

---



# DOGS: Distributed-Oriented Gaussian Splatting for Large-Scale 3D Reconstruction Via Gaussian Consensus

---

**Yu Chen**

National University of Singapore

yuchen01@u.nus.edu

**Gim Hee Lee**

National University of Singapore

gimhee.lee@nus.edu.sg

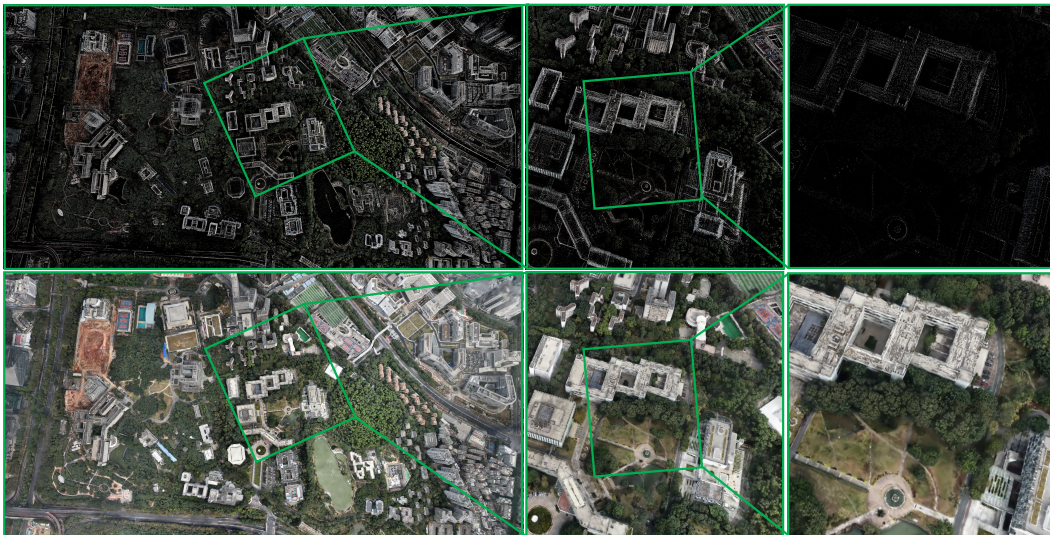


Figure 1: DOGS accelerates 3D GS training on large-scale UrbanScene3D dataset [29] by 6+ times with better rendering quality. Top: 3D Gaussian primitives (8.27M); Bottom: rendered images.

## Abstract

The recent advances in 3D Gaussian Splatting (3DGS) show promising results on the novel view synthesis (NVS) task. With its superior rendering performance and high-fidelity rendering quality, 3DGS is excelling at its previous NeRF counterparts. The most recent 3DGS method focuses either on improving the instability of rendering efficiency or reducing the model size. On the other hand, the training efficiency of 3DGS on large-scale scenes has not gained much attention. In this work, we propose DOGS, a method that trains 3DGS distributedly. Our method first decomposes a scene into  $K$  blocks and then introduces the Alternating Direction Method of Multipliers (ADMM) into the training procedure of 3DGS. During training, our DOGS maintains one global 3DGS model on the master node and  $K$  local 3DGS models on the slave nodes. The  $K$  local 3DGS models are dropped after training and we only query the global 3DGS model during inference. The training time is reduced by scene decomposition, and the training convergence and stability are guaranteed through the consensus on the shared 3D Gaussians. Our method accelerates the training of 3DGS by 6+ times when evaluated on large-scale scenes while concurrently achieving state-of-the-art rendering quality. Our code is publicly available at <https://github.com/AIBluefisher/DOGS>.

# 1 Introduction

Neural 3D scene reconstruction has taken a giant step beyond the limitations of traditional photogrammetry tools. Neural radiance fields (NeRFs) [36, 17, 47], which encode scenes implicitly in MLP or explicitly in voxels, exhibit superior resilience to non-Lambertian effects, appearance changes, dynamic scenes, *etc.* However, most NeRF-based methods are inefficient in rendering scenes due to the need to query massive points for volume rendering to infer scene geometry and color. Recently, 3D Gaussian Splatting (3DGS) [21] has shown promising results on real-time applications and inspires many follow-up works. 3DGS encodes scenes into a set of 3D anisotropic Gaussians, where each 3D Gaussian is represented by a covariance matrix, a center position, opacity, and the latent features to encode color information. Pixel colors by projecting 3D Gaussians into 2D image space can be efficiently computed via rasterization, which is highly optimized on modern graphics processors. However, 3DGS often requires larger memory during training compared to NeRF methods. This is because 3DGS needs millions of 3D Gaussians to represent a scene to recover high-fidelity scene details. Consequently, the memory footprint increases drastically for training 3DGS on larger scenes, *e.g.* city-scale scenes. Moreover, the training of a huge number of 3D Gaussians on larger scenes leads to longer training time. Unfortunately, in comparison to NeRF where the rays can be conveniently distributed into different compute nodes, dispatching 3D Gaussians into different compute nodes is much more difficult due to the highly customized rasterization procedure of 3DGS. In summary, the **two challenges** for reconstructing large-scale scenes with 3DGS are: 1) High GPU memory to hold the large 3D model during training; 2) Long training time due to the large areas of scenes.

Previous large-scale NeRF methods [49, 51, 35] solve the above-mentioned two issues by embracing a divide-and-conquer approach, where scenes are split into individual blocks with smaller models fitted into each block. However, these methods require querying multiple sub-models during inference, which slows down the rendering efficiency. This leads us to the following question:

*“Can we apply a similar methodology to 3DGS during training while querying only a global consistent model during inference?”*

In this work, we propose DOGS to answer the aforementioned question. Following previous large-scale NeRF methods, our DOGS splits scene structures into blocks for distributed training. Inspired by previously distributed bundle adjustment methods [15, 62], we apply the Alternating Direction Method of Multipliers (ADMM) [5] to ensure the consistency of the shared 3D Gaussians between different blocks. Specifically, we first estimate a tight bounding box for each scene. Subsequently, we split training views and point clouds into blocks. To guarantee each block has a similar size, we split scenes into two blocks each time along the longer axis of the bounding box. Scenes are split recursively in the same way until we obtain the desired number of blocks. Finally, we re-estimate a bounding box for each block and expand the bounding box to construct shared 3D Gaussians between blocks. During training, we maintain a global 3D Gaussian model on the master node and dispatch local 3D Gaussians into other slave nodes. We further guarantee the consistency of the shared 3D Gaussians and the convergence of the training through 3D Gaussian consensus. Specifically, the local 3D Gaussians are collected onto the master node and averaged to update the current global 3D Gaussian model at the end of each training iteration, and then the updated global 3D Gaussian model is shared to all slave nodes to regularize the training of local 3D Gaussians. In this way, our method guarantees that the local 3D Gaussians converge to the global 3D Gaussian model during training. By training 3DGS in a distributed way, our DOGS can reduce the training time by 6+ times compared to the original 3DGS. Furthermore, our DOGS guarantees training convergence and therefore achieves better rendering quality than its counterparts with the 3D Gaussians consensus in the distributed training. After training, we can drop all local 3D Gaussians and maintain only the global 3D Gaussians. During inference, we only need to query the global model while maintaining the rendering performance of 3DGS.

Our contributions are summarized as follows:

- We propose a recursive approach to split scenes into blocks with balanced sizes.
- We introduce a distributed approach to train 3D Gaussians for large-scale scenes. The training time is reduced by 6+ times compared to the original 3DGS.
- We conduct exhaustive experiments on standard large-scale datasets to validate the effectiveness of our method.

## 2 Related Work

**Neural Radiance Fields.** Neural radiance fields [36] enable rendering from novel viewpoints with encoded frequency features [50]. To improve training and rendering efficiency, follow-up works [27, 59, 17] either encodes scenes into sparse voxels, or a multi-resolution hash table [37] where the hash collision is implicitly handled during optimization. TensorRF [6] uses CP-decomposition or VM-decomposition to encode scenes into three orthogonal axes and planes. While the previously mentioned work focuses on per-scene reconstruction, other methods also focus on the generalizability of NeRF [54, 7, 19, 46, 28], bundle-adjusting camera poses and NeRF [25, 34, 8], and leveraging sparse or dense depth to supervise the training of NeRF [56, 14, 42], *etc.* To address the aliasing issue in vanilla NeRF, Mip-NeRF [1] proposed to use Gaussian to approximate the cone sampling, the integrated positional encodings are therefore scale-aware and can be used to address the aliasing issue of NeRF. Mip-NeRF360 [2] further uses space contraction to model unbounded scenes. Zip-NeRF [3] adopted a hexagonal sampling strategy to handle the aliasing issue for Instant-NGP [37]. NeRF2NeRF [18] and DReg-NeRF [9] assumes images are only available during training in each block, and they propose methods to register NeRF blocks together.

**Gaussian Splatting.** Gaussian Splatting [21] initializes 3D Gaussians from a sparse point cloud. The 3D Gaussians are used as explicit scene representation and dynamically merged and split during training. Real-time novel view synthesis is enabled by the fast splatting operation through rasterization. Scaffold-GS [31] initializes a sparse voxel grid from the initial point cloud and encodes the features of 3D Gaussians into corresponding feature vectors. The introduction of the sparse voxel reduced the Gaussian densities by avoiding unnecessary densification on the scene surface. Octree-GS [41] introduces the level-of-details to dynamically select the appropriate level from the set of multi-resolution anchor points, which ensures consistent rendering performance with adaptive LOD adjustment and maintains high-fidelity rendering results. To reduce the model size, methods [16, 38] also remove redundant 3D Gaussians through exhaustive quantization. Other methods also focus on alleviating the aliasing issue of 3D Gaussians [60, 58], or leveraging the efficient rasterizer of point rendering for real-time indoor reconstruction [20, 33].

