MaRINeR: Enhancing Novel Views by Matching Rendered Images with Nearby References

– Supplementary Material –

Lukas Bösiger[1], Mihai Dusmanu[2], Marc Pollefeys[1,2], and Zuria Bauer[1]

[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Microsoft Mixed Reality & AI Lab, Zurich, Switzerland
https://boelukas.github.io/mariner/

The supplementary material contains the following sections:

- ◇ Sec. A provides details about the generation of the training dataset.
- ◇ Sec. B presents additional qualitative results.
- ◇ Sec. C brings additional ablation studies, notably regarding:
  - Weights of perceptual and adversarial losses
  - Architecture
  - Data augmentation
  - Iterations
- ◇ Finally, Sec. D describes the metrics used for evaluation:
  - Peak Signal-to-Noise Ratio (PSNR)
  - Structural Similarity Index Measure (SSIM)
  - Edge Restoration Quality Assessment (ERQA)
  - Learned Perceptual Image Patch Similarity (LPIPS)

A Training dataset

To render the images from the meshes we use RayBender [4]. Because the dataset created by LaMAR [13] contains large rendering artifacts we filter the data first to remove renderings that are only artifact or not recognizable. We do this by calculating a homography error in a similar way as SuperPoint [3]. While this is not an accurate way of assessing whether the localization was successful, it is enough to filter out the renderings with many artifacts. We estimate the homography based on SuperPoint [3] features with SuperGlue [12] matches. Ideally the homography should be identity. The homography error uses the estimated homography to remap the corners of the image. If the corners end up at their original position, the homography is close to identity and the error is close to zero. Using this method we filter out 32 % of the data.

B Additional visual results

We show additional results for the qualitative comparison in Fig. Sup. 1 and Fig. Sup. 2. Fig. Sup. 3 shows further results of the validation of localization

Table Sup. 1: Left - Encoders Results of using different encoders. **Middle - Decoders** Results of replacing SAMs and DRAMs in the decoder. **Right - Inference time** Impact of the number of iterations on the inference time.

Encoder	CAB	
	PSNR	SSIM
Learned 64	**19.88**	**0.687**
Learned 64 128 256	19.66	0.685
VGG	19.08	0.653

Decoder	CAB	
	PSNR	SSIM
SAM + DRAM	**19.88**	**0.687**
No SAM	19.73	0.685
No DRAM	19.71	0.676

iterations #	Runtime (ms)
1	49.2
2	66.3
3	88.7
4	110.5

pseudo-ground-truth. Fig. Sup. 4 shows more enhanced synthetic trajectories and Fig. Sup. 5 shows further results of enhanced NeRF renderings. Fig. Sup. 6 shows additional results on the 12 Scenes [16] dataset and Fig. Sup. 7 shows additional results of enhancing greyscale renderings using references captured by a HoloLens 2 device. Fig. Sup. 8 shows zero-shot prediction results for renderings of the image-based rendering method IBRnet [17].

C Additional ablation studies

Further ablation studies are performed on the influence of the perceptual and adversarial loss weights and the impact of different encoders and decoders. Finally, we provide further results extending the data-augmentation and iterative refinement ablations.

Influence of the perceptual loss. Because the task of novel view enhancement is different from RefSR, we investigate the effectiveness of the commonly used perceptual loss on our task. Fig. Sup. 10 $\lambda_{\text{per}} = 0$ shows that without the perceptual loss, fine geometric structure like the texture of the box are not correctly transferred. Increasing the weight to $\lambda_{\text{per}} = 0.02$ and $\lambda_{\text{per}} = 0.1$, we observe increased texture details. A higher perceptual weight $\lambda_{\text{per}} = 0.5$ leads to grid like artifacts [8] which are more visible in image regions where the correspondence matching is less confident. The extreme case can be observed for $\lambda_{\text{per}}^{\text{MASA}}$ using the same perceptual loss as MASA-SR [11]. Fig. Sup. 10 shows that $\lambda_{\text{per}} = 0.1$ increases the details optimally while introducing minimal artifacts which is also confirmed numerically in Fig. Sup. 9a.

Influence of the adversarial loss. Using the perceptual loss can lead to grid-like artifacts [8]. To remove those and make the images more visually pleasing [2, 11] we use the adversarial loss. Fig. Sup. 11 shows the impact of different weights for the loss. $\lambda_{\text{adv}} = 0$ contains the artifacts from the perceptual loss. $\lambda_{\text{adv}} = 0.005$ removes those artifacts completely but introduces high frequency details not present in the reference. Fig. Sup. 9b shows that also the scores decrease with higher adversarial loss weight. We found that with $\lambda_{\text{adv}} = 0.001$ the perceptual loss artifacts are removed while minimal new details are wrongly introduced.

Encoder. An important part of the model performance is whether the matching between rendering and real image is successful. This matching is performed on the features from the encoder. MASA-SR [11] uses features trained end-to-end

with the super resolution task which has the advantage that the features are tailored for the task. Another option is to use pre-trained features [10, 20]. If we use for example VGG features, we can leverage that those models were trained on a much larger dataset and the features potentially generalize better. Fig. Sup. 12 shows an overview over alternative encoders. The first encoder is trained end-to-end like ours but increases the feature dimension with each stage. The second one uses a pre-trained VGG16 [14] encoder where we use the `relu1_1`, `relu2_2` and `relu3_3` features. Tab. Sup. 1 validates the choice of the encoder used in our architecture.

