




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

Lukas Bösiger¹, Mihai Dusmanu², Marc Pollefeys^{1,2}, and Zuria Bauer¹

¹ Department of Computer Science, ETH Zurich, Switzerland

² Microsoft Mixed Reality & AI Lab, Zurich, Switzerland

<https://boelukas.github.io/mariner/>

Abstract. Rendering realistic images from 3D reconstruction is an essential task of many Computer Vision and Robotics pipelines, notably for mixed-reality applications as well as training autonomous agents in simulated environments. However, the quality of novel views heavily depends of the source reconstruction which is often imperfect due to noisy or missing geometry and appearance. Inspired by the recent success of reference-based super-resolution networks, we propose MaRINeR, a refinement method that leverages information of a nearby mapping image to improve the rendering of a target viewpoint. We first establish matches between the raw rendered image of the scene geometry from the target viewpoint and the nearby reference based on deep features, followed by hierarchical detail transfer. We show improved renderings in quantitative metrics and qualitative examples from both explicit and implicit scene representations. We further employ our method on the downstream tasks of pseudo-ground-truth validation, synthetic data enhancement and detail recovery for renderings of reduced 3D reconstructions.

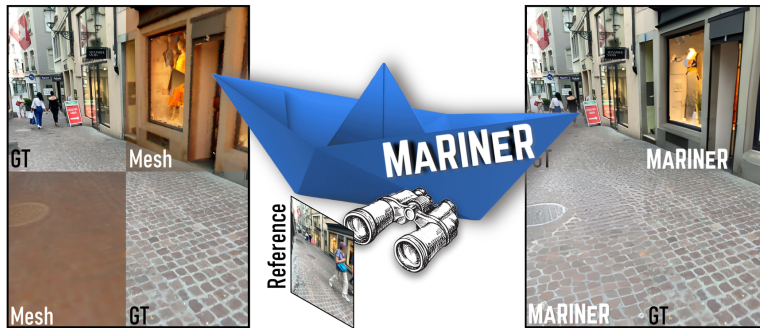


Fig. 1: We introduce **MaRINeR**: a pipeline taking as input a novel-view obtained from a 3D reconstruction exhibiting geometric and / or appearance artifacts and inaccuracies as well as a nearby reference used during the reconstruction process, and outputting an enhanced version of the novel-view through feature matching and transfer.

1 Introduction

One of the fundamental problems of computer vision and robotics is reconstructing the environment from sensorial data such as color or depth cameras or LiDAR scanners. These pipelines produce a computer-friendly representation of the space which can be either explicit (e.g., point-clouds, meshes), implicit

(e.g., occupancy nets [30, 32], neural radiance fields – NeRF [31]), or hybrid (e.g., Gaussian splats [19]) which serve as starting point for many subsequent tasks, notably novel-view synthesis, environment understanding, planning, and navigation. All existing methods have limitations: point-clouds are highly dependent on the sensor quality [57], often contain artifacts due to moving objects [26], and are not suitable for occlusion checking [2] or pattern rendering. Meshing algorithms often create both appearance and geometric artifacts and inconsistencies while connecting the vertices and coloring / texturing the polygons [4, 6, 14, 35, 40, 41]. More modern implicit methods show exemplary rendering performance but often require very densely sampled frames or even depth maps which are not always available [36]. Furthermore, these methods also need extensive per-scene training. The performance decreases drastically as the frame-rate and the input modalities are reduced. Any artifacts or inconsistencies produced by the reconstruction pipeline can lead to significant impact in downstream tasks. To create renderings of novel views without 3D scene representation, image based rendering methods, such as IBRnet [48], learn to interpolate between existing views of the 3D scene. However, the novel view renderings can also contain artifacts such as blurry image parts or noisy geometry. While methods exist to remove such artifacts for a specific type of pipeline, such as NeRFLiX [67] for neural radiance fields, these techniques often lack the ability to remove artifacts produced by other types of pipelines.

To address these limitations, we propose a post-processing step of novel rendered views entitled **MaRINeR** by *Matching the Rendered Images with Nearby References*. To this end, we make further use of input images to the reconstruction process as reference data. Our task is strongly connected with Reference-based Super-Resolution (RefSR) since similar to renderings from low-quality or noisy 3D reconstructions, a naively up-scaled version of a low-resolution (low-res.) image lacks details. RefSR methods use details present in a closely related high-resolution (high-res.) reference image to help super-resolve the low-res. image. We notice that the methods used to match between low-res. and high-res. image domains for information transfer and fusion can more generally be used to transfer details from a reference to a related image of any nature. However, the classically used CUFED5 [62] dataset is not suitable for our task of novel view enhancement. We therefore generate new training and test datasets building upon the recently released LaMAR [38] dataset.

The enhanced novel views show promising results for different downstream tasks. First, our method quantitatively and qualitatively narrows the gap between renderings and real images, including these of implicit representations. As a by-product, this improves the quality of data obtained when using digital twins for training reinforcement agents, following the recent advances in ego-centric human synthetic data generation [23]. Second, many recent datasets (12 Scenes [46], RIO10 [47], LaMAR [38]) designed pseudo-ground-truth pipelines to automatically and accurately register trajectories from various devices, notably in the context of mixed reality experiences. One common way to validate the accuracy of these pipeline is through qualitative checks between the aligned images

