

# GMT: Enhancing Generalizable Neural Rendering via Geometry-Driven Multi-Reference Texture Transfer

Youngho Yoon<sup>✉\*</sup>, Hyun-Kurl Jang<sup>✉\*</sup>, and Kuk-Jin Yoon<sup>✉</sup>

Visual Intelligence Lab., KAIST  
{dudgh1732, jhg0001, kjyoon}@kaist.ac.kr  
<https://github.com/yh-yoon/GMT>

**Abstract.** Novel view synthesis (NVS) aims to generate images at arbitrary viewpoints using multi-view images, and recent insights from neural radiance fields (NeRF) have contributed to remarkable improvements. Recently, studies on generalizable NeRF (G-NeRF) have addressed the challenge of per-scene optimization in NeRFs. The construction of radiance fields on-the-fly in G-NeRF simplifies the NVS process, making it well-suited for real-world applications. Meanwhile, G-NeRF still struggles in representing fine details for a specific scene due to the absence of per-scene optimization, even with texture-rich multi-view source inputs. As a remedy, we propose a **Geometry-driven Multi-reference Texture transfer network (GMT)** available as a plug-and-play module designed for G-NeRF. Specifically, we propose ray-imposed deformable convolution (RayDCN), which aligns input and reference features reflecting scene geometry. Additionally, the proposed texture preserving transformer (TPFormer) aggregates multi-view source features while preserving texture information. Consequently, our module enables direct interaction between adjacent pixels during the image enhancement process, which is deficient in G-NeRF models with an independent rendering process per pixel. This addresses constraints that hinder the ability to capture high-frequency details. Experiments show that our plug-and-play module consistently improves G-NeRF models on various benchmark datasets.

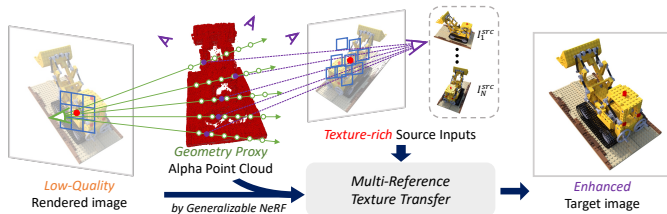
**Keywords:** Generalizable neural radiance fields · Image enhancement

## 1 Introduction

Novel view synthesis (NVS) is an approach that synthesizes an image of an arbitrary viewpoint from multi-view source images. Early studies on NVS have primarily explored image-based rendering [17, 26, 39, 57]. Recently, neural radiance fields [33] (NeRF) proposed a volume rendering method through 5D radiance field optimization that extracts densities and colors from 5D inputs (3D locations and 2D directions). NeRF-based approach has inspired numerous studies in NVS tasks, resulting in remarkable performance enhancements.

---

\* Equal contribution.



**Fig. 1:** The proposed Geometry-driven Multi-reference Texture transfer (GMT) model.

Recent studies have pointed out that NeRF requires per-scene optimization for NVS and proposed various generalizable NeRF (G-NeRF) methods [41, 46–48, 52, 60]. These studies enable NVS on-the-fly without per-scene optimization through cross-scene generalization. Since then, several studies have attempted to improve rendering quality and speed based on G-NeRF. Neuray [30] and GeoNeRF [23] improve understanding of scene geometry and occlusion by utilizing multi-view stereo methods. Additionally, recent research leveraging 3D Gaussian splatting [25] in generalizable neural rendering has significantly accelerated rendering speed [3]. Despite these advances, generalizable NVS approaches commonly encounter difficulties in accurately representing high-frequency details. The methods proposed in G-NeRF studies often struggle to capture the local textures for a specific scene due to the absence of per-scene optimization, even with texture-rich multi-view source inputs.

To solve this problem, as shown in Fig. 1, we aim to enhance rendered images by 1) geometric priors derived from the rendering process of G-NeRF and 2) transferring high-frequency details of texture-rich source images. Inspired by prior studies in reference-based image enhancement [22, 61, 64], we develop a network that improves NVS performance by transferring textures acquired from multi-view source images onto rendered images from G-NeRF. Enhancement is done in a plug-and-play manner, demonstrating improved results within seconds.

To achieve the goal, we propose a ray-imposed deformable convolution network, RayDCN, that conducts geometry-driven reference feature alignment with the target view feature map. RayDCN determines the spatial location on source images for deformable convolution by leveraging the alpha values of sampled points obtained during the G-NeRF rendering process. Through this approach, we make deformable convolution incorporate the scene geometry by utilizing the estimated spatial location of source views of each ray. While utilizing estimated geometry for source-target feature alignment, we additionally employ correspondence matching to handle occlusion and inaccuracies in geometry estimation.

We also introduce a texture-preserving transformer, called TPFormer, designed for multi-reference feature aggregation. TPFormer transfers the texture from multi-reference images to the target view image through two steps. Firstly, view-dependent attention performs self-attention with input features of view differences and the corresponding view to obtain preliminary information for reference-texture selection. Secondly, reference-texture selection process aggre-

gates features from multiple reference images via feature selection, considering the relationships between multiple references obtained during the view-dependent attention step. TPFormer seamlessly integrates multi-view source features extracted by RayDCN without impairing texture information.

Consequently, our model handles multi-ray features with a receptive field on the target viewpoint, while G-NeRF models handle single-ray. G-NeRF models estimate the color of each pixel independently in a batch-wise manner, which causes a deficiency in direct interaction among adjacent pixels in the target image. This deficiency hampers the model from representing subtle variations or intricate patterns in the image. However, our module induces interactions among adjacent rays and generates a superb NVS result. We experimentally demonstrate that our method consistently enhances the performance of existing G-NeRF models on various NVS benchmark datasets. In summary, our contributions are as follows:

- We introduce a Geometry-driven Multi-reference Texture transfer network (GMT) for generalizable neural rendering.
- We propose ray-imposed deformable convolution, which performs feature alignment reflecting scene geometry.
- We propose a texture-preserving transformer for source features aggregation with preserving texture features.
- Our plug-and-play module consistently improves generalizable NeRF model performances on various benchmark datasets without additional training.

## 2 Related Works

### 2.1 Image based rendering

Image-based rendering (IBR) methods render novel views directly from input images without the necessity of 3D scene representations. Early research in IBR [17, 26, 39] rendered novel views from dense input image sets using the 4D plenoptic function. Studies adopting geometry proxies [4, 9, 17, 20, 35, 44, 57] have demonstrated that satisfactory rendering quality could be attained using depth or mesh data obtained from input images. As deep learning has progressed, numerous works utilizing deep neural networks have come to the forefront [7, 18, 24, 37, 45, 54]. LFNR [42] has leveraged the strength of classic IBR technique [26] which is resistant to reflections. NeRF-based models like IBR-Net [48] and Neuray [30] applied volume rendering techniques akin to NeRF using features from input images and geometric information.

