CaLiSa-NeRF: Neural Radiance Field with Pinhole Camera Images, LiDAR point clouds and Satellite Imagery for Urban Scene Representation

Juyeop Han[1], Guilherme V. Cavalheiro[1], Josef Biberstein[1], Elham Alkabawi[2], Shahad Alqhatni[2], Fadwa Alaskar[2], Eman Bin Khunayn[2], Sertac Karaman[1*]
[1]MIT, USA [2]KACST, Saudi Arabia

Abstract

*Research on Neural Radiance Fields (NeRF) has advanced rapidly due to their ability to impressively reconstruct 3D scenes from perspective camera images alone. Recently, other modalities, such as LiDAR point clouds and satellite imagery, have also been successfully explored for NeRF models. Despite the potential to create accurate reconstructions from each of these data sources, it can only be realized when the available data sufficiently covers a scene of interest, a condition that is hard to satisfy in practice for these sensor modalities in isolation. To tackle this issue, this work studies the unexplored task of training NeRFs by combining ground-based and satellite-based data, two data sources with complementary coverage attributes. We propose **CaLiSa-NeRF**, a novel NeRF model that simultaneously integrates perspective camera images, satellite images with Rational Polynomial Coefficients (RPCs), and LiDAR point clouds to represent urban environments better. Various techniques are introduced to harmonize these heterogeneous sensor inputs for NeRF training and the resulting methods are able to represent both side and top views, unlike the methods restricted to a particular data origin. We demonstrate the effectiveness of the proposed methods by training and evaluating them on a real dataset collected from Riyadh.*

1. Introduction

The first Neural Radiance Field (NeRF) model was proposed in [11] to create photorealistic representations of a 3D scene from a collection of perspective images with known poses by means of learning optical properties of scene with a neural network. It has garnered significant attention from researchers due to its impressive reconstruction quality and conceptual flexiblity, though still having significant limitations. Subsequent research has sought to address these limitations by improving its computational performance [12] and extending its applicability to more uncontrolled and unbounded environments [1, 9].

One improvement direction involves incorporating different types of data into the method, in addition to or as a replacement for the ubiquitous use of perspective camera images. In particular, depth information can facilitate training by providing direct information about the occupancy on a region, thus improving novel view synthesis quality with fewer training images. This information can be obtained from diverse sources, such as structure from motion [3], LiDAR point clouds [2, 13], and depth completion methods [14]. Another valuable source of information is satellite imagery, as this sensor modality can complement data coverage limitations from ground sensors (e.g., viewing a scene from the top), while also providing logistical benefits in comparison to both ground and aerial collection. However, satellite images, besides having their own resolution and data coverage limitations (e.g., viewing a scene laterally), present additional challenges in their usage with NeRF models. They are often captured over multiple dates, making factors like solar direction [4] and transient objects [7] important considerations and they also require more complex camera models than the typical pinhole one, such as the rational polynomial coefficient (RPC) model. Despite these complexities, satellite-based NeRF models [8] have been shown to reconstruct urban environments more effectively than conventional multi-view stereo methods [5].

Several studies have augmented NeRF models to improve their results on urban environments, potentially at city-scale level. Some studies partition 3D regions of the environment [16] or regions within training images [18] and train multiple NeRF models in a more scalable manner. Other works [13,19] incorporate LiDAR data with camera images to identify scene elements (such as dynamic objects and the sky) that can be more efficiently handled when treated separately from other elements. Drone-captured images [18] and satellite images [8] have also each been used in isolation to model an urban environment with a NeRF. Nevertheless, whether satellite, aerial, or ground data, these

*Corresponding author, sertac@mit.edu

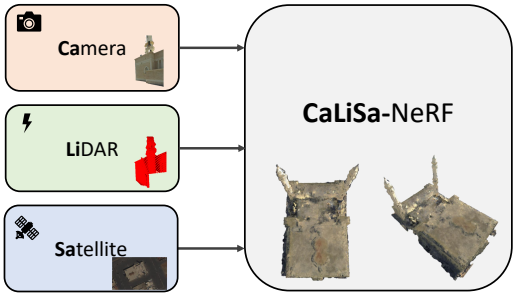


Figure 1. Overview of the CaLiSa-NeRF Framework: Integrating RGB pinhole camera images, LiDAR point clouds, and satellite imagery to create a unified NeRF model.

methods are limited by the restrictions inherent in their respective sensor modalities. Aerial images may not be able to be used due to their cost and regulatory constraints. Ground images alone cannot offer top-down views of the urban environment, while satellite images cannot offer side views.