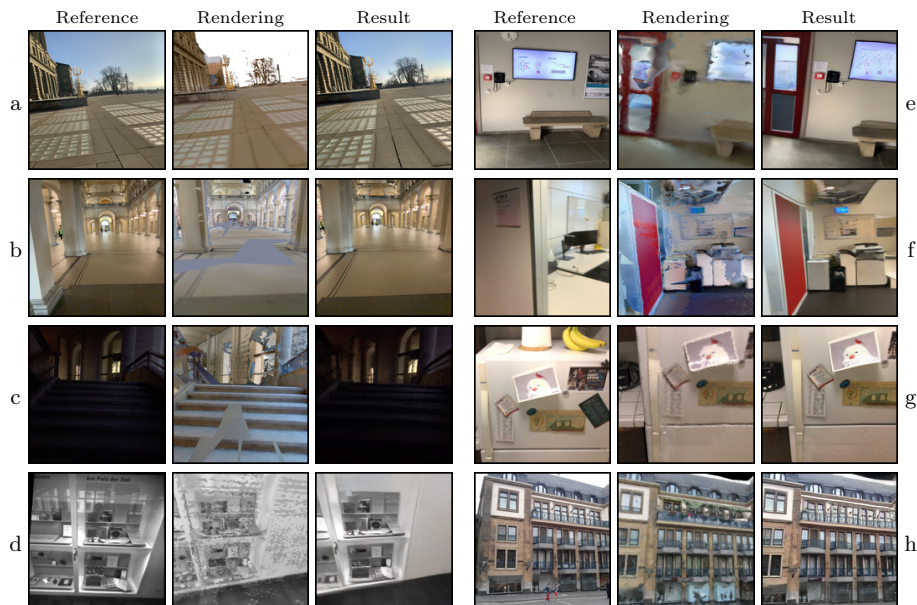


Fig. 2: Robustness of MaRINeR. Our model recovers missing parts that appear due to rendering artifacts **a**, **b**. It adopts the illumination from the reference **c**, is device agnostic generalizing to gray-scale images **d**. The model enhances renderings of low triangle meshes **e** and also improves the rendering even if the reference has little content in common **f**. It can be applied to unseen scenes such as 12 Scenes [46] **g** or Aachen Day-Night [39] **h** without retraining.

and an associated rendering from a 3D reconstruction. Narrowing the render-to-real gap by removing the artifacts and improving the textural accuracy and realism opens the door to automating this process by taking advantage of existing geometric methods to estimate the accuracy of the ground-truth (GT). Third, for any downstream application, decisions have to be made to reduce the size of 3D representations in order to efficiently process them at the cost of reduced detail accuracy and realism. Our method is capable of recovering the lost details and realism of such 3D reconstructions as illustrated in Fig. 2. We will release the source code of our paper upon acceptance.

To summarize, our contributions are:

- We introduce **MaRINeR** which, to the best of our knowledge, is the first method enhancing novel views by using a close-by reference image that is applicable for renderings from a wide range of 3D reconstruction pipelines.
- An extensive evaluation of the proposed method is performed, providing not only a qualitative and quantitative analysis of the method but also an overview of the robustness of **MaRINeR** to different datasets, temporal conditions and temporal changes in scenes, while discussing limitations.
- We showcase the excellent performance of our model in several applications: elimination of manual checks in pseudo-GT pipelines, improvement of synthetic AR trajectories, and enhancing the output of neural renderings.

2 Related work

We provide an overview of related research fields which also incorporate information from a reference image into a target image, notably reference-based image super-resolution and style transfer.

Reference-based image Super-Resolution (RefSR). The goal of RefSR is to recover high-res. from low-res. images by transferring missing details from high-res. reference images. The methods usually work by aligning and fusing features extracted from low-res. and ref. images. While early work uses hand crafted features [56], more recent works use either pre-trained features [25, 62] or train the feature extraction end-to-end with the task [11, 16, 28, 42, 52, 54, 64, 65]. The alignment of the ref. and low-res. features proposes a challenge because of the resolution difference. Some methods use implicit alignment: CrossNet [65] estimates the optical flow between ref. and low-res. images. Because optical flow fails at capturing long distance correspondences, SSEN [42] utilizes deformable convolutions which ensure a large receptive field. Other work uses explicit alignment by feature or patch matching. SRNTT [62] uses feature similarity and transfers textures from the ref. images at different scales. To reduce the computational complexity, MASA [28] proposes a coarse-to-fine correspondence matching module. C^2 -Matching [16] introduces knowledge distillation and contrastive learning methods to improve the matching between low-res. and ref. despite the resolution gap. WTRN [25] uses wavelets to separate high and low frequency parts of the images, which helps to more transfer more visually plausible texture patterns. DATSR [3] uses Swin-Transformers [27] to replace the commonly used residual blocks [13], leading to more robust matches and texture transfer. HMCF [49] improves the matching between low-res. and ref. of similar objects with different texture by using high-to-low-level feature matching and complementary information fusion. RRSGAN [11] uses generative adversarial networks and deformable convolutions. FRFSR [29] notes that the commonly used perceptual and adversarial loss have an adverse effect on texture transfer and reconstruction. As a solution, they propose the use of a texture reuse framework. RRSR [59] uses a reciprocal learning strategy to strengthen the training process by using the super resolution result as reference to help super-resolve a low-res. variant of the original high-res. reference. CMRSR [66] notes that due to the gap between inputs and reference, the super resolution image often yields distortions and ghosting artifacts and they propose a contrastive attention-guided multi-level feature registration module to mitigate those. There are also methods that use multiple references as input such as CIMR-SR [53], AMRSR [33], AMSA [63] or LMR [58]. We notice that many of ideas to align ref. and low-res. features are not limited to align images with resolution differences but can be used more broadly to also align rendered images to real images.

Style Transfer (ST). Artistic style-transfer methods transfer the style of a style image to a content image. A subcategory is the universal ST methods [1, 15, 24], which transfer any style to the content image. This can be done by separating content and style information in the images. AdaIN [15] transfers

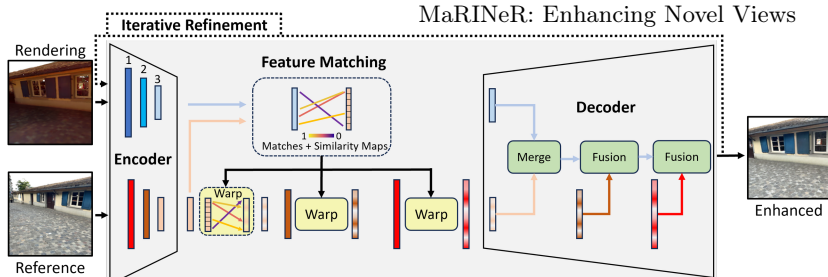


Fig. 3: MaRINeR architecture. The learned features of the encoder are used to for correspondence matching and warping of the reference features. They are fused with the rendering features to create a enhanced rendering, which is iteratively refined.

channel-wise mean and variance feature statistics. WCT [24] uses whitening and coloring transformations, where the whitening transformation can remove the style of the content image and the coloring transformation can incorporate the style of the style image. However the separation of content and style is challenging and some content can be corrupted. ArtFlow [1] calls this issue content leak and introduces a reversible neural flow-based network to avoid it. StyTr2 [9] uses transformers to extract and maintain global image information, which then help with the content leak problem. For universal ST, the style images usually have little content in common with the content image. The results look like an artistic version of the content image which is however far from being realistic. Semantic ST methods [21, 61] work with style images that contain similar objects as the content image. The goal is to build semantic correspondences between similar objects and map the style region only to the semantically similar content regions [17]. NNST [21] matches VGG [43] features between content and style and replaces the content features with the nearest style features. MST [61] uses graph cuts for matching between content and style features. While those methods work well at transferring the semantic correspondences, they can introduce distortions and don't produce photo-realistic images. Photo-realistic ST methods aim at transferring the style of the color distribution while preserving the structures of the content image [17]. WCT2 [55] adds a wavelet based correction to the whitening and coloring transforms of WCT [24]. This helps to preserve the structural and statistical properties of the VGG features during stylization. The result is a more photo-realistic image without distortions. However, photo-realistic ST assumes that the content image is already photo-realistic. If this image contains artifacts, then those are also carried over to the stylized image.