### 2.2 Generalizable NeRF

Synthesizing photo-realistic images has been a long-standing area of research interest. Neural scene representations, exemplified by the Neural Radiance Fields (NeRF) [33] is effective and impressive solution for view synthesis. Following works of NeRF improved rendering quality [10, 38, 49] and optimization, and

rendering speed [5, 14, 34, 43, 53, 59]. However, NeRF retains the limitation that per-scene optimization is required to perform novel view synthesis. Various works have explored generalizable NeRF (G-NeRF) [1, 6, 23, 30, 41, 47, 48, 56, 60] to overcome this limitation. G-NeRF learns a view interpolation function from source images, enabling cross-scene generalization. In G-NeRF, the typical approach involves using volume rendering to aggregate information obtained from images, such as deep features, depth maps, and cost volumes [6, 23, 30, 48, 52]. GNT [47] and GPNR [41] leverage transformers as feature aggregators to enhance the interaction of information within a single ray, resulting in the direct acquisition of RGB values for each pixel. PixelSplat [3] introduces generalizable volume rendering using scene parameterization based on 3D Gaussian primitives [25]. Despite the advance, G-NeRF still maintains the independence of the rendering process for each pixel, leading to the failure of transferring fine textures in source images. To solve this problem, we propose a novel method to integrate information from reference images using multi-ray aggregation. Furthermore, our approach applies to various G-NeRF models and collectively enhances their performance.

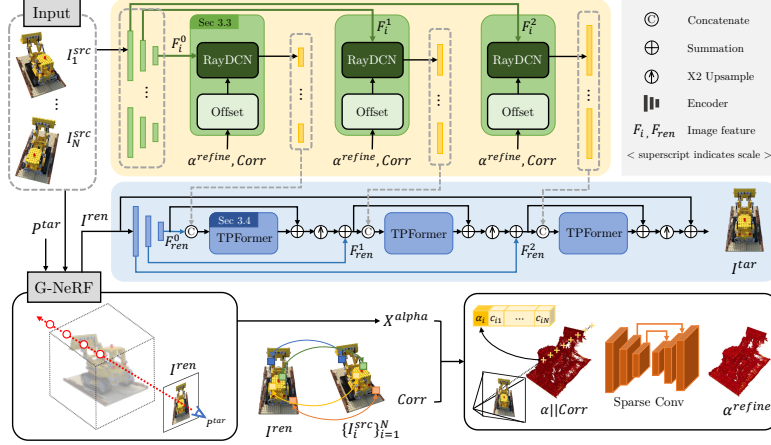
### 2.3 Reference-based Image Enhancement

Image enhancement tasks aim to rectify degradation or elevate overall visual quality of given image. Some studies also incorporate reference images to retain the fine textures and intricate details found within reference images [27, 29, 64, 65, 68]. In the context of reference-based image super-resolution [2, 22, 31, 51, 55, 61, 62, 64, 65], the approach involves transferring additional details from high-resolution reference images to low-resolution input images. The common practice in reference-based super-resolution is to utilize a single reference image, but some approaches employ multiple reference images [36, 58, 61]. In reference-based deblurring tasks [27, 29, 68], high-quality features extracted from reference images enhance the quality of blurred input images. In reference-based restoration tasks, it is typical to establish image-based correspondences between input images and reference images to identify applicable reference features. The present study introduces a novel approach that leverages geometric priors to estimate correspondences while retaining multi-view consistency. The correspondences, derived from sampled points and the associated alpha values generated during the rendering process, are jointly utilized with the image pair correlation to enhance the accuracy of locating relevant features.

## 3 Method

### 3.1 Preliminary

In common, Generalizable Neural Radiance Fields (G-NeRF) generate a novel view image  $I^{ren}$  from  $N$  source-view images  $\{I_i^{src}\}_{i=1}^N$  taken from sparse views, without the process of scene optimization performed in NeRF. A ray  $p(r)$  with direction vector  $r$  emitted from the target view camera center  $o$  can be expressed



**Fig. 2:** Overall framework of the Geometry-driven Multi-reference Texture transfer (GMT) model. When generalizable NeRFs (G-NeRF) renders novel view image  $I^{ren}$  with  $N$  source images  $\{I_i^{src}\}_{i=1}^N$  and a target camera pose  $P^{tar}$ , the process inherently generates alpha point cloud  $X^{alpha}$  for volume rendering process. Using  $\alpha^{refine}$  extracted from  $\alpha$  and correlation values  $Corr$ , RayDCN enables feature alignment considering scene geometry. Subsequently, TPFormer conducts multi-reference feature aggregation and the model generates final output  $I^{tar}$ .

as  $p(r) = o + zr$ . After sampling  $K$  points among the points existing on ray  $p(r)$ , alpha  $\{\alpha_i\}_{i=1}^K$  and colors  $\{c_i\}_{i=1}^K$  of each point are estimated. In this process, G-NeRF learns a network that aggregates source features projected from  $\{I_i^{src}\}_{i=1}^N$  to estimate  $\alpha_i$  and  $c_i$  without per-scene optimization. Finally, the color  $c$  of ray  $p(r)$  is calculated using the following equation:

$$c = \prod_{i=1}^K h_i c_i = \prod_{i=1}^K T_i \alpha_i c_i \quad (1)$$

where the hitting probability  $h_i = T_i \alpha_i$  and the accumulated transmittance  $T_i = \prod_{j=1}^{i-1} (1 - \alpha_j) \alpha_i$ .

### 3.2 Multi-Reference Texture Transfer Network

We present a geometry-driven multi-reference texture transfer network aiming to overcome the methodological constraints observed in previous studies of G-NeRF—particularly those characterized by independent ray sampling and color rendering for each pixel. Inspired by the potential for texture transfer from unblemished local textures and high-frequency details in the source images  $\{I_i^{src}\}_{i=1}^N$ , we leverage these images for our image enhancement model. Our model addresses the common texture scarcity issue and blurry artifacts in G-NeRF by leveraging valuable information in source images, ultimately extracting

improved rendering results denoted as  $I^{tar}$ . As in Fig. 2, we formulate our proposed approach through the following stages.

