High-Resolution and Few-shot View Synthesis from Asymmetric Dual-lens Inputs

Ruikang Xu[1], Mingde Yao[2], Yue Li[1], Yueyi Zhang[1], Zhiwei Xiong[1]✉

[1] MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
xurk@mail.ustc.edu.cn, zwxiong@ustc.edu.cn
[2] The Chinese University of Hong Kong

Abstract. Novel view synthesis has achieved remarkable quality and efficiency by the paradigm of 3D Gaussian Splatting (3D-GS), but still faces two challenges: 1) significant performance degradation when trained with only few-shot samples due to a lack of geometry constraint, and 2) incapability of rendering at a higher resolution that is beyond the input resolution of training samples. In this paper, we propose Dual-Lens 3D-GS (DL-GS) to achieve high-resolution (HR) and few-shot view synthesis, by leveraging the characteristics of the asymmetric dual-lens system commonly equipped on mobile devices. This kind of system captures the same scene with different focal lengths (*i.e.*, wide-angle and telephoto) under an asymmetric stereo configuration, which naturally provides geometric hints for few-shot training and HR guidance for resolution improvement. Nevertheless, there remain two major technical problems to achieving this goal. First, how to effectively exploit the geometry information from the asymmetric stereo configuration? To this end, we propose a consistency-aware training strategy, which integrates a dual-lens-consistent loss to regularize the 3D-GS optimization. Second, how to make the best use of the dual-lens training samples to effectively improve the resolution of newly synthesized views? To this end, we design a multi-reference-guided refinement module to select proper telephoto and wide-angle guided images from training samples based on the camera pose distances, and then exploit their information for high-frequency detail enhancement. Extensive experiments on simulated and real-captured datasets validate the distinct superiority of our DL-GS over various competitors on the task of HR and few-shot view synthesis. The implementation code is available at https://github.com/XrKang/DL-GS.

Keywords: Novel View Synthesis · 3D Gaussian Splatting · Asymmetric Dual-lens System · Few-shot Training · Super-resolution

1 Introduction

Novel view synthesis (NVS) aims to generate images at arbitrary viewpoints of a 3D scene, which is a fundamental task in computer vision and finds widespread applications in mobile scenarios [6,11,26], such as virtual/augmented reality [14, 36] and immersive telepresence [15, 35]. Neural Radiance Field (NeRF)-based methods [2, 2, 3, 30] have made great success on this task, using the implicit radiance field for high-fidelity rendering. However, the high training and rendering costs of NeRF-based methods limit their applications [17,32,59]. To achieve

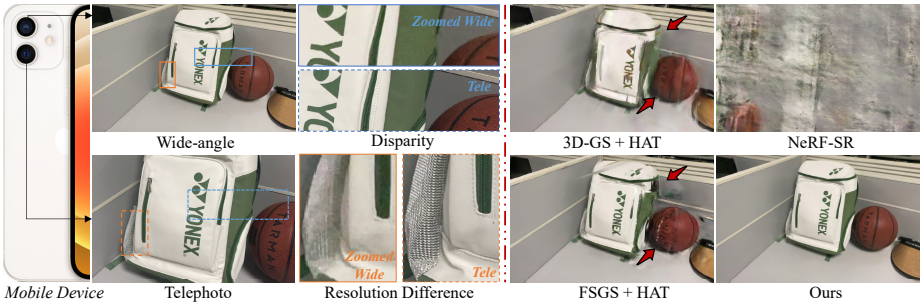


Fig. 1: Left: An asymmetric dual-lens system naturally provides geometric information and additional HR details for high-quality rendering. Right: Visual comparison of view synthesis with 5 training samples for 2× SR, where 3D-GS [19] and FSGS [65] employ HAT [9] for resolution improvement. NeRF-SR [49] directly synthesizes HR views but fails in this few-shot case. Our method can synthesize HR views with coherent geometry.

real-time and high-fidelity rendering, 3D Gaussian Splatting (3D-GS) [19] has recently emerged as a powerful paradigm, which replaces the implicit radiance field of NeRF with an explicit representation based on 3D Gaussians. Despite achieving effectiveness and efficiency in photo-realistic rendering, two challenges still hinder the practicality of NVS: 1) A large set of training samples is typically required, and performance would significantly drop as the number of samples is reduced [13, 42]. 2) The rendering resolution is dependent on the collected samples, which may not be competent for ultra-high-resolution applications [24, 40].

For the first challenge, the few-shot training samples would lead to an incorrect convergence or optimization failure [57, 59] due to insufficient geometric constraints. To tackle this challenge, introducing additional geometric information for regularization is a promising way [12, 13, 48, 55, 65], such as depth from COLMAP [37] or depth from a single view. However, the former tends to be too sparse while the latter could be inconsistent within multiple views due to the inherent ambiguity of monocular depth estimation [4, 5]. As for the second challenge, existing methods adopt the idea of super-resolution (SR) to recover the high-frequency details of synthesized views [49, 58] or introduce additional high-resolution (HR) images as references to enhance details [18]. However, the performance of using SR models tends to be limited due to a lack of direct in-domain guidance, while collecting additional HR reference images could be inconvenient in practice [23, 51].

In this paper, we propose a new solution based on 3D-GS, termed DL-GS, which leverages the characteristics of the asymmetric dual-lens system for few-shot view synthesis while overcoming the resolution limitation. Specifically, the dual-lens system is widely equipped on mobile devices (*e.g.*, smartphones), which consists of a wide-angle lens with a short focal length and a telephoto lens with a large focal length, capturing the same scene with different field-of-views (FoVs). As shown in Fig. 1, the characteristics of this kind of system are well-suited to address the aforementioned challenges that obstruct the applications of NVS: 1) Combining the wide-angle and telephoto images forms an asymmetric stereo configuration, which stores the geometric information to facilitate the few-shot

training. 2) The telephoto images have higher resolution than the wide-angle ones within the overlapped FoV, naturally providing additional HR guidance to improve the resolution of newly synthesized views.

Nevertheless, there remain two major technical problems to achieving the success of DL-GS. At first, exploiting the geometric information from the asymmetric stereo configuration is non-trivial, since different focal lengths make explicit disparity estimation difficult. To this end, we design a consistency-aware training strategy to implicitly exploit the geometry information. Specifically, we first pre-upsample the wide-angle images to match the resolution of the telephoto images within the overlapped FoV. This processing also offers better 3D Gaussians initialization due to inputting a denser point cloud. Then, we propose a dual-lens-consistent loss to regularize 3D Gaussians optimization by enforcing the view consistency between the rendered wide-angle views and the corresponding telephoto views. Meanwhile, we introduce a depth-wise loss to regularize the non-overlapped area of two lenses, which supervises the distribution difference between the rendered depth and the estimated depth. As such, we can synthesize images with accurate geometric structures even using few-shot training samples.

