Aug3D: Augmenting large scale outdoor datasets for Generalizable Novel View Synthesis

Omar Alama *
Carnegie Mellon University
oalama@andrew.cmu.edu

Aditya Rauniyar *
Carnegie Mellon University
rauniyar@cmu.edu

Abstract

Recent works in novel view synthesis have seen tremendous progress and methods like radiance fields and Gaussian splatting have made good progress in the area. However, these methods are still limited when performing novel view synthesis on larger outdoor scenes. While works have attempted to address this using optimization-based NVS models, the area is still underexplored for generalizable feed-forward methods which have many advantages over their optimization-based counterparts. In this work, we explore training pixelNerf a feed-forward NVS model on the large-scale urbanscene3D dataset. We provide multiple training strategies to cluster the dataset and train, we show that performance still lacks due to the big translational component in the different views in the dataset. Through experiments, we find that limiting the number of images in a cluster from 20 views to 10 views improves the PSNR by 10% yet is still below par. To address that, we propose Aug3D. Aug3D reconstructs the scene using traditional SfM and samples novel views in a well-conditioned manner based on grid and semantic approaches to facilitate feed-forward NVS model learning. We show through experimentation that the aforementioned approach improves PSNR significantly and improves reconstruction quality. We further evaluate combining the newly generated novel views with the original dataset but show that the combination does not help the model predict novel views.

1. Introduction

Photorealistic Novel View Synthesis (NVS) is crucial for AR/VR, providing immersive experiences, and enabling deeper understanding of objects or scenes from limited 2D captures in 3D content creation. Densifying initial datasets using NVS methods enhances the perceptual capabilities of autonomous vehicles [26]. Its feasibility in large outdoor scenes is increasingly intriguing.

Generalizable models, exemplified by works like PixelNeRF [39] and Splatter-Image [25], render photorealistic novel views applicable to a wider range of inputs. However, these models are typically trained on smaller, object-centric scenes or indoor environments. In this work, we extend the application of NVS to large outdoor environments, aiming to broaden the scope of these methods for novel view synthesis in such settings.

Contrastingly, we take inspiration from scene-specific NeRF approaches in the research community, such as MegaNeRF [31] and VastGaussian [16], which fine-tune the NeRF model for NVS on specific scenes. These provide insights into selecting large outdoor scenes for training generalizable models to synthesize novel views.

Challenges: Utilizing large outdoor scenes for generalizing NVS models presents several hurdles. One challenge arises from how these scenes are typically captured using drones, often employing constant-altitude grid scans over regions of interest [22, 31]. This results in captures that vary predominantly in a translated direction, introducing novel features to the scene between consecutive shots, and posing difficulties for NVS methods to operate effectively. Additionally, most existing NVS works focus on object-centric scene captures, whether for objects or indoor/outdoor environments. Such captures are vital as the models rely on correlated features across input images to render novel views. Furthermore, generalizable NVS models typically train on datasets with minimal variation across input images (e.g., DTU dataset [12]), where input images are object-centrically placed and exhibit controlled changes in elevation and azimuth. As a result, novel views are interpolated rather than extrapolated. Therefore, large outdoor scene environments used for scene-specific NVS models must (1) align with existing generalizable NVS model training setups, introducing fewer new elements across input images, and (2) feature input images that are closely spaced with controlled variations in view poses (e.g., DTU [12], Shapenet [5] dataset).

Aug3D: Addressing the challenges, we introduce Aug3D 1, an augmentation camera sampling strategy to adapt large outdoor scene datasets such as UrbanScene3D

