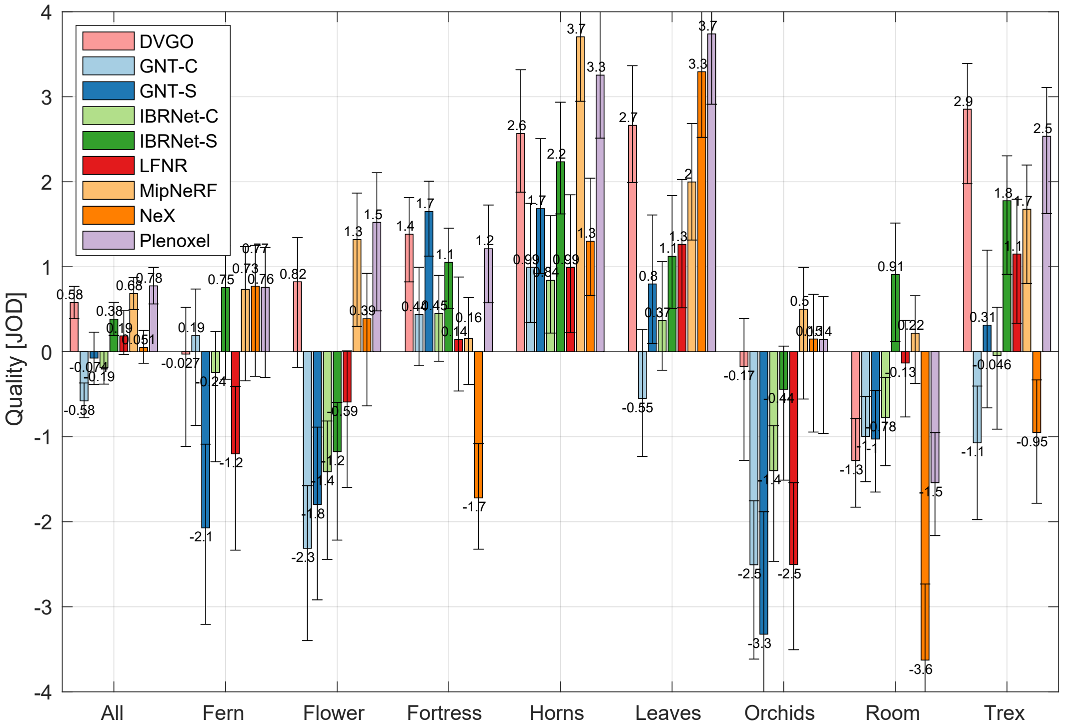

Figure 3: *Perceptual preference of NeRF methods on the LLFF dataset.*

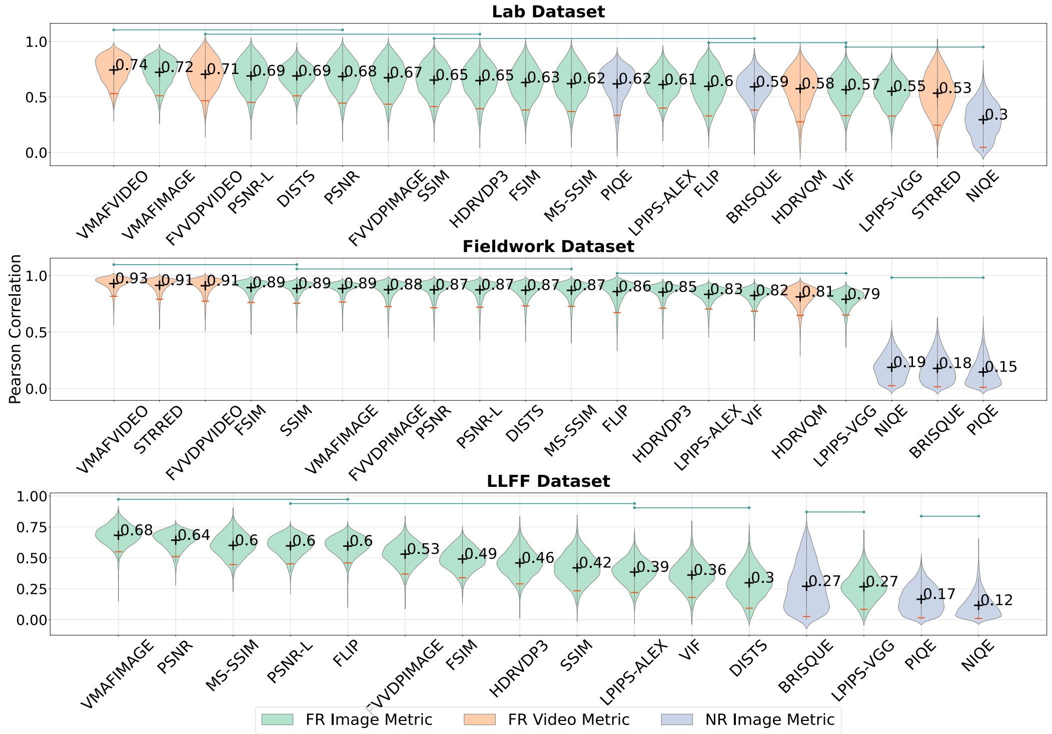

Figure 4: *Bootstrapped distributions of Pearson Linear Correlation Coefficients (PLCC) for all metrics, reported separately for each dataset. The higher the number, the better is metric's performance. The