In this work, we present **CaLiSa-NeRF**, a unified NeRF model that combines data from perspective camera images, LiDAR point clouds, and RPC satellite imagery to represent outdoor urban scenes. We design specialized data processing and training procedures for CaLiSa-NeRF, including ray sampling for satellite images using RPC coefficients, to effectively integrate these heterogeneous data sources. We further optimize the NeRF architecture and loss function to handle the unique characteristics of these inputs.

The main contributions of this research are as follows. First, we propose a NeRF model training framework that integrates perspective camera images, LiDAR point clouds, and satellite imagery. To the best of our knowledge, this is the first work to combine satellite-level data with any ground-level data to train a NeRF, thus also being the first to incorporate these three modalities simultaneously. Second, CaLiSa-NeRF performs a novel view synthesis of outdoor urban environments from random camera poses. Specifically, training with both ground data and satellite imagery enables the representation of both side views and top views of urban areas. Lastly, We validate the performance and versatility of CaLiSa-NeRF through experiments using data collected in Riyadh.

2. Related Work

2.1. Neural Radiance Fields

Neural Radiance field (NeRF) [11] is typically used as an umbrella term to describe methods that use machine learning techniques to model the opacity and color information at each spatial position in a scene. This information is aggregated along arbitrary rays through a process called volume rendering, resulting in the ability to synthesis novel views from arbitrary locations within the scene. Given a ray, $\mathbf{r}$, from the desired novel view, described by an origin $\mathbf{o} \in \mathbb{R}^3$ and a direction $\mathbf{d} \in \mathbb{R}^3$, the position of any point $\mathbf{x}$ along it can be described by $\mathbf{x} = \mathbf{o} + t\mathbf{d}$ for some scalar t that ensure appropriate ray start and termination locations induced by t_{start} and t_{end} respectively. The radiance information along a ray is aggregated into the ray color $\mathbf{C}$ according to:

$$C(r) = \int_{t_{start}}^{t_{end}} W(r(t), t)c(r(t), d)dt \tag{1}$$

where $\mathbf{c}$ is the color at a specific 3D location and direction (which can be encoded with 2 coordinates) and W is a weight factor that can be computed from the opacity σ along the ray:

$$W(r(t), t) = T(r(t), t)\sigma\left(r(t)\right) \tag{2}$$

$$T(r(t), t) = \exp\left(-\int_{t_{start}}^{t} \sigma(r(s))ds\right) \tag{3}$$

The quantity T is the transparency or transmittance along a ray and quantifies the extent to which light can propagate between the start of the ray and the sample position. The opacity or denisty σ reflects the rate of light absorption and, conversely, the rate of light emission.

In practice, these integrals are discretized along sampling points, while the opacity and color information per point is represented as a learnable functions of arbitrary locations and directions in the scene, typically using neural networks. This allows the training of these mappings by minimizing a photometric loss between the color $\mathbf{C}$ of rendered rays and the ground-truth color $\overline{\mathbf{C}}$ associated with a training image for the same ray:

$$\mathcal{L}_{RGB}(\mathcal{R}) = \sum_{\mathbf{r}\in\mathcal{R}} ||\overline{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})||^2. \tag{4}$$

where $\mathcal{R}$ is the set of the sampled ray used for training.

2.2. Depth Priors and LiDAR Integration in Neural Radiance Fields

In recent years, various approaches have been developed to integrate depth priors into NeRF for enhanced scene representation. This extra information typically improves the learned geometric structure of the NeRF, leading to quality on non-training images, especially on the few-shot scenarios.