Decoder. We show that the SAM and DRAM blocks are also applicable for the task of novel view enhancement. For this we train two models, where in the first one the decoder has no SAMs. In the second model, the DRAMs are replaced by simply concatenating the features and merging them using a convolution. Tab. Sup. 1 shows that the scores are the best using both DRAMs and SAMs.

Data augmentation. We show the visual impact of the random reference level data-augmentation in Fig. Sup. 13. The impact on the visual results of the mesh quality data-augmentation is shown in Fig. Sup. 14. This leads to increased robustness against meshes of various qualities, as v visualized in Fig. Sup. 15.

Iterations. We show the effect on the PSNR and SSIM scores of different numbers of iterations in the iterative refinement process in Fig. Sup. 16. The impact on the inference time is shown in Tab. Sup. 1.

D Metrics

We provide the definitions of the metrics used to evaluate our model. The metrics are calculated between the ground truth image $\mathbf{I}_{\mathrm{GT}}$ and the enhanced rendering $\mathbf{I}_{\mathrm{ER}}$.

Peak Signal-to-Noise Ratio (PSNR). The PSNR [5] is defined as

$$\begin{aligned} \mathrm{PSNR}(\mathbf{I}_{\mathrm{GT}}, \mathbf{I}_{\mathrm{ER}}) &= 10\log_{10}\left(\frac{255^2}{\mathrm{MSE}(\mathbf{I}_{\mathrm{GT}}, \mathbf{I}_{\mathrm{ER}})}\right) \\ \mathrm{MSE}(\mathbf{I}_{\mathrm{GT}}, \mathbf{I}_{\mathrm{ER}}) &= \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(\mathbf{I}_{\mathrm{GT}}(i,j) - \mathbf{I}_{\mathrm{ER}}(i,j))^2 \end{aligned} \tag{1}$$

where the MSE is the mean squared error.

Structural Similarity Index Measure (SSIM). The SSIM [18] is calculated on two equally sized windows $x \subset \mathbf{I}_{\mathrm{GT}}$ and $y \subset \mathbf{I}_{\mathrm{ER}}$

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2}$$

where $c_1 = (0.01 \cdot 255)^2$ and $c_2 = (0.03 \cdot 255)^2$. The formula is based on three components that measure the difference between x and y in terms of luminace, constrast and structure.

Edge Restoration Quality Assessment (ERQA). The ERQA [6] finds edges in $\mathbf{I}_{\mathrm{GT}}$ and $\mathbf{I}_{\mathrm{ER}}$ using the Canny algorithm [1]. Those edges are compared using the F1 score

$$\text{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \text{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{F}_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

where TP (True Positive) are the number of pixels detected as edge in both $\mathbf{I}_{\mathrm{GT}}$ and $\mathbf{I}_{\mathrm{ER}}$. FP (False Positive) is the number of pixels detected only in $\mathbf{I}_{\mathrm{ER}}$, FN (False Negative) are pixels only detected in $\mathbf{I}_{\mathrm{GT}}$. To account for networks that produce small edge shifts either globally over the entire image or locally, ERQA builds in compensations to match the pixels of those edges before calculating the $F1$ score.

Learned Perceptual Image Patch Similarity (LPIPS). The LPIPS [19] uses deep neural networks as feature extractor and trains a similarity predictor network based on the feature difference of the images at several layers.

$$\mathrm{LPIPS}(\mathbf{I}_{\mathrm{GT}}, \mathbf{I}_{\mathrm{ER}}) = \sum_l \mathcal{G}_l \left(\frac{1}{H_l W_l} \sum_i^{H_l} \sum_j^{W_l} \| w_l \odot (\phi_l(\mathbf{I}_{\mathrm{GT}})_{i,j} - \phi_l(\mathbf{I}_{\mathrm{ER}})_{i,j} \|_2^2) \right) \tag{5}$$

where ϕ_l denotes the output of layer l of the pretrained AlexNet [9]. LPIPS uses layers `conv_1` to `conv_5`. $\mathcal{G}_l$ is the trained prediction network for layer l, $\odot$ stands for scaling the activations channel-wise by a vector w_l.

Fig. Sup. 1: Qualitative comparison – 1. Additional results of comparing MASA [11], DATSR [2] (RefSR) and NNST [7] (ST) with **MaRINeR** on the task of novel view enhancement.