notation is the same as for Figure 6 in the main paper.*

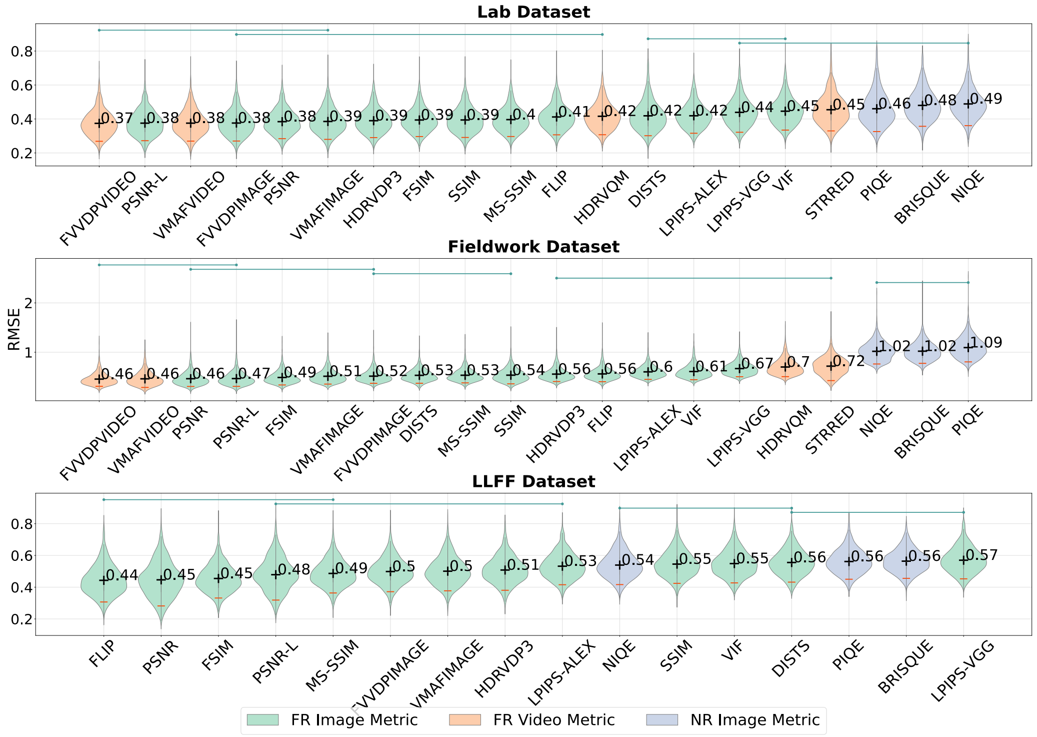

Figure 5: *Bootstrapped distributions of Room Mean Squared Errors (RMSE) for all metrics computed separately for each dataset. The lower the number, the better is metric's performance. The notation is the same as for Figure 6 in the main paper.*