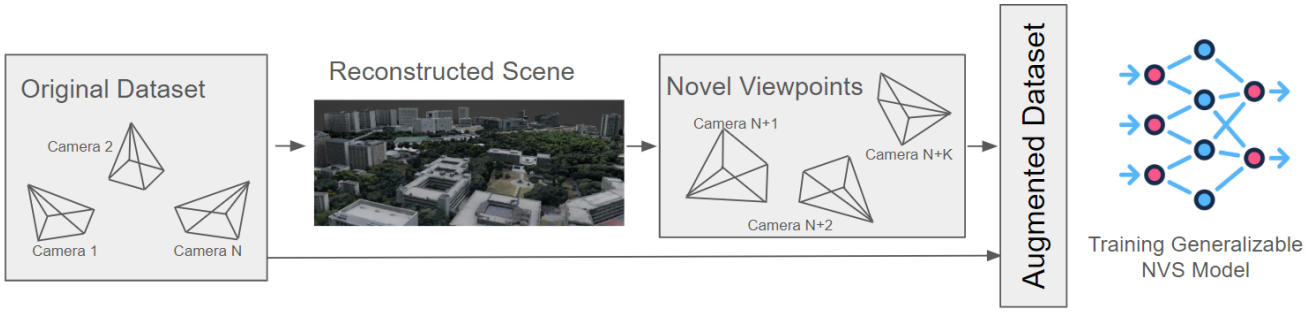

Figure 1. The outdoor scene dataset is used to reconstruct a scene and then view samples are extracted and mixed with the main dataset for training the NVS Model.

[17] and Mill-19 [31] for training generalizable novel view synthesis models. To mitigate sensitivity to input image poses, we cluster them into N views, maximizing shared points through Structure from Motion (SfM). However, sparse dataset collection via drone flight necessitates further measures to enhance feature correlation among input images. To accommodate poses beyond original locations and ensure scale invariance, we sample camera poses by geometric reconstruction of large scenes. While reconstruction quality impacts these views, advancements in photorealistic scene-specific NVS models like Mega-NeRF [31], Block-NeRF [26], and VastGaussians [16] suggest sufficient development within the research community for our proposed method.

Contributions: Our work addresses the question: "How can we effectively train existing Generalizable NVS models for large-scale outdoor datasets?" Here are our key contributions:

- We cluster outdoor datasets using high point matching, aligning them with the DTU format for compatibility with any NVS model designed for the DTU dataset. We validate this approach with PixelNeRF [39].
- Our multi-scaling camera sampling method generates additional viewpoints not present in the original dataset. These new viewpoints, derived from mesh-based scene capture, produce synthetic renders whose quality relies on reconstruction accuracy.
- We optimize the augmentation process with semantically aware sampling, enhancing the diversity of novel viewpoints added to the dataset. This pipeline combines geometric and feature-wise segmentation techniques.

2. Related Work

Novel View Synthesis. Novel view synthesis (NVS), tackles the challenge of synthesizing novel RGB views given a set of RGB input views (without necessarily constructing explicit 3D geometry). NVS has seen a rapid growth of interest with the advent of many recent breakthroughs in learning/neural based methods. These neural methods can be broadly classified into surface or volumetric based approaches [27]. Neural surface approaches reason about the surfaces in the scene either representing them implicitly with zero level set functions [13,42], with continuous parametric methods [3,30], or explicitly using meshes [23], or points/surfels [2,34]. Neural volumetric approaches reason about the volumes occupied by elements in the scene representing them implicitly [20,24], as a Neural Radiance Field (NeRF) [18], as a set of volumetric primitives [14], or explicitly as voxel grids [**?**, 11, 38] or multi-plane images [35].

In this work, we focus our evaluation on neural volumetric methods namely, the NeRF [18] lines of work due to their large success in high fidelity novel view synthesis and the existence of works attempting to extend such approaches to large-scale urban settings.

Generalizable NVS. Generalizable, image-based, or feed-forward NVS refers to models that can predict novel views at test time without having to re optimize any learnable parameters. This is done by conditioning the architecture on sets of input views describing different scenes while training. In contrast to the optimization-based single-scene networks, feed-forward models can learn semantic priors that make them superior in sparse input NVS.

Works like PixelNeRF [39] conditions NeRF on pixel aligned features recovered by projecting a query point onto feature maps of the input views. IBRNet [33] uses a similar approach but uses transformers. MVSNeRF [7] uses 3D convolutions on top of a plane sweep of input images to get per voxel image features and uses that to condition NeRF per query point. MuRF [37] constructs a frustum volume aligned with the target view allowing them to utilize 3D convolutions to predict the volume. Similar recent works [6, 9, 25] have worked on generalizing 3D Gaussian splatting through input image conditioning.

Nevertheless, all mentioned works focus on small to medium scale scenes with very limited target view ranges mainly due to the absence of city scale datasets amenable to feed-forward NVS. Our objective is to offer a training and data augmentation strategy to allow such works to learn large-scale urban scene priors efficiently.