The second problem is how to make the best use of the dual-lens training samples to refine the newly synthesized views. Although we can directly synthesize wide-angle views at the target resolution through the pre-upsampling, they still lack realistic high-frequency details. To address this problem, we design a multi-reference-guided refinement module, which first selects proper telephoto and wide-angle guided images from training samples based on the distance of camera poses. Then, we exploit the HR guidance from selected telephoto images through similarity-aware attention, while simultaneously utilizing the multi-view information of selected wide-angle images with pixel-wise attention. Notably, the multi-reference-guided refinement is supervised by the self-training scheme without the ground-truth data, making it suitable for practical deployment.

Based on the above designs to effectively utilize information from the dual-lens system, our DL-GS generates significantly improved view synthesis results over representative competitors, as shown in Fig. 1. To comprehensively evaluate the performance of this new solution for NVS, we build a simulated dataset and a real-captured dataset. For the simulated dataset, we utilize the pseudo stereo pairs from [47] to generate dual-lens image pairs of forward-facing scenes. For the real-captured dataset, we collect a set of dual-lens image pairs of static scenes from different views using an iPhone12, which includes both inward-facing (360°) and forward-facing scenes. Extensive experiments demonstrate the distinct superiority of our DL-GS over various solutions.

Contributions of this paper are summarized as follows:

- We propose a new 3D-GS-based solution for HR and few-shot views synthesis by leveraging the characteristics of the asymmetric dual-lens system.
- We propose a consistency-aware training strategy to exploit the geometric information of dual-lens pairs for regularizing 3D Gaussians optimization.
- We propose a multi-reference-guided refinement module to enhance newly synthesized views by making the best use of dual-lens training samples.

• Extensive experiments on both simulated and real-captured datasets validate the superiority of our solution over various competitors.

2 Related Work

2.1 Novel View Synthesis

With the fast development of NVS, existing works have achieved real-time and photorealistic rendering [2,11,19,32]. However, most methods still require plenty of training samples to render a single scene, while their rendered resolution is constrained by that of training samples, thereby limiting their practical applications. Recently, several methods have been proposed to address either of the two challenges, and here we provide a brief overview of their development.

Few-shot Novel View Synthesis. Few-shot NVS aims to render a scene with limited training samples [39, 41, 57], and it can be roughly divided into three classes: Existing works for few-shot NVS can be roughly divided into three classes: (a) Early methods utilize a large dataset to train a general model, and then fine-tune the pre-trained model to the target scene, using a few training samples [8, 59]. However, these methods require large efforts to collect accurate and diverse 3D scenes, and might suffer from the domain gap problem [64] between the collected dataset and the target scene. (b) Another category of methods utilizes the geometric or semantic constraints in the appearances of synthesized views [20,33], but this regularization cannot guarantee the complete 3D geometric reconstruction due to multiple layouts of scenes [50]. (c) Recently methods estimate the depth map (*i.e.*, using COLMAP [37] or monocular depth estimation [13,42,50,55,65]) to provides additional geometric information for regularization and made promising progress. However, depth maps from COLMAP are too sparse, and those from monocular estimation are inconsistent across multiple views due to inherent ambiguity, thereby limiting their performance. Different from previous methods, the asymmetric dual-lens system naturally stores the geometric information to enhance the few-shot training.

High-Resolution Novel View Synthesis. Previous methods have explored generating HR views through various strategies, such as super-sampling [49], exploiting external priors from SR models [58], and using additional HR images as references [18]. Despite the remarkable progress achieved, there are still challenges in efficiently acquiring and effectively utilizing high-frequency information, whether from pre-trained SR models or additional HR images. Specifically, the pre-trained SR models only provide high-frequency hints from extra scenes and additional HR reference image collection is typically difficult in practical applications. Moreover, existing works primarily focus on the HR NVS based on the NeRF paradigm, with few efforts to explore 3D-GS. In this paper, we leverage the characteristic of asymmetric dual-lens systems, which naturally introduce HR information for NVS from telephoto images.

2.2 Dual-lens System

The asymmetric dual-lens system consists of two lenses with different focal lengths to capture the same scene with different FoVs, equipped on most mobile

devices such as smartphones. This system, also known as the dual-camera/zoomed system, is typically configured with a wide-angle lens and a telephoto lens. Specifically, the wide-angle lens is the main lens with a short focal length and a large FoV, while the telephoto lens has a large focal length and a narrow FoV. In the overlapped FoV, the telephoto lens exhibits higher resolution than the wide-angle lens, opening possibilities for various applications. Existing works exploit the characteristics of this kind of system to address different tasks [1,27,31]. For instance, some works focus on estimating correspondences between the two lenses within the overlapped FoV [10,43,61]. Additionally, dual-lens SR [51,56,60,63] aims to utilize telephoto images as HR guidance to improve the resolution of wide-angle images. To the best of our knowledge, our work is the first effort to leverage the asymmetric dual-lens system to address the task of NVS.

3 Preliminaries

3.1 Background: 3D Gaussian Splatting

3D-GS represents a scene in an explicit paradigm using a set of 3D Gaussians, where each 3D Gaussian is defined by its mean value $\boldsymbol{\mu} \in \mathbb{R}^3$ and covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$ in the position $\boldsymbol{x} \in \mathbb{R}^3$ in the 3D space, as

$$G(\boldsymbol{x}) = \boldsymbol{e}^{(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))}. \tag{1}$$

Specifically, Σ is defined by a rotation matrix R and a scale matrix S to ensure it is positive semi-definite following the physical constraints, as $\Sigma = RSS^TR^T$. Then, each 3D Gaussian is projected to 2D space for rendering, generating the corresponding 2D Gaussian with covariance matrix $\Sigma^{'} \in \mathbb{R}^{2\times2}$. Specifically, $\Sigma^{'}$ is calculated by the world-to-camera matrix W and the Jacobian J of the affine approximation of perspective projection transformation, as $\Sigma^{'} = JW\Sigma W^TJ^T$. In addition, each Gaussian stores an opacity $o \in \mathbb{R}$ and a view-dependent color $\boldsymbol{c} \in \mathbb{R}^3$ represented by the spherical harmonic coefficients.

Each pixel color $\boldsymbol{C} \in \mathbb{R}^3$ of the synthesized 2D image $\hat{I}$ can be computed by blending N sequential 2D Gaussians overlapping on the pixel, denoted as

$$\boldsymbol{C} = \sum_{i=1}^{N} \boldsymbol{c}_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \tag{2}$$

where α_i is estimated by its corresponding the opacity and 2D Gaussian, following the principle of point-based rendering [19]. After completing the rendering of all pixels, the 3D Gaussians are optimized by supervising the loss function $\mathcal{L}_{GS}$, which is calculated by a $\mathcal{L}_1$ regularization combined with a D-SSIM [54] term $\mathcal{L}_{SSIM}$, denoted as

$$\mathcal{L}_{GS} = (1-\lambda)\mathcal{L}_1(I,\hat{I}) + \lambda\mathcal{L}_{SSIM}(I,\hat{I}), \tag{3}$$

where $\hat{I}$ is the synthesized view and I is the ground-truth data, and $\lambda = 0.2$.