**Large-Scale 3D Reconstruction.** The classical photogrammetry methods utilized Structure-from-Motion (SfM) [43] and keypoint extraction with SIFT [30] to reconstruct sparse scene structures. One of such foundational software is Phototourism [45]. To handle city-scale scenes, the ‘divide-and-conquer’ strategy is widely adopted for the extensibility and scalability of the 3D reconstruction system. SfM methods [4, 63, 10, 12, 11] splitting scenes based on the view graph, where images with strong connections are divided into the same block. By estimating similarity transformations and merging tracks, all local SfM points and camera poses are fused into a global reference system. Existing NeRF methods also follow a similar strategy for reconstructing city-scale scenes. When camera poses are known, the scene can be split into grid cells. Block-NeRF [49] focus on the day-night or even cross-season street views. It utilizes the appearance encoding in NeRF-W [32] to fix the appearance inconsistency issue between different blocks, while Mega-NeRF [51] aims at encouraging the sparsity of the network under aerial scenes. Urban-NeRF [40] leverages lidar points to supervise the depth of NeRF in outdoor scenes. SUDS [52] further extended Mega-NeRF into dynamic scenes. Different from previous large-scale NeRF methods, Switch-NeRF [35] uses a switch transformer that learns to assign rays to different blocks during training. Grid-NeRF [57] designed a two-branch network architecture, where the NeRF branch can encourage the feature plane [6] branch recover more scene details under large-scale scenes. However, the two-branch training scheme is trivial and needs a long time to train. Concurrent works to our method are VastGaussian [26] and Hierarchy-GS [22], which also utilize 3D Gaussians for large-scale scene reconstruction. VastGaussian and Hierarchy-GS split scenes into independent chunks and train independent chunks simultaneously. However, VastGaussian relies on exhaustive searching of the training views and initial points to guarantee the convergence of training, and each block is trained without data sharing. Hierarchy-GS consolidates independent chunks into intermediate nodes for further rendering. However, the hierarchical approach needs to preserve redundant models and it is specially designed for street view scenes. Our method, on the other hand, focuses on the distributed training of 3DGS and built upon the consensus of shared 3D Gaussians between different blocks has a guaranteed convergence rate that achieves better performance.

### 3 Our Method

#### 3.1 Preliminary

**Gaussian Splatting.** 3D Gaussian Splatting represents a scene with a set of anisotropic 3D Gaussians  $\mathcal{G} = \{\mathbf{G}_i \mid i \in N\}$ . Each 3D Gaussian primitive  $\mathbf{G}_i$  has a center  $\mathbf{u} \in \mathbb{R}^3$  and a covariance  $\Sigma \in \mathbb{R}^{3 \times 3}$  and can be described by:

$$\mathbf{G}_i(\mathbf{p}) = \exp\left\{-\frac{1}{2}(\mathbf{p} - \mathbf{u}_i)^\top \Sigma_i^{-1}(\mathbf{p} - \mathbf{u}_i)\right\}. \quad (1)$$

During training, the covariance is decomposed into a rotation matrix  $\mathbf{R} \in \mathbb{R}^{3 \times 3}$  and a diagonal scaling matrix  $\mathbf{S} \in \mathbb{R}^{3 \times 3}$ , *i.e.*  $\Sigma_i = \mathbf{R}\mathbf{S}\mathbf{S}^\top\mathbf{R}^\top$  to ensure the covariance matrix is positive semi-definite. To render the color for a pixel  $\mathbf{p}$ , the 3D Gaussians are projected into the image space for alpha blending:

$$\mathbf{C} = \sum_i \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (2)$$

where  $\alpha_i$  is the rendering opacity and is computed by  $\alpha = o \cdot \mathbf{X}^{\text{proj}}(\mathbf{p})$ .

When training 3D Gaussian Splatting, we minimize the loss function below:

$$\mathcal{L}(\mathbf{x}) = \mathcal{L}_{\text{rgb}} + \lambda \mathcal{L}_{\text{ssim}}, \quad (3)$$

where  $\mathbf{x}_i = \{\mathbf{u}_i, \mathbf{q}_i, \mathbf{s}_i, \mathbf{f}_i, o_i\}$ ,  $\mathbf{q}$  are quaternions corresponds to the rotation matrix  $\mathbf{R}$ ,  $\mathbf{s}$  are vectors corresponds to the three diagonal elements of  $\mathbf{S}$ , and  $\mathbf{f}$  are coefficients of the spherical harmonics.

#### 3.2 Distributed 3D Gaussian Consensus

The ‘divide-and-conquer’ method is a common paradigm for large-scale 3D reconstruction, which we also adopt in our framework. Different from previous methods such as Block-NeRF [49] and VastGaussian [26] which are pipeline parallelized, our method is optimization parallelized with guaranteed convergence. The pipeline of our algorithm is shown in Fig. 2. Firstly, we split a scene (training views and point clouds) into  $K$  intersected blocks. Secondly, we assign training views and points into different blocks. By introducing the ADMM into the training process, we also maintain a globally consistent 3DGS model on a master node. Thirdly, during training, we collect and average the local 3D Gaussians to update the global 3DGS model in each consensus step. The global 3D Gaussians are also shared with each block before we distributedly train the local 3D Gaussians in each block. Finally, we drop all local 3D Gaussians while only using the global 3D Gaussians to render novel views.

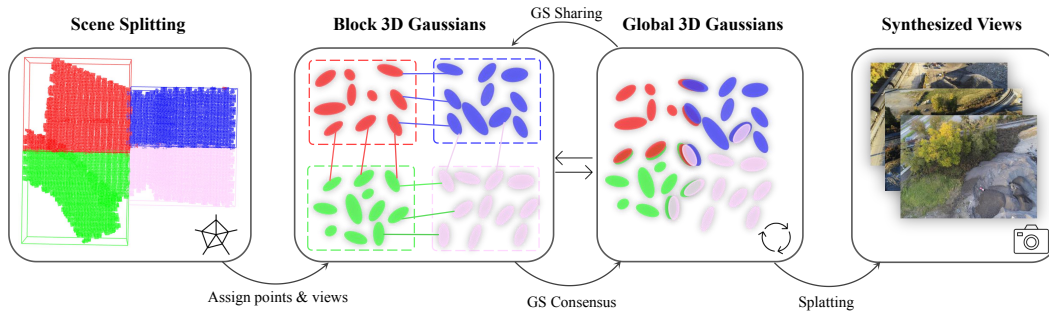


Figure 2: **The pipeline of our distributed 3D Gaussian Splatting method.** 1) We first split the scene into  $K$  blocks with similar sizes. Each block is extended to a larger size to construct overlapping parts. 2) Subsequently, we assign views and points into different blocks. The shared local 3D Gaussians (connected by solid lines in the figure) are a copy of the global 3D Gaussians. 3) The local 3D Gaussians are then collected and averaged to the global 3D Gaussians in each consensus step, and the global 3D Gaussians are shared with each block before training all blocks. 4) Finally, we use the final global 3D Gaussians to synthesize novel views.

In this section, we first introduce the ADMM algorithm. Subsequently, we derive the distributed 3DGS training algorithm. We also present a scene splitting algorithm, which recursively and evenly splits the scene into two blocks each time.

**ADMM.** A general form for consensus ADMM is given by:

$$\text{minimize } \sum_{i=1}^N f_i(\mathbf{x}_i), \quad \text{s.t. } \mathbf{x}_i - \mathbf{z} = 0, \quad i \in [1, N]. \quad (4)$$

By definition, the constraints are applied such that all the local variables  $\mathbf{x}_i$  agree with the global variable  $\mathbf{z}$ . By applying the augmented Lagrangian method, we have:

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = \sum_{i=1}^N (f_i(\mathbf{x}_i) + \mathbf{y}_i^\top (\mathbf{x}_i - \mathbf{z}) + \frac{\rho}{2} \|\mathbf{x}_i - \mathbf{z}\|_2^2), \quad (5)$$

where  $\mathbf{y}_i$  is the dual variable,  $\rho$  is the penalty parameter. During optimization, ADMM alternatively updates the local variables  $\mathbf{x}_i$ , global variable  $\mathbf{z}$  and the dual variables  $\mathbf{y}_i$  at the  $t + 1$  iteration by:

$$\mathbf{x}_i^{t+1} := \arg \min (f_i(\mathbf{x}_i) + \mathbf{y}_i^\top (\mathbf{x}_i - \mathbf{z}^t) + \frac{\rho}{2} \|\mathbf{x}_i - \mathbf{z}^t\|_2^2), \quad (6a)$$

$$\mathbf{z}^{t+1} := \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i^{t+1} + \frac{1}{\rho} \mathbf{y}_i^t), \quad (6b)$$

$$\mathbf{y}_i^{t+1} := \mathbf{y}_i^t + \rho(\mathbf{x}_i^{t+1} - \mathbf{z}^{t+1}). \quad (6c)$$

**Distributed Training.** We apply the ADMM method to distributedly train a large-scale 3D Gaussian Splatting model. In our problem,  $f_i(\cdot)$  in Eq. (4) corresponds to the loss function  $\mathcal{L}(\cdot)$  in Eq. (3). To simplify the implementation for Eq. (6a), we adopt a scaled form of ADMM by defining  $\mathbf{u}_i = \frac{1}{\rho} \mathbf{y}_i$ . We can then rewrite Eq. (5) into (see supplementary for the derivation):