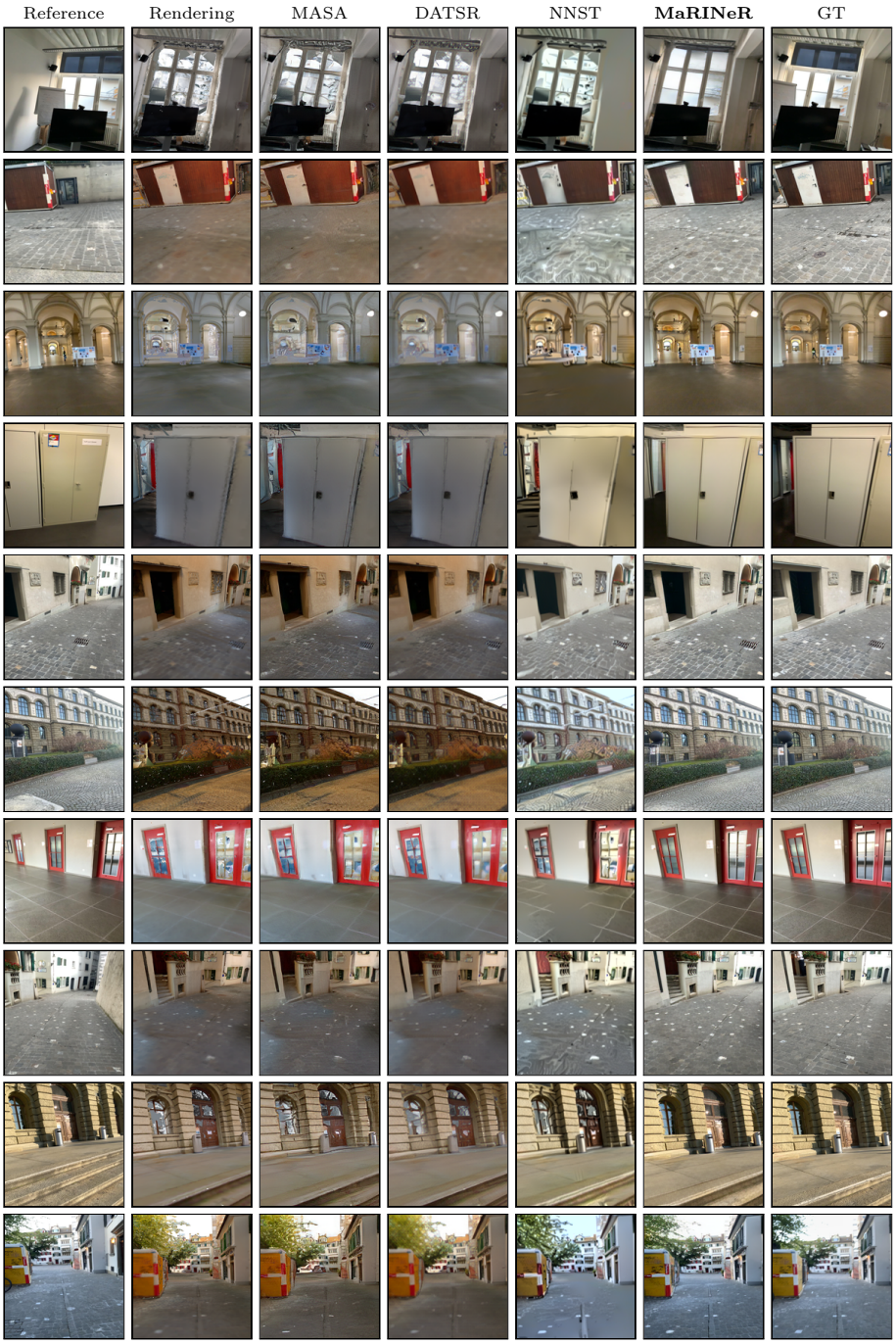


Fig. Sup. 2: Qualitative comparison – 2. Additional results of comparing MASA [11], DATSR [2] (RefSR) and NNST [7] (ST) with **MaRINeR** on the task of novel view enhancement.

Fig. Sup. 3: Further homography estimation results. Using enhanced renderings of MARINeR, estimating a homography to the aligned source image is more accurate and can be used to automate manual sanity checks in the LaMAR [13] pipeline.

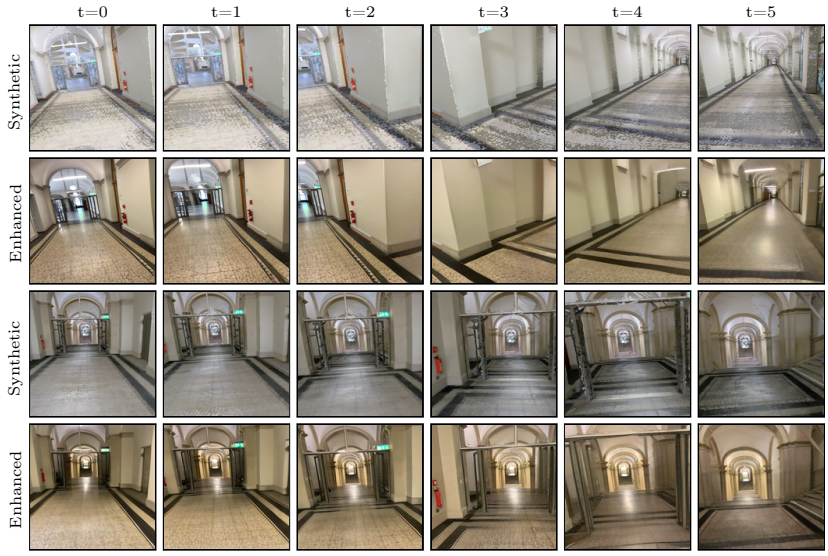


Fig. Sup. 4: Enhancing synthetic trajectories with nearby localized images. Additional results showing that because of the increased realism, the results from **MaRINeR** can extend the current dataset without introducing a gap between synthetic and human recorded trajectories.