3.2 Problem Formulation

Problem 1: Few-shot Training Samples. To represent a scene, vanilla 3D-GS requires a dense set of training samples to optimize the 3D Gaussians [19, 65]. However, collecting plenty of training samples is difficult and inconvenient for

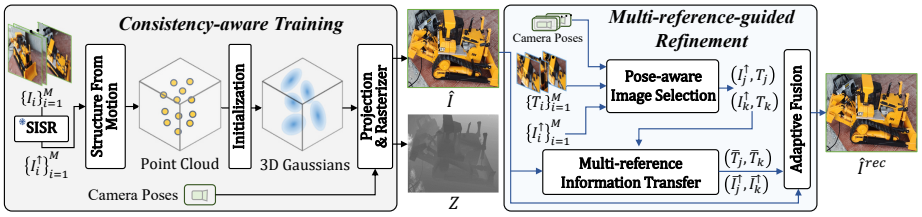


Fig. 2: Overview of DL-GS. Consistency-aware training is designed to optimize 3D Gaussians with few-shot training samples by elaborated regularization terms. Multi-reference-guided refinement aims to exploit the guidance information from the training sample for enhancing the newly synthesized views, overcoming the resolution limitation.

users. The performance will significantly drop when the number of samples is reduced to a few (≤ 20), since the lack of geometric constraints would lead the model to optimization failure or overfit on limited data [33, 42, 59]. To overcome this problem, an alternative way is incorporating the additional geometric information for regularization in the optimization process [13, 50, 65]. In this paper, we leverage the inherent geometric information from the dual-lens system as regularization to facilitate the few-shot training.

Problem 2: Resolution Limitation. To obtain the 3D representation, vanilla 3D-GS is optimized by the image-level reconstruction loss as Eq. 3. This form of supervision constrains the resolution of synthesized views to the same as that of the training samples, which may not be competent for ultra-high-resolution applications [24, 40]. An effective formulation to address this problem is introducing additional HR guidance to reconstruct the synthesized views [18, 23, 49, 51]. In this paper, we select guided images from dual-lens training samples, and then exploit their information to reconstruct the synthesized views.

4 Dual-lens Gaussian Splatting

4.1 Motivation and Overview

We propose a new solution based on 3D-GS, termed DL-GS, which leverages the characteristics of the asymmetric dual-lens system to achieve few-shot view synthesis while exceeding the resolution limitation of training samples. It consists of two key components: a consistency-aware training strategy to regularize the 3D-GS optimization, and a multi-reference-guided refinement module to reconstruct the newly synthesized views. The overview pipeline of DL-GS is in Fig. 2.

Given a set of wide-angle images $\{I_i\}_{i=1}^{M}$ along with their corresponding telephoto images $\{T_i\}_{i=1}^{M}$ (M is the number of training samples), we first optimize 3D-GS by our consistency-aware training strategy to address the few-shot training. Specifically, our training strategy utilizes the inherent geometric information of the dual-lens system as regularization for few-shot training. However, exploiting the geometric information is challenging, since the different focal lengths of the two lenses result in the stereo configuration exhibiting resolution asymmetry and being limited to the overlapped FoV area. To address this challenge, our proposed training strategy optimizes 3D-GS by introducing dual-lens-consistent and depth-wise losses. Further details are provided in Sec. 4.2.

After the 3D-GS optimization, we can synthesize a new wide-angle view $\hat{I}$ at the target resolution through pre-upsampling, but it still lacks high-frequency details. Therefore, the multi-reference-guided refinement module is proposed to reconstruct the details of $\hat{I}$. Specifically, it first selects telephoto and wide-angle guided images from the training samples ($\{I_i\}_{i=1}^{M}$, $\{T_i\}_{i=1}^{M}$), then exploits their information to generate the final results $\hat{I}^{rec}$. The module is supervised in a self-training scheme without requiring ground-truth data, advancing it to practical deployment. We provide a detailed illustration of the module in Sec. 4.3.

4.2 Consistency-aware Training Strategy

Two lenses of the dual-lens system can form a stereo configuration, which offers geometric information to serve as the regularization for the few-shot training. However, the wide-angle and telephoto lenses have different focal lengths, hence, the stereo configuration is resolution asymmetry and is limited to the overlapped FoV area. As a result, explicit disparity estimation is difficult. To address this challenge, we propose a consistency-aware training strategy to optimize 3D-GS by implicitly leveraging the geometric constraint of the two lenses.

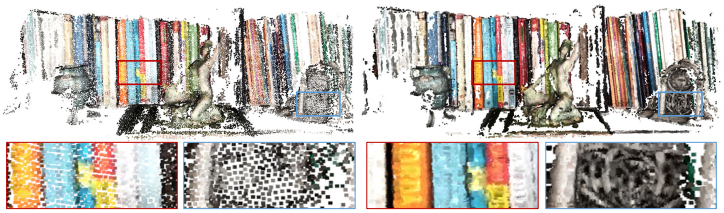
Pre-upsampling. We first pre-upsample the wide-angle images to address the problem of resolution asymmetry. Given a set of wide-angle images $\{I_i\}_{i=1}^{M}$, we utilize a pre-trained single image SR (SISR) network [9] to generate HR counterparts $\{I_i^{\uparrow}\}_{i=1}^{M}$. As a result, the wide-angle and telephoto images are resolution-matching within the overlapped FoV area, facilitating subsequent utilization of geometric information. Meanwhile, the pre-upsampling processing also provides a better 3D Gaussians initialization, since the inputting point cloud is denser and more accurate, as shown in Fig. 3.

Dual-lens-consistent loss. Due to the different focal lengths of the two lenses, explicit disparity estimation cannot be employed to utilize their geometric information directly. Therefore, we propose a dual-lens-consistent loss to implicitly exploit the geometric information, which serves as a regularization term for the 3D-GS optimization. The proposed loss aims to enforce the view consistency between the newly synthesized wide-angle view $\hat{I}$ and the corresponding telephoto image $T \in \{T_i\}_{i=1}^{M}$, denoted as

$$\mathcal{R}_c = V\|\mathrm{Warp}_c(\hat{I}) - T\|_1, \tag{4}$$

where $\mathrm{Warp}_c(\cdot)$ represents the warping operation for the center area of $\hat{I}$ using the corresponding optical flow from the target image $I^{\uparrow}$. V is the visibility mask of two views [21]. We employ a pre-trained estimator (*e.g.*, RAFT [45]) for efficient optical flow computation. Specifically, the estimated optical flow is the correspondence of the two lenses with different focal lengths, served as a coarse disparity without stereo-calibration [16, 22]. Hence, the dual-lens-consistent loss can implicitly exploit geometric information to facilitate the few-shot training, avoiding additional capture efforts on dual-lens devices.