Some methods leverage depth information from the available image data or pre-existing models. DS-NeRF [3] utilizes sparse 3D features obtained from Structure-from-Motion (SfM) for depth-aware scene reconstruction. This sparse depth information is transformed into dense depth data in [14] by leveraging depth completion networks. Point-NeRF [22] uses cost-volume based neural networks

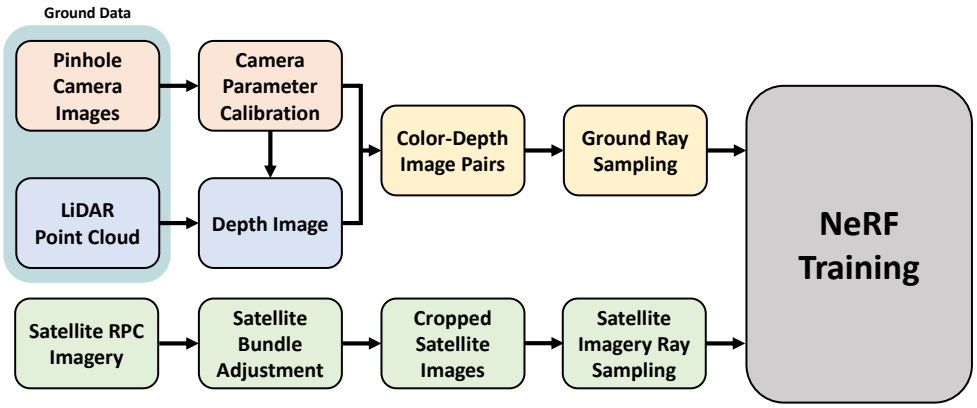


Figure 2. Data processing workflow for sampling rays from multi-modal data sources for NeRF training.

to predict depth and reproject 2D images points into 3D in order to facilitate an alternate point-based scene representation.

Other methods focus on utilizing LiDAR measurements as source of depth information to supplement camera data or as target for learning. For example, Urban Radiance Fields [13] integrates expected depth and line-of-sight losses to refine scene representation, while CLONER [2] decouples learning of geometry and color by training a network for occupancy prediction exclusively on LiDAR information and a separate network for color prediction exclusively on color images. Another method, called NFL [6], only utilize LiDAR point clouds. It focuses on synthesizing realistic LiDAR scans from novel viewpoints by a volume rendering procedure suitable for active sensors.

2.3. Neural Radiance Fields for Large Scale Outdoor Environment

Training a NeRF outside of controlled small scale environments poses additional challenges. It is necessary to account for data collection taken over long period of times, with uncontrolled factors such as transient objects, varying illumination and unbounded scenes. Furthermore, scalability can become a concern, depending on the dimensions of the scene and the desired level of detail in the reconstruction.

Unbounded scenes are often handled with warping techniques that approximate a virtually infinite background space with a finite one. NeRF++ [23] and Mip-NeRF 360 [1] contract unbounded scenes into bounded scenes using sphere-like space warping for more efficient training and rendering. Moreover, F2-NeRF [20] generalizes these space-warping techniques, enabling NeRFs to be trained with images captured along random trajectories in unbounded scenes. To represent a city-wide space in detail, Block-NeRF [16] and Mega-NeRF [18] partition large spaces into smaller subspaces, assigning individual NeRF models to render each subspace, thereby allowing better scaling beyond a single neural network. The problem of handling aerial images from earth-scale to city-scale is explored in [21].

NeRFs have been trained with satellite imagery to reconstruct the 3D surface geometry and appearance of urban areas. Shadow-NeRF [4] was the first to train a NeRF using multi-view, multi-date satellite imagery, accounting for variations in solar lighting conditions and diffuse light sources from the sky. Additionally, SatNeRF [7] and EO-NeRF [8] replace the traditional pinhole camera model with the RPC camera model, which is better suited for processing satellite imagery and is a standard choice by suppliers. These models utilize loss functions related to solar and sky lighting, as well as transient phenomena introduced in Shadow-NeRF and NeRF-W [9], respectively.

Besides satellite imagery, other sensor modalities have been explored to train NeRFs for large outdoor environments. LiDAR measurements were used in [13] as depth priors to maximize the efficiency of the sparse training data. SUDS [19] uses optical flow and 2D descriptors on multiple videos, in addition to LiDAR information.

This work aims to demonstrate the effectiveness of fusing real data from ground-level collections in the form of LiDAR and camera images with real satellite imagery. To the best of our knowledge, no prior research has combined these distinct types of information or incorporated perspectives as significantly different as side views from up close and top views from afar.

3. Method

3.1. Data processing

To train the NeRF model using the given multi-modal data (ground RGB images, LiDAR point clouds, and RPC

satellite imagery) each type of raw data must be preprocessed as described in Fig 2.