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{u}) = \sum_{i=1}^N (f_i(\mathbf{x}_i) + \frac{\rho}{2} \|\mathbf{x}_i - \mathbf{z} + \mathbf{u}_i\|_2^2 - \frac{\rho}{2} \|\mathbf{u}_i\|_2^2). \quad (7)$$

Compared to Eq. (5), Eq. (7) can be made easier to implement by expressing all terms in the squared difference errors. Suppose the variables are decomposed into  $K$  blocks, we then denote the  $i$ th 3D Gaussians in the  $k$ th block as  $\mathbf{x}_i^k$ . Accordingly, we revise the ADMM updating rule by:

$$(\mathbf{x}_i^k)^{t+1} := \arg \min (f(\mathbf{x}_i^k) + \frac{\rho}{2} \|(\mathbf{x}_i^k)^t - \mathbf{z}^t + (\mathbf{u}_i^k)^t\|_2^2), \quad (8a)$$

$$\mathbf{z}^{t+1} := \frac{1}{K} \sum_{k=1}^K (\mathbf{x}_i^k)^{t+1}, \quad (8b)$$

$$(\mathbf{u}_i^k)^{t+1} := (\mathbf{u}_i^k)^t + (\mathbf{x}_i^k)^{t+1} - \mathbf{z}^{t+1}. \quad (8c)$$

Eq. (8a) is the original loss function in Gaussian Splatting with an additional regularizer term  $\frac{\rho}{2} \|(\mathbf{x}_i^k)^t - \mathbf{z}^t + (\mathbf{u}_i^k)^t\|_2^2$ . Note that Eq. (8b) should be  $\mathbf{z}^{t+1} := \frac{1}{K} \sum_{k=1}^K ((\mathbf{x}_i^k)^{t+1} + (\mathbf{u}_i^k)^t)$ . However, the dual variables have an average value of zero after the first iteration. Consequently, Eq. (8b) can be simplified as the global 3D Gaussians are formed by the average of the local 3D Gaussians from all blocks. Moreover, Eq. (8b) is called a ‘consensus’ step and it requires collecting the local 3D Gaussians from all blocks. After updating the global model  $\mathbf{z}$ , we update the dual variables  $\mathbf{u}_i$  in Eq. (8c) and share the global 3D Gaussians  $\mathbf{z}$  to each block for optimizing the local 3D Gaussians in Eq. (8a). Note that each 3D Gaussian  $\mathbf{x}_i$  has different properties  $\{\mathbf{u}_i, \mathbf{q}_i, \mathbf{s}_i, \mathbf{f}_i, O_i\}$ . As a result, the penalty terms and dual variables should be represented separately according to these properties. The detailed form of Eq. (8) is given in the supplementary.

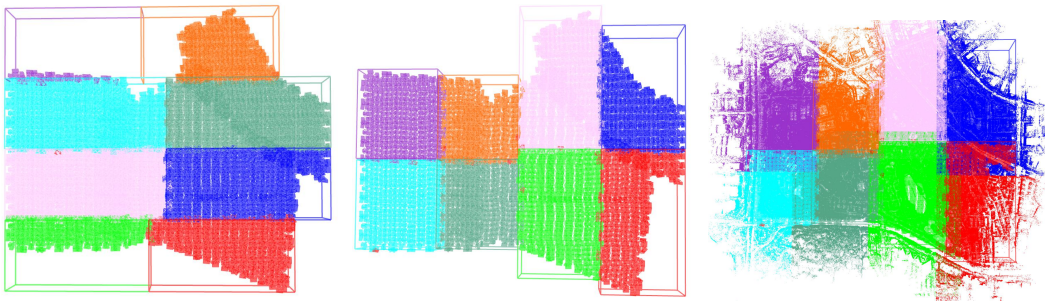
**Scene Splitting.** We decompose the scene into  $K$  blocks before applying the updating rule in Eq. (8). Unlike VastGaussian [26], which focuses mostly on splitting large-scale scenes and needs exhaustive search on the training views and point clouds to ensure the consistency of 3D Gaussians in different blocks, our method relies on the consensus step to ensure the consistency of the 3DGS. However, scene splitting is still important to the convergence of our method. We propose two constraints for the scene-splitting method to best balance the training efficiency and rendering quality:

1. Individual blocks should have a similar size.

2. Adjacent blocks should have enough overlaps to boost the convergence of training.

The first constraint is proposed to ensure that: 1) Each block can be fed into GPUs with the same capacity. This is important since a larger block can cause an out-of-memory of the GPU during training due to the imbalanced splitting results. 2) Each block has a similar training time at each iteration. After every  $t$  iteration, we collect all local 3D Gaussians from each block. Intuitively, larger blocks require a longer time to train. Consequently, the master node and all other slave nodes have to wait for the nodes with larger blocks to finish, which increases the training time unnecessarily.

The second constraint is used to boost the convergence of ADMM. From Eq. (8b), the local 3D Gaussians would converge to the global 3D Gaussians by averaging the shared local 3D Gaussians during training. Sufficient shared 3D Gaussians encourage reconstruction consistency between different blocks. Too many shared local 3D Gaussians can bring more communication overhead, which inevitably slows down the training while a lack of shared 3D Gaussians leads to divergence of the algorithm. Although there is no theoretical analysis to show the optimal value of overlapping parts, we empirically use a constant value which we will introduce later in our experiments.



(a) Splitting results from VastGaussian (camera trajectories) (b) Splitting results from our method (camera trajectories) (c) Splitting results from our method (camera trajectories and 3D points)

Figure 3: **Scene splitting results of our method v.s. VastGaussian [26].** (a) VastGaussian can result in imbalanced blocks. (b) Our recursive bipartite strategy solves the imbalanced splitting issue. (c) Points and views with the same grid coordinate are assigned to the same block.

We assume one of the axes of the scene is aligned to physical ground, which can usually be done under the Manhattan world assumption. VastGaussian [26] adopts a grid-splitting method that first splits the scene into  $m$  cells along the  $x$ -axis, and then splits each of the  $m$  cells into  $n$  sub-cells along the  $y$ -axis. As we show in Fig. 3a, this strategy can result in imbalanced blocks. Our splitting method is inherited from VastGaussian while adopting a recursive spitting method to resolve the imbalanced issue. Specifically, we first estimate a tight bounding box for the initial scene. We then split the scene into two parts along the longer axis of the scene. Splitting the scene along the longer axis can prevent the blocks from becoming too shallow on one axis. We re-estimate a tighter bounding box for each of the two cells and split them into smaller blocks. This step is repeated until the number of blocks reaches our requirement. We present the result of this recursive method in Fig. 3b. Compared to Fig. 3a, we produce more balanced blocks. To construct overlapping areas, we expand the bounding box of each block by a scale factor  $s$ , any points/views that are covered by the same bounding box are grouped into the same block. The training views and point clouds are split in the same way as is shown in Fig. 3c.

### 3.3 Improving Convergence Rate

ADMM is known to be sensitive to the initialization of penalty parameters. Since improper initial penalty parameters can slow down the training, we introduce the *adaptive penalty parameters* and *over-relaxation* to improve the convergence rate.

**Primal Residual and Dual Residual.** We define the primal residual  $\mathbf{r}^t$  and the dual residual  $\mathbf{s}^t$  as

$$\mathbf{r}^t = \mathbf{x}_i^t - \mathbf{z}^t, \quad \mathbf{s}^t = \rho(\mathbf{z}^t - \mathbf{z}^{t-1}). \quad (9)$$

In ADMM, the primal residual and dual residual are used as the stopping criteria which terminate the optimization. In our method, we use a hard threshold of training iteration to terminate the algorithm. The primal residual and dual residual are used to adaptively adjust the penalty parameters.

**Adaptive Penalty Parameters.** We adopt a similar scheme from [5] to adaptively adjust the penalty parameters:

$$\rho^{t+1} = \begin{cases} \tau^{\text{inc}} \rho^t, & \|\mathbf{r}^t\|_2 > \mu \|\mathbf{s}^t\|_2, \\ \frac{\rho^t}{\tau^{\text{dec}}}, & \|\mathbf{s}^t\|_2 > \mu \|\mathbf{r}^k\|_2, \\ \rho^t, & \|\mathbf{s}^t\|_2 = \mu \|\mathbf{r}^k\|_2, \end{cases} \quad (10)$$

where  $\mu > 1$ ,  $\tau^{\text{inc}}$ ,  $\tau^{\text{dec}}$  are hyper-parameters. The existing convergence proof of the ADMM algorithm is based on the fixed penalty parameters [5]. To guarantee the convergence of our algorithm, we stop adjusting the penalty parameters after 2000 iterations in all of our experiments.

**Over Relaxation.** Similar to [5], we replace  $\mathbf{x}^{t+1}$  with  $\alpha^t \mathbf{x}^{t+1} - (1 - \alpha^t)(-\mathbf{z}^t)$  in Eq. (8b) and Eq. (8c), where  $\alpha^t \in (0, 2)$  is the relaxation parameter and experiments show that the over-relaxation with  $\alpha^t \in [1.5, 1.8]$  can improve convergence.

## 4 Experiments

**Datasets.** We evaluate our method on the two large-scale urban datasets, the Mill19 [51]<sup>1</sup> dataset, the UrbanScene3D [29]<sup>2</sup> dataset, and the MatrixCity [24] dataset. Both datasets are captured by drones and each scene contains thousands of high-resolution images. During training and evaluation, we adopt the original image splitting in Mega-NeRF [51], and downsample the images by 6+ times from the original resolution.