3 Method

The **MaRINeR** pipeline, illustrated in Fig. 3, takes as input a rendering with noisy appearance and geometry as well as a nearby reference and outputs an enhanced version of the rendering by transferring relevant information from the reference. The pipeline starts by densely extracting features at multiple levels from both input images with a shared convolutional encoder. Next, the deepest features extracted from the rendering are matched to those of the reference to retrieve similar content. These matches are then used to warp the reference

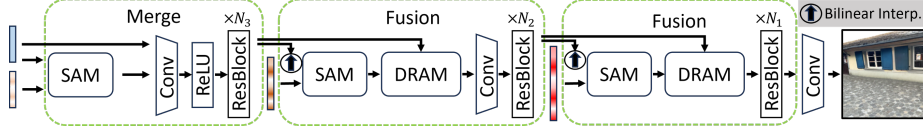


Fig. 4: Architecture of the decoder. We fuse the rendering and warped reference features using SAM [28], DRAM [28] and residual blocks [13].

features at different levels. The warped reference features are fused with those of the rendering in the decoder. Given the severe artifacts sometimes present in novel views, we employ an iterative refinement approach that repeats the process by replacing the input rendering with the enhanced output of the previous iteration. We start from the MASA RefSR [28] pipeline and implement several changes in architecture, loss function as well as data augmentation to make it amenable to the novel task of reference-based rendering enhancement.

Encoder. As mentioned above, we use a shared convolutional encoder to extract features at multiple levels from both the rendered image I and the reference R , for simplicity assumed both of size $H \times W$. We use three levels, each halving the resolution of the previous one, yielding two sets of dense tensors: $\{\mathcal{F}_1^I \in \mathbb{R}^{H \times W \times F_1}, \mathcal{F}_2^I \in \mathbb{R}^{H/2 \times W/2 \times F_2}, \mathcal{F}_3^I \in \mathbb{R}^{H/4 \times W/4 \times F_3}\}$ and $\{\mathcal{F}_1^R, \mathcal{F}_2^R, \mathcal{F}_3^R\}$ for the rendering and the reference, respectively, where F_i is the number of channels of the output of level i . These features will next be used to find corresponding patches between the rendering and the reference in a coarse-to-fine fashion.

Feature matching. We use the Matching and Extraction Module (MEM) from MASA [28] to match the deepest features of both input images, \mathcal{F}_3^R and \mathcal{F}_3^I , using cosine similarity. The MEM performs matching first on a coarse grid with a stride and then densely within a fixed-size window around the resulting matches. This step yields a mapping m of indices from the level 3 features of the rendering to those of the reference and associated matching scores s :

$$m_{I \rightarrow R} : (x, y) \in \mathcal{F}_3^I \rightarrow \{(u, v) \in \mathcal{F}_3^R, s \in \mathbb{R}\} . \quad (1)$$

This mapping is used to warp and weight the reference features at each of the three levels i , where blocks of features with size relative to the spatial resolution of the current level are cropped and moved together resulting in warped feature maps $\{\mathcal{F}_1^{R \rightarrow I}, \mathcal{F}_2^{R \rightarrow I}, \mathcal{F}_3^{R \rightarrow I}\}$. In contrast to RefSR methods which have an input with lower resolution and thus need to perform the matching on the \mathcal{F}_1 features of the low-res. input and down-scaled reference [28], we use deeper features, allowing us to leverage the increased robustness to find better quality matches. Weighting the warped features based on the matching scores reduces the impact of features with low confidence matches. This enables the model to only use the reference features if they have a confident match and otherwise use the rendering features when fusing them in the decoder.

Decoder. Using the deepest features of the rendering \mathcal{F}_3^I and the warped reference features $\mathcal{F}_3^{R \rightarrow I}$, $\mathcal{F}_2^{R \rightarrow I}$ and $\mathcal{F}_1^{R \rightarrow I}$, we fuse them using Spatial Adaptation Modules (SAM) [28], Dual Residual Aggregation Modules (DRAM) [28] and residual blocks [13], as shown in Fig. 4. SAM learns to remap the distribution of the reference features to the one of the rendering features. DRAM fuses features of different spatial resolution aiming to refine and aggregate the details of both

branches and up-sample the low-res. features with a transposed convolution. The decoder procedure can be summarized as follows:

$$\begin{aligned}\mathcal{O}_3 &= P_3(\text{SAM}(\mathcal{F}_3^I, F_3^{R \rightarrow I}) \oplus \mathcal{F}_3^I) \\ \mathcal{O}_2 &= P_2(\text{DRAM}(\text{SAM}(\mathcal{O}_3^\uparrow, \mathcal{F}_2^{R \rightarrow I}), \mathcal{O}_3)) \\ \mathcal{O}_1 &= P_1(\text{DRAM}(\text{SAM}(\mathcal{O}_2^\uparrow, \mathcal{F}_1^{R \rightarrow I}), \mathcal{O}_2))\end{aligned}\tag{2}$$

where P_i stands for processing the features using a convolution and N_i residual blocks [13], \oplus for a convolution to merge the features and \uparrow for bilinear interpolation. We merge the level 3 rendering features \mathcal{F}_3^I with the warped level 3 reference features $\mathcal{F}_3^{R \rightarrow I}$ by concatenating and processing them using a convolution, leveraging that the rendering and warped reference features are of similar spatial resolution. The SAM aligns the rendering and warped reference feature distributions such that the DRAM can successfully merge the features. The output image is then created from \mathcal{O}_1 using a convolution to reduce the feature dimension. In contrast, RefSR methods such as MASA deal with features with a different spatial resolution that can not directly be merged. MASA first fuses the low resolution features of level 3 to 1 together before merging the reference features. For the task of RefSR this is beneficial because the result is encouraged to be structurally similar to the input with only additional details from the reference. For the task of rendering enhancement where the rendering can contain structures that come from artifacts, merging the features of similar spatial resolution enables the model to also take structural information from the reference. This is beneficial to remove rendering artifacts or fill in missing image parts caused by gaps in the source 3D reconstruction.