First, the data captured on the ground is preprocessed jointly. As both camera and LiDAR operate simultaneously, a point cloud with the collected LiDAR data can be synchronized with each image based on their respective capture times The extrinsic and intrinsic parameters of the perspective camera images, including lens distortion, are estimated through bundle adjustment using tools such as COLMAP [15]. Based on these estimated parameters, a depth image is generated by measuring the depth between the camera and the scene based on the LiDAR point cloud and projecting it onto the image plane. The pairs of RGB images and depth images are then both used to sample training rays from the ground.

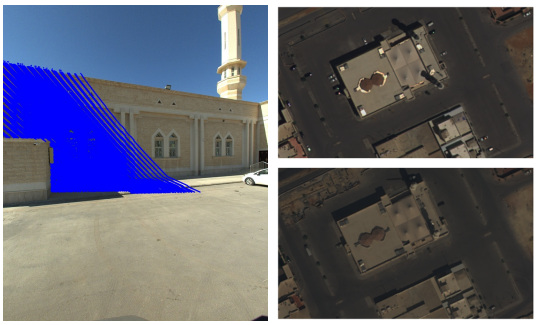
Similarly, Each Rational Polynomial Coefficient (RPC) in the satellite imagery is refined through bundle adjustment [10]. After refinement, the satellite images are cropped according to the desired region of interest (ROI) to allow for more efficient sampling of rays. Since the number of satellite images is smaller compared to the number of ground RGB images, we fix the ratio of ray samples between satellite and RGB images.

To combine ground data and satellite imagery, we represent both types in a common global coordinate system given by UTM coordinates augmented with altitude values. As mentioned in EO-NeRF [8], such system is locally Cartesian and has the advantaged of being aligned with the vertical axis, which facilitate efficient use and specification of the scene bounds. Global pose information for the ground data is obtained with the help of GPS sensors, while georeferencing is already part of the satellite data product.

3.2. RPC ray casting

A custom camera is defined in nerfstudio [17] to generate rays for each pixel of images whose cameras are modeled by RPC cameras, similar to the approach used in EO-NeRF [8]. Unlike pinhole camera models, RPC camera models define mappings between 3D coordinates and their corresponding 2D pixel coordinates as the ratio of polynomials. Rays from such mappings are determined from the lines connecting two extreme points who share the same 2D coordinates within a region of interest.

More specifically, a bounding box for the scene is determined as the smallest bounding box that contains the bounds of all the images associated with RPCs and expressed in the common system of UTM coordinates, as explained in subsection 3.1. Within this bounding box, rays can then be defined for a pair of pixel coordinates as starting from the point of highest altitude to the one of lowest altitude. Proper scaling and normalization are also applied to match rays with the bounding boxe used by NeRF itself.



(a) Ground data (b) Satellite imagery

Figure 3. (a) The LiDAR point cloud is projected onto the pinhole camera images. (b) Examples satellite images cropped to the region of interest of the scene.

3.3. Network Architecture and Loss

We utilize depth-nerfacto model available in nerfstudio [17]. Depth-nerfacto is a depth-supervised version of nerfacto, a nerftudio model meant to be a standard choice that combines feature from several methods. It includes components such as appearance embedding and hash encoding. Hash encoding enables fast training and rendering of a NeRF model. Appearance embedding handles variations in the appearance of the scene such as illumination changes, camera exposure, and transient objects. In particular, the appearance embedding is crucial for this problem since the capture date, lighting condition and other view specific characteristics of the ground RGB images and the satellite imagery differ significantly.

For depth-supervision, we use the depth-loss function based on Kullback-Leibler (KL) divergence described in DS-NeRF [3]:

$$\mathcal{L}_{depth}(\mathcal{R}) = \sum_{\mathbf{r}\in\mathcal{R}} \sum_{i=1}^{N_s} -\log w_i \exp\Big(-\frac{(t_i - \mathbf{D}(\mathbf{r}))^2}{2\sigma^2}\Big)\delta_i \tag{5}$$

where $\mathbf{D}(\mathbf{r})$ is the measured depth, and σ represents its standard deviation. The other notations are described in the subsection 2.1.

The total loss of the network model is represented as the weighted sum of the color loss (4) and the depth loss (5):

$$\mathcal{L}_{total}(\mathcal{R}) = \mathcal{L}_{RGB}(\mathcal{R}) + \lambda \mathcal{L}_{depth}(\mathcal{R}_{depth}) \tag{6}$$

where $\mathcal{R}_{depth} \subset \mathcal{R}$ is the set of rays having depth measurement information, and λ is the weight of the depth loss.