**Implementation Details.** For Mill19 and UrbanScene3D, we use the camera poses provided by the official site of Mega-NeRF [51]. The y-axis of the scene is aligned to the horizontal plane by COLMAP [43] under the Manhattan world assumption. We use the CPU version of SIFT (SIFT-CPU) in COLMAP [43] to extract keypoints and match each image to its nearest 100 images using vocabulary trees. With known camera poses and keypoint matches, we further triangulate 3D points and bundle adjust them. The SIFT-CPU can extract more points than the SIFT-GPU, which can benefit the initialization of 3D Gaussians. For the original 3D Gaussian Splatting (denoted as ‘3DGS’), we train it in 500,000 iterations and densify it for every 200 iterations until it reaches 30,000 steps. We train both VastGaussian and our method in 80,000 iterations. The densification intervals and termination steps are the same as the 3DGS. We reimplement VastGaussian [26] since its code was not released during this work. Note that we did not implement the decoupled appearance embedding in VastGaussian, which can be used to remove floaters caused by inconsistent image appearance. We argue that we still provided a fair comparison since this module can be applied to all 3DGS-based methods. For our method, consensus and sharing are enabled every 100 iterations. We leverage the remote procedure call (RPC) framework of PyTorch [39] to implement our distributed algorithm and transmit data across different compute nodes.

**Results.** We employ PSNR, SSIM [55] and LPIPS [61] as metrics for novel view synthesis. We compare our method against the state-of-the-art large-scale NeRF-based methods: Mega-NeRF [51], Switch-NeRF [35]<sup>3</sup>, and 3DGS-based methods: 3DGS [21]<sup>4</sup>, Fed3DGS [48], VastGaussian [26], Hierarchy-GS [22]. For Mega-NeRF and Switch-NeRF, we use the officially provided checkpoints on 8 blocks for evaluation. The results of Hierarchy-GS are cited from the original paper since its code has not been released during this work.

We present the quantitative visual quality results in Table 1. The training time and rendering efficiency are also provided in Table 2. As shown in the tables, methods based on 3D Gaussians achieved better results than NeRF-based methods. Although NeRF-based methods are comparable to 3DGS methods in PSNR, the rendered images lack details in such large-scale scenes. This is also validated from

<sup>1</sup><https://github.com/cmusatyalab/mega-nerf>

<sup>2</sup><https://vcc.tech/UrbanScene3D>

<sup>3</sup><https://github.com/MiZhenxing/Switch-NeRF>

<sup>4</sup><https://github.com/graphdeco-inria/gaussian-splatting>

Fig. 4 and Fig. 5. Moreover, NeRF-based methods are much slower than 3DGS-based methods and take longer time to train – even if they are trained distributedly. Notably, our method achieves the best results in almost all the scenes. The original 3DGS has results comparable to ours. However, it takes  $6 \sim 8\times$  more training time than our method. We also emphasize that we build a strong baseline of 3DGS for fair comparison: the densification interval is 200 iterations, which is 8 times more frequent than the 3DGS baseline in VastGaussian; the training iteration is 500K (in comparison, the training iteration of the 3DGS baseline is 450K in VastGaussian). VastGaussian trains 3DGS faster than our method. This is because our method requires additional time for data transmission. However, our method achieves better rendering quality than VastGaussian. Moreover, the data transmission time does not become the bottleneck of our method due to our balanced scene splitting method. Particularly, each block has a similar training time and the master node does not need to wait for a long time for the fat nodes to finish its job.

To further show the applicability of our method to larger-scale scenes, we evaluated our method on the  $2.7\text{km}^2$  Small City scene in the MatrixCity [24] dataset, which contains 5,620 training views and 741 validation views. We early terminated the training of the original 3DGS since it did not finish the training within two days. VastGaussian failed on this dataset since two blocks produce no 3D Gaussian primitives due to the imbalanced splitting. From Table 3, our method achieved the best results in rendering quality. The visual qualitative results are shown in Fig. 6.

Table 1: **Quantitative results of novel view synthesis on Mill19 [51] dataset and UrbanScene3D [29] dataset.**  $\uparrow$ : higher is better,  $\downarrow$ : lower is better. The red, orange and yellow colors respectively denote the best, the second best, and the third best results.  $\dagger$  denotes without applying the decoupled appearance encoding.

Scenes	Building			Rubble			Campus			Residence			Sci-Art		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Mega-NeRF [51]	20.92	0.547	0.454	24.06	0.553	0.508	23.42	0.537	0.636	22.08	0.628	0.401	25.60	0.770	0.312
Switch-NeRF [35]	21.54	0.579	0.397	24.31	0.562	0.478	23.62	0.541	0.616	22.57	0.654	0.352	26.51	0.795	0.271
3DGS [21]	22.53	0.738	0.214	25.51	0.725	0.316	23.67	0.688	0.347	22.36	0.745	0.247	24.13	0.791	0.262
Fed3DGS [48]	18.66	0.602	0.362	20.62	0.588	0.437	21.64	0.635	0.436	20.00	0.665	0.344	21.03	0.730	0.335
VastGaussian <sup>†</sup> [26]	21.80	0.728	0.225	25.20	0.742	0.264	23.82	0.695	0.329	21.01	0.699	0.261	22.64	0.761	0.261
Hierarchy-GS [22]	21.52	0.723	0.297	24.64	0.755	0.284	–	–	–	–	–	–	–	–	–
DOGS	22.73	0.759	0.204	25.78	0.765	0.257	24.01	0.681	0.377	21.94	0.740	0.244	24.42	0.804	0.219

Table 2: **Quantitative results of novel view synthesis on Mill19 dataset and UrbanScene3D dataset.** We present the training time (hh:mm), the number of final points ( $10^6$ ), the allocated memory (GB), and the framerate (FPS) during evaluation.  $\dagger$  denotes without applying the decoupled appearance encoding.

Scenes	Building				Rubble				Campus				Residence				Sci-Art			
	Train $\downarrow$	Points	Mem	FPS $\uparrow$	Train $\downarrow$	Points	Mem	FPS $\uparrow$	Train $\downarrow$	Points	Mem	FPS $\uparrow$	Train $\downarrow$	Points	Mem	FPS $\uparrow$	Train $\downarrow$	Points	Mem	FPS $\uparrow$
Mega-NeRF [51]	19:49	–	5.84	0.009	30:48	–	5.88	0.009	29:03	–	5.86	0.008	27:20	–	5.99	0.006	27:39	–	5.97	0.006
Switch-NeRF [35]	24:46	–	5.84	0.009	38:30	–	5.87	0.009	36:19	–	5.85	0.007	35:11	–	5.94	0.007	34:34	–	5.92	0.008
3DGS [21]	21:37	7.99	4.62	90.09	18:40	3.85	2.18	166.67	23:03	13.6	7.69	59.52	23:13	5.35	3.23	142.86	21:33	2.31	1.61	240.96
VastGaussian <sup>†</sup> [26]	03:26	5.60	3.07	121.35	02:30	4.71	2.74	163.93	03:33	17.6	9.61	47.84	03:12	6.26	3.67	118.48	02:33	4.21	3.54	120.33
DOGS	03:51	6.89	3.39	122.33	02:25	4.74	2.54	147.06	04:15	8.27	4.29	99.85	04:33	7.64	6.11	82.34	04:23	5.67	3.53	107.87

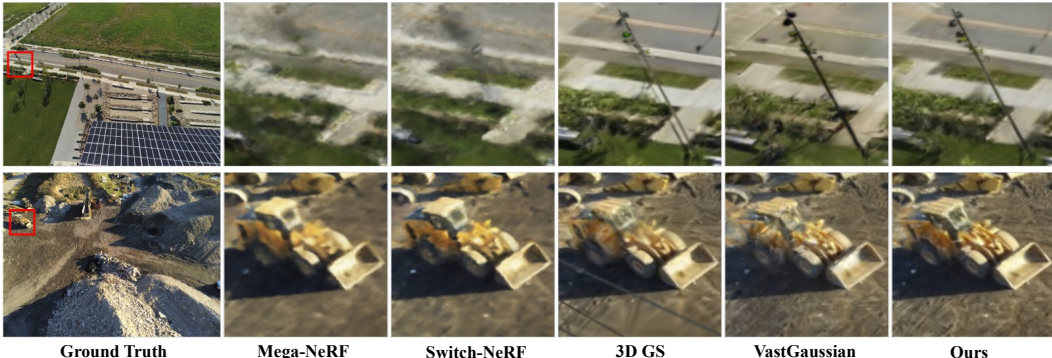


Figure 4: **Qualitative comparisons of our method and others on the Mill19 dataset.** The first row and second row are respectively the results of scene ‘building’ and ‘rubble’.

**Ablation Study.** We ablate the effectiveness of our method and present the results in Table 4. Our method without applying the 3D Gaussian consensus is denoted as *w.o. CS*, our method without adopting the self-adaptation of penalty parameters is denoted as *w.o. SD*, our method without adopting





Figure 5: **Qualitative comparisons of our method and others on the UrbanScene3D dataset.** From top to bottom are respectively the results of scenes ‘campus’, ‘residence’, and ‘sci-art’.

Table 3: **Quantitative results of novel view synthesis on the MatrixCity [24] dataset.**  $\uparrow$ : higher is better,  $\downarrow$ : lower is better. The red, orange and yellow colors respectively denote the best, the second best, and the third best results.