Iterative refinement. Since the gap between the rendering and the reference can be large due to rendering artifacts that occlude the underlying geometry, we found it beneficial to apply the model several times in an iterative fashion. The first iteration can thus focus on removing artifacts and enhancing the image. The following iterations are then more successful in establishing correspondences and transferring the missing details to the enhanced rendering. To this end, we supervise the model after each iteration, thus obtaining a more general model that can deal with a wide variety of rendering qualities.

3.1 Loss function

Our goal is to preserve the spatial information of the rendering while removing artifacts, adding details from the reference, and producing a visually pleasing result. To this end, we combine a reconstruction loss, perceptual loss, and adversarial loss with associated weights λ_{rec} , λ_{per} , and λ_{adv} , written as:

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} .\tag{3}$$

Reconstruction loss. The enhanced rendering I_{ER} should be close to the GT image taken at the same pose as the rendering by using the information present in the close-by reference. We adopt the following reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{I}_{\text{GT}} - \mathbf{I}_{\text{ER}}\|_1 ,\tag{4}$$

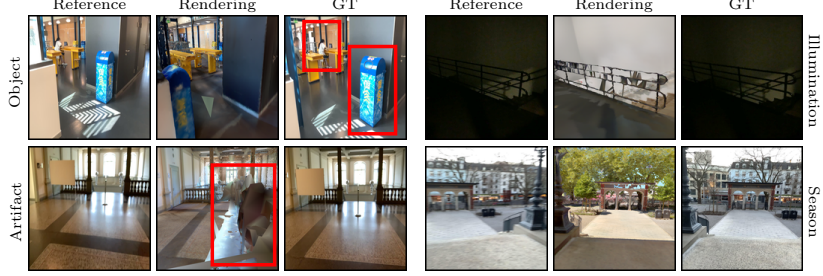


Fig. 5: Common dataset challenges. There can be different objects present between rendering and GT, some of which can be artifacts. The illumination can also be different because of day time or seasonal changes.

where $\|\cdot\|_1$ is the ℓ_1 norm.

Perceptual loss. The perceptual loss is widely used by RefSR models [3, 16, 28, 29] to enhance the visual quality of the result by guiding the resulting image to be more semantically similar to the GT. This loss is formulated as:

$$\mathcal{L}_{\text{per}} = \frac{1}{3} \sum_{i=1}^3 \|\phi_i(\mathbf{I}_{\text{GT}}) - \phi_i(\mathbf{I}_{\text{ER}})\|_2^2, \quad (5)$$

where $\phi_i(\cdot)$ denotes the outputs of ImageNet [8]-pretrained VGG16 [43] at layers `relu1_1`, `relu2_2` and `relu3_3`. Contrary to RefSR methods, we chose to use more shallow features [34] since the domain gap between rendering and reference leads to a mismatch between the deeper features and therefore causes increased artifact generation. We show qualitative results in the supplementary.

Adversarial loss. The drawback of the perceptual loss is that it tends to generate grid like artifacts [22]. The adversarial loss [12] helps to remove those artifacts and generate visually pleasing images:

$$\mathcal{L}_{\text{disc}} = -\mathbb{E}_{\mathbf{I}_{\text{GT}}}[\log(D(\mathbf{I}_{\text{GT}}, \mathbf{I}_{\text{ER}}))] - \mathbb{E}_{\mathbf{I}_{\text{ER}}}[\log(1 - D(\mathbf{I}_{\text{ER}}, \mathbf{I}_{\text{GT}}))] , \quad (6)$$

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{\mathbf{I}_{\text{GT}}}[\log(1 - D(\mathbf{I}_{\text{GT}}, \mathbf{I}_{\text{ER}}))] - \mathbb{E}_{\mathbf{I}_{\text{ER}}}[\log(D(\mathbf{I}_{\text{ER}}, \mathbf{I}_{\text{GT}}))] , \quad (7)$$

where $\mathcal{L}_{\text{disc}}$ represents the discriminator loss and \mathcal{L}_{adv} is the generator loss. We adopted the Relativistic GAN [18] formulation following MASA [28].

4 Experiments

Datasets. We use the recently introduced LaMAR dataset [38] to create training and test datasets. LaMAR [38] consists of scenes represented by 3D scans and localized AR device trajectories within those scenes. The devices used were iPhones/iPads and HoloLens 2. We use the RGB trajectories from the iPhones and iPads. The dataset consists of three different scenes: CAB, LIN and HGE. CAB is a multi-floor office building, LIN is a few blocks of an old town and HGE the ground floor of a historical university building. We create a training set from CAB and LIN, where CAB contains mostly indoor and LIN outdoor images. Test sets are created from CAB, LIN and HGE. HGE is used to test generalization to novel scenes. For this we take the trajectories, which are a sequence of RGB image and camera pose pairs. The poses are used to render an image from the 3D scan, which we use as the input rendering to be enhanced.

The RGB image is used as the GT. As reference we use a nearby image with a different pose than the GT. Example pairs can be seen in Fig. 5. The renderings can contain artifacts due to the scan quality. Because the trajectories and the 3D scan were not recorded at the same time, different objects, illumination and seasonal changes can be present within the rendering and the GT. The CAB and LIN training set contains 21350 image pairs. The CAB, LIN and HGE test sets consist of 329, 608 and 492 image pairs. The datasets contain different references of various levels for each rendering: a low level indicates that the reference pose is close to the GT pose (easier) and a high level indicates that the reference is further away (harder). Because RefSR methods usually train on the CUFED [62] dataset which consist of images with resolution 160x160, we also rescale our dataset images to this resolution.