Large Scale Scene Reconstruction: Large city-scale reconstruction has been a long-standing field of research. Many works attempt to reconstruct large scenes using traditional methods such as Lidar point clouds [15], meshes [32], or signed distance functions [21]. However, there is an increased interest in using neural volumetric representations for their high-fidelity reconstructions. [26, 31, 41] recognize NeRF's capacity limitations and propose forms of spatial decomposition and train many NeRF's to represent different parts of the large scene. Mega-NeRF [31] and BirdNeRF [41] focus on bird view reconstruction, while Block-NeRF [26] focuses on street view. BungeeNeRF [36] takes a different approach focusing on satellite view reconstruction, recognizes the need for multi-scale reconstruction, and progressively trains from big to small scales while increasing network capacity. Urban Radiance Fields [22] presents a multi-modal approach of combining lidar information with RGB signals to address exposure differences in outdoor scenes. VastGaussian [16] introduces spatial decomposition approaches to 3D Gaussian splatting for large-scale bird view scene reconstruction.

However, The aforementioned works develop optimization-based models that need extensive training time and are not suitable for online reconstruction during navigation or data acquisition. Whereas we explore the capabilities of feed-forward approaches to reconstruct large scale urban scenes allowing on the fly reconstruction times.

Augmentation for scene understanding: Data augmentation is a proven technique for improving ML model generalizability. Numerous augmentation methods have been developed in the 2D vision space. We take inspiration from CutOut [10] and CutMix [40] that cut 2D images out and mix cuts respectively. These methods however cannot be directly applied on input images for 3D NVS as they compromise cross-view consistency. Recently, 3D augmentation techniques started to be developed. Notably, Mix3D [19] mixes elements/meshes from different synthetic indoor scenes to compose new scenes that are not necessarily semanticaly reasonable to improve generalizability following the effective techniques of domain randomization [28, 29]. Their work however is done in a limited indoor setting for 3D semantic segmentation. There exists very few works [4, 8] that tackle augmentation for feed-forward NVS yet they only augment in 2D image space severely limiting the variations introduced.

3. Approach

3.1. Preparing Data for Generalizable NVS

Large scale urban scene data is not readily amenable for generalizable NVS as the data covers a huge baseline. For example, urbanscene3d [17] real datasets can cover more than $1km^2$ areas spanning multiple high rise and low rise buildings. Hence, an image in the scan may not necessarily contribute meaningfully to the reconstruction of another view. Thus, clustering images meaningfully is critical. We test different algorithms as shown in 2 to achieve that targeting the following criteria: First, it is pivotal to cluster images in the scan that are related to each other (ie. looking more or less at the same structures in the scene). Second, the selection of the group size is crucial as too small of a group size will give very little information to the model whilst a very big group size would give confusing and unrelated information to the model. Third, the group size should be constant to allow efficient batching when training. We show a qualitative output of clustering images in our appendix section at 8.

Capture Sequence grouping: Using the capture sequence to cluster images is the simplest approach we can use. However sudden turns in the capture sequence can leave images in the same group looking at very different areas of the scene. Furthermore, this method fails to include valuable images from later time steps that are looking at the same area.

Grid based grouping In another approach, we can super impose a grid on the ground plane clustering cameras that are the K nearest neighbors to the grid cell center. However this approach may cluster cameras that are close by in euclidean space yet very far apart in terms of their viewing frustums. We further experimented by adding an angle constraint to ensure that cameras are not only close but roughly looking at the same direction. Still, this approach fails to capture images looking at the same area from different angles.

Ray intersection with ground plane In an attempt to capture both euclidean distance and viewing distance, we tried unprojecting the center pixel of each image such that it intersects with the ground plane. We then use the distances between the intersection point as our clustering metric. A drawback of this approach is that you need to estimate the distance of the ground plane to each camera. To do this we use Metashape to run SfM on a small hand picked images and calculate the height of the cameras relative to the ground plane then use that to calculate all other camera distances to the ground plane with the assumption that the ground is flat. This approach improved clustering performance but still failed in many cases near high rise buildings as cameras could be looking at different areas even though their rays intersect close to each other at the ground plane