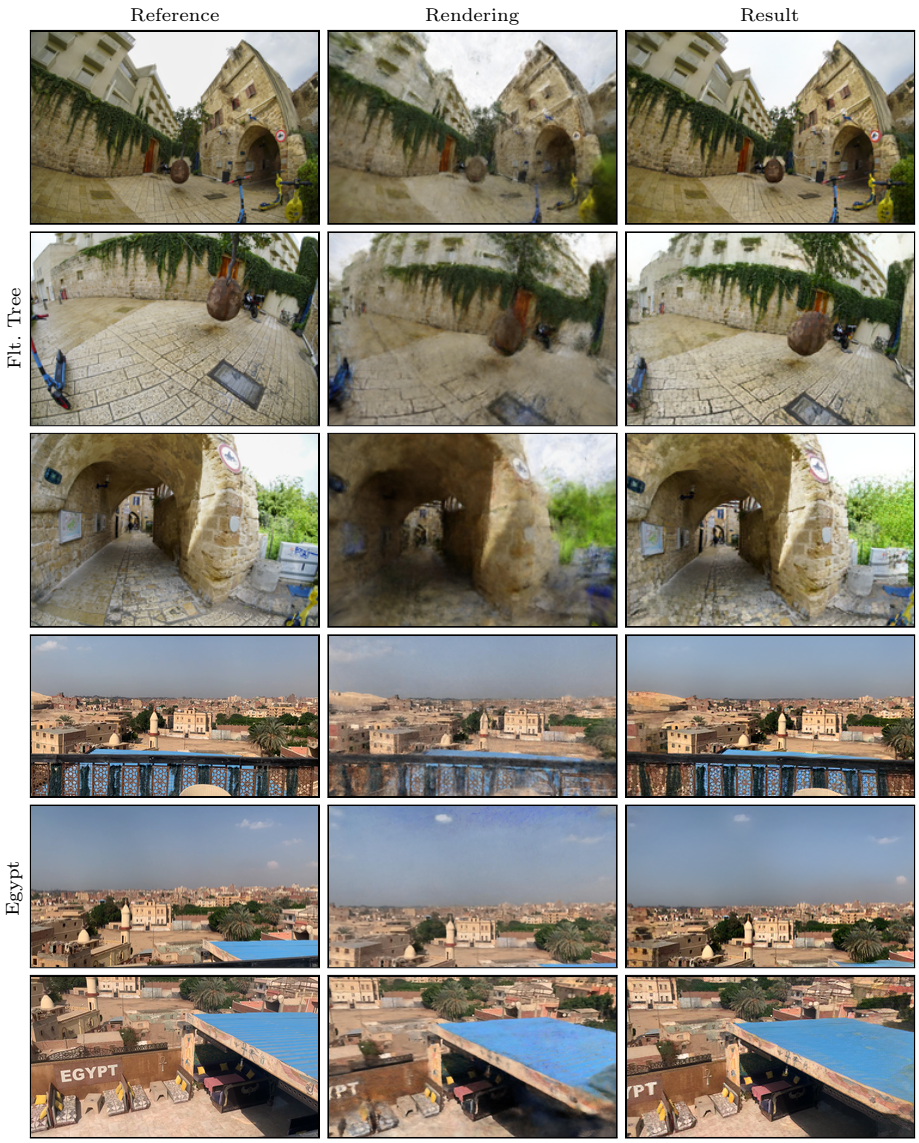


Fig. Sup. 5: Additional NeRF postprocessing results. Training a nerfacto [15] model on the Floating tree and Egypt data. We use the smallest nerfacto model and the result contains artifacts which our model can successfully remove.