Implementation details. The encoder consists of 3 levels where each level is connected to the next one and consists of 1 convolutional layer and 4 residual blocks [13]. In our experiments, we keep the number of feature channels fixed to $F_i = 64$ for all levels. We train on 160x160 images following the convention of recent RefSR methods [3, 28, 29]. In the decoder we use $N_3 = 12$, $N_2 = 8$ and $N_1 = 4$ residual blocks in the merge and fusion layers. We train our model for 60 epochs using only the reconstruction and perceptual loss and fine-tune the model for 20 epochs using additionally the adversarial loss. In our experiments, the weight coefficients λ_{rec} , λ_{per} and λ_{adv} are 1, 1 and 0.001. For training we use a NVIDIA Tesla A100 40GB GPU with a batch size of 9 for 37 hours. We use two data-augmentation strategies specifically targeted to our task. For generalization to a wide range of rendering qualities, we augment the training data with renderings from down-sampled versions of the meshes containing only 10% of the original triangles. To ensure that the model removes artifacts and enhances the rendering even if the reference image is far away or has little content in common, we pick the training reference images randomly from within a 5s temporal window in the sequence.

Evaluation metrics. Because we start from a RefSR method, we use the same metrics for evaluation, notably: Peak Signal Noise Ration (PSNR \uparrow) and Structural Similarity Index Measure (SSIM \uparrow). To follow the convention [3, 28], all PSNR and SSIM results are evaluated on the Y channel of the YCbCr color space. Because PSNR and SSIM can not determine visual quality we also report the Learned Perceptual Image Patch Similarity (LPIPS \downarrow) [60] and the Edge Restoration Quality Assessment (ERQA \uparrow) [20]. LPIPS represents the visual quality with respect to the human perception. Because PSNR and SSIM do not align with human perception when it comes to value blurry images against images with details [60], we also use ERQA that measures how well a method performs at restoring edge details.

4.1 Comparison with RefSR and ST methods

We conduct quantitative and qualitative comparisons between our method and existing RefSR and ST methods. The RefSR methods we compare with are MASA [28] and DATSR [3]. The ST methods are the universal method Art-flow [1], the photorealistic method WCT2 [55] and the semantic method NNST [21].

Table 1: Quantitative evaluation. Our model enhances the rendering in common image quality metrics. It does so using the optimal reference Ref. = GT as an upper bound or using a close-by reference. It generalizes to the unseen HGE scene and performs better than existing RefSR (RSR) and style transfer (ST) methods.

	Method	CAB _{Ref.=GT}				CAB				LIN				HGE			
		PSNR	SSIM	ERQA	LPIPS	PSNR	SSIM	ERQA	LPIPS	PSNR	SSIM	ERQA	LPIPS	PSNR	SSIM	ERQA	LPIPS
	Render	15.60	0.559	0.564	0.380	15.60	0.559	0.564	0.380	14.39	0.529	0.549	0.392	15.84	0.575	0.619	0.364
RSR	MASA [28]	15.55	0.555	0.568	0.347	15.47	0.524	0.544	0.367	14.17	0.419	0.523	0.397	15.62	0.478	0.576	0.360
	DATSR [3]	15.65	0.568	0.553	0.349	15.63	0.557	0.530	0.364	14.34	0.468	0.483	0.438	15.80	0.536	0.553	0.376
ST	Artflow [1]	16.30	0.472	0.533	0.334	15.39	0.414	0.489	0.393	17.00	0.503	0.622	0.321	16.97	0.500	0.602	0.341
	WCT2 [55]	16.48	0.559	0.569	0.357	16.08	0.554	0.567	0.367	17.44	0.569	0.565	0.332	17.52	0.589	0.623	0.324
	NNST [21]	18.52	0.643	0.620	0.303	16.33	0.559	0.566	0.370	18.53	0.568	0.661	0.315	18.48	0.591	0.646	0.315
	MaRINeR	23.89	0.799	0.722	0.089	20.03	0.697	0.643	0.180	21.73	0.668	0.705	0.155	20.96	0.673	0.684	0.176

Tab. 1 shows the results of the quantitative comparison. The metrics are calculated between the GT and the enhanced rendering. The row *Render* is the baseline and shows the scores of the not enhanced rendering and the GT. In the column CAB_{Ref.=GT} the GT is used as the reference showing the methods’ performance when using the optimal reference image. MaRINeR successfully enhances the rendering leading to a significant improvement in all metrics. It also performs better at the task than existing RefSR and ST methods. Even though HGE is a novel scene, our model performs similarly well as on the CAB and LIN scenes, demonstrating that our model exhibits a strong generalization ability. Because in our dataset different objects can be present in rendering and GT, the enhanced rendering will not necessarily exactly look like the GT. This is also reflected in the scores, which are generally lower then when comparing RefSR methods where the GT and low-res. match content-wise.

Qualitative comparison. Fig. 6 shows a visual comparison with RefSR and ST methods. RefSR methods stay close to their low-res. input structure and color wise. They add details from the reference which are first transformed to the low-res. color distribution. Therefore they can not remove artifacts and the color distribution is not adapted to the one of the reference. Style-transfer methods on the other hand have a built in trade-off between content preservation and style transfer. Photo-realistic methods successfully adapt the color distribution of the reference while not changing the content of the rendering, which inadvertently also preserve the artifacts. Universal style transfer methods transfer both the color and content from the reference to the output so some artifacts can disappear. However, they do not distinguish between real content and artifacts and therefore introduce unrealistic distortions into the image. Semantic style transfer methods match between content and style and successfully transfer the style of matching objects. If no matches are found, then the methods also introduce distortions. Our model successfully transfers the colors, removes rendering artifacts while preserving the underlying content and fills in missing parts.

4.2 Applications

Novel view enhancement can be applied to a variety of situations. We showcase the benefits of using our model for eliminating manual sanity checks, enhancing synthetic trajectories and as post-processing tool for renderings of NeRFs.