**Rendering by G-NeRF.** We initially conduct a novel view synthesis from  $\{I_i^{src}\}_{i=1}^N$  using one of the G-NeRF models to generate the rendered image  $I^{ren}$ , which will be used as a query image of our model. The neural rendering process inherently generates alpha point cloud  $X^{alpha}$ , which includes alpha values from the volume rendering process. We utilize  $X^{alpha}$  as proxy geometry and the initial input for the following offset estimation network.

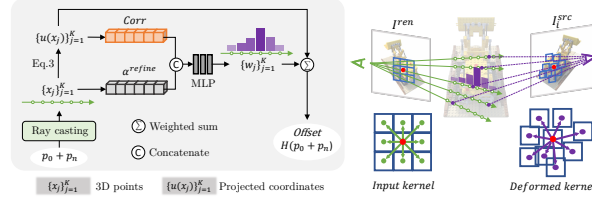
**Feature Extraction.** We initiate feature extraction with the VGG model to derive multi-scale texture features  $\{\mathbf{F}_i\}_{i=1}^N$  from input source images  $\{I_i^{src}\}_{i=1}^N$  and  $\mathbf{F}_{ren}$  from rendered image.

**Offset Estimation and Feature Alignment.** In this stage, we perform feature alignment by finding correspondence between the source and target features. This correspondence position features beneficial to the target region at their corresponding coordinates. To achieve this, we propose a ray-imposed deformable convolution network (RayDCN) (Sec. 3.3). RayDCN leverages refined alpha and correlation for offset estimation while considering multi-view constraints. Subsequently, feature alignment is executed through deformable convolution.

**Reference Feature Aggregation.** Using the aligned features, we conduct reference feature aggregation to merge them with the rendered features extracted from  $I^{ren}$ . In this process, we introduce texture-preserving transformer (TPFormer), an aggregation module designed to maintain the local textures of the reference images (Sec. 3.4). We synthesize the final result for the target viewpoint image  $I^{tar}$  by fusion and upscaling three times.

### 3.3 Ray-Imposed Deformable Convolution (RayDCN)

The general framework of multi-reference-based enhancement (MRE) first involves finding the corresponding points between images through optical flow estimation [65] or correspondence matching [22, 61]. In MRE, source images are considered referenceable, but in most cases, reference images are not obtained from scenes identical to the target image. In contrast, in G-NeRF, the source and target images are always obtained from the same scene. Hence, multi-view geometry is applicable while enhancing the rendered image  $I^{ren}$  from G-NeRF. In light of this circumstance, we propose RayDCN for multi-reference texture transfer which finds corresponding points utilizing  $\{I_i^{src}\}_{i=1}^N$  and  $I^{ren}$  obtained through G-NeRF, and  $X^{alpha}$  for proxy representation of scene geometry. Raw alpha values in G-NeRF include unreliable noise due to inaccurately estimated geometry. Thus, we employed inter-image correlation to obtain a reliable and refined alpha. we first generate point cloud features of  $N + 1$  channels by concatenating the alpha values of sampled points and the correlation values  $Corr$  between  $I^{ren}$  and  $\{I_i^{src}\}_{i=1}^N$ . Generated point cloud feature is processed through sparse convolution, MinkowskiEngine [8], and returns refined alpha value  $\alpha^{refine}$  of the points. Following this, as depicted in Fig. 3, we estimate offset which decides reference point in source image  $I_i^{src}$  of view  $i$ . This estimation is performed based on the  $Corr$  and  $\alpha^{refine}$ . When single 2D coordinate  $p$  on  $I^{ren}$  is decided,



**Fig. 3:** Ray-imposed Deformable Convolution (RayDCN). It has a deformed kernel shape considering scene geometry and aggregates the source features of multiple rays.

we can specify the points  $\{x_j\}_{j=1}^K$  on the ray passing through  $p$ . To estimate the offset of  $p$ , we calculate weight  $\{w_j\}_{j=1}^K$  from  $\{x_j\}_{j=1}^K$  and its allotted  $\alpha^{refine}$  and  $Corr$ .  $\{w_j\}_{j=1}^K$  is obtained through the following equation.

$$\{w_j\}_{j=1}^K = \mathbf{S}(MLP(M_{val} \odot \{\alpha^{refine}\} || Corr)) \quad (2)$$

where  $M_{val}$  is a valid projection mask that determines whether 3D point  $x_j$  can be projected onto  $I_i^{src}$  and  $\mathbf{S}$  indicates softmax function. Then,  $u(x_j)$  is the projected coordinate on the image plane of  $I_i^{src}$  corresponding to  $x_j$ . To compute  $u(x_j)$ , the equation is as follows:

$$u(x_j) = K_i^s (R_i^r D_j (K^t)^{-1} x_j + T_i^r) \quad (3)$$

where  $K_i^s$  and  $K^t$  are the intrinsic matrices of  $I_i^{src}$  and  $I^{tar}$ , respectively.  $R_i^r$  and  $T_i^r$  are the relative rotation and relative translation between  $I_i^{src}$  and  $I^{tar}$ .  $D_j$  is the depth of  $x_j$ . The obtained  $w_j$  and  $u(x_j)$  are combined through a weighted summation to generate  $\mathbf{H}(p)$  which is the offset of  $p$ .  $\mathbf{H}(p)$  is calculated by this equation:

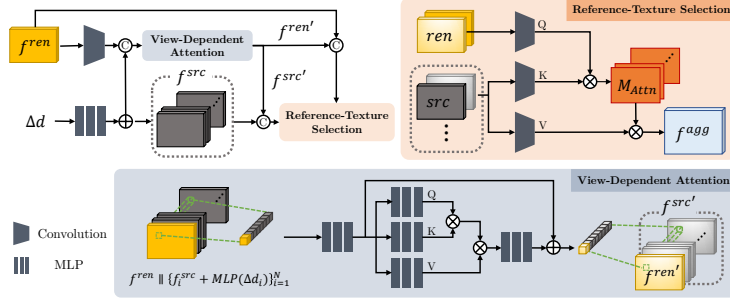
$$\mathbf{H}(p) = \left\{ \sum_j w_j \cdot u(x_j) / \sum_j w_j \right\} - p \quad (4)$$

where  $w_j$  is used as a weight, and  $u(x_j)$  is the value. Because  $p + \mathbf{H}(p)$  is an affine combination with the coordinates  $\{u(x_j)\}_{j=1}^K$  as an element, it always exists on the epipolar line. Therefore, we can narrow the candidate of offsets through epipolar constraints. Calculated  $p + \mathbf{H}(p)$  is the reference point of  $p$  derived from a single-ray. Then, RayDCN performs feature alignment of the source image features to the target viewpoint. The introduced RayDCN deforms the convolution kernel shape, reflecting scene geometry and aggregating multi-ray derived features. As shown in Fig. 3, the offsets  $\{\mathbf{H}(p_0 + p_n)\}_{\mathcal{R}}$  of rays passing through  $p_0$  and neighboring pixels on target view will be goes through convolution filter:

$$\mathbf{y}(p_0) = \sum_{p_n \in \mathcal{R}} \mathbf{w}(p_n) \cdot \mathbf{F}(p_0 + p_n + \mathbf{H}(p_0 + p_n)) \quad (5)$$

where  $\mathbf{y}(p_0)$  is the value of  $p_0$  on the output feature map  $\mathbf{y}$ ,  $\mathbf{w}$  is weights of convolution filter, and  $\mathbf{F}$  is the input feature map.  $\mathcal{R}$  is regular grids of the convolution filter and defined as,

$$\mathcal{R} = \{(-1, -1), (-1, 0), \dots, (0, 1), (1, 1)\} \quad (6)$$



**Fig. 4:** Texture-Preserving Transformer (TPFormer). TPFormer aggregates features from multiple source views while preserving textures from the source image.

### 3.4 Texture-Preserving Transformer (TPFormer)

The architectural configurations commonly employed in G-NeRF models involve aggregating input features considering ray-direction. Notably, IBRNet [48] and Neuray [30] utilize a multi-layer perceptron (MLP), whereas GeoNeRF [23] and GNT [47] incorporate a multi-head attention (MHA) layer. However, we found that the MLP and MHA structures tend to blend detailed texture features derived from input images, leading to blurry artifacts. As illustrated in Fig. 4, we introduce a novel component termed the texture-preserving transformer (TPFormer) to address this issue. TPFormer aims to perform appropriate aggregation according to the ray-direction of the source view image features while preserving their high-frequency details. TPFormer includes a view-dependent attention (VA) module and a reference-texture selection (RS) module. Before operating TPFormer, RayDCN conduct alignment and generate aligned source image feature  $f_i^{src}$ , which has a texture information of  $I_i^{src}$ . After that, we conduct the VA module in TPFormer to perform pixel-wise self-attention on the features of  $I^{ren}$  and  $I_i^{src}$ .  $f_i^{src}$  is added by relative view direction embedding and combined with the target view feature before self-attention is applied. This process aggregates source view features based on relative viewing direction, as described by the following equation:

$$f^{ren'}, \{f_i^{src'}\}_i = Net_{attn}(f^{ren} \parallel \{f_i^{src} + Net_{mlp}(\Delta d_i^{src})\}_i) \quad (7)$$

where  $\Delta d_i$  is view direction of  $i$ -th source view relative to the target view. The RS module receives input features  $f^{ren''}, f^{src''}$ , which is a fusion of resulting features of VA and aligned features from RayDCN.

$$f^{ren''} = f^{ren} \parallel f^{ren'} \quad (8)$$

$$f_i^{src''} = f_i^{src} \parallel f_i^{src'} \quad (9)$$

RS module aggregates  $N$  features from source view  $\{f_i^{src''}\}_{i=1}^N$  with  $f^{ren''}$ . The attention map for  $i$ -th view is calculated using the query from  $f^{ren''}$  and key from



$f_i^{src''}$ . Final result of TPFormer  $f^{agg}$  is derived from aggregation of attention map and value from  $f_i^{src''}$ , equation is as follows:

$$f^{agg} = \sum_i^N (Attn(Q(f^{ren''}), K(f_i^{src''})) \cdot V(f_i^{src''})) \quad (10)$$

TPFormer performs feature aggregation using the relationship between each feature and viewing direction obtained through VA. Therefore, textures in the reference view can be seamlessly transferred to the target view image.

## 4 Experiments

### 4.1 Implementation Details

**Training Datasets.** Our training datasets consist of synthetic and real datasets. For synthetic data, we use Google Scanned Objects dataset [11], which contains 1023 objects, and we employ 10 images from each object. For real datasets, we used 109 scenes from the DTU dataset [21], 35 scenes from the Real Forward-Facing dataset [32] and 89 scenes from the Spaces dataset [13]. We use a black background for Google Scanned Objects and DTU dataset.

**Test Datasets.** In evaluation, we use DTU dataset [21], Synthetic NeRF [33], and Real Forward-Facing [32]. Within the DTU dataset, we use four scenes (birds, tools, bricks, and snowman) at  $800 \times 600$  resolution. Real Forward-Facing dataset and Synthetic NeRF dataset are both consist of 8 scenes. The evaluation resolutions are  $1008 \times 756$  for Real Forward-Facing dataset, and  $800 \times 800$  for Synthetic NeRF dataset. DTU dataset and Synthetic NeRF are evaluated with a black background.

**Network Details.** Offset estimation for the proposed RayDCN integrates an image-based correspondence matching module with a geometry-driven approach. The correspondence matching module is based on C2-Matching [22], which only uses features from rendered and reference images. We use the VGG extractor [40] to extract image features, which is known for its ability to represent texture [15, 16]. Additionally, to leverage multi-scale features, relu1\_1, relu2\_1, and relu3\_1 layers in VGG extractor are used as encoders. When processing the alpha values generated by G-NeRF, which are sparse 3D points, we utilized MinkowskiNet [8], specifically MinkUNet14.

**Dataset Generation.** Our model is an enhancer that improves the performance of G-NeRF models. It takes novel view images and the corresponding alpha values generated during the rendering process as input. In order to accelerate the training and evaluation process, we pre-rendered novel view images from the G-NeRF model. We stored these images, their corresponding alpha values, and 3D point positions. The alpha values were sorted for each ray based on the  $h_i$ , and we selected the top 12 sampled points and corresponding alpha. For constructing the training dataset, we utilized Neuray [30], which follows the same train-test dataset split as our experimental setup and one of the state-of-the-art (SOTA) models.

**Table 1:** Novel view synthesis results on DTU, Synthetic NeRF, and Real Forward-Facing datasets. Ours consistently exhibits a significant improvement across IBRNet, GNT, GeoNeRF, Neuray, and MuRF. **Bold** indicates the best, and underline indicates the second best results.