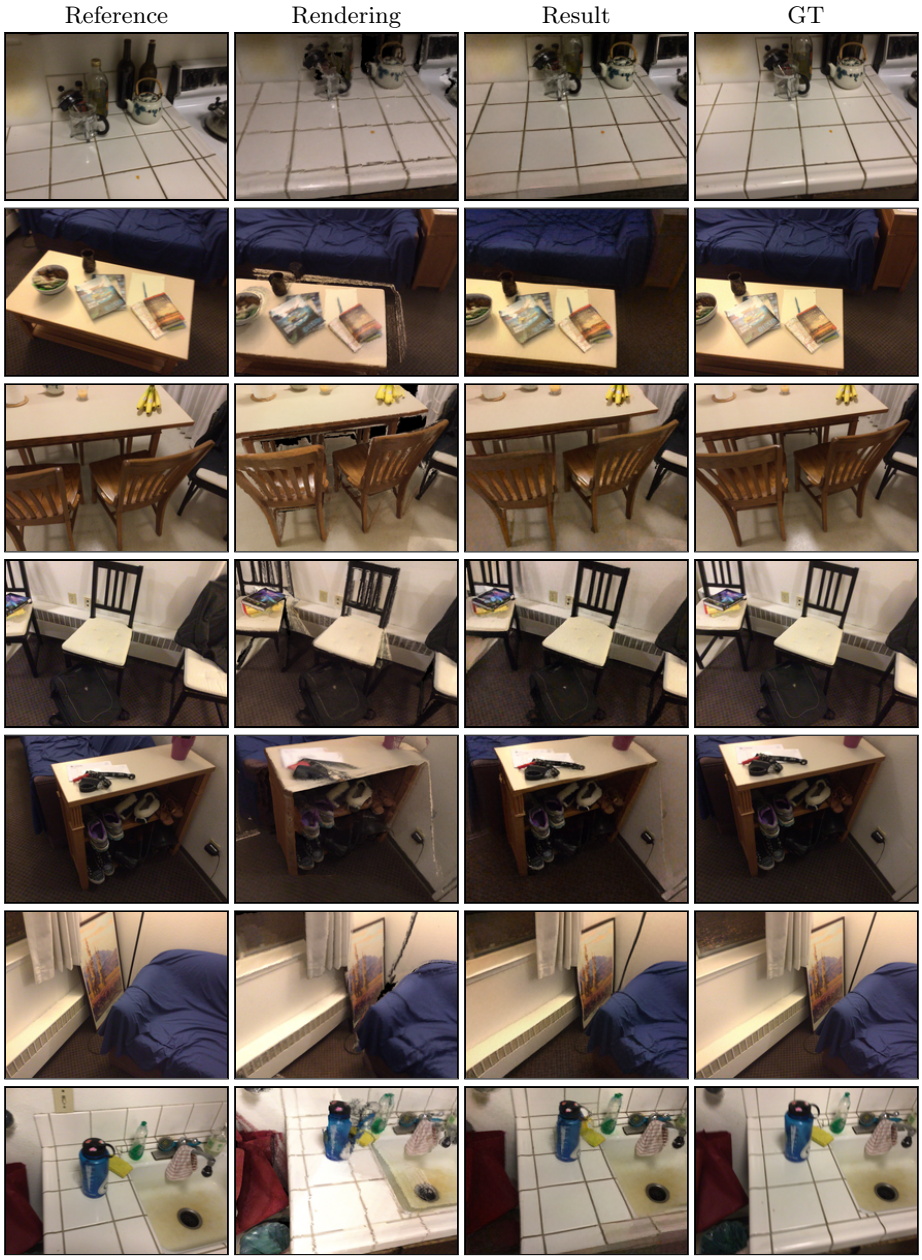


Fig. Sup. 6: 12 Scenes results. Evaluating the model on the apartment 1 kitchen and living scenes of the 12 Scenes [16] dataset shows that **MaRINeR** also enhances renderings of 3D reconstructions created by a different algorithm than the one used by LaMAR [13].

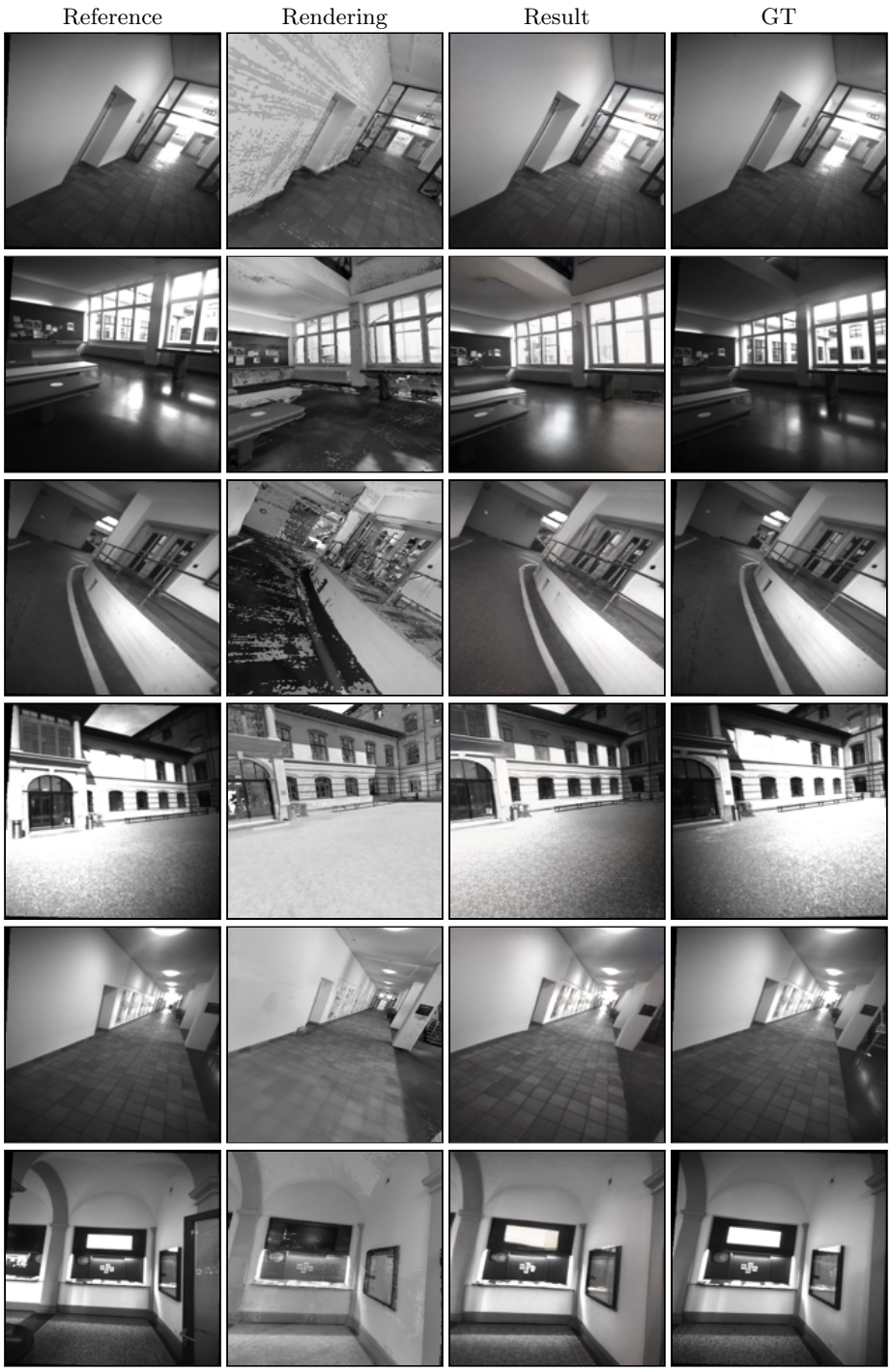


Fig. Sup. 7: HoloLens 2 results. Enhancing greyscale renderings using references recorded by a HoloLens 2 device.