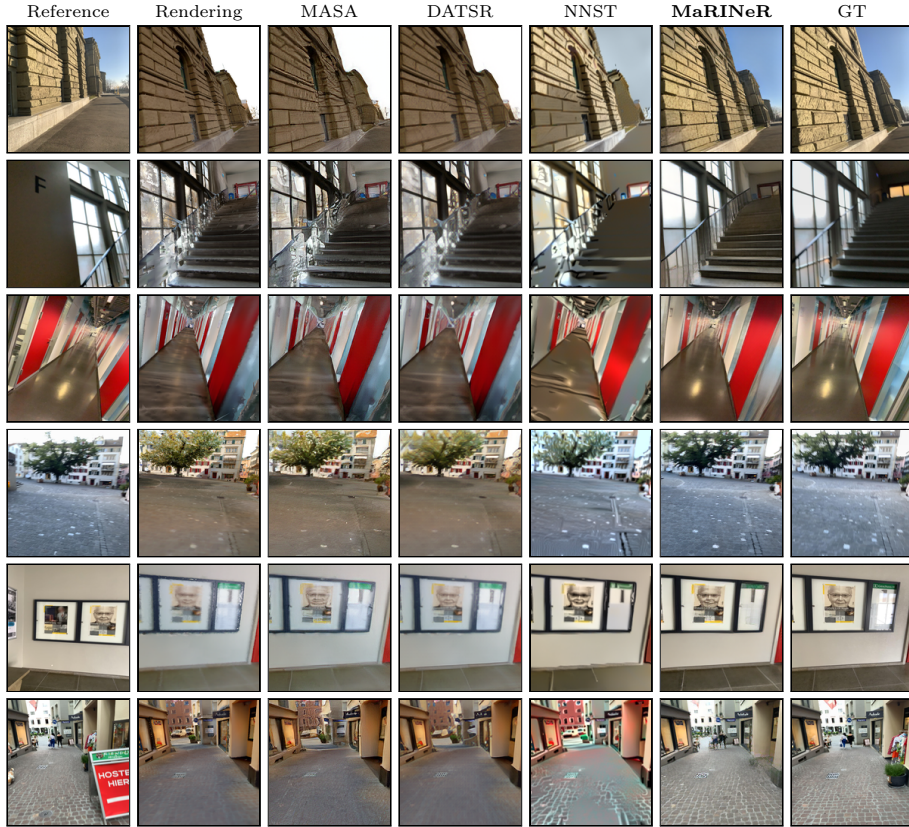


Fig. 6: Qualitative comparison. We compare MASA [28], DATSR [3] (RefSR) and NNST [21] (ST) with **MaRINeR** on the task of novel view enhancement.

Validation of localization pseudo-ground-truth. A limitation of automatic pseudo-GT pipelines for localization is that they often require manual validation. For instance, the LaMAR [38] pipeline registers sequences of images recorded by AR devices, into a common 3D reconstruction based on a high quality LiDAR scanner. To check if the generated alignment is accurate, manual checks between renderings from mesh and input images are needed. To automate those, an option is to estimate a homography between the rendering of the scene at the estimated pose and the associated image of the input sequence. The homography should be identity if the localization of the pipeline was successful. However, estimating a homography between the rendered and real image is not accurate because of the domain gap. Using our method, we can enhance the renderings and improve the accuracy of the estimated homography. We use SuperPoint [10] for feature extraction and SuperGlue [37] for matching. Fig. 7 and Tab. 2 show that with our method we increase the number of matches and the inlier ratio supporting the homography. This leads to a more accurate homography estimation, characterized by the homography error [10] in the table which is optimally zero. Because of the increased accuracy of the homography we can therefore eliminate the manual sanity checks and replace them with an automated tool.

Table 2: Applications. Left – Improved homography estimation. Using enhanced renderings improves the matching and homography estimation between the rendering and the raw image. **Right – NeRFs.** Post-processing NeRF renderings leads to improved visual image quality reflected by better ERQA and LPIPS scores.

	# matches inlier ratio homography error			PSNR SSIM ERQA LPIPS			
Render \leftrightarrow Image	39.21	61.24%	4.86	NeRF output	21.14	0.622	0.646 0.238
+ MaRINeR \leftrightarrow Image	58.89	78.16%	1.88	+ MaRINeR	20.45	0.592	0.701 0.167

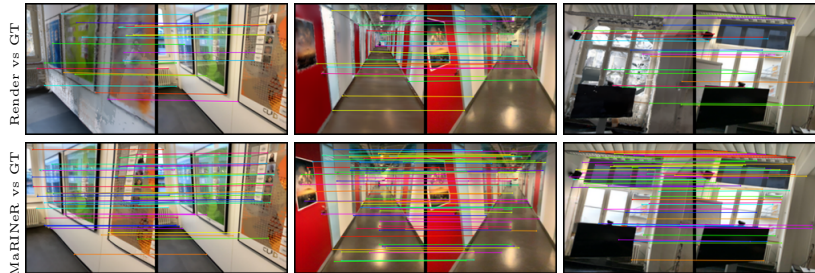


Fig. 7: Better homography estimation. Using enhanced renderings of MaRINeR, estimating a homography to the aligned source image is more accurate and can be used to automate manual sanity checks in the LaMAR [38] pipeline.

Enhancing synthetic trajectories. When creating large AR datasets, substantial human effort is needed to record AR devices trajectories. With recent advances in simulating natural human body movements in 3D scenes such as EgoGen [23], synthetic trajectories can be generated effortlessly to extend the existing datasets. However, because of the quality of the underlying 3D representations, there is a gap between synthetic and real images. Fig. 8 shows that with our method we can take a synthetic trajectory and enhance it using a nearby existing reference image from previously recorded trajectories.

NeRF postprocessing. Training NeRFs can be computationally expensive and requires a large number of images to generate accurate 3D representations [36]. A sufficient number of images may not always be available and training a small model on a large scene with not enough data can lead to noisy representations. Fig. 9 shows how our off-the-shelf model removes the artifacts created by the nerfacto [45] model on the Floating tree and Egypt scenes. Both scenes are large and detail-rich outdoor scenes. We use the smallest nerfacto model

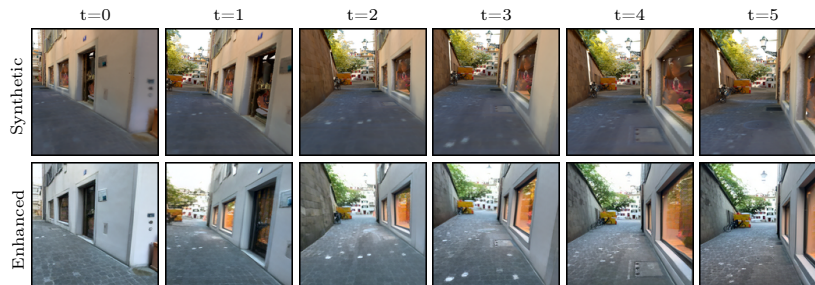


Fig. 8: Enhancing synthetic trajectories with nearby localized images. The result exhibits increased realism and can extend the current dataset without introducing a gap between synthetic and human recorded trajectories.