Method	DTU			Synthetic NeRF			Real Forward-Facing		
	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )
pixelNeRF [60]	19.40	0.463	0.447	22.65	0.808	0.202	18.66	0.588	0.463
IBRNet [48]	26.76	0.879	0.136	25.03	0.900	0.102	25.19	0.822	0.173
MVSNeRF [6]	23.83	0.723	0.286	25.15	0.853	0.159	21.18	0.691	0.301
GNT [47]	25.46	0.818	0.171	23.11	0.763	0.141	25.54	0.835	0.177
GeoNeRF [23]	-	-	-	<u>29.59</u>	0.933	0.071	25.64	0.847	0.150
ContraNeRF [56]	27.69	0.904	0.129	27.92	0.930	<u>0.060</u>	25.44	0.842	0.178
Neuray [30]	<u>28.37</u>	<u>0.906</u>	<u>0.112</u>	28.36	0.928	0.071	25.43	0.833	0.161
MuRF [52]	24.87	0.870	0.183	22.26	0.612	0.225	<u>26.49</u>	<u>0.909</u>	0.143
Ours+IBRNet [48]	26.96 (0.20 $\uparrow$ )	0.893 (0.014 $\uparrow$ )	0.124 (0.012 $\downarrow$ )	25.29 (0.26 $\uparrow$ )	0.909 (0.009 $\uparrow$ )	0.089 (0.013 $\downarrow$ )	25.57 (0.38 $\uparrow$ )	0.839 (0.017 $\uparrow$ )	0.154 (0.019 $\downarrow$ )
Ours+GNT [47]	25.60 (0.14 $\uparrow$ )	0.832 (0.014 $\uparrow$ )	0.155 (0.016 $\downarrow$ )	23.23 (0.12 $\uparrow$ )	0.773 (0.010 $\uparrow$ )	0.124 (0.017 $\downarrow$ )	25.81 (0.27 $\uparrow$ )	0.850 (0.015 $\uparrow$ )	0.157 (0.020 $\downarrow$ )
Ours+GeoNeRF [23]	-	-	-	<b>29.81</b> (0.22 $\uparrow$ )	<b>0.942</b> (0.009 $\uparrow$ )	<b>0.052</b> (0.019 $\downarrow$ )	25.80 (0.16 $\uparrow$ )	0.857 (0.010 $\uparrow$ )	<b>0.137</b> (0.013 $\downarrow$ )
Ours+Neuray [30]	<b>28.72</b> (0.35 $\uparrow$ )	<b>0.920</b> (0.014 $\uparrow$ )	<b>0.101</b> (0.011 $\downarrow$ )	28.96 (0.60 $\uparrow$ )	<u>0.936</u> (0.008 $\uparrow$ )	0.062 (0.010 $\downarrow$ )	25.82 (0.39 $\uparrow$ )	0.850 (0.017 $\uparrow$ )	0.142 (0.019 $\downarrow$ )
Ours+MuRF [52]	24.91 (0.04 $\uparrow$ )	0.872 (0.002 $\uparrow$ )	0.180 (0.003 $\downarrow$ )	22.36 (0.10 $\uparrow$ )	0.614 (0.002 $\uparrow$ )	0.217 (0.008 $\downarrow$ )	<b>26.81</b> (0.32 $\uparrow$ )	<b>0.915</b> (0.016 $\uparrow$ )	<u>0.139</u> (0.004 $\downarrow$ )

**Table 2:** Novel view synthesis results on ACID and RealEstate10k datasets. **Bold** indicates the best, and underline indicates the second best results.

Method	ACID			RealEstate10k		
	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )
pixelNeRF [60]	20.97	0.547	0.533	20.43	0.589	0.550
GPNR [41]	25.28	0.764	0.332	24.11	0.793	0.255
Du et al. [12]	26.88	0.799	0.218	24.78	0.820	0.213
pixelSplat [3]	<u>28.10</u>	<u>0.846</u>	<u>0.122</u>	<u>25.86</u>	<u>0.865</u>	<u>0.110</u>
Ours+pixelSplat [3]	<b>28.87</b> (0.77 $\uparrow$ )	<b>0.855</b> (0.009 $\uparrow$ )	<b>0.107</b> (0.015 $\downarrow$ )	<b>26.40</b> (0.54 $\uparrow$ )	<b>0.871</b> (0.006 $\uparrow$ )	<b>0.101</b> (0.009 $\downarrow$ )

**Table 3:** Novel view synthesis with per-scene finetuning on Real Forward-Facing dataset.

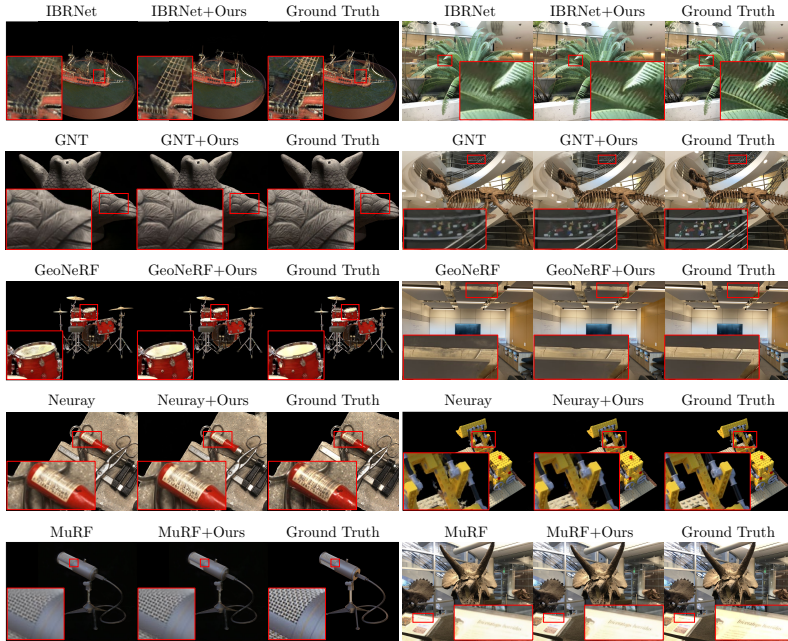
Method	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )
LLFF [32]	23.93	0.798	0.212
NeRF [33]	26.36	0.811	0.250
NeX [50]	27.03	0.890	0.182
NLF [42]	28.03	0.917	0.129
Neuray-ft [30]	27.40	0.869	0.129
GNT-ft [47]	<u>30.73</u>	<u>0.943</u>	<b>0.081</b>
Ours+Neuray-ft [30]	27.56 (0.16 $\uparrow$ )	0.876 (0.007 $\uparrow$ )	0.125 (0.004 $\downarrow$ )
Ours+GNT-ft [47]	<b>31.03</b> (0.30 $\uparrow$ )	<b>0.946</b> (0.003 $\uparrow$ )	<u>0.082</u> (0.001 $\uparrow$ )