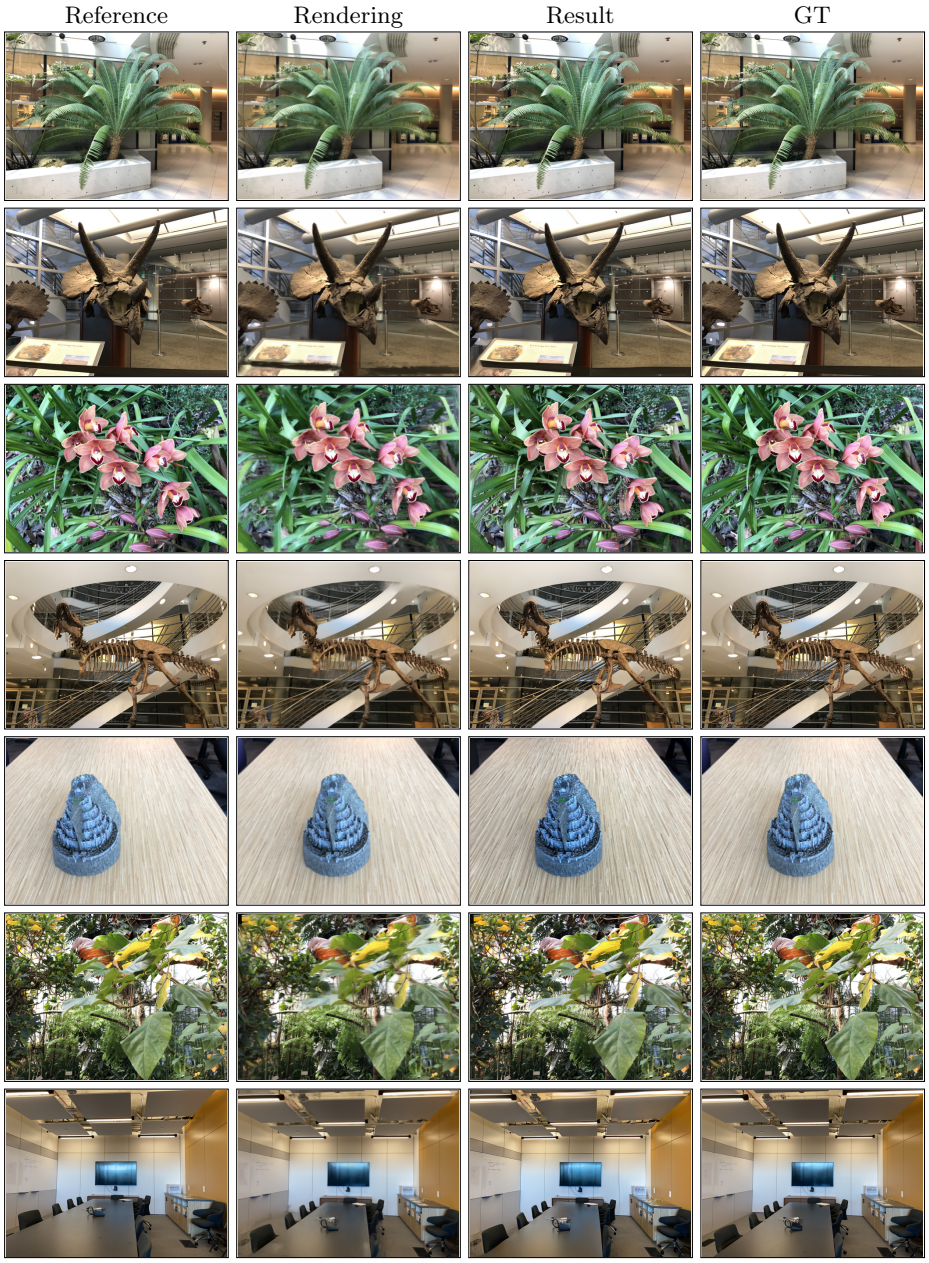
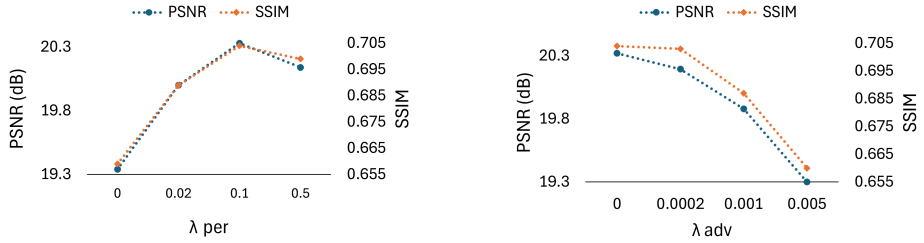


Fig. Sup. 8: Enhancing novel view renderings created by the image-based rendering method IBRNet [17] using our model without retraining.



(a) Perceptual loss weight influence

(b) Adversarial loss weight influence

Fig. Sup. 9: Influence of the loss weights. The results of our experiments finding the optimal weights for a the perceptual loss and b the adversarial loss.