Depth-wise loss. We further introduce a depth-wise loss as another regularization term to supervise the non-overlapped FoV area of two lenses. Specifically, we obtain the depth D from the image $I^{\uparrow} \in \{I_i^{\uparrow}\}_{i=1}^{M}$ using a pre-trained monocular depth estimator [34], and render the depth Z from 3D-GS using a



(a) w/o Pre-upsampling (b) w/ Pre-upsampling

Fig. 3: Comparison on inputting point clouds. (a) Estimated point cloud without pre-upsampling. (b) Estimated point cloud with pre-upsampling.

differentiable depth rasterizer [44]. To mitigate the scale ambiguity caused by monocular depth estimation, we utilize Pearson correlation [38] to construct the depth-wise loss, which minimizes the normalized distribution difference between Z and D rather than the absolute error, as

$$\mathcal{R}_d = \frac{\mathrm{Cov}\,(Z, D)}{\sqrt{\mathrm{Var}\,(Z)}\sqrt{\mathrm{Var}\,(D)}}, \tag{5}$$

where $\mathrm{Cov}(\cdot,\cdot)$ and $\mathrm{Var}(\cdot)$ represent the covariance and variance, respectively.

The overall loss function of our training strategy is formulated as

$$\mathcal{L} = \mathcal{L}_{GS}(\hat{I}, I^{\uparrow}) + \beta_1 \mathcal{R}_c + \beta_2 \mathcal{R}_d, \tag{6}$$

where β_1 and β_2 are the weighting factors. The effectiveness of each regularization term is validated in Sec. 7.

4.3 Multi-reference-guided Refinement

After 3D-GS optimization, we can directly synthesize wide-angle views at the target resolution through pre-upsampling, but they still lack realistic high-frequency details due to the resolution limitation. To address this problem, we select guided images from dual-lens training samples to refine the newly synthesized view, given that the corresponding telephoto image is unavailable. To this end, we design a multi-reference-guided refinement module, consisting of the pose-aware image selection and the multi-reference information transfer. Specifically, the selection part selects guided images from the dual-lens training samples based on the camera pose distances. The transfer part exploits HR and multi-view information from the guided images to enhance the synthesized view through dedicated attention mechanisms. The flow diagram is depicted in Fig. 4.

Pose-aware image selection. We first select guided images from the training samples to enhance the newly synthesized view $\hat{I}$. Specifically, we calculate the camera pose distance between different views as $\mathrm{Dis}_{pose} = \|\mathcal{P}_i - \mathcal{P}\|_2$, where $\{\mathcal{P}_i\}_{i=1}^{M}$ are the camera poses of wide-angle images $\{I_i^{\uparrow}\}_{i=1}^{M}$ and $\mathcal{P}$ represents the camera pose of the synthesized view $\hat{I}$. Based on this criterion, we can identify the two nearest wide-angle/telephoto pairs $(I_j^{\uparrow}, T_j)$ and $(I_k^{\uparrow}, T_k)$ to the synthesized view $\hat{I}$, where $j \neq k \in [1, M]$.

Multi-reference information transfer. After image selection, we transfer HR information from the selected telephoto images and multi-view information from the selected wide-angle images to enhance the details of synthesized views.

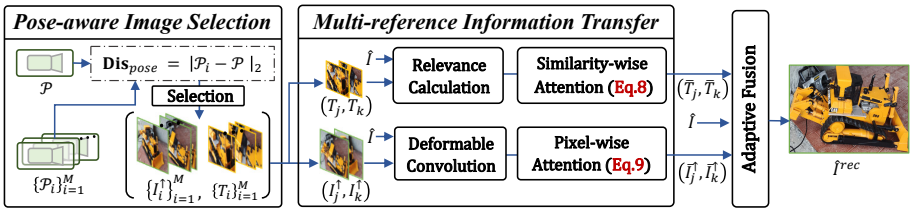


Fig. 4: Pipeline of multi-reference-guided refinement. It first selects the guided images from the dual-lens training samples based on the camera pose distances, then exploits their information by similarity-wise attention and pixel-wise attention mechanisms.

Specifically, we propose a similarity-aware attention mechanism to exploit the HR information from the two nearest telephoto images (T_j, T_k). Here, we take the processing of T_j as an illustration. We first map T_j and $\hat{I}$ to an embedding space for efficient and robust similarity calculation. Then, we unfold the embedded features to a set of patches, and calculate the cosine similarity $r_{n,m}$ between the n-th patch of embedded T_j and the m-th patch of embedded $\hat{I}$. Based on the similarity, we can obtain the index map H to select similar patches from the original T_j, and calculate the similarity-wise attention map S to weight the selected patches, denoted as

$$h_n = \underset{m}{\operatorname{argmax}}\ r_{n,m}, \quad s_n = \max_m r_{n,m}, \tag{7}$$

where h_n is the n-th element of H, representing the index of the most similar patch in the telephoto image T_j to the m-th patch in the synthesized view $\hat{I}$. s_n is the n-th element of S, representing the corresponding confidence of the selected patch. These two matrices (H and S) model the correspondence between the synthesized view $\hat{I}$ and the neighbor telephoto image T_j.

Once obtained H and S, we utilize them to transfer the HR information from T_j to $\hat{I}$, and generate a *pseudo* telephoto image $\bar{T}_j$ as

$$\bar{T}_j = \text{Index}(T_j, H) \odot S, \tag{8}$$

where $\text{Index}(\cdot,\cdot)$ and $\odot$ represent index selection and element-wise multiplication, respectively. The other telephoto image T_k also performs the same operation to generate the *pseudo* telephoto image $\bar{T}_k$. As a result, both *pseudo* telephoto images, $\bar{T}_j$ and $\bar{T}_k$, provide HR information spatially aligned with the synthesized view $\hat{I}$ for the final adaptive fusion (see Fig. 4).