4. Experiment

The purpose of the experiment is to validate whether our CaLiSa-NeRF can render the region of interest (ROI) from wider range of novel views. For the experiments, we collect and process the dataset. Given the dataset of the ROI,

Figure 4. Rendered views of the mosque from various perspectives, including the rooftop, for all methods except 'Sat.'.

we compare the rendering performance of NeRF models trained with various data sources: RGB camera images, LiDAR point clouds, and Satellite imagery.

4.1. Dataset

The dataset for training the NeRF models and testing their performance was collected in Riyadh, Saudi Arabia. We restrict the ROI as a mosque in Riyadh.

The pinhole camera images and LiDAR point clouds

were recorded together with latitude, longitude, and altitude coordinates via GPS. The dimensions of each camera image are 2048 $\times$ 2448 pixels, and a total of 157 images are used in the dataset. We transform the camera poses and LiDAR point clouds into UTM coordinates and altitude, and the camera intrinsic parameters are estimated using COLMAP [15]. The pose values measured by GPS are used to provide camera extrinsics in place of the values estimated with structure-from-motion due to their superior accuracy. We then undistort the raw camera images using the estimated intrinsic camera parameters. Furthermore, LiDAR point clouds captured within 0.5 seconds before or after each camera image are projected to generate depth image as shown in Fig 3 (a). Note that the LiDAR point clouds only covers a small part of the scene due to the displacement between the LiDAR and the camera.

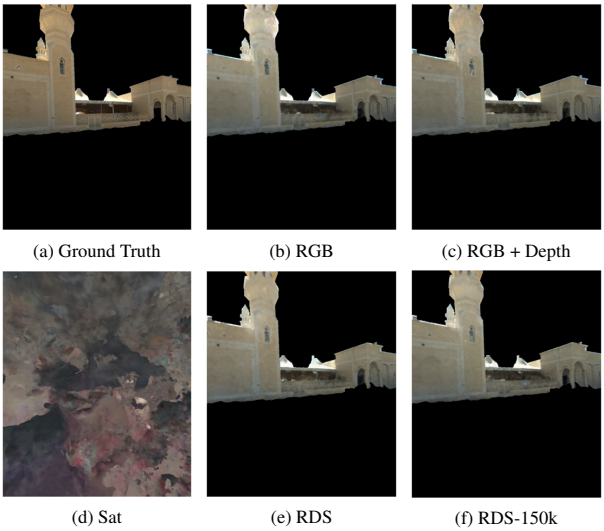
Besides ground-based data, a total of six satellite images captured over the span of six months are used as part of the dataset, as illustrated in Fig 3 (b). The RPCs of each images are optimized through bundle adjustment as described in Section 3. We manually calibrate the UTM coordinate offset of the rays from the satellite imagery to account for the coordinate discrepancy between the ground data and the satellite imagery.

4.2. Setup and Training Details

In this experiment, we compare (depth-)nerfacto models trained with multiple sources of data. The models are as follows: **Sat.**, **RGB**, **RGB + Depth**, **RGB + Depth + Sat.**, and **RGB + Depth + Sat.-150k**. 'RGB', 'Depth', and 'Sat.' indicate the use of RGB camera images, LiDAR depth, and satellite imagery as training data for NeRF, respectively.

All models are trained for 100k iterations except for 'RGB + Depth + Sat.-150k', which is trained for 150k iterations. For the 'RGB + Depth + Sat.' model, half of the rays are sampled from satellite imagery during training, whereas in 'RGB + Depth + Sat.-150k', 34% of rays are sampled from satellite imagery. The purpose of 'RGB + Depth + Sat.-150k' is to provide a fair comparison with models trained solely on ground data. Specifically, the total number of rays sampled from ground data in 'RGB + Depth + Sat.' is half of that sampled in 'RGB' and 'RGB + Depth'. For every model, the number of rays per batch is 2048, and the learning rate for each NeRF field starts at 10^{-2} and exponentially decreases to 10^{-4}.

We compare the rendering performance of each model from two perspectives: (1) how well each model represents the scene from the ground view, and (2) how well each model represents the scene from the rooftop view. For the ground view, we select eight held-out ground RGB images for evaluation and generate a mask for each image to only evaluate the rendering and estimated depth of the main building of interest in the region. For the rooftop view, we conduct only a qualitative study by comparing the rendered rooftop scene of each model due to limitations in the amount of available data.