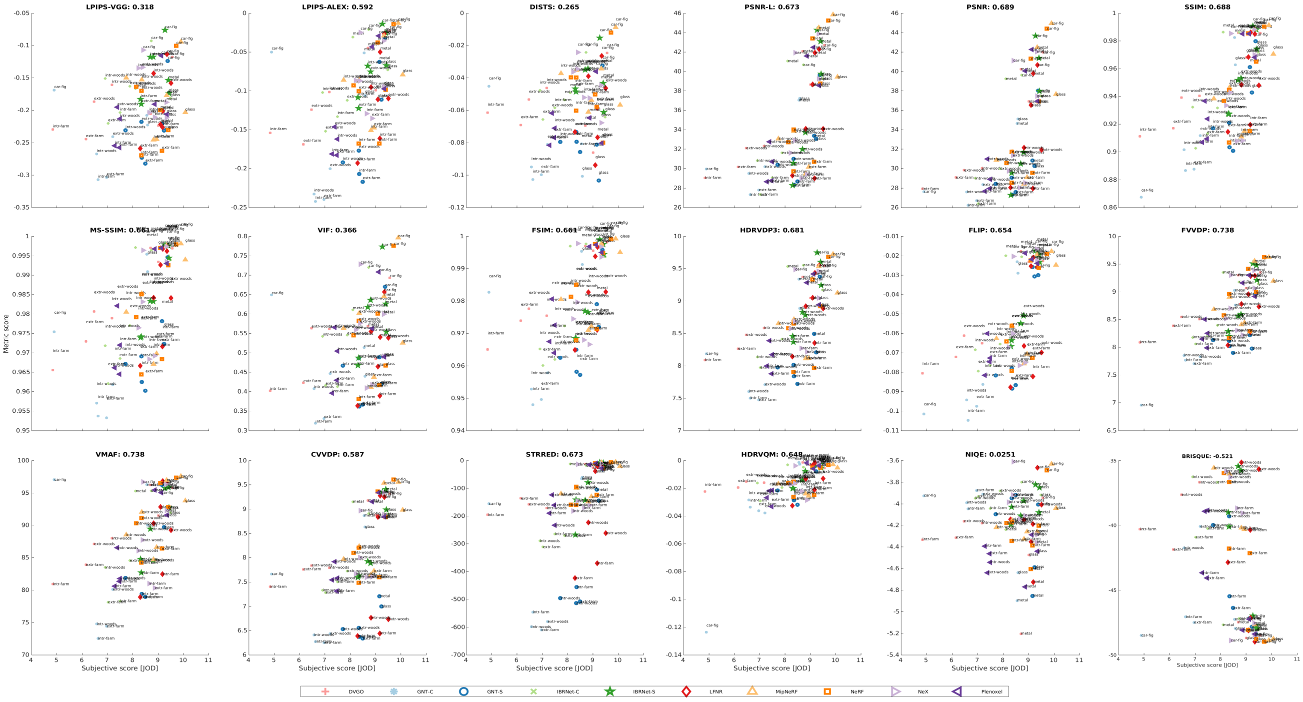

Figure 6: *Scatter plots of per-scene metric predictions vs. subjective scores for the* Lab *dataset. The subjective scores of each scene are shifted such that Reference videos have jod values equal to 10. The numbers above each plot show Spearman correlation. Note that the correlation reported in other plots has been computed on the metric predictions and subjective scores averaged across all scenes.*