Meanwhile, inspired by multi-frame reconstruction [7, 23, 53], we propose to exploit multi-view information from selected wide-angle images $(I_j^{\uparrow}, I_k^{\uparrow})$. Specifically, we first introduce the deformable convolution [46, 52] to align $(I_j^{\uparrow}, I_k^{\uparrow})$ toward the synthesized view $\hat{I}$. Then, we utilize pixel-wise attention to leverage their information to generate the *pseudo* wide-angle images $(\bar{I}_j^{\uparrow}, \bar{I}_k^{\uparrow})$ as

$$(\bar{I}_j^{\uparrow}, \bar{I}_k^{\uparrow}) = value \odot \text{Sigmoid}(query \odot key), \tag{9}$$

where *key* and *value* are mapped from the selected image $(I_j^{\uparrow}, I_k^{\uparrow})$, and *query* is mapped from the synthesized view $\hat{I}$. The sigmoid function $\text{Sigmoid}(\cdot)$ is used to stabilize gradient back-propagation. In this way, the *pseudo* wide-angle im-

ages ($\bar{I}_j^{\uparrow}, \bar{I}_k^{\uparrow}$) contain the multi-view information transferred from other views, facilitating the final adaptive fusion.

Finally, the final reconstructed image $\hat{I}^{rec}$ is generated by fusing the information from *pseudo* telephoto images ($\bar{T}_j, \bar{T}_k$) and *pseudo* wide-angle images ($\bar{I}_j^{\uparrow}, \bar{I}_k^{\uparrow}$) with the synthesized view $\hat{I}$, as shown in Fig. 4. Benefiting from the effectiveness of the proposed multi-reference information transfer, the final adaptive fusion only requires several convolution layers to achieve a promising performance.

Since ground-truth data is inaccessible in practical development, we propose a self-training loss to optimize the module, denoted as

$$\mathcal{L}_{DL} = \lambda_1 \|\text{Crop}(\hat{I}^{rec}) - T^{align}\|_2 + \lambda_2 \|\hat{I}^{rec} - I^{\uparrow}\|_2 + \lambda_3 \mathcal{L}_{cx}(\text{Crop}(\hat{I}^{rec}), T). \quad (10)$$

Specifically, the first term leverages HR information from the spatially aligned image T^{align}, which is pre-aligned from $T \in \{T_i\}_{i=1}^{M}$ using optical flow estimation. Crop$(\cdot)$ denotes the cropping operation to obtain the overlapped FoV area of two lenses. The second term provides an essential content constraint using the pre-upsampled image $I^{\uparrow} \in \{I_i^{\uparrow}\}_{i=1}^{M}$. The third term directly minimizes differences between $\hat{I}^{rec}$ and the un-aligned telephoto image $T \in \{T_i\}_{i=1}^{M}$, since the contextual loss $\mathcal{L}_{cx}$ [28, 29] measures the similarity without considering the spatial positions. It is worth noting that, all the images in Eq. 10 are divided into patches for efficient training while full-size images are used during inference.

5 Experiments on Simulated Data

5.1 Dataset

We utilize the pseudo stereo image pairs from [47] to simulate dual-lens image pairs. Specifically, scenes in [47] comprise 100 images captured from various viewpoints, and then these images are used to train a NeRF-based method for rendering stereo image pairs. We randomly choose 8 forward-facing scenes. We downsample the left images with a scale factor of 2× to severe as the wide-angle images, while cropping the central area of the right images to obtain the telephoto images. The original left images are preserved as the ground-truth data for evaluation purposes. For each scene, we select every 10-th image as the test set, while we sample 10/20/90 shots from the remaining images for training.

5.2 Implementation Details and Baselines

All components of DL-GS are implemented by using PyTorch 2.1. For optimizing the 3D Gaussians with our training strategy, we initialize the camera poses and point clouds by COLMAP [37]. The weight factors β_1 and β_2 are both set to 0.05 for calculating the loss $\mathcal{L}$. For training the multi-reference-guided refinement, we set λ_1=0.8, λ_2=0.5, and λ_3=0.05 for calculating the loss $\mathcal{L}_{DL}$. We crop image patches with the size of 160×160, and the mini-batch size is set to 16.

To verify the superiority of DL-GS, we compare it with several representative methods including four categories: 1) vanilla 3D-GS [19] followed with SISR: Bicubic, SwinIR [25], HAT [9]; 2) vanilla 3D-GS [19] followed with dual-lens SR: DCSR [51]; 3) HR NVS method: NeRF-SR [49]; 4) few-shot NVS methods followed with HAT [9]: DS-NeRF [13], RegNeRF [33], SparseNeRF [50], FSGS [65].

Table 1: Quantitative comparison with previous state-of-the-art methods on our simulated dataset with different training sample numbers and a scale factor of 2×. **Bold** and underline indicate the best and second best performance, respectively.

Method	10-shot			20-shot			90-shot		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
3D-GS [19] + Bicubic	17.63	0.4979	0.4411	20.75	0.5905	0.3624	24.03	0.7113	0.2677
3D-GS [19] + SwinIR [25]	17.84	0.4988	0.4387	20.87	0.5924	0.3613	24.52	0.7192	0.2628
3D-GS [19] + HAT [9]	17.89	0.4995	0.4402	20.89	0.5931	0.3615	24.57	0.7196	0.2650
3D-GS [19] + DCSR [51]	17.92	0.5037	0.4314	20.92	0.5964	0.3592	24.60	0.7265	0.2582
NeRF-SR [49]	17.40	0.5032	0.4830	20.84	0.5967	0.3726	24.89	0.7394	0.2422
DS-NeRF [13] + HAT [9]	19.05	0.5546	0.4598	21.54	0.5801	0.4366	22.47	0.6002	0.4395
RegNeRF [33] + HAT [9]	18.78	0.5539	0.4573	20.18	0.5613	0.4476	22.34	0.6259	0.4040
SparseNeRF [50] + HAT [9]	19.12	0.5441	0.4482	21.31	0.5724	0.4398	22.49	0.6329	0.4006
FSGS [65] + HAT [9]	19.09	0.5511	0.4321	20.68	0.5897	0.3637	24.42	0.7183	0.2526
Ours	**19.67**	**0.5772**	**0.3877**	**21.77**	**0.6366**	**0.3366**	**25.61**	**0.7692**	**0.2076**

Table 2: Training and rendering times of methods for 10-shot training with a scale factor of 2×. * represents using HAT [9] for SR.

Method	3D-GS* [19]	NeRF-SR [49]	DS-NeRF* [13]	RegNeRF* [33]	SparseNeRF* [50]	FSGS* [65]	Ours
Training	9min 18s	5h 8min	6h 32min	16h 12min	5h 22min	12min 20s	14min 35s
Rendering	0.584s	45.574s	20.619s	36.457s	35.808**s**	0.584s	0.728s

Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [54] and LPIPS [62] are used as the evaluation metrics.