**Table 4:** Quantitative comparisons of reference-based image enhancement models.

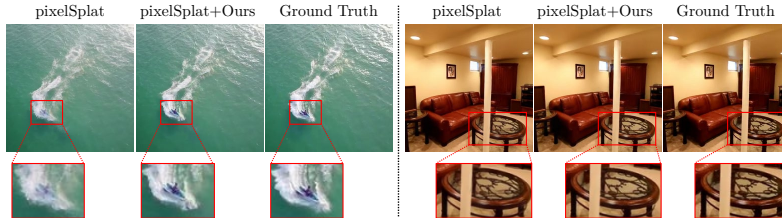
Model	Ref. Type	Speed (sec)			Params (M)	# Ref.
		DTU	Syn.	RFF.		
C2-Matching [22]	Single	0.32	1.22	0.49	8.9	1
MReFSR [61]	Multi	1.99	8.98	3.25	23.7	8
NeRFlix [66]		1.48	2.35	1.95	35.2	2
Ours		<b>1.07</b>	<b>1.68</b>	<b>1.61</b>	<b>9.1</b>	<b>8</b>

## 4.2 Comparison with Generalizable NeRFs

**Experimental Setup.** We conduct comparative experiments with state-of-the-art generalizable NeRF methods. The models include IBRNet [48], MVSNeRF [6], Neuray [30], GeoNeRF [23], GPNR [41], ContraNeRF [56], GNT [47], and MuRF [52]. Our model aims to enhance the rendering performance of the G-NeRF models. Therefore, we selected baseline models for comparison, and evaluated the enhanced results in comparison to the performance of the baseline models. IBRNet, GNT, GeoNeRF, Neuray, and MuRF are chosen as the baseline model for this purpose. The evaluation is conducted using a total of three datasets: DTU [21], Synthetic NeRF [33], and Real Forward-Facing [32]. Eval-



**Fig. 5:** Qualitative comparisons of generalizable NeRF models on DTU, Real Forward-Facing, and Synthetic NeRF datasets.



**Fig. 6:** Qualitative comparisons of pixelSplat results on ACID(left) and RealEstate10k(right) datasets.

uation metrics include PSNR, SSIM [19], and LPIPS [63]. Additionally, input images that are used for NVS are utilized as reference images for enhancement purposes. Base models use individually learned pretrained models. Pretrained GeoNeRF was trained on a dataset that includes part of the DTU test dataset; therefore, GeoNeRF is excluded from the evaluation of the DTU dataset.

**Quantitative Analysis.** As shown in Table 1, our models lead to consistent improvements across all metrics and all datasets. In the DTU dataset, performance improved up to 0.35 from the baseline model. In the Synthetic NeRF dataset, performance improved up to 0.6 from the baseline model. In the real forward-facing dataset, performance improved up to 0.39 from the baseline model.

**Qualitative Comparison and Analysis.** As shown in Fig. 5, our models have more apparent textures and high-frequency details than IBRNet, GeoNeRF, GNT, Neuray, and MuRF. In addition, our models can remove blurry and foggy effects generated by incorrectly estimated scene geometry.

**Experiment on pixelSplat [3].** We quantitatively and qualitatively demonstrate the performance improvement using our method in pixelSplat [3], a generalizable rendering model based on 3D gaussian splatting [25]. For this experiment, our model is trained on the same ACID [28] and RealEstate10k [67] datasets as pixelSplat. As shown in Table 2, it has the highest performance compared to existing models in both datasets. In addition, there is a performance improvement when using ours compared to the results of pixelSplat. As shown in Fig. 6, it can be qualitatively confirmed that our model perfectly synthesizes complex objects and textures without artifacts.

**Novel view synthesis with per-scene finetuning.** For comparison with recent novel view synthesis models, we perform per-scene finetuning on G-NeRF models and image enhancement on the finetuned G-NeRF models. As shown in Table 3, G-NeRF models with ours (Neuray-ft [30]+Ours, GNT-ft [47]+Ours) are higher than NeRF-based models [33, 50] light-field based models [32, 42] on Real Forward-Facing datasets. In addition, G-NeRF models with ours (Neuray-ft+Ours, GNT-ft+Ours) produce high performance improvements for most metrics than those without.

### 4.3 Comparison with Image Enhancement Models

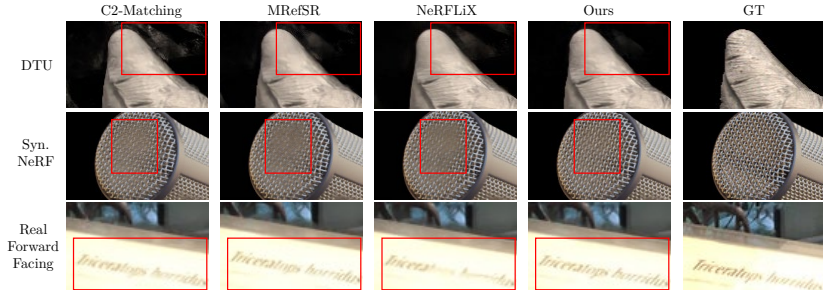
**Experimental Setup.** In this section, we compare performance with existing reference-based or frame-based image enhancement models. Among state-of-the-art models, we select C2-matching [22] to represent single reference-based models and MRefSR [61] to represent multi reference-based models. We also select NeRFLiX [66], a frame-based model recently applied to NeRF by applying video frame interpolation. C2-matching and MRefSR are trained in our training dataset. Nerflix is fine-tuned on our training dataset using a pretrained model. For all models, including ours, G-NeRF base model is Neuray [30].

**Quantitative Analysis.** As shown in Table 5, our model outperforms C2-matching, MRefSR, and NeRFLiX for all metrics and datasets. We also compare the inference time, model trainable parameters, and the number of reference view inputs. As shown in Table 4, C2-Matching has the least parameters and uses only one reference view, so it has the fastest inference time but the lowest PSNR value. On the other hand, compared to other models that use multi-reference views (MRefSR, NeRFLiX), our model has the smallest number of parameters at 9.1 million and the fastest inference time for all datasets. Additionally, our model has the highest PSNR value compared to all models.

**Qualitative Comparison and Analysis.** As shown in Fig. 7, our model provides image clarity and texture details in comparison to other models. Fine textures such as mic patterns and letters are well transferred from source images. In addition, our model effectively removes fog effects and generates clear textures, thereby promoting high-quality representation of images and minimizing noise.

**Table 5:** Novel view synthesis results with various reference-based image enhancement models. For a fair comparison, we use Neuray as the G-NeRF baseline for all models.