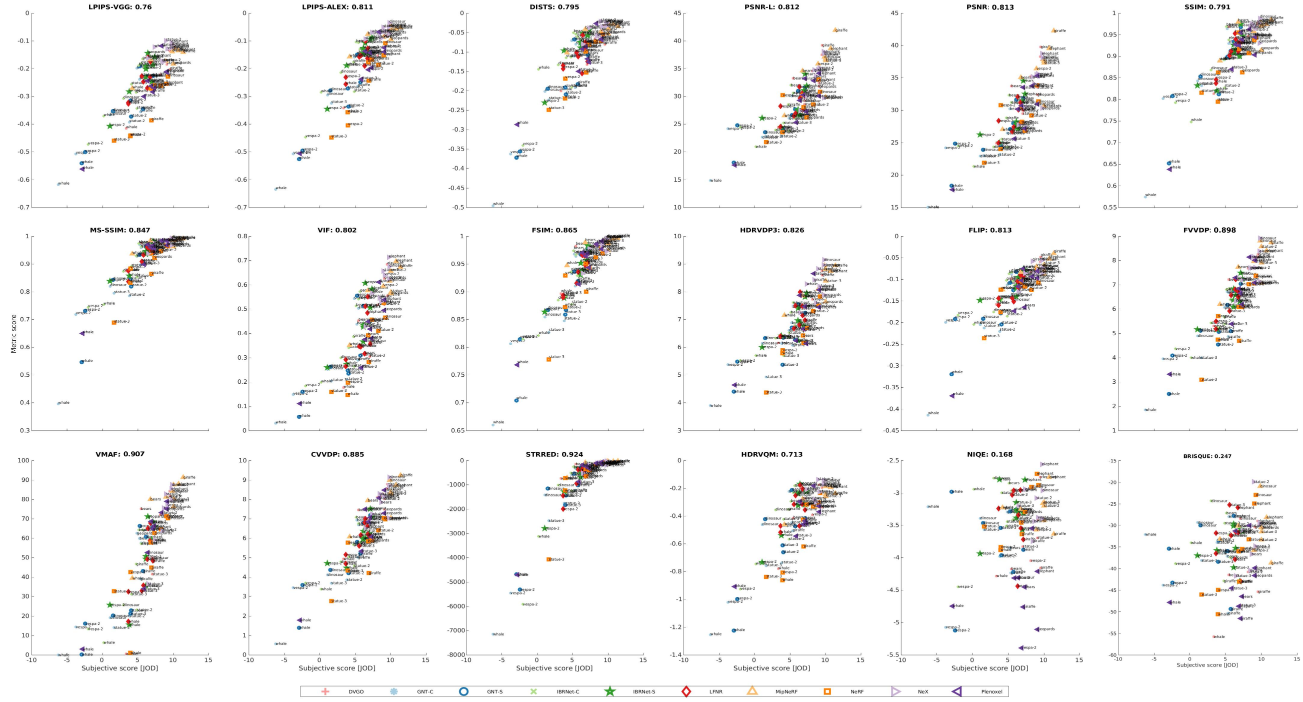

Figure 7: *Scatter plots of per-scene metric predictions vs. subjective scores for the* Fieldwork *dataset. The notation is the same as in Figure 6.*

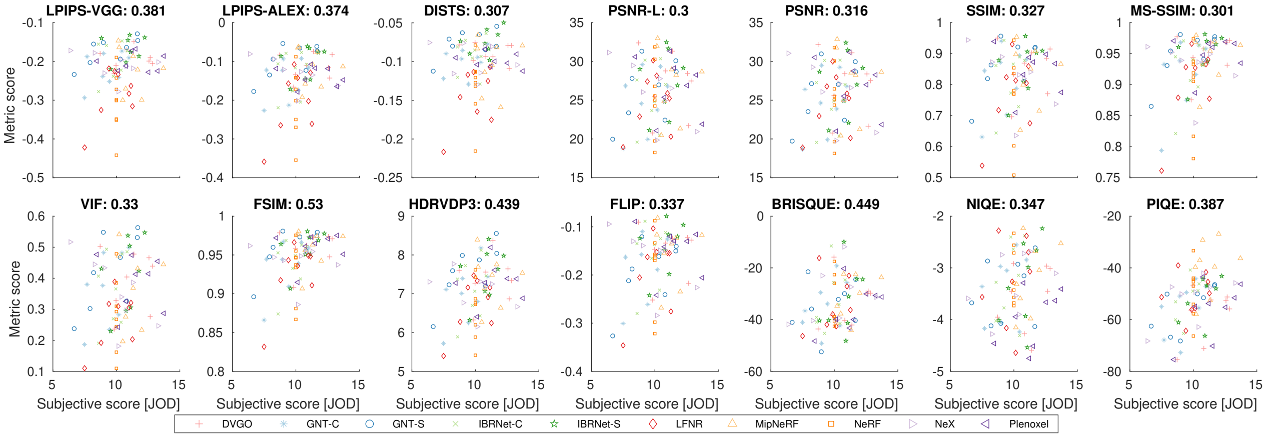

Figure 8: *Scatter plots of per-scene metric predictions vs. subjective scores for the LLFF dataset. The subjective scores of each scene are shifted such that NeRF results have jod values equal to 10. The notation is the same as in Figure 6.*