Scenes	aerial							street						
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Time $\downarrow$	Points	Mem $\downarrow$	FPS $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Time $\downarrow$	Points	Mem $\downarrow$	FPS $\uparrow$
3DGS	27.36	0.818	0.237	47:40	11.9	6.31	45.57	20.03	0.643	0.650	14:24	1.85	2.33	193.32
VastGaussian <sup>†</sup>	28.33	0.835	0.220	05:53	12.5	6.99	40.04	-	-	-	-	-	-	-
DOGS	28.58	0.847	0.219	06:34	10.3	5.82	48.34	21.61	0.652	0.649	02:33	2.37	2.89	180.51

the over-relaxation is denoted as *w.o. OR*. As shown in the table, the performance drastically drops without the ADMM consensus step. Furthermore, the results without applying the self-adaptation of penalty parameters is about 1.5 dB lower than the full model in PSNR. The model without applying over-relaxation is comparable to the full model in SSIM and LPIPS but has lower PSNR. We thus employ over-relaxation in our method. We also present the qualitative differences in Fig. 7, and it clearly shows the full model has better quality in the rendered images and geometries. We include more qualitative results in Fig. 8 to show the importance of the consensus step. We can observe that the distributed training presents noisy results without the consensus step. From the two bottom-right figures, we can observe obvious artifacts along the block boundary without the consensus step.

Moreover, we ablate the scale factor in constructing the overlapping areas in Table 5. We can find that the performance of our method is improved with a larger scale factor. Our method has similar performance when the scale factor is 1.4 and 1.5. However, we select 1.4 in our experiments since a larger scale factor brings a longer time and more memory requirement.

## 5 Conclusion

In this paper, we proposed DOGS, a scalable method for training 3DGS distributedly under large-scale scenes. Our method first splits scenes into multiple intersected blocks recursively and then trains each block distributedly on different workers. Our method additionally maintains a global 3DGS model on the master node. The global 3DGS model is shared with each block to encourage each block to converge to the global model. The local 3DGS of all blocks is collected to update the global 3DGS model. When evaluated on large-scale datasets, our method accelerates the 3DGS training time by  $6\times \sim 8\times$  while achieving the best rendering quality in novel view synthesis.

Table 4: **Ablation study** of our method.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
w.o. CS	22.80	0.677	0.326
w.o. SD	24.30	0.729	0.285
w.o. OR	24.45	0.766	0.259
full model	25.78	0.765	0.257

Table 5: **Ablation study** of the scale factor in our method.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Times $\downarrow$	Points	FPS $\uparrow$
1.2	24.25	0.739	0.276	02:27	4.95	129.87
1.3	24.86	0.750	0.270	02:27	5.02	128.73
1.4	25.78	0.765	0.257	02:25	4.74	147.06
1.5	25.97	0.767	0.257	02:39	4.84	130.28

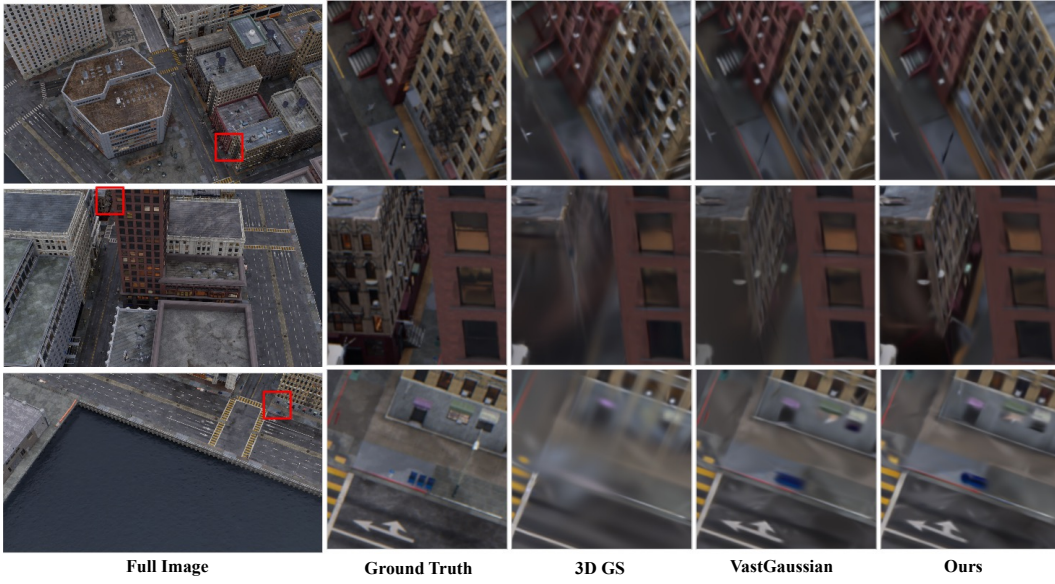


Figure 6: Qualitative results on the MatrixCity [24] dataset.

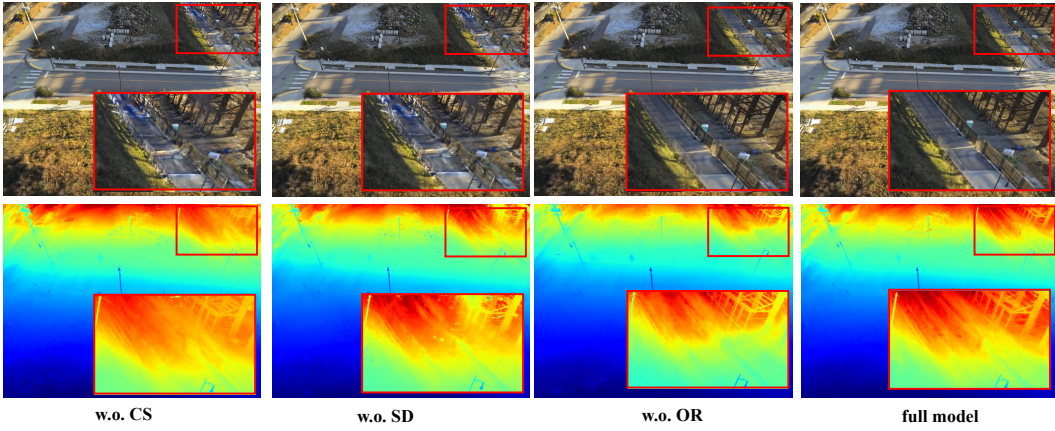


Figure 7: **Ablation study** of our method. Top: rendered images; Bottom: rendered depths.

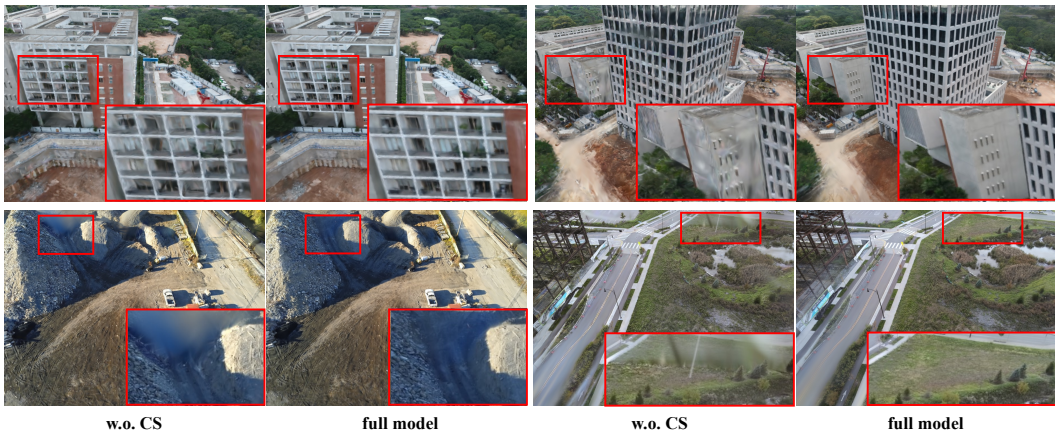


Figure 8: Importance of the consensus step.

**Acknowledgement.** This research / project is supported by the National Research Foundation (NRF) Singapore, under its NRF-Investigatorship Programme (Award ID. NRF-NRFI09-0008). Yu Chen is also partially supported by a Google PhD Fellowship.

## References

- [1] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *IEEE/CVF International Conference on Computer Vision*, pages 5835–5844, 2021.
- [2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5460–5469, 2022.
- [3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *CoRR*, abs/2304.06706, 2023.
- [4] Brojeshwar Bhowmick, Suvam Patra, Avishek Chatterjee, Venu Madhav Govindu, and Subhashis Banerjee. Divide and conquer: Efficient large-scale structure from motion using graph partitioning. In *Asian Conference on Computer Vision*, pages 273–287, 2014.
- [5] Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
- [6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Computer Vision - ECCV 2022 - 17th European Conference*, volume 13692, pages 333–350, 2022.
- [7] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 14104–14113. IEEE, 2021.
- [8] Yu Chen and Gim Hee Lee. Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24–34, June 2023.
- [9] Yu Chen and Gim Hee Lee. Dreg-nerf: Deep registration for neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22703–22713, October 2023.
- [10] Yu Chen, Shuhan Shen, Yisong Chen, and Guoping Wang. Graph-based parallel large scale structure from motion. *Pattern Recognition*, 107:107537, 2020.
- [11] Yu Chen, Zihao Yu, Shu Song, Tianning Yu, Jianming Li, and Gim Hee Lee. Adasfm: From coarse global to fine incremental adaptive structure from motion. In *IEEE International Conference on Robotics and Automation*, pages 2054–2061. IEEE, 2023.
- [12] Yu Chen, Ji Zhao, and Laurent Kneip. Hybrid rotation averaging: A fast and robust rotation averaging approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10358–10367, 2021.
- [13] Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. *arXiv preprint arXiv:2402.14650*, 2024.
- [14] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.
- [15] Anders P. Eriksson, John Bastian, Tat-Jun Chin, and Mats Isaksson. A consensus-based framework for distributed bundle adjustment. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1754–1762, 2016.
- [16] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, DeJia Xu, and Zhangyang Wang. Light-gaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps, 2023.
- [17] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5491–5500, 2022.
- [18] Lily Goli, Daniel Rebain, Sara Sabour, Animesh Garg, and Andrea Tagliasacchi. nerf2nerf: Pairwise registration of neural radiance fields. *CoRR*, abs/2211.01600, 2022.