(a) Ground Truth (b) RGB (c) RGB + Depth

(d) Sat (e) RDS (f) RDS-150k

Figure 5. Test Image and the NeRF-rendered images generated from models trained on different combinations of data. Note that no masking is applied (d) for better visualization. For conciseness, we abbreviate 'RGB + Depth + Sat.' and 'RGB + Depth + Sat.-150k' in (e) and (f) to 'RDS' and 'RDS-150k,' respectively.

4.3. Results on Ground View

Fig. 5 (d) and Fig. 6 (a) show one of the ground truth images alongside the rendered images generated by each NeRF model. It can be observed that 'Sat.' fails to render the scene accurately from the ground-based pespective. This may be because the rays from the satellite images have small parallax, which makes the positions of ray intersections highly sensitive to small perturbations in ray directions and due to the small data coverage for the side of the building.

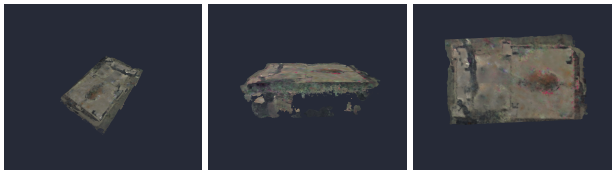
The other models render the scene mostly correctly. However, a discrepancy in camera pose between the ground truth and the rendered images is observed, particularly through the unexpected inclusion of the sky within the mask. This indicates that the ground truth camera pose, which was measured via GPS, is not entirely accurate. Therefore, it is important to note that this discrepancy may introduce additional error into the quantitative results.

Table 1 presents quantitative results for image quality and depth estimation. 'Sat.' is excluded from quantitative comparison due to its explicitly poor scene reconstruction performance. PSNR and SSIM are computed excluding the masked pixels, whereas LPIPS is computed including the masked pixels.

In terms of image quality, all the metrics for each method

Table 1. Mean and standard deviation of metrics used to evaluate image quality and depth. Standard deviations are presented in parentheses. Bold values indicate the best performance.

	PSNR ↑	SSIM ↑	LPIPS ↓	MAE ↓ (m)	RMSE ↓ (m)
RGB	20.16 (2.29)	0.804 (0.057)	0.123 (0.067)	2.268 (0.650)	2.980 (0.610)
RGB + Depth	**20.41 (2.28)**	**0.806 (0.061)**	**0.122 (0.068)**	**2.156 (0.632)**	**2.880 (0.578)**
RGB + Depth + Sat.	20.17 (2.14)	0.801 (0.064)	0.127 (0.068)	2.386 (0.608)	3.121 (0.604)
RGB + Depth + Sat.-150k	20.25 (2.23)	0.796 (0.066)	0.125 (0.067)	2.336 (0.611)	3.062 (0.607)



(a) Oblique view (b) Side view (c) Top view

Figure 6. NeRF-rendered images trained only with satellite imagery.

are very similar, although 'RGB + Depth' shows slightly better performance. The slight quality reductions observed in 'RGB + Depth + Sat.' and 'RGB + Depth + Sat.-150k' compared to 'RGB + Depth' can be attributed to biases in the available evaluation data. First, the evaluation metrics are computed only on side views, while the model is required to also reconstruct rooftop views. This division of the model's capacity between two tasks may reduce its performance on the evaluated task. Second, by including satellite images as part of the training data, the appearance embedding could enhance the NeRF model's generalizability to time-variant factors, such as multi-date imagery and varying lighting conditions. However, since all test images were ground images captured at the same time, this may create an unfair comparison for NeRF models trained with satellite imagery.

With respect to depth estimation, we compute the mean absolute error (MAE) and root mean square error (RMSE). The depth estimation performance of each method is similar to that of 'RGB' alone. The potential reasons for this are as follows: (1) The point clouds do not sufficiently cover the ROI, as shown in Fig 5 (a). (2) The pose discrepancy mentioned in the previous paragraph affects the accuracy of the estimation. Although extremely large depth errors were removed to account for this factor and other potential outliers, it is not guaranteed that they have been fully eliminated.

The experiments demonstrate that including satellite images along with other ground data in the training process does not significantly degrade the image rendering quality for ground views, though a finer characterization requires further studies with improved data sources.