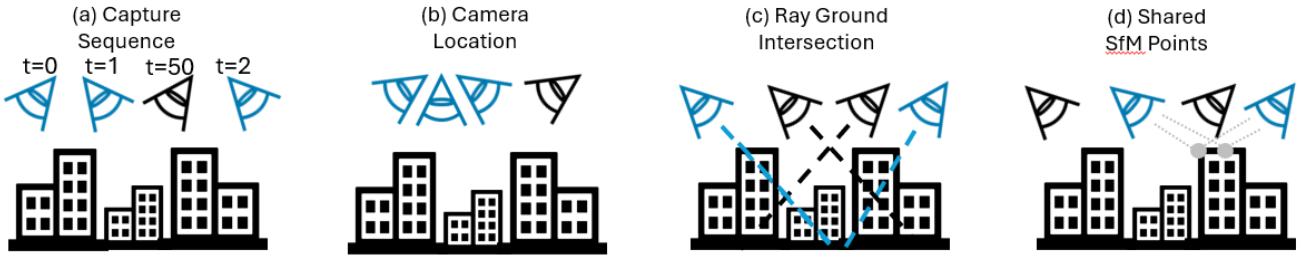

Figure 2. Approaches to cluster images from a capture. Colored cameras refer to cameras in the same cluster/group

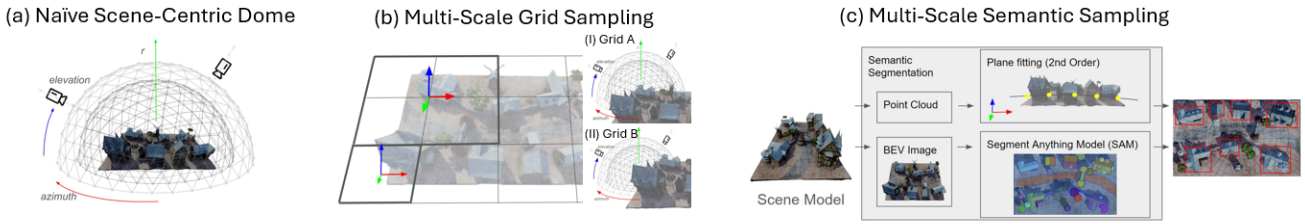

Figure 3. Different sampling strategies given a representation of the scene

level.

SfM shared points To enforce that images in the cluster are looking at the same structures, we resort to performing full SfM for each scene then using the number of shared points across different cameras as our distance metric for clustering. This approach yielded the best results as shown in 8 and avoided the edge cases of all of the aforementioned approaches. Nevertheless, the approach like all others, requires careful tuning of the cluster size.

3.2. Augmentation

Recognizing the difficulty of training feed-forward NVS models directly on the real data with unconstrained capture trajectories, we further propose to augment such scenes with constrained sampling methods by first reconstructing the scene using traditional structure from motion and multi-view stereo approaches, then sampling novel views in an object-centric manner to augment the training of the feed-forward model. We discuss various approaches to sampling in what follows.

3.2.1 Naive Scene Centric Dome

The naive approach given a reconstructed mesh of the scene, is to use an archimedeal spiral or a dome 3.a. like space on top of the whole mesh to sample images from the scene as is done traditionally in most feed-forward models [39]. However, since scenes can be arbitrarily large, this approach would push the model to only be able to reconstruct views from a large height where the scene is almost flat. The model may regress to simply learn a homography transformations relating the different views. In fact, we show in 5.2 that such a sampling approach can be trivially trained to achieve good results yet be unable to generalize to height varying scenes. Hence, we propose better sampling strategies that can cover the scene efficiently at varying scales.

3.2.2 Multiscale Grid Sampling

Another simple approach is to divide the scene into cells at different grid scales 3.b. We use multiple scales to avoid over-fitting the model on a single scale. We then place a virtual dome on top of each cell and sample cameras uniformly given a limited azimuth and elevation range. Furthermore, to avoid fine-tuning grid scales for each scene, we utilize the height to width/length ratios of the scene to dynamically choose the grid scales. This allows us to dynamically have finer grids for meshes representing large scenes and coarser grids for meshes representing small scenes as illustrated in 3.b..

3.2.3 Semantic Building Sampling

To avoid spending valuable training cycles uniformly across the scene as in multi-scale grid, we propose another approach to instead, detect buildings as regions of interest and sample cameras around those areas only 3.b.

Plane fitting To simplify the detection process, we detect buildings using a geometric approach fitting a second order plane using least squares over the top Kth percentile of

the points (Sorted by Z height) comprising the scene mesh. We then use that plane to cut out the mesh and render a top down orthographic view of the cut scene. This allows us to easily convert the render to binary masks to convert subsequently to bounding boxes. These bounding boxes are then used to initialize dome placement for camera sampling.