5.3 Quantitative Comparison

We compare our method with the aforementioned methods under the number of training samples of 10-shot, 20-shot, and 90-shot in Table 1. For the cases of few-shot training (10/20-shot), vanilla 3D-GS [19] followed with SR methods show limited performance due to lacking regularization in the training process. The same reason also constrains the capability of NeRF-SR [49]. Vanilla 3D-GS [19] followed with dual-lens SR outperforms that followed with SISR in the dense training case, since using the HR guidance for resolution improvement. The performance of few-shot methods (*e.g.*, SparseNeRF [50] and FSGS [65]) followed with HAT [9] is limited due to lacking the guidance information. It can be seen that our method shows superior performance over the previous methods by leveraging the characteristics of the dual-lens system.

We also compare the training and rendering times of different methods in Table 2, where 3D-GS* [19], DS-NeRF* [13], RegNeRF* [33], SparseNeRF* [50] and FSGS* [65] represent these methods employ HAT [9] for SR. Specifically, the image size is 1160×522 and the rendering time is measured for a single image. We can observe that NeRF-based methods need long training and rendering times, while our method shows a comparable time cost with 3D-GS [19].

5.4 Qualitative Comparison

Qualitative comparisons between DL-GS and other methods under different training sample numbers are shown in Fig. 5, Fig. 6 and Fig. 7. We can observe that our DL-GS can reconstruct more accurate and fine-grained details, while previous methods suffer from blurry or unrealistic artifacts.

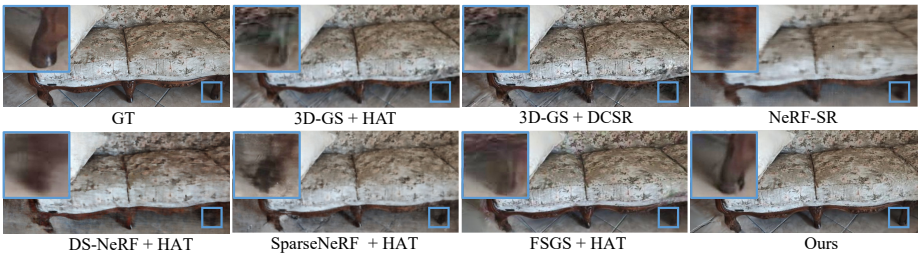


Fig. 5: Visual comparison with different methods for 10-shot training samples.

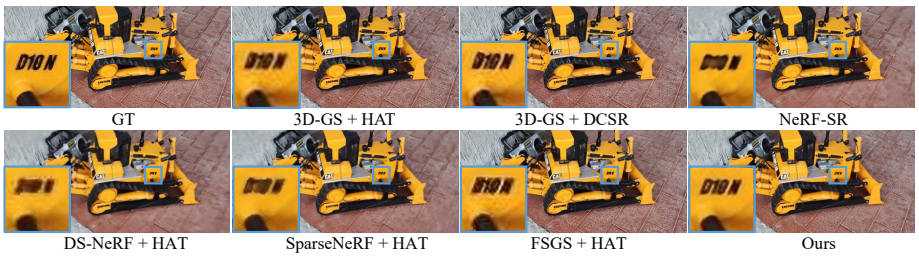


Fig. 6: Visual comparison with different methods for 20-shot training samples.

Fig. 7: Visual comparison with different methods for 90-shot training samples.

6 Real-captured Experiments

6.1 Dataset

To evaluate the performance of our DL-GS in the real-capture data, we collect a set of dual-lens image pairs, captured from different viewpoints of static scenes. Specifically, the real-capture dataset consists of 4 forward-facing scenes and 4 inward-facing (360°) scenes by an off-the-shelf smartphone *i.e.*, iPhone12. We downscale the dual-lens image pairs with a scale factor of 2×, while the original wide-angle images are used as the ground-truth data. For the few-shot training, we utilize 5 shots for the forward-facing scenes, while taking 15 shots for the inward-facing scenes. Meanwhile, we take 10 shots from each scene as the test set, and we utilize 50 shots for the dense training.

6.2 Comparison Results

Quantitative comparison on the real-captured dataset is shown in Tabel 3. We compare our method with previous methods under different training sample

Table 3: Quantitative comparison with previous state-of-the-art methods on our real-captured dataset with different training sample numbers and a scale factor is 2×. **Bold** and underline indicate the best and second best performance, respectively.

Method	Forward-facing						Inward-facing (360°)					
	5-shot			50-shot			15-shot			50-shot		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
3D-GS [19] + Bicubic	20.87	0.7064	0.3690	28.38	0.8467	0.2769	21.00	0.7036	0.4570	29.15	0.8278	0.3625
3D-GS [19] + SwinIR [25]	20.94	0.7088	0.3681	28.46	0.8485	0.2750	21.02	0.7043	0.4479	29.19	0.8280	0.3609
3D-GS [19] + HAT [9]	20.91	0.7086	0.3670	28.52	0.8492	0.2763	21.01	0.7050	0.4477	29.21	0.8283	0.3611
3D-GS [19] + DCSR [51]	21.03	0.7099	0.3636	28.29	0.8421	0.2674	21.00	0.7051	0.4540	29.05	0.8248	0.3646
NeRF-SR [49]	12.53	0.5389	0.5606	30.53	0.8687	**0.2372**	15.80	0.6300	0.5638	29.76	0.8419	0.3479
DS-NeRF [13] + HAT [9]	19.18	0.7019	0.4516	27.02	0.7951	0.3594	22.38	0.7307	0.4685	26.81	0.7711	0.4502
RegNeRF [33] + HAT [9]	22.39	0.7075	0.3679	24.31	0.7541	0.4160	20.20	0.6958	0.4949	23.55	0.7321	0.4829
SparseNeRF [50] + HAT [9]	22.98	0.7127	0.3787	24.57	0.7641	0.4091	20.31	0.7048	0.4767	23.72	0.7497	0.4794
FSGS [65] + HAT [9]	23.08	0.7322	0.3595	29.90	0.8319	0.2996	22.96	0.7461	0.4415	27.77	0.8023	0.4038
Ours	**24.05**	**0.7525**	**0.3249**	**31.28**	**0.8823**	0.2435	**24.07**	**0.7601**	**0.4172**	**30.72**	**0.8597**	**0.3224**

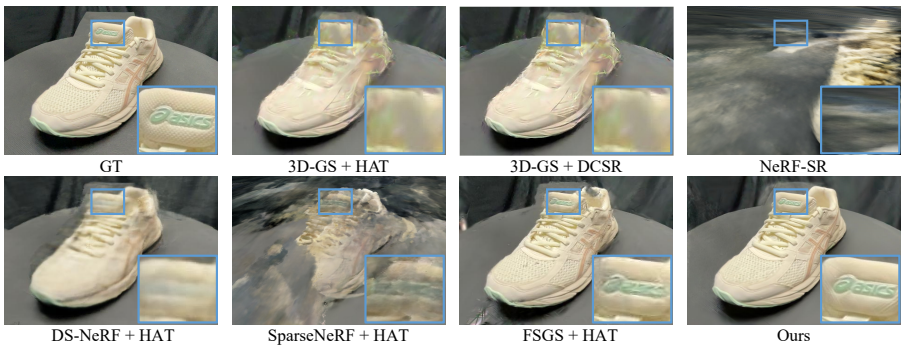


Fig. 8: Visual comparison with different methods for 15-shot training samples on the inward-facing (360°) scene.