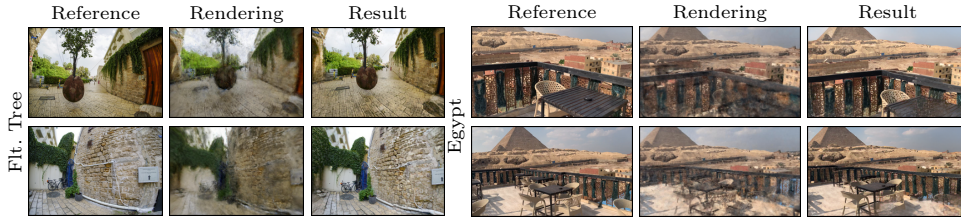


Fig. 9: NeRF postprocessing results. Training a nerfacto [45] model on the Floating tree and Egypt data. We use the smallest nerfacto model and the result contains artifacts which our model can successfully remove.

with default parameters and enhance the evaluation images using the closest training image as reference. Tab. 2 shows that our model successfully enhances the nerfacto rendering with respect to ERQA and LPIPS.

4.3 Ablation study

We refer to the supplemental for more ablations of the architecture and loss function weights.

Reconstruction method. Contrary to LaMAR [38] which uses a NavVis scanner running LiDAR-inertial SLAM followed by the Advancing Front [5] algorithm for meshing, we report results on the 12 Scenes [46] dataset which uses RGB-D SLAM on Kinect data and BundleFusion [7] in Tab. 3.

Table 3: 12 Scenes evaluation.

Method	12 Scenes			
	PSNR	SSIM	ERQA	LPIPS
Render	20.59	0.732	0.640	0.164
MaRINeR	22.99	0.775	0.703	0.071

Iterative refinement. Fig. 10 shows the effect of refining the rendering over subsequent iterations. With one iteration the model fails matching regions with large artifacts. When using two or more iterations, the first iteration removes the artifacts, allowing the correspondence matching to succeed in the next ones. We see the largest improvement when using 2 iterations. More iterations gives only a small improvement but linearly increases the inference time.

Data augmentation. We analyze the effectiveness of our data augmentation strategies. Tab. 4 left shows the model trained without the augmentation per-

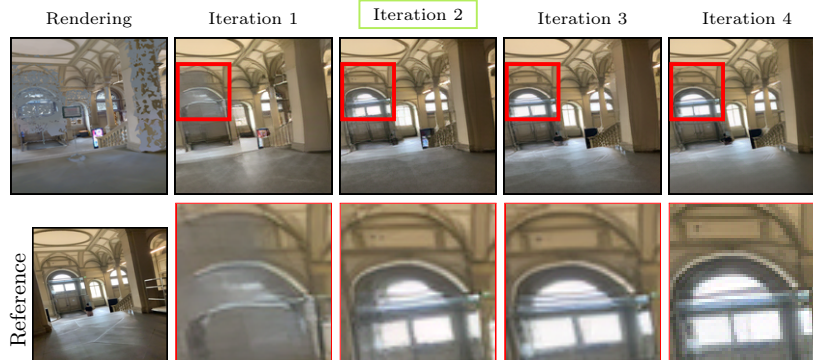


Fig. 10: Visual improvement with iterative refinement. Refining the result over several iterations helps the correspondence matching in presence of large artifacts.

Table 4: Robustness. Left – Mesh quality. Augmenting the data with renderings of a down-sampled mesh increases robustness to changes in the mesh resolution. **Right – Reference level.** A higher level indicates a larger temporal distance to the rendering within a sequence, roughly one meter / 20 degrees / one second per level, which generally correlates with less content in common.

Mesh size	CABaug		CAB	
	PSNR	SSIM	PSNR	SSIM
100%	19.80	0.687	19.45	0.686
75%	19.46	0.680	18.92	0.678
50%	19.41	0.677	18.83	0.673
25%	19.33	0.669	18.56	0.660
10%	19.10	0.650	17.98	0.626

Ref. Level	CABaug		CAB	
	PSNR	SSIM	PSNR	SSIM
0 = GT	22.91	0.777	23.51	0.783
1	19.88	0.687	19.85	0.681
2	18.99	0.664	18.48	0.644
5	18.25	0.646	17.13	0.607
8	18.04	0.643	16.79	0.600

forms worse on meshes with different resolutions. Tab. 4 right shows that for the case without random reference augmentation, the performance is better when the reference is the GT image. As soon as the reference level increases, the performance drastically drops compared to the augmented model.

5 Limitations

While the model detects and removes rendering artifacts, it is also possible that some content is wrongly detected as an artifact and removed. This can lead to blurry or smeared out image parts. The model preserves the content of the rendering, but may transfer additional content from the reference. Currently the model works best on images with resolution in the order of 160 with any aspect ratio. Larger resolutions are only indirectly supported, by first down-scaling the rendering, running our model and then up-scaling the image again using a super resolution method, such as Real-ERSGAN [50]. This could be addressed in the future by transitioning to a more advanced matching pipeline such as LoFTR [44] or CroCro [51] which would come at the cost of more inference time. The model is targeted to enhance low quality renderings, thus high quality renderings are only improved with very close references. The method matches objects on a texture level and not on a semantic level. This means that the objects should have similar texture, where the rendering is a low quality version. The current model may introduce flickering between neighboring frames of a sequence. For video prediction, the pipeline may be further extended to ensure temporal consistency between the generated frames.

6 Conclusions

In this work, we propose a novel method to enhance renderings of 3D reconstructions. Specifically, we use localized images in the 3D scene to enhance renderings from the 3D reconstruction. Our experiments verify that our model enhances the rendering better than existing models in the domains of RefSR or style transfer. It is scene and device agnostic, robust to mesh resolution changes, generalizes to grayscale and reliably removes 3D reconstruction artifacts. Possible applications include automatization of manual sanity checks in ground-truthing pipelines, enhancement of synthetic training data and improvement of neural renderings trained with limited data or resources.

Acknowledgements

This project is partially funded by the Swiss National Science Foundation (SNSF) Advanced Grant number **TMAG-2_216260**.