Model	DTU			Synthetic NeRF			Real Forward-Facing		
	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )
C2-Matching [22]	28.38	0.910	0.105	28.49	0.928	0.068	25.55	0.839	0.154
MRefSR [61]	28.52	0.903	0.104	28.64	0.917	0.071	25.64	0.844	0.144
NeRFLiX [66]	28.54	0.906	0.103	28.79	0.926	0.062	25.65	0.844	0.143
Ours	<b>28.72</b>	<b>0.920</b>	<b>0.101</b>	<b>28.96</b>	<b>0.936</b>	<b>0.062</b>	<b>25.82</b>	<b>0.850</b>	<b>0.142</b>



**Fig. 7:** Qualitative comparisons of reference-based image enhancement on DTU, Real Forward-Facing, and Synthetic NeRF datasets.

#### 4.4 Ablation Studies

In this section, we conduct two ablation studies. First, we analyze the influence of alpha and correlation values on the RayDCN. Second, we examine the effect of varying model architectures on performance in the feature aggregation process. **RayDCN.** We perform an ablation study to determine the effectiveness of geometry proxy  $X^{\alpha}$  and image-based correlation  $Corr$  as inputs to RayDCN. We categorize the input types for RayDCN into three cases: the first utilizes only  $Corr$ , the second employs only  $\alpha$ , and the third incorporates both inputs. The first case represents the original DCN, while the third case corresponds to the proposed complete RayDCN. Table 6 shows that performance diminishes when alpha and correlation are used singly; the proposed RayDCN, which utilizes alpha and correlation, shows substantial performance improvement.

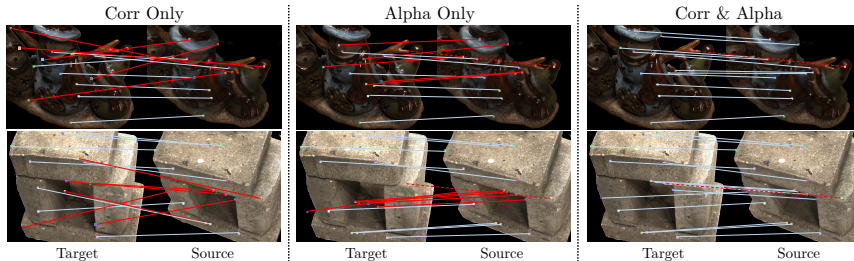
For qualitative comparison, Fig. 8 present offset estimation visualizations under varied input conditions. For a randomly sampled single coordinate  $p_0$  in the rendered image, offset estimation yields a projected point  $p' = p + \mathbf{H}(p)$  on the source image. As a result, employing complete RayDCN(Corr & alpha) ensures that  $p'$  and  $p$  correspond to the same 3D point, effectively constraining the position of  $p'$  within the epipolar line and demonstrating improved accuracy in estimating 3D point on the surface of an object.

**Table 6:** Ablation study of different input types for RayDCN with *Corr* and *Alpha* as inputs. **Bold** indicates the best, and underline indicates the second best results.

Input Type	DTU			Synthetic NeRF			Real Forward-Facing		
	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )
<i>Corr</i> Only	28.56	0.916	0.105	28.68	<b>0.936</b>	0.065	25.67	0.832	0.156
<i>Alpha</i> Only	28.57	<u>0.917</u>	0.109	28.77	0.937	<u>0.065</u>	25.73	<u>0.846</u>	<u>0.155</u>
<i>Corr &amp; Alpha</i>	<b>28.72</b>	<b>0.920</b>	<b>0.101</b>	<b>28.96</b>	<b>0.936</b>	<b>0.062</b>	<b>25.82</b>	<b>0.850</b>	<b>0.142</b>

**Table 7:** Ablation study of the aggregation types. **Bold** indicates the best, and underline indicates the second best results.

Agg. Type	DTU			Synthetic NeRF			Real Forward-Facing		
	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )	PSNR( $\uparrow$ )	SSIM( $\uparrow$ )	LPIPS( $\downarrow$ )
MLP	28.54	0.915	0.110	28.61	0.934	0.068	25.63	0.842	0.153
Multi-Head Attention	<u>28.58</u>	<u>0.916</u>	<u>0.106</u>	<u>28.69</u>	<u>0.935</u>	<u>0.066</u>	<u>25.67</u>	<u>0.844</u>	0.154
TPFormer	<b>28.72</b>	<b>0.920</b>	<b>0.101</b>	<b>28.96</b>	<b>0.936</b>	<b>0.062</b>	<b>25.82</b>	<b>0.850</b>	<b>0.142</b>

**Fig. 8:** Offset estimation results on the DTU dataset: **Blue** lines indicate correct estimates, **Red** lines indicate incorrect estimates, and dotted lines signify missing offsets.

**TPFormer.** We conduct an ablation study on model architecture for source image feature aggregation. The first is the MLP layer structure used in IBRNet [48] and Neuray [30], the second is the Multi-Head Attention (MHA) proposed by GNT [47], and the third is TPFormer proposed in this paper. As shown in Table 7, TPFormer shows the best performance for the three metrics for all three datasets. Specifically, TPFormer outperforms the MLP layer, as well as MHA. The superior performance indicates the effectiveness of TPFormer in aggregating source image features, making it a promising choice for enhancing model architecture in tasks related to feature fusion and aggregation.

## 5 Conclusion

In this paper, we propose a geometry-driven multi-reference texture transfer network, called GMT, to enhance generalizable neural rendering. We also propose novel modules, RayDCN and TPFormer, for aggregating the local texture of source images using a rendered image and an alpha point cloud generated from the volume densities of G-NeRF. We demonstrate consistent improvement for various datasets using most G-NeRF as based models.

### Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636).