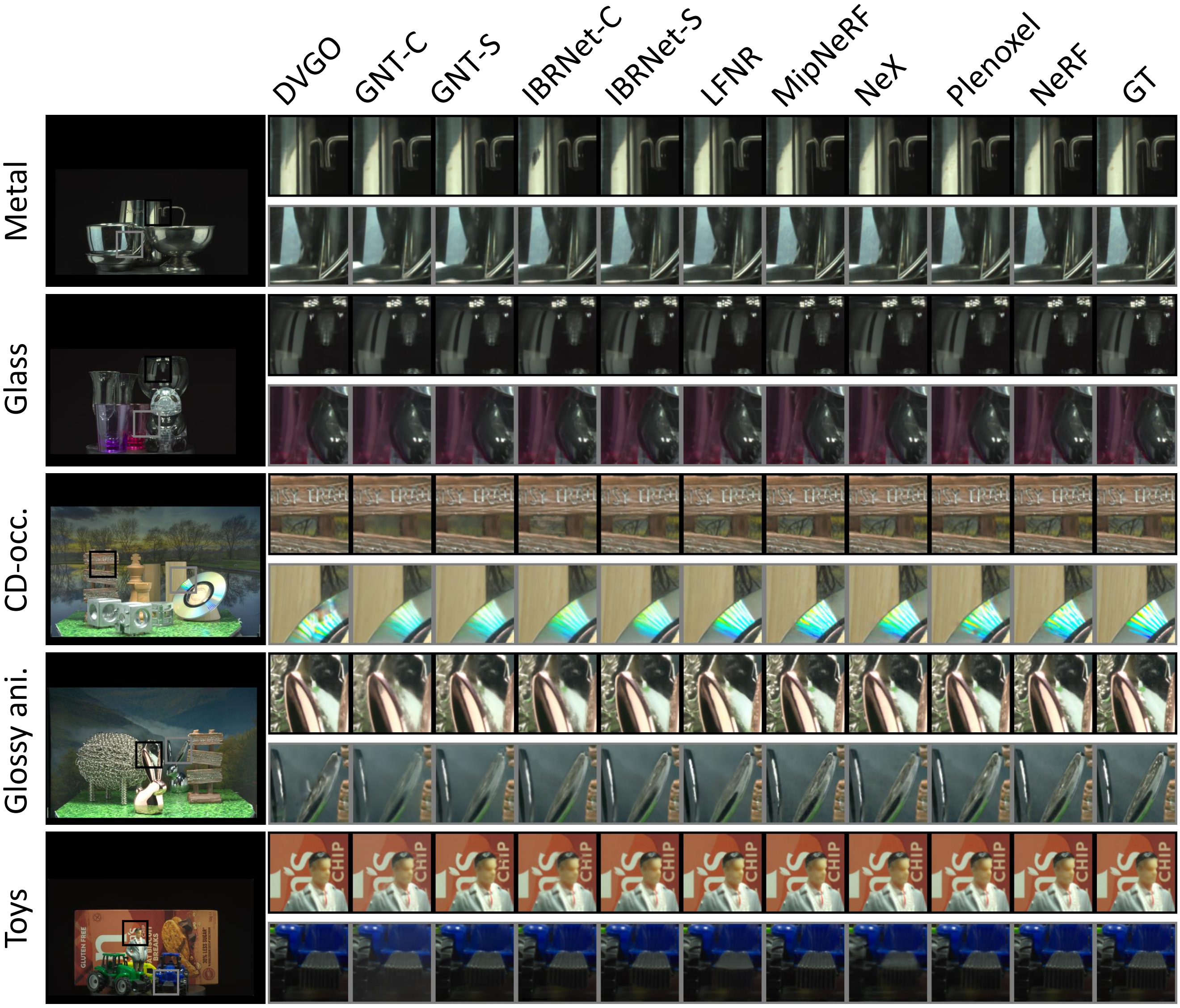

Figure 9: *Example inference results for Lab dataset.*