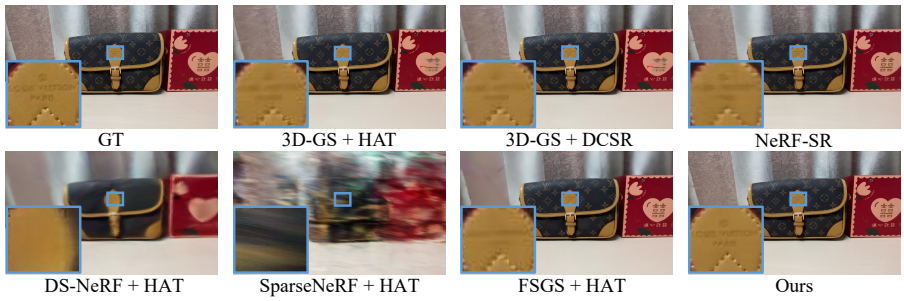


Fig. 9: Visual comparison with different methods for 50-shot training samples on the forward-facing scene. Please zoom in for a better visual experience.

numbers for forward-facing and inward-facing scenes. We can observe that our DL-GS shows superior performance over the previous methods in most cases, which verifies the effectiveness of our method on the real-captured data. To further demonstrate the superiority of our method, we provide qualitative comparisons under different cases in Fig. 8 and Fig. 9. It can be observed that our method reconstructs more realistic details and accurate geometry compared with other competitors. More results are provided in the supplement.

Table 4: Ablation study on the key components of our DL-GS, using the simulated dataset with different numbers of training samples while a scale factor is 2×.

Method	10-shot			20-shot			90-shot		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Baseline	17.89	0.4995	0.4402	20.89	0.5931	0.3615	24.57	0.7196	0.2650
+ Pre-upsampling	18.65	0.5385	0.4226	21.11	0.6088	0.3491	24.73	0.7309	0.2444
+Pre-upsampling +$\mathcal{R}_d$	19.03	0.5468	0.4006	21.25	0.6122	0.3459	24.89	0.7445	0.2332
+Pre-upsampling + $\mathcal{R}_d$ + $\mathcal{R}_c$	**19.40**	**0.5590**	**0.3893**	**21.42**	**0.6216**	**0.3405**	**25.08**	**0.7570**	**0.2286**
+ $(I_j^{\uparrow}, I_k^{\uparrow})$	19.51	0.5646	0.3888	21.55	0.6273	0.3389	25.35	0.7614	0.2173
+ $(I_j^{\uparrow}, I_k^{\uparrow})$ +(T_j, T_k)	**19.67**	**0.5772**	**0.3877**	**21.77**	**0.6366**	**0.3366**	**25.61**	**0.7692**	**0.2076**

7 Ablation Study

Components of Consistency-aware Training Strategy. To investigate the effectiveness of the pre-upsampling and regularization terms ($\mathcal{R}_c$ and $\mathcal{R}_d$) in our consistency-aware training strategy, we conduct an ablation study on the simulated dataset with different training sample numbers. Specifically, we take vanilla 3D-GS [19] followed by HAT [9] as the baseline method. The results are presented in the top part of Table 4. It can be seen that the evaluation metrics gradually increase by adding the three components to the processing of 3D-GS optimization, which demonstrates that they are effective in performance improvement, especially for the few-shot training.

Components of Multi-reference-guided Refinement. We also investigate the effectiveness of two branches in the multi-reference-guided refinement module (*i.e.*, the branch of wide-angle images $(I_j^{\uparrow}, I_k^{\uparrow})$ and the branch of telephoto images (T_j, T_k)). As can be seen at the bottom of Table 4, the evaluation metrics increase as incorporating the two branches for different training sample numbers. The results indicate that the two branches both are important for reconstruction fidelity improvement. We also provide ablations of its loss function terms and selected image numbers in the supplement.

8 Conclusion

In this paper, we propose DL-GS, a new 3G-GS solution to leverage the characteristics of the asymmetric dual-lens system for few-shot view synthesis while overcoming the resolution limitation of training samples. Specifically, we design a consistency-aware training strategy to explore the geometric information inherent in the dual-lens system for facilitating the few-shot training. After that, we propose a multi-reference-guided refinement module to enhance the details of newly synthesized views by effectively exploiting the guidance information of the dual-lens training samples. Experiments on simulated and real-captured datasets demonstrate the superiority of DL-GS over various solutions. Experimental results also show the effectiveness and potential of leveraging the characteristics of the dual-lens system for high-quality view synthesis. We believe that this paradigm can address other challenging cases in NVS, *e.g.*, focus control using the different depth-of-field of two lenses.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62131003 and 62021001.

References

1. Alzayer, H., Abuolaim, A., Chan, L.C., Yang, Y., Lou, Y.C., Huang, J.B., Kar, A.: Dc2: Dual-camera defocus control by learning to refocus. In: CVPR (2023) 5
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: CVPR (2022) 1, 4
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. In: ICCV (2023) 1
4. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023) 2
5. Bhoi, A.: Monocular depth estimation: A survey. arXiv preprint arXiv:1901.09402 (2019) 2
6. Cao, J., Wang, H., Chemerys, P., Shakhrai, V., Hu, J., Fu, Y., Makoviichuk, D., Tulyakov, S., Ren, J.: Real-time neural light field on mobile devices. In: CVPR (2023) 1
7. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: CVPR (2021) 9
8. 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: CVPR (2021) 4
9. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: CVPR (2023) 2, 7, 10, 11, 13, 14
10. Chen, X., Xiong, Z., Cheng, Z., Peng, J., Zhang, Y., Zha, Z.J.: Degradation-agnostic correspondence from resolution-asymmetric stereo. In: CVPR (2022) 5
11. Chen, Z., Funkhouser, T., Hedman, P., Tagliasacchi, A.: Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In: CVPR (2023) 1, 4
12. Chung, J., Oh, J., Lee, K.M.: Depth-regularized optimization for 3d gaussian splatting in few-shot images. arXiv preprint arXiv:2311.13398 (2023) 2
13. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: CVPR (2022) 2, 4, 6, 10, 11, 13
14. Deng, N., He, Z., Ye, J., Duinkharjav, B., Chakravarthula, P., Yang, X., Sun, Q.: Fov-nerf: Foveated neural radiance fields for virtual reality. IEEE Transactions on Visualization and Computer Graphics **28**(11), 3854–3864 (2022) 1
15. Dong, J., Fang, Q., Yang, T., Shuai, Q., Qiao, C., Peng, S.: ivs-net: Learning human view synthesis from internet videos. In: ICCV (2023) 1
16. Hattori, H., Maki, A.: Stereo without depth search and metric calibration. In: CVPR (2000) 7
17. Hu, T., Liu, S., Chen, Y., Shen, T., Jia, J.: Efficientnerf efficient neural radiance fields. In: CVPR (2022) 1
18. Huang, X., Li, W., Hu, J., Chen, H., Wang, Y.: Refsr-nerf: Towards high fidelity and super resolution view synthesis. In: CVPR (2023) 2, 4, 6
19. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023) 2, 4, 5, 10, 11, 13, 14
20. Kim, M., Seo, S., Han, B.: Infonerf: Ray entropy minimization for few-shot neural volume rendering. In: CVPR (2022) 4
21. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: ECCV (2018) 7