- [19] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18344–18347. IEEE, 2022.
- [20] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat, track and map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023.
- [22] Bernhard Kerbl, Andreas Meuleman, Georgios Kopanas, Michael Wimmer, Alexandre Lanvin, and George Drettakis. A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM Transactions on Graphics*, 43(4), July 2024.
- [23] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. *arXiv preprint arXiv:2403.06912*, 2024.
- [24] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215, 2023.
- [25] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: bundle-adjusting neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 5721–5731. IEEE, 2021.
- [26] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, and Wenming Yang. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *CVPR*, 2024.
- [27] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems 33*, 2020.
- [28] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7814–7823. IEEE, 2022.
- [29] Yilin Liu, Fuyou Xue, and Hui Huang. Urbanscene3d: A large scale urban scene dataset and simulator. 2021.
- [30] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal Computer Vision*, 60(2):91–110, 2004.
- [31] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffoldsg: Structured 3d gaussians for view-adaptive rendering, 2023.
- [32] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021.
- [33] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian Splatting SLAM. 2024.
- [34] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 6331–6341. IEEE, 2021.
- [35] Zhenxing Mi and Dan Xu. Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In *The Eleventh International Conference on Learning Representations*, 2023.
- [36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12346, pages 405–421, 2020.
- [37] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022.
- [38] KL Navaneet, Kossar Pourahmadi Meibodi, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Compact3d: Compressing gaussian splat radiance field models with vector quantization. *arXiv preprint arXiv:2311.18159*, 2023.

- [39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 8024–8035, 2019.
- [40] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas A. Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12922–12932, 2022.
- [41] Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Linning Xu, Zhangkai Ni, and Bo Dai. Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians, 2024.
- [42] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.
- [43] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.
- [44] Qing Shuai, Haoyu Guo, Zhen Xu, Haotong Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Real-time view synthesis for large scenes with millions of square meters. 2024.
- [45] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, 2006.
- [46] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8259–8269. IEEE, 2022.
- [47] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5449–5459, 2022.
- [48] Teppei Suzuki. Fed3dgs: Scalable 3d gaussian splatting with federated learning. *CoRR*, abs/2403.11460, 2024.
- [49] Matthew Tancik, Vincent Casser, Xinchun Yan, Sabeek Pradhan, Ben P. Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretschmar. Block-nerf: Scalable large scene neural view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8248, 2022.
- [50] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, 2020.
- [51] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12912–12921, 2022.
- [52] Haithem Turki, Jason Y. Zhang, Francesco Ferroni, and Deva Ramanan. SUDS: scalable urban dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12375–12385, 2023.
- [53] Evangelos Ververas, Rolandos Alexandros Potamias, Jifei Song, Jiankang Deng, and Stefanos Zafeiriou. Sags: Structure-aware 3d gaussian splatting. *arXiv:2404.19149*, 2024.
- [54] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699. Computer Vision Foundation / IEEE, 2021.
- [55] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.
- [56] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 5590–5599. IEEE, 2021.
- [57] Linning Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, and Dahua Lin. Grid-guided neural radiance fields for large urban scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8306, 2023.

- [58] Zhiwen Yan, Weng Fei Low, Yu Chen, and Gim Hee Lee. Multi-scale 3d gaussian splatting for anti-aliased rendering. *CoRR*, abs/2311.17089, 2023.
- [59] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenotrees for real-time rendering of neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 5732–5741, 2021.
- [60] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. *arXiv:2311.16493*, 2023.
- [61] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018.
- [62] Runze Zhang, Siyu Zhu, Tian Fang, and Long Quan. Distributed very large scale bundle adjustment by global camera consensus. In *IEEE International Conference on Computer Vision*, pages 29–38, 2017.
- [63] Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan. Very large-scale global sfm by distributed motion averaging. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2018.
- [64] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting, 2023.

## A Appendix

### A.1 Derivation of Eq. (7)

Given Eq. (5)  $\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = \sum_{i=1}^N (f_i(\mathbf{x}_i) + \mathbf{y}_i^\top (\mathbf{x}_i - \mathbf{z}) + \frac{\rho}{2} \|\mathbf{x}_i - \mathbf{z}\|_2^2)$ ,  $\mathbf{r}_i = \mathbf{x}_i - \mathbf{z}$  and  $\mathbf{u}_i = \frac{1}{\rho} \mathbf{y}_i$ , we have

$$\begin{aligned}
 & \mathbf{y}_i^\top (\mathbf{x}_i - \mathbf{z}) + \frac{\rho}{2} \|\mathbf{x}_i - \mathbf{z}\|_2^2 \\
 &= \mathbf{y}_i^\top \mathbf{r}_i + \frac{\rho}{2} \|\mathbf{r}_i\|_2^2 \\
 &= \frac{\rho}{2} \|\mathbf{r}_i\|_2^2 + \frac{\rho}{2} \|\mathbf{y}_i\|_2^2 - \frac{1}{2\rho} \|\mathbf{y}_i\|_2^2 \\
 &= \frac{\rho}{2} \|\mathbf{r}_i + \mathbf{u}_i\|_2^2 - \frac{\rho}{2} \|\mathbf{u}_i\|_2^2.
 \end{aligned} \tag{11}$$

Substitute Eq. (11) into Eq. (5), we have the scaled form ADMM:

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{u}) = \sum_{i=1}^N (f_i(\mathbf{x}_i) + \frac{\rho}{2} \|\mathbf{x}_i - \mathbf{z} + \mathbf{u}_i\|_2^2 - \frac{\rho}{2} \|\mathbf{u}_i\|_2^2)$$

### A.2 Proof of Eq. (8b)

The complete form of Eq. (8b) of is:

$$\mathbf{z}^{t+1} = \frac{1}{K} \sum_{i=1}^K (\mathbf{x}_i^{t+1} + \mathbf{u}_i^t). \tag{12}$$

We denote the average of a vector with an overline. Then the consensus step for Eq. (8b) can be rewritten by:

$$\mathbf{z}^{t+1} = \bar{\mathbf{x}}^{t+1} + \bar{\mathbf{u}}^t. \tag{13}$$

Moreover, averaging both sides of Eq. (8c) for the dual variables gives:

$$\bar{\mathbf{u}}^{t+1} = \bar{\mathbf{u}}^t + \bar{\mathbf{x}}^{t+1} - \mathbf{z}^{t+1}. \tag{14}$$

Substituting Eq. (13) into Eq. (14) gives that  $\bar{\mathbf{u}}^{t+1} = \mathbf{0}$ , which shows that the dual variables have average value zeros after the first iteration. Therefore, we proved  $\mathbf{z}^{t+1} := \frac{1}{K} \sum_{i=1}^K (\mathbf{x}_i^{t+1})$ .

### A.3 Detailed form of Eq. (8)

We provide the detailed form of Eq. (8) in this section. Note that each 3D Gaussian can be represented by  $\mathbf{X}_i = \{\mathbf{u}_i, \mathbf{q}_i, \mathbf{s}_i, \mathbf{f}_i, o_i\}$ , we need to apply different penalty terms to different properties of 3D GS. Specifically, the dual variables correspond to different properties of 3D GS are denoted as:

$$\mathbf{u}_i = \{\mathbf{u}_p, \mathbf{u}_q, \mathbf{u}_s, \mathbf{u}_f, \mathbf{u}_o\}, \tag{15}$$

where  $\mathbf{u}_p$ ,  $\mathbf{u}_q$ ,  $\mathbf{u}_s$ ,  $\mathbf{u}_f$ ,  $\mathbf{u}_o$  are respectively denote the dual variable corresponds to the mean  $\mathbf{p}$ , the quaternion  $\mathbf{u}_q$ , the scaling matrix  $\mathbf{u}_s$ , the feature vectors to encode color information  $\mathbf{u}_f$ , and the

opacity  $\mathbf{u}_o$ . Therefore, we can expand Eq. (8) as:

$$\begin{aligned}
(\mathbf{x}_i^k)^{t+1} &:= \arg \min (f(\mathbf{x}_i^k) + \frac{\rho_p}{2} \|(\mathbf{p}_i^k)^t - \mathbf{z}_p^t + (\mathbf{u}_{p,i}^k)^t\|_2^2 + \frac{\rho_q}{2} \|(\mathbf{q}_i^k)^t - \mathbf{z}_q^t + (\mathbf{u}_{q,i}^k)^t\|_2^2 \\
&+ \frac{\rho_s}{2} \|(\mathbf{s}_i^k)^t - \mathbf{z}_s^t + (\mathbf{u}_{s,i}^k)^t\|_2^2 + \frac{\rho_f}{2} \|(\mathbf{f}_i^k)^t - \mathbf{z}_f^t + (\mathbf{u}_{f,i}^k)^t\|_2^2 \\
&+ \frac{\rho_o}{2} \|(\mathbf{o}_i^k)^t - \mathbf{z}_o^t + (\mathbf{u}_{o,i}^k)^t\|_2^2),
\end{aligned} \tag{16a}$$

$$\begin{aligned}
\mathbf{z}_p^{t+1} &:= \frac{1}{K} \sum_{k=1}^K (\mathbf{p}_i^k)^{t+1}, \mathbf{z}_q^{t+1} := \frac{1}{K} \sum_{k=1}^K (\mathbf{q}_i^k)^{t+1}, \mathbf{z}_s^{t+1} := \frac{1}{K} \sum_{k=1}^K (\mathbf{s}_i^k)^{t+1}, \\
\mathbf{z}_f^{t+1} &:= \frac{1}{K} \sum_{k=1}^K (\mathbf{f}_i^k)^{t+1}, \mathbf{z}_o^{t+1} := \frac{1}{K} \sum_{k=1}^K (\mathbf{o}_i^k)^{t+1},
\end{aligned} \tag{16b}$$