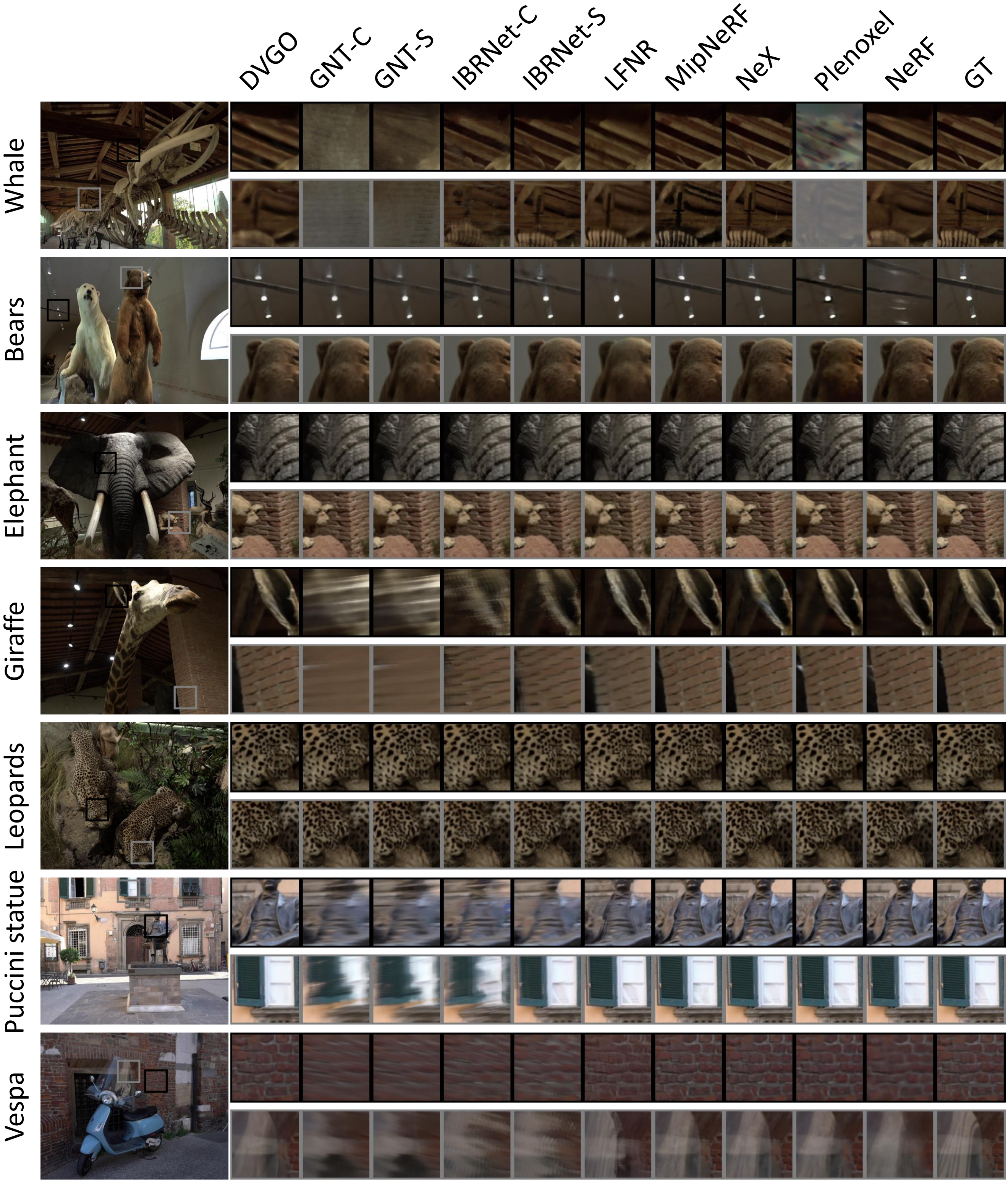

Figure 10: *Example inference results for Fieldwork dataset.*