References

1. An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: Artflow: Unbiased image style transfer via reversible neural flows (2021)
2. Bassier, M., Vergauwen, M., Poux, F.: Point cloud vs. mesh features for building interior classification. *Remote Sensing* (2020)
3. 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)
4. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)* (2017)
5. Cohen-Steiner, D., Da, F.: A greedy delaunay-based surface reconstruction algorithm. *The Visual Computer* **20**, 4–16 (2004)
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: *Proc. Computer Vision and Pattern Recognition (CVPR)*, IEEE (2017)
7. Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration (2017)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: *CVPR09* (2009)
9. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr²: Image style transfer with transformers (2022)
10. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description (2018)
11. Dong, R., Zhang, L., Fu, H.: Rrsgan: Reference-based super-resolution for remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing* (2022)
12. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
14. Huang, R., Peng, S., Takmaz, A., Tombari, F., Pollefeys, M., Song, S., Huang, G., Engelmann, F.: Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels (2023)
15. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization (2017)
16. Jiang, Y., Chan, K.C.K., Wang, X., Loy, C.C., Liu, Z.: Robust reference-based super-resolution via c2-matching (2021)
17. Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., Song, M.: Neural style transfer: A review (2018)
18. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard gan (2018)
19. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering (2023)

20. 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)
21. Kolkin, N., Kucera, M., Paris, S., Sykora, D., Shechtman, E., Shakhnarovich, G.: Neural neighbor style transfer (2022)
22. Krawczyk, P., Gaertner, M., Jansche, A., Bernthaler, T., Schneider, G.: Artifact generation when using perceptual loss for image deblurring (2023)
23. Li, G., Zhao, K., Zhang, S., Lyu, X., Dusmanu, M., Zhang, Y., Pollefeys, M., Tang, S.: Egogen: An egocentric synthetic data generator (2024)
24. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms (2017)
25. 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)
26. Litomisky, K., Bhanu, B.: Removing moving objects from point cloud scenes. In: *International Workshop on Advances in Depth Image Analysis and Applications* (2012)
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021)
28. Lu, L., Li, W., Tao, X., Lu, J., Jia, J.: Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution (2021)
29. Mei, X., Yang, Y., Li, M., Huang, C., Zhang, K., Lió, P.: A feature reuse framework with texture-adaptive aggregation for reference-based super-resolution (2023)
30. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space (2019)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis (2020)
32. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks (2020)
33. Pesavento, M., Volino, M., Hilton, A.: Attention-based multi-reference learning for image super-resolution (2021)
34. Pittaluga, F., Koppal, S.J., Kang, S.B., Sinha, S.N.: Revealing scenes by inverting structure from motion reconstructions. In: *CVPR* (2019)
35. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., Savva, M., Zhao, Y., Batra, D.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai (2021)
36. Remondino, F., Karami, A., Yan, Z., Mazzacca, G., Rigon, S., Qin, R.: A critical analysis of nerf-based 3d reconstruction. *Remote Sensing* (2023)
37. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks (2020)
38. 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)
39. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T.: Benchmarking 6dof outdoor visual localization in changing conditions (2018)
40. Schöps, T., Sattler, T., Pollefeys, M.: BAD SLAM: Bundle adjusted direct RGB-D SLAM. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2019)

41. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
42. Shim, G., Park, J., Kweon, I.S.: Robust reference-based super-resolution with similarity-aware deformable convolution. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
44. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers (2021)
45. 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)
46. Valentin, J., Dai, A., Nießner, M., Kohli, P., Torr, P., Izadi, S., Keskin, C.: Learning to navigate the energy landscape (2016)
47. Wald, J., Sattler, T., Golodetz, S., Cavallari, T., Tombari, F.: Beyond controlled environments: 3d camera re-localization in changing indoor scenes. In: European Conference on Computer Vision (ECCV) (2020)
48. 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>
49. Wang, S., Sun, Z., Li, Q.: High-to-low-level feature matching and complementary information fusion for reference-based image super-resolution. *The Visual Computer* **40** (2023)
50. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: International Conference on Computer Vision Workshops (ICCVW)
51. Weinzaepfel, Philippe and Leroy, Vincent and Lucas, Thomas and Brégier, Romain and Cabon, Yohann and Arora, Vaibhav and Antsfeld, Leonid and Chidlovskii, Boris and Csurka, Gabriela and Revaud Jérôme: CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. In: NeurIPS (2022)
52. Xie, Y., Xiao, J., Sun, M., Yao, C., Huang, K.: Feature representation matters: End-to-end learning for reference-based image super-resolution. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) *Computer Vision – ECCV 2020*. pp. 230–245. Springer International Publishing, Cham (2020)
53. Yan, X., Zhao, W., Yuan, K., Zhang, R., Li, Z., Cui, S.: Towards content-independent multi-reference super-resolution: Adaptive pattern matching and feature aggregation. In: European Conference on Computer Vision (2020)
54. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution (2020)
55. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms (2019)
56. Yue, H., Sun, X., Yang, J., Wu, F.: Landmark image super-resolution by retrieving web images. *IEEE Transactions on Image Processing* (2013)
57. Zhang, H., Wang, C., Tian, S., Lu, B., Zhang, L., Ning, X., Bai, X.: Deep learning-based 3d point cloud classification: A systematic survey and outlook. *Displays* **79** (2023)

58. Zhang, L., Li, X., He, D., Ding, E., Zhang, Z.: Lmr: A large-scale multi-reference dataset for reference-based super-resolution (2023)
59. Zhang, L., Li, X., He, D., Li, F., Wang, Y., Zhang, Z.: Rrsr:reciprocal reference-based image super-resolution with progressive feature alignment and selection (2022)
60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
61. Zhang, Y., Fang, C., Wang, Y., Wang, Z., Lin, Z., Fu, Y., Yang, J.: Multimodal style transfer via graph cuts (2020)
62. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer (2019)
63. Zhao, K., Tan, H., Yau, T.F.: Multi-reference image super-resolution: A posterior fusion approach (2022)
64. Zheng, H., Ji, M., Han, L., Xu, Z., Wang, H., Liu, Y., Fang, L.: Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution. In: British Machine Vision Conference (2017)
65. Zheng, H., Ji, M., Wang, H., Liu, Y., Fang, L.: Crossnet: An end-to-end reference-based super resolution network using cross-scale warping (2018)
66. Zheng, J., Liu, Y., Feng, Y., Xu, H., Zhang, M.: Contrastive attention-guided multi-level feature registration for reference-based super-resolution. *ACM Trans. Multimedia Comput. Commun. Appl.* (2023)
67. Zhou, K., Li, W., Wang, Y., Hu, T., Jiang, N., Han, X., Lu, J.: Nerflix: High-quality neural view synthesis by learning a degradation-driven inter-viewpoint mixer (2023), <https://arxiv.org/abs/2303.06919>