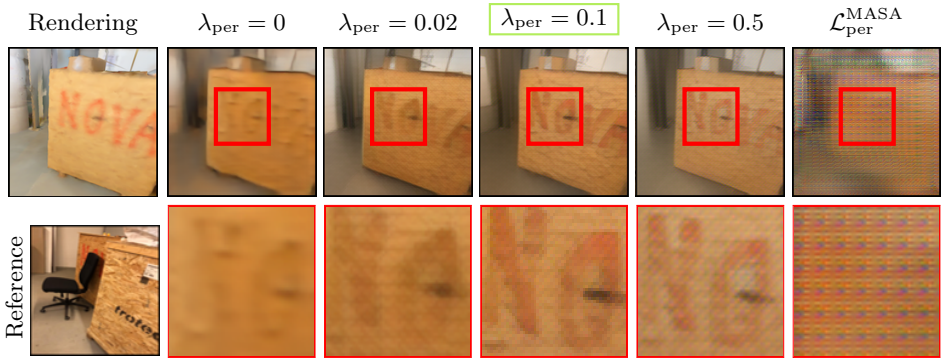


Fig. Sup. 10: Impact of the perceptual loss. Increased weight enhances details and the visual quality but also introduces perceptual loss specific grid-like artifacts.

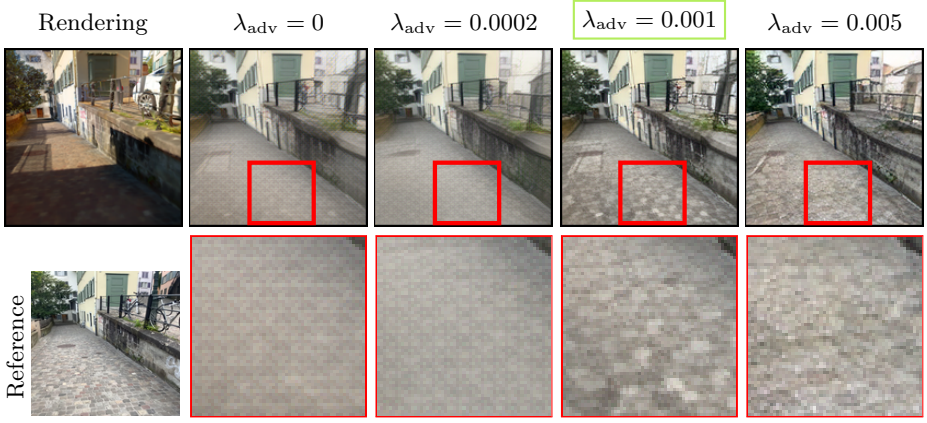


Fig. Sup. 11: Impact of the adversarial loss. Increased weight removes the perceptual loss artifacts and keeps the underlying texture. Increasing the weight too much leads to the introduction of hallucinated details not present in the reference.

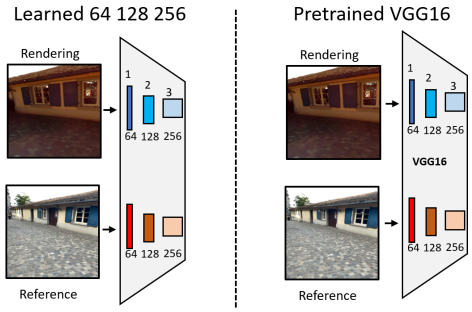


Fig. Sup. 12: Architecture of alternative encoders. We validate the choice of our encoder by comparing it against an end-to-end trained encoder with larger feature channels and an encoder using pre-trained VGG16 [14] features.

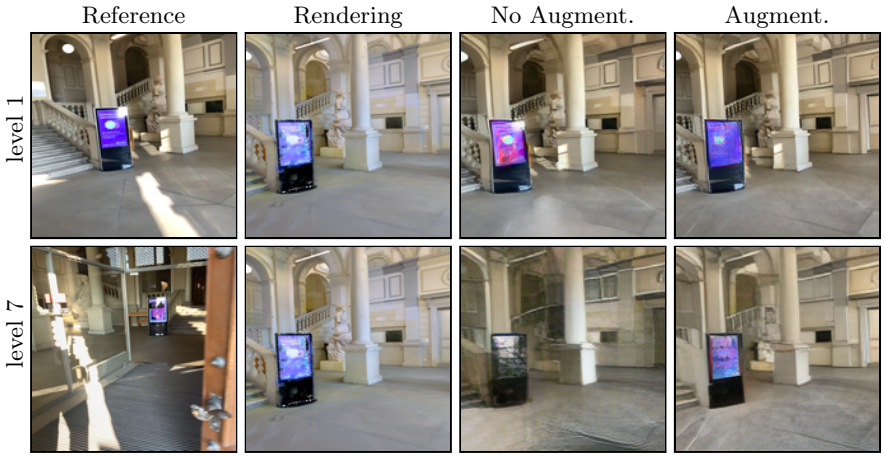


Fig. Sup. 13: Ref. level data augmentation. Impact of using random close-by images as reference instead of only the closest one. While the model performs similar for ref. level 1, the correspondence matching fails for ref. level 7 leading to worse results.