$$\begin{aligned}
(\mathbf{u}_{p,i}^k)^{t+1} &:= (\mathbf{u}_{p,i}^k)^t + (\mathbf{p}_i^k)^{t+1} - \mathbf{z}_p^{t+1}, (\mathbf{u}_{q,i}^k)^{t+1} := (\mathbf{u}_{q,i}^k)^t + (\mathbf{q}_i^k)^{t+1} - \mathbf{z}_q^{t+1}, \\
(\mathbf{u}_{s,i}^k)^{t+1} &:= (\mathbf{u}_{s,i}^k)^t + (\mathbf{s}_i^k)^{t+1} - \mathbf{z}_s^{t+1}, (\mathbf{u}_{f,i}^k)^{t+1} := (\mathbf{u}_{f,i}^k)^t + (\mathbf{f}_i^k)^{t+1} - \mathbf{z}_f^{t+1}, \\
(\mathbf{u}_{o,i}^k)^{t+1} &:= (\mathbf{u}_{o,i}^k)^t + (\mathbf{o}_i^k)^{t+1} - \mathbf{z}_o^{t+1}.
\end{aligned} \tag{16c}$$

#### A.4 Algorithm for Recursive Scene Splitting

We present the algorithm for the recursive scene splitting in Alg. 1.

---

##### Algorithm 1 Recursive Scene Splitting Algorithm

---

**Input:** 3D points  $\{\mathbf{X}_i\}$ , number of blocks  $K$ .

**Output:** Local 3D points in each block  $\mathcal{X} = \{\mathbf{X}_{i,k}\}$

- 1: Estimate a bounding box aabb which tightly covers all 3D points.
  - 2: Initialize cells  $\mathcal{C} = \{\text{aabb}\}$ , local 3D points  $\mathcal{X} = \emptyset$
  - 3: **while**  $|\mathcal{C}| < K$  **do**
  - 4:     Let current cells  $\mathcal{C}_{\text{cur}} \leftarrow \mathcal{C}$
  - 5:     **while**  $|\mathcal{C}_{\text{cur}}| > 0$  **do**
  - 6:         Pop a bounding box in current cells  $\text{aabb} := \mathcal{C}_{\text{cur}}.\text{pop}(0)$
  - 7:         Remove the bounding box from cells  $\mathcal{C} := \mathcal{C} - \{\text{aabb}\}$
  - 8:         Split the bounding box into two sub-cells  $\text{aabb}_1, \text{aabb}_2$  along the longer axis
  - 9:         Group points into two blocks  $\{\mathbf{X}_1, \mathbf{X}_2\}$
  - 10:         Re-estimate two tighter bounding boxes  $\text{aabb}'_1, \text{aabb}'_2$  for  $\{\mathbf{X}_1, \mathbf{X}_2\}$
  - 11:         Push bounding boxes  $\text{aabb}'_1, \text{aabb}'_2$  into cells:  $\mathcal{C} := \mathcal{C} + \{\text{aabb}'_1, \text{aabb}'_2\}$
  - 12: **for** block  $k \in |\mathcal{C}|$  **do**
  - 13:     Group points  $\mathcal{X}_k$  located in the same cell  $\mathcal{C}_k$  into a same block
  - 14:     Update  $\mathcal{X}$  by  $\mathcal{X} := \mathcal{X} + \mathcal{X}_k$
- 

At line 5,  $|\cdot|$  denotes the capacity of a set. Algorithm 1 is adopted to split both the training views and sparse point clouds from SfM.

#### A.5 Algorithm for Distributed Training of 3D GS

We present the algorithm for the distributed training of 3D Gaussian Splatting in Alg. 2.

Note that we adopt Eq. (16) in our implementation.

#### A.6 More Implementation Details

All experiments are conducted on the Nvidia RTX 6000 GPUs with 48 GB memory. For our method, we initialize the dual variables  $\mathbf{u}_p, \mathbf{u}_q, \mathbf{u}_s, \mathbf{u}_f, \mathbf{u}_o$  to zeros. For the penalty parameters, we set  $\rho_p, \rho_q, \rho_s, \rho_o$  to  $1e4$  and  $\rho_f$  to  $1e3$  empirically. Though there are other choices of the initial values



---

**Algorithm 2** Distributed 3D Gaussian Training Algorithm

---

**Input:** Initial 3D Gaussians in each block  $\{\mathbf{X}_k \mid k \in [1, K]\}$ , consensus interval  $\text{intv}$

**Output:** Global 3D Gaussians  $\mathbf{z} = \{\mathbf{u}_i, \mathbf{q}_i, \mathbf{s}_i, \mathbf{f}_i, o_i\}$

```
1: Initialize  $\mathbf{u}_i$  as 0
2: for  $t < T$  do
3:   for block  $k < K$  distributedly do
4:      $\{\mathbf{u}_{i,k}^{t+1}, \mathbf{q}_{i,k}^{t+1}, \mathbf{s}_{i,k}^{t+1}, \mathbf{f}_{i,k}^{t+1}, o_{i,k}^{t+1}\} := \arg \min (f_i(\mathbf{x}_{i,k}) + \frac{\rho}{2} \|\mathbf{x}_{i,k}^t - \mathbf{z}^t + \mathbf{u}_{i,k}^t\|_2^2)$ 
5:   if  $t \bmod \text{intv} == 0$  then
6:     for block  $k < K$  distributedly do
7:       Send local 3D Gaussians  $\mathbf{X}_k^{t+1}$  to the master node
8:       Apply the consensus step  $\mathbf{z}^{t+1} := \frac{1}{K} \sum_{i=1}^K (\mathbf{x}_{i,k}^{t+1})$ 
9:       Broadcast the global 3D Gaussians  $\mathbf{z}^{t+1}$  to all slave nodes
10:    for block  $k < K$  distributedly do
11:      Update the dual variables  $\mathbf{u}_{i,k}^{t+1} := \mathbf{u}_{i,k}^t + \mathbf{x}_{i,k}^{t+1} - \mathbf{z}^{t+1}$ 
```

---

that could improve the results of our method, we found this set of values is good enough for all scenes in our experiments and we did not do more ablations on the initial values. Due to computational resources limitation, we test our method on only 5 GPU nodes, where one is the master node that maintains the global model and the others are slave nodes for training local 3D Gaussians. The performance of our method can be improved further with more GPU nodes.

### A.7 More Qualitative Results

We present more qualitative results of our method in Fig. 9 and Fig. 10. We provide qualitative comparisons with VastGaussian in areas where blocks overlap in Fig. 10. Both methods produce fairly consistent results. However, our method presents higher fidelity rendering results than VastGaussian near the splitting boundary, which also validated the effectiveness and importance of the consensus step.

### A.8 Limitations and Future Work

Our method can distributedly train 3D GS on large-scale scenes. However, it brings additional communication overheads to the system. Fortunately, we found the communication overhead did not slow down the training performance. This is due to our balanced splitting algorithm minimizing the synchronization time when there is a need to consensus the local 3D Gaussians from all blocks. Moreover, we pruned unnecessary small 3D Gaussians to further reduce the number of 3D GS, which also reduced the communication overhead.

Though our method can train large-scale scenes efficiently, the GPU memory requirement can still be an issue. This is because when we zoom out to capture larger areas, more 3D Gaussians are included in the rasterization step, which consumes more GPU memory. Our future work will consider introducing the level-of-details (LOD) technique into our distributed training pipeline. Similar to existing LOD GS methods [41, 44], LOD can be utilized to reduce the number of 3D Gaussians that are far away from the cameras.

### A.9 More Discussions

**Training and Waiting Time of Each Block.** We tested the time cost from transferring the data to the master node to receiving data from the master node for each slave node on the Campus dataset. The mean and variance of time are respectively 5.63 seconds and 0.75 seconds each time. The low variance indicates that our method can balance the training time well. We argue that the data transfer time of our method can be kept constant since we can always control the number of local Gaussians to a constant number (*e.g.*  $\leq 6000,000$  3D Gaussian primitives) with enough GPUs, no matter how large the scene is, since the data transfer between different slave nodes and the master node is executed distributedly instead of sequentially.