### References

1. Cao, A., Rockwell, C., Johnson, J.: Fwd: Real-time novel view synthesis with forward warping and depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15713–15724 (2022)
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. In: European conference on computer vision. pp. 325–342. Springer (2022)
3. Charatan, D., Li, S., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. arXiv preprint arXiv:2312.12337 (2023)
4. Chaurasia, G., Duchene, S., Sorkine-Hornung, O., Drettakis, G.: Depth synthesis and local warps for plausible image-based navigation. ACM Transactions on Graphics (TOG) **32**(3), 1–12 (2013)
5. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision. pp. 333–350. Springer (2022)
6. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021)
7. Choi, I., Gallo, O., Troccoli, A., Kim, M.H., Kautz, J.: Extreme view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7781–7790 (2019)
8. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3075–3084 (2019)
9. Debevec, P., Yu, Y., Borshukov, G.: Efficient view-dependent image-based rendering with projective texture-mapping. In: Rendering Techniques’ 98: Proceedings of the Eurographics Workshop in Vienna, Austria, June 29–July 1, 1998 9. pp. 105–116. Springer (1998)
10. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
11. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022)
12. Du, Y., Smith, C., Tewari, A., Sitzmann, V.: Learning to render novel views from wide-baseline stereo pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4970–4980 (2023)
13. Flynn, J., Broxton, M., Debevec, P., DuVall, M., Fyffe, G., Overbeck, R., Snavely, N., Tucker, R.: Deepview: View synthesis with learned gradient descent. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2367–2376 (2019)

14. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5501–5510 (2022)
15. Gatys, L., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. *Advances in neural information processing systems* **28** (2015)
16. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
17. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: *The Lumigraph*. Association for Computing Machinery, New York, NY, USA, 1 edn. (2023), <https://doi.org/10.1145/3596711.3596760>
18. Hedman, P., Philip, J., Price, T., Frahm, J.M., Drettakis, G., Brostow, G.: Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (ToG)* **37**(6), 1–15 (2018)
19. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
20. Huang, J., Thies, J., Dai, A., Kundu, A., Jiang, C., Guibas, L.J., Nießner, M., Funkhouser, T., et al.: Adversarial texture optimization from rgb-d scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1559–1568 (2020)
21. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanaes, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 406–413 (2014)
22. Jiang, Y., Chan, K.C., Wang, X., Loy, C.C., Liu, Z.: Robust reference-based super-resolution via c2-matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2103–2112 (2021)
23. Johari, M.M., Lepoittevin, Y., Fleuret, F.: Geonerf: Generalizing nerf with geometry priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18365–18375 (2022)
24. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)* **35**(6), 1–10 (2016)
25. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* **42**(4) (2023)
26. Levoy, M., Hanrahan, P.: Light field rendering. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. p. 31–42. SIGGRAPH '96, Association for Computing Machinery, New York, NY, USA (1996). <https://doi.org/10.1145/237170.237199>, <https://doi.org/10.1145/237170.237199>
27. Li, Y., Luo, Y., Lu, J.: Reference-guided deep deblurring via a selective attention network. *Applied Intelligence* pp. 1–13 (2022)
28. Liu, A., Tucker, R., Jampani, V., Makadia, A., Snavely, N., Kanazawa, A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14458–14467 (2021)
29. Liu, C., Hua, Z., Li, J.: Reference-based dual-task framework for motion deblurring. *The Visual Computer* pp. 1–15 (2023)
30. Liu, Y., Peng, S., Liu, L., Wang, Q., Wang, P., Theobalt, C., Zhou, X., Wang, W.: Neural rays for occlusion-aware image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7824–7833 (2022)



31. Lu, L., Li, W., Tao, X., Lu, J., Jia, J.: Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6368–6377 (2021)
32. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)* **38**(4), 1–14 (2019)
33. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**(1), 99–106 (2021)
34. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989* (2022)
35. Penner, E., Zhang, L.: Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG)* **36**(6), 1–11 (2017)
36. Pesavento, M., Volino, M., Hilton, A.: Attention-based multi-reference learning for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14697–14706 (2021)
37. Riegler, G., Koltun, V.: Free view synthesis. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16. pp. 623–640. Springer (2020)
38. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12892–12901 (2022)
39. Shum, H., Kang, S.B.: Review of image-based rendering techniques. In: Visual Communications and Image Processing 2000. vol. 4067, pp. 2–13. SPIE (2000)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
41. Suhail, M., Esteves, C., Sigal, L., Makadia, A.: Generalizable patch-based neural rendering. In: European Conference on Computer Vision. pp. 156–174. Springer (2022)
42. Suhail, M., Esteves, C., Sigal, L., Makadia, A.: Light field neural rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8269–8279 (2022)
43. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5459–5469 (2022)
44. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)* **38**(4), 1–12 (2019)
45. Thies, J., Zollhöfer, M., Theobalt, C., Stamminger, M., Nießner, M.: Ignor: Image-guided neural object rendering. *arXiv preprint arXiv:1811.10720* (2018)
46. Trevithick, A., Yang, B.: Grf: Learning a general radiance field for 3d representation and rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15182–15192 (2021)
47. Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z., et al.: Is attention all nerf needs? *arXiv preprint arXiv:2207.13298* (2022)
48. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021)

49. Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5610–5619 (2021)
50. Wizarawongsa, S., Phongthawee, P., Yenphraphai, J., Suwajanakorn, S.: Nex: Real-time view synthesis with neural basis expansion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8534–8543 (2021)
51. Xia, B., Tian, Y., Hang, Y., Yang, W., Liao, Q., Zhou, J.: Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2768–2776 (2022)
52. Xu, H., Chen, A., Chen, Y., Sakaridis, C., Zhang, Y., Pollefeys, M., Geiger, A., Yu, F.: Murf: Multi-baseline radiance fields. arXiv preprint arXiv:2312.04565 (2023)
53. Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., Neumann, U.: Point-nerf: Point-based neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5438–5448 (2022)
54. Xu, Z., Bi, S., Sunkavalli, K., Hadap, S., Su, H., Ramamoorthi, R.: Deep view synthesis from sparse photometric images. ACM Transactions on Graphics (ToG) **38**(4), 1–13 (2019)
55. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5791–5800 (2020)
56. Yang, H., Hong, L., Li, A., Hu, T., Li, Z., Lee, G.H., Wang, L.: Contranerf: Generalizable neural radiance fields for synthetic-to-real novel view synthesis via contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16508–16517 (2023)
57. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 767–783 (2018)
58. Youngho, Y., Kuk-Jin, Y.: Cross-guided optimization of radiance fields with multi-view image super-resolution for high-resolution novel view synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. p. 12428–12438 (2023)
59. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenotrees for real-time rendering of neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5752–5761 (2021)
60. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
61. Zhang, L., Li, X., He, D., Li, F., Ding, E., Zhang, Z.: Lmr: A large-scale multi-reference dataset for reference-based super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13118–13127 (2023)
62. 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. In: European Conference on Computer Vision. pp. 648–664. Springer (2022)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
64. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7982–7991 (2019)

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. In: Proceedings of the European conference on computer vision (ECCV). pp. 88–104 (2018)
66. 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12363–12374 (2023)
67. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
68. Zou, H., Suganuma, M., Okatani, T.: Reference-based motion blur removal: Learning to utilize sharpness in the reference image. arXiv preprint arXiv:2307.02875 (2023)