4.4. Results on Rooftop View

Fig. 6 shows the rendered images of the building of interest for the 'Sat.' model. This figure demonstrates that our model training with satellite images alone is unable to represent side views of the building, but is able to capture its top views with some fidelity. Despite capturing the overall structure of the mosque, certain features, particularly the spires or minarets, were not well represented.

A comparison of side and top views for other variants of the method can be seen in Fig. 4. 'RGB + Depth + Sat.' and 'RGB + Depth + Sat.-150k' were able to improve the aforementioned issues with the satellite-only model. It is also notable that the addition of ground data led to rooftop rendering at a more accurate altitude, similarly to what was observed in ground views in Subsection 4.3

When viewed against ground-only models ('RGB' and 'RGB + Depth'), 'RGB + Depth + Sat.' and 'RGB + Depth + Sat.-150k' were able to more successfully render rooftop views, since the former lacked data from this perspective. Besides geometrical differences, blue artifacts from the sky and other high-frequency color artifacts were more present in the models trained only with ground data and can be attributed to the same cause.

In summary, this qualitative study showed that models combining ground and satellite images displayed improved results on different aspects than either ground-only or satellite-only models. When also considering the analysis from Subsection 4.3, we can conclude that these benefits in rooftop representation can be obtained at only a small degradation in quality for grounds views.

5. Conclusion

In this paper, we introduced CaLiSa-NeRF, a novel framework that integrates ground RGB images, LiDAR point clouds, and satellite imagery to represent urban environments. By combining these data types, CaLiSa-NeRF enables a more comprehensive omnidirectional scene representation that effectively both side and top views. Its results are on par or superior to variants restricted to either ground or satellite data, thus showcasing the benefits of the proposed approach. Particularly, improvements were more substantial on areas where a singular source of data lacked

visibility (e.g., ground-based sensors being unable to view rooftops) or had other inherent limitations (e.g., the parallax for satellite images).

CaLiSa-NeRF shows significant potential for applications where it's desired to generate 3D representations that are accurate from multiple perspectives for large urban scene in a scalable manner, such as in urban planning, virtual tourism, or autonomous navigation. In these settings, acquiring aerial data to complement the limitations of ground data can pose practical and financial barriers, so the usage of satellite imagery as an alternative is an enticing possibility. Future work may address some accuracy limitations of the dataset and expand results to city-level scene representations, as is done in Block-NeRF [16].

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, June 2022. 1, 3

[2] Alexandra Carlson, Manikandasriram S. Ramanagopal, Nathan Tseng, Matthew Johnson-Roberson, Ram Vasudevan, and Katherine A. Skinner. Cloner: Camera-lidar fusion for occupancy grid-aided neural representations. *IEEE Robotics and Automation Letters*, 8(5):2812–2819, 2023. 1, 3

[3] 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)*, pages 12882–12891, June 2022. 1, 2, 4

[4] Dawa Derksen and Dario Izzo. Shadow neural radiance fields for multi-view satellite photogrammetry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1152–1161, June 2021. 1, 3

[5] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt. MGM: A Significantly More Global Matching for Stereovision. In BMVA Press, editor, *BMVC 2015*, Swansea, United Kingdom, 2015. 1

[6] Shengyu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidler, Konrad Schindler, and Or Litany. Neural lidar fields for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18236–18246, October 2023. 3

[7] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Sat-nerf: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using rpc cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1311–1321, June 2022. 1, 3

[8] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Multi-date earth observation nerf: The detail is in the shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2034–2044, June 2023. 1, 3, 4

[9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7210–7219, June 2021. 1, 3

[10] Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, Jérémy Anger, and Gabriele Facciolo. A Generic Bundle Adjustment Methodology for Indirect RPC Model Refinement of Satellite Imagery. *Image Processing On Line*, 11:344–373, 2021. https://doi.org/10.5201/ipol.2021.352. 4

[11] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing. 1, 2

[12] 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, July 2022. 1

[13] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12932–12942, June 2022. 1, 3

[14] 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)*, pages 12892–12901, June 2022. 1, 2

[15] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 6

[16] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8248–8258, June 2022. 1, 3, 8

[17] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 2023. 4

[18] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12922–12931, June 2022. 1, 3

[19] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3

[20] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4150–4159, June 2023. 3

[21] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *European conference on computer vision*, pages 106–122. Springer, 2022. 3

[22] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5438–5448, June 2022. 2

[23] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields, 2020. 3