22. Larsson, V., Zobernig, N., Taskin, K., Pollefeys, M.: Calibration-free structure-from-motion with calibrated radial trifocal tensors. In: ECCV (2020) 7
23. Lee, J., Lee, M., Cho, S., Lee, S.: Reference-based video super-resolution using multi-camera video triplets. In: CVPR (2022) 2, 6, 9
24. Li, Q., Li, F., Guo, J., Guo, Y.: Uhdnerf: Ultra-high-definition neural radiance fields. In: ICCV (2023) 2, 6
25. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: CVPRW (2021) 10, 11, 13
26. Lin, C.Y., Fu, Q., Merth, T., Yang, K., Ranjan, A.: Fastsr-nerf: Improving nerf efficiency on consumer devices with a simple super-resolution pipeline. In: WACV (2024) 1
27. Manne, S.K.R., Prasad, B., Rosh, K.: Asymmetric wide tele camera fusion for high fidelity digital zoom. In: ICCVIP (2019) 5
28. Mechrez, R., Talmi, I., Shama, F., Zelnik-Manor, L.: Maintaining natural image statistics with the contextual loss. In: ACCV (2019) 10
29. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: ECCV (2018) 10
30. 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) 1
31. Mohan, M.M., Nithin, G., Rajagopalan, A.: Deep dynamic scene deblurring for unconstrained dual-lens cameras. IEEE Transactions on Image Processing **30**, 4479–4491 (2021) 5
32. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022) 1, 4
33. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: CVPR (2022) 4, 6, 10, 11, 13
34. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021) 7
35. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: CVPR (2020) 1
36. Santesteban, I., Otaduy, M., Thuerey, N., Casas, D.: Ulnef: Untangled layered neural fields for mix-and-match virtual try-on. NIPS (2022) 1
37. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016) 2, 4, 10
38. Sedgwick, P.: Pearson's correlation coefficient. Bmj **345** (2012) 8
39. Seo, S., Chang, Y., Kwak, N.: Flipnerf: Flipped reflection rays for few-shot novel view synthesis. In: ICCV (2023) 4
40. Shao, R., Zhang, H., Zhang, H., Chen, M., Cao, Y.P., Yu, T., Liu, Y.: Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In: CVPR (2022) 2, 6
41. Somraj, N., Soundararajan, R.: Vip-nerf: Visibility prior for sparse input neural radiance fields. In: ACM SIGGRAPH (2023) 4
42. Song, J., Park, S., An, H., Cho, S., Kwak, M.S., Cho, S., Kim, S.: Därf: Boosting radiance fields from sparse input views with monocular depth adaptation. In: NIPS (2023) 2, 4, 6
43. Song, T., Kim, S., Sohn, K.: Unsupervised deep asymmetric stereo matching with spatially-adaptive self-similarity. In: CVPR (2023) 5

44. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023) 8
45. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV (2020) 7
46. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally-deformable alignment network for video super-resolution. In: CVPR (2020) 9
47. Tosi, F., Tonioni, A., De Gregorio, D., Poggi, M.: Nerf-supervised deep stereo. In: CVPR (2023) 3, 10
48. Uy, M.A., Martin-Brualla, R., Guibas, L., Li, K.: Scade: Nerfs from space carving with ambiguity-aware depth estimates. In: CVPR (2023) 2
49. Wang, C., Wu, X., Guo, Y.C., Zhang, S.H., Tai, Y.W., Hu, S.M.: Nerf-sr: High quality neural radiance fields using supersampling. In: ACM MM (2022) 2, 4, 6, 10, 11, 13
50. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In: ICCV (2023) 4, 6, 10, 11, 13
51. Wang, T., Xie, J., Sun, W., Yan, Q., Chen, Q.: Dual-camera super-resolution with aligned attention modules. In: ICCV (2021) 2, 5, 6, 10, 11, 13
52. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: CVPRW (2019) 9
53. Wang, Y., Wang, L., Wu, G., Yang, J., An, W., Yu, J., Guo, Y.: Disentangling light fields for super-resolution and disparity estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 425–443 (2022) 9
54. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003) 5, 11
55. Xiong, H., Muttukuru, S., Upadhyay, R., Chari, P., Kadambi, A.: Sparsegs: Real-time 360 sparse view synthesis using gaussian splatting. arXiv preprint arXiv:2312.00206 (2023) 2, 4
56. Xu, R., Yao, M., Xiong, Z.: Zero-shot dual-lens super-resolution. In: CVPR (2023) 5
57. Yang, J., Pavone, M., Wang, Y.: Freenerf: Improving few-shot neural rendering with free frequency regularization. In: CVPR (2023) 2, 4
58. Yoon, Y., Yoon, K.J.: Cross-guided optimization of radiance fields with multi-view image super-resolution for high-resolution novel view synthesis. In: CVPR (2023) 2, 4
59. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: CVPR (2021) 1, 2, 4, 6
60. Yue, H., Cui, Z., Li, K., Yang, J.: Kedusr: Real-world dual-lens super-resolution via kernel-free matching. arXiv preprint arXiv:2312.17050 (2023) 5
61. Zhang, J., Yang, H., Ren, J., Zhang, D., He, B., Cao, T., Li, Y., Zhang, Y., Liu, Y.: Mobidepth: real-time depth estimation using on-device dual cameras. In: Proceedings of the 28th Annual International Conference on Mobile Computing And Networking. pp. 528–541 (2022) 5
62. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 11
63. Zhang, Z., Wang, R., Zhang, H., Chen, Y., Zuo, W.: Self-supervised learning for real-world super-resolution from dual zoomed observations. In: ECCV (2022) 5
64. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 4

65. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting. arXiv preprint arXiv:2312.00451 (2023) 2, 4, 5, 6, 10, 11, 13