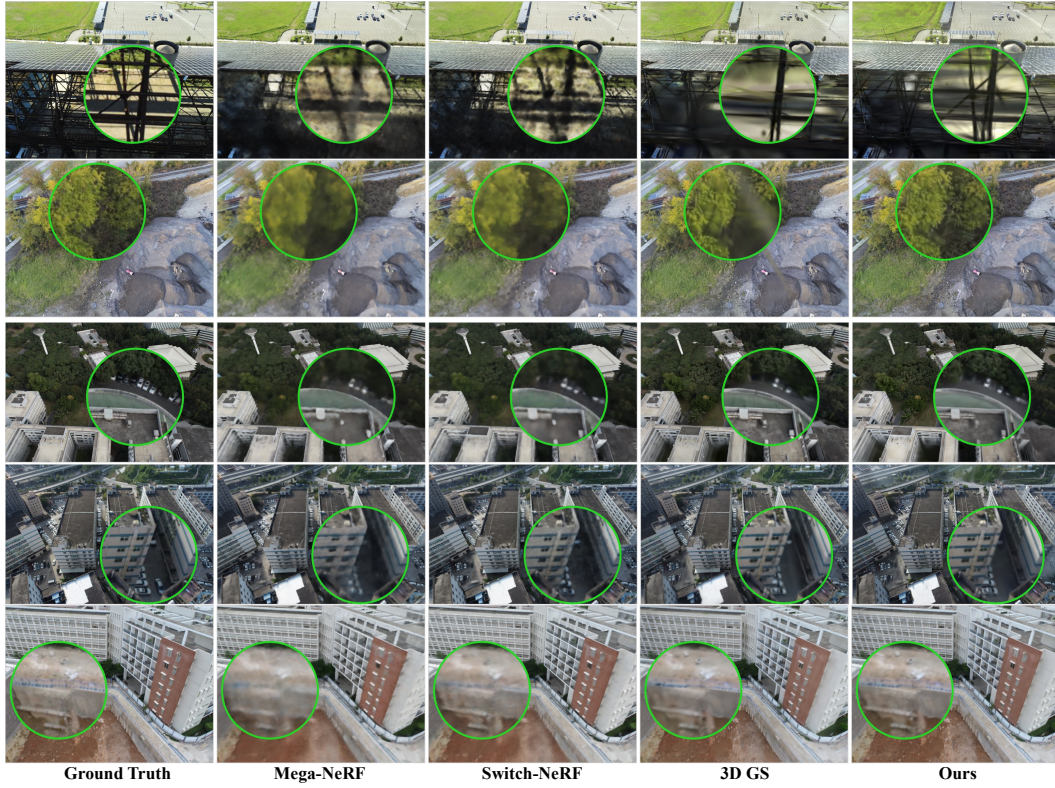


Figure 9: More qualitative results on the mill19 dataset and the UrbanScene3D dataset. Zoom in for the best view.

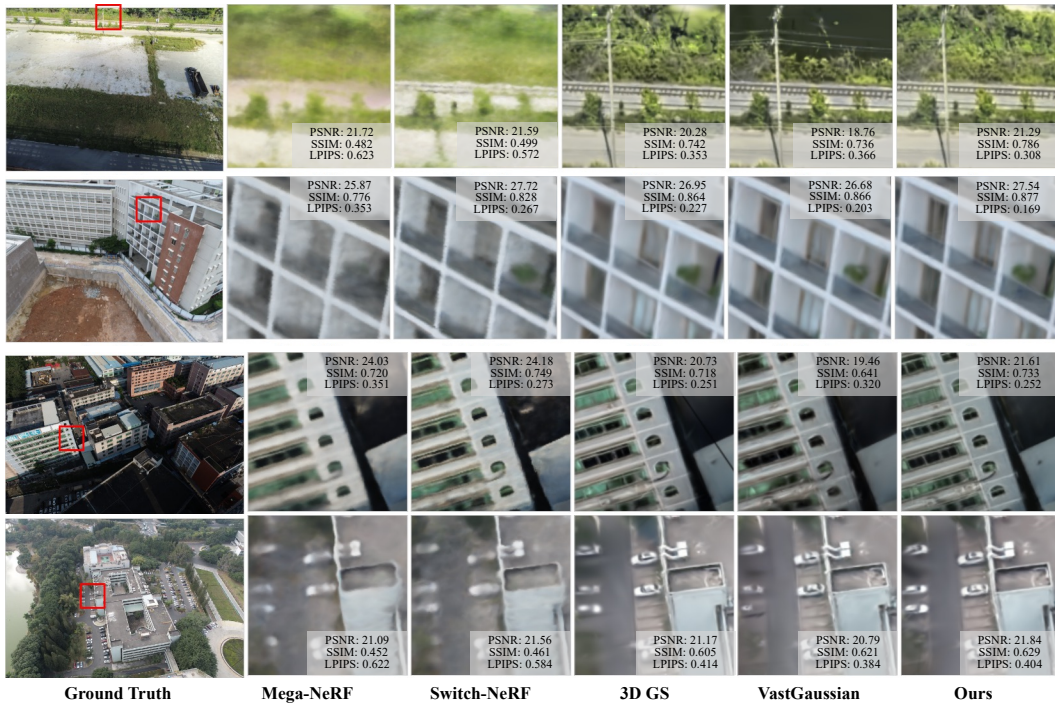


Figure 10: More qualitative comparisons of our method and state-of-the-art methods.

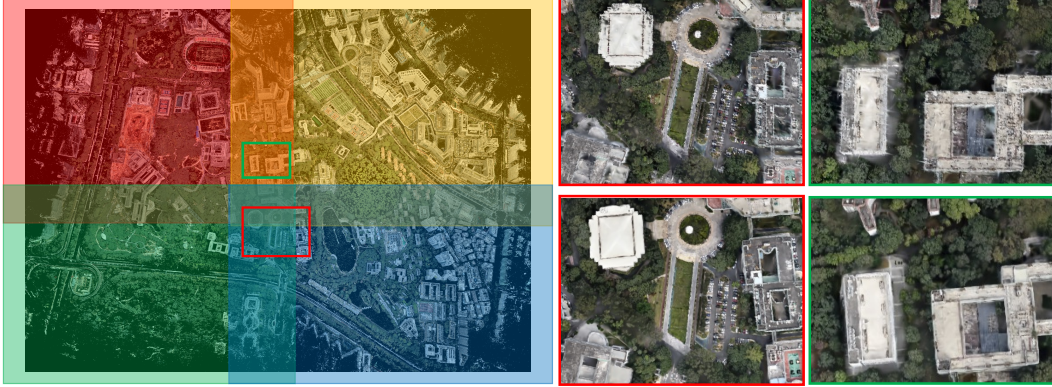


Figure 11: Qualitative results near scene boundary on the UrbanScene3D dataset. Top-Right: VastGaussian. Bottom-Right: our method.

**Implementation Flexibility.** One another issue with our method is that the implementation of our method is not as flexible as VastGaussian. Our method relies on an RPC module for the data transmission and communication between the master node and the slave nodes. On the other hand, VastGaussian can be implemented decentralized without the master node. However, our method can render higher-quality images and the communication overheads can be neglected compared to the long training time of the original 3D GS. Therefore, this is not a limitation of our work.

**Shift to Other 3D GS Representations.** Many follow-up works improved the original 3D GS. Some works focus on compressing the size of 3D GS [16, 38] and some of the other works focus on constraining the training without changing the optimization parameters [64, 23, 13]. These methods can also be applied to the training of each block in our method. Some other works also adopted intermediate representations to improve the original 3D GS, *i.e.*, OctreeGS [41] decoded the properties of 3D GS from the feature embedding in each anchor node, and SAGS [53] adopts hash encodings for each 3D Gaussian, a GNN encoder, and a corresponding decoder to generate the properties of 3D Gaussians. In this case, we can change the optimization parameters and the corresponding penalty parameters and dual variables to these intermediate feature embeddings/encodings. In our future work, we will consider a more consolidated implementation that can easily shift to these 3D GS representations.

**Comparison to Concurrent Works.** 1) **VastGaussian** [26] focuses mostly on the data splitting strategy and guaranteeing the consistency of different 3D Gaussians can also be a challenge in different scenes. However, our method ensured the consistency of the shared 3D Gaussians through the 3D Gaussian consensus with only a quite simple splitting approach. Nonetheless, the data splitting approach in VastGaussian can be introduced into our framework to enhance the robustness. Moreover, the decoupled appearance encoding can also be applied to our method to reduce floaters. Our work and VastGaussian are thus complementary to each other. 2) **Hierarchy-GS** [22] also trains 3DGS distributedly in different grid cells. However, Hierarchy-GS focuses more on the rendering speed by designing a hierarchical tree structure for 3D Gaussians. During training, the generated hierarchy is loaded and optimized by randomly selecting a training view and a target granularity. After training all individual chunks, Hierarchy-GS first runs a coarse initial optimization on the full dataset and adds an auxiliary skybox. This coarse model serves as a minimal basis for providing backdrop details for parts of the scene that are outside of a chunk. When consolidating different chunks, a 3D Gaussian primitive is deleted if it is associated to-but outside-chunk and is closer to another chunk to maintain the consistency of 3D Gaussians. The naive pruning approach neglects the gradient flow of the 3D Gaussians to other 3D Gaussians that are inside only in an associated chunk. Therefore, its rendering quality can be worse than VastGaussian and our method. Nonetheless, its hierarchical LOD approach can still serve as a good complement to our method and VastGaussian.

**Social Impact.** DOGS, VastGaussian and Hierarchy-GS focus on different parts of training 3D GS on large-scale scenes and therefore are complementary to each other. These methods can be consolidated into a more robust and scalable system to open a new world in city-scale 3D reconstruction. Our method, however, requires an additional master node for controlling the training of the 3DGS, which consumes more GPU resources.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Refer to the abstract and the last paragraph of Sec. 1 for our contribution.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Please refer to Sec. A.8 in the appendix for the limitations.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Proofs can be found in Sec. A.1 and Sec. A.2 in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Implementation details are included in Sec. 4 in the main paper. Additional implementation details are included in Sec. A.6 in the supplementary. Pseudo algorithms are included Sec. A.4 and Sec. A.5 in the supplementary for better understanding and reproducing the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release related code and checkpoints upon the acceptance of the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Training and testing details are included in Sec. 4 in the main paper. Additional details are included in Sec. A.6 in the supplementary.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We follow related works for the setting of error bar.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Compute resources and running time are included in Sec. 4 in the main paper and Sec. A.6 in the supplementary.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We strictly follow the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Societal impacts is included in Sec. A.9 in the supplementary.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper has no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We included the urls and cited the papers for code and datasets used in this paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.



- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We use and cite existing datasets in this work. Other assets including related code/models will be released upon acceptance.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.