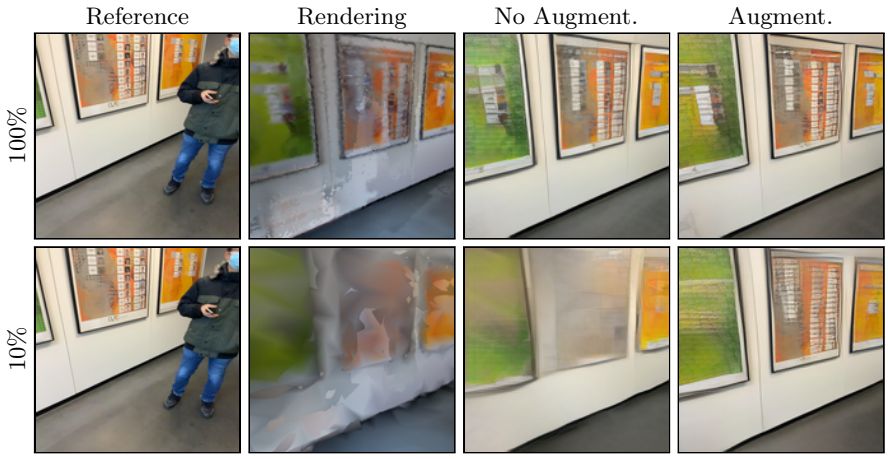


Fig. Sup. 14: Mesh quality data-augmentation Comparison of the model with and without augmenting the data with renderings from a down-sampled mesh. While for a mesh size of 100% the results are visually similar, for the mesh size of 10% the non augmented model fails to find correspondences and the result lacks the details from the reference.

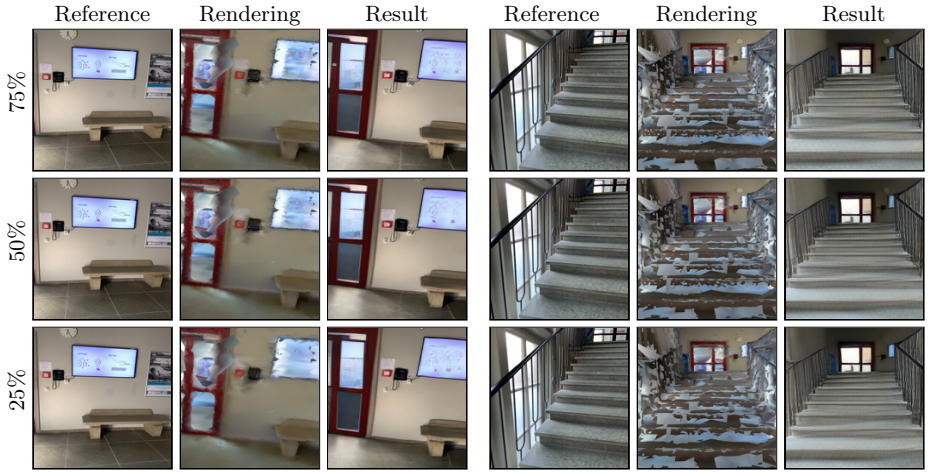


Fig. Sup. 15: Results of our model on low quality meshes. The visual quality of the results stays high even if the mesh size is reduced to 75% and 25% of the original mesh triangles.

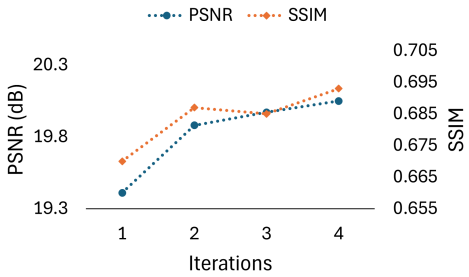


Fig. Sup. 16: Impact of the number of iterations. Increasing the number of iterations leads to better results. The largest improvement can be seen between 1 and 2 iterations.

References

1. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-8**(6), 679–698 (1986). https://doi.org/10.1109/TPAMI.1986.4767851
2. Cao, J., Liang, J., Zhang, K., Li, Y., Zhang, Y., Wang, W., Gool, L.V.: Reference-based image super-resolution with deformable attention transformer (2022)
3. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description (2018)
4. Dusmanu, M., Sarlin, P.E., Speciale, P.: Raybender - fast python ray-tracing. https://github.com/cvg/raybender (2021)
5. Horé, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. pp. 2366–2369 (08 2010). https://doi.org/10.1109/ICPR.2010.579
6. Kirillova., A., Lyapustin., E., Antsiferova., A., Vatolin., D.: Erqa: Edge-restoration quality assessment for video super-resolution. In: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2022)
7. Kolkin, N., Kucera, M., Paris, S., Sykora, D., Shechtman, E., Shakhnarovich, G.: Neural neighbor style transfer (2022)
8. Krawczyk, P., Gaertner, M., Jansche, A., Bernthaler, T., Schneider, G.: Artifact generation when using perceptual loss for image deblurring (2023)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (may 2017). https://doi.org/10.1145/3065386, https://doi.org/10.1145/3065386
10. Li, Z., Kuang, Z.S., Zhu, Z.L., Wang, H.P., Shao, X.L.: Wavelet-based texture reformation network for image super-resolution. IEEE Transactions on Image Processing (2022)
11. Lu, L., Li, W., Tao, X., Lu, J., Jia, J.: Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution (2021)
12. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks (2020)
13. Sarlin, P.E., Dusmanu, M., Schönberger, J.L., Speciale, P., Gruber, L., Larsson, V., Miksik, O., Pollefeys, M.: Lamar: Benchmarking localization and mapping for augmented reality (2022)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
15. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., Mcallister, D., Kerr, J., Kanazawa, A.: Nerfstudio: A modular framework for neural radiance field development. In: Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings. SIGGRAPH '23, ACM (Jul 2023)
16. Valentin, J., Dai, A., Nießner, M., Kohli, P., Torr, P., Izadi, S., Keskin, C.: Learning to navigate the energy landscape (2016)
17. Wang, Q., Wang, Z., Genova, K., Srinivasan, P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering (2021), https://arxiv.org/abs/2102.13090
18. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402 Vol.2 (2003). https://doi.org/10.1109/ACSSC.2003.1292216

19. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
20. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer (2019)