Furthermore, inline with our desire for multiscale NVS, we use a bounding box combination algorithm to incorporate multiple buildings in the same scene. For each detected bounding box, we choose 1 to M nearest boxes to combine with to generate a list of boxes encompassing single and multiple buildings.

Segment Anything Model We further experimented with using Segment anything model (SAM) to detect buildings from a top down view however we found that SAM was quite sensitive to shadows, and found that the geometric plane fitting approach was giving better results.

4. Experimental Setup

Dataset: For our experimental analysis, we utilize the comprehensive UrbanScene3D [17] dataset, which encompasses a variety of real (Campus, Hospital, Residence, SciArt, and Polytech) and virtual environments (School, Town, Bridge, and Castle). These scenes provide thousands of high-resolution images, capturing the intricate details and diverse structures. We also work on real scenes from the Mill-19 [31] dataset, which has two environments: building and rubble. We use all these datasets for dome-based object-centric sampling 3.2.1, and grid-based sampling 3.2.2 with its corresponding mesh reconstruction using Agisoft Metashape [1].

Metric: In evaluating our model, we will apply a combination of quantitative metrics and qualitative assessments to ensure a robust analysis. Quantitatively, we will utilize the Peak Signal-to-Noise Ratio (PSNR) to measure the fidelity of the reconstructions against the corrupting noise. Qualitatively, our approach will include visual inspections to assess the realistic rendering of the scenes. Together, these methods will provide a comprehensive evaluation of our model's performance from both statistical accuracy and user experience perspectives. Our baseline is training on UrbanScene3d's Campus environment.

Comparison: We first setup our baseline. Since there is no existing work that does G-NVS for the mentioned dataset. We train PixelNeRF [39] on real Campus environments from UrbanScene3D to fit 10 and 20-view clusters using SfM clustering using the SfM-based share points mentioned in 3.1. We run three experiments with 20-view clusters on Campus where the input images are 3, 6, and 9. This is done to analyze what cluster size is most suitable and the dependency of input images over the novel views. For the camera sampling approach, we conduct a dome 3.2.1 vs grid-based sampling 3.2.2 experiment. And for the final evaluation of augmentation we experiment with 9 input views on original 20-view-clustered Campus environment, 20-view-clustered Campus environment with grid-based sampling, and 20-view-clustered Campus environment with plane fitting 3.2.3.

Compute Setup: We run PixelNeRF [39] with 256 hidden layers with fixed encoder weights as to its default configuration to fit low computational requirements. We use 2 32Gb Tesla V100 GPUs to train and evaluate original campus, original campus plus plane experiment - training took nearly 23 hours each. We used 1 24Gb NVIDIA RTX3090Ti for dome-centric training which took early 60 hours, and the original campus plus grid for 9 input views which took nearly 16 hours. All other mentioned experiments use 10Gb NVIDIA RTX3080.

5. Results and Discussions

Our experimental analysis focuses on evaluating the performance of our approach on both real and synthetic datasets, as well as augmented datasets. In the real dataset experiments, we explore the impact of varying the number of input views (3, 6, and 9) and cluster sizes (10 and 20) on reconstruction quality. Additionally, we investigate the effectiveness of different camera sampling approaches, including dome-based and grid-based sampling. Furthermore, we assess the performance of our model on augmented datasets, specifically examining the contributions of semantic-aware augmentation and grid-based augmentation. The results presented in this section provide insights into the effectiveness of our approach across different dataset configurations and augmentation strategies.

5.1. Real dataset

Here 4.a., as we vary the number of input images for the campus real dataset clustered with 20 views, we observe that the best PSNR values for input views 3, 6, and 9, after 35k iterations, are 20.03, 19.95, and 19.59, respectively. Surprisingly, this suggests that PixelNeRF performs better with fewer input views, contrary to the typical behavior expected from GNVS models, where increasing the number of input images usually leads to better RGB predictions for novel viewpoints. Our analysis reveals that adding more input images introduces significant variance and new features among them, posing challenges for correlation. This key finding underscores PixelNeRF's sensitivity to novel features and high pose fluctuations among input images.

This leads to another experiment 4.b. where we conduct experiment to check the size of cluster we form for training PixelNeRF. As we form the clusters based SfM shared points, and find the N images closest to each other that shares the most tie points. As we set a high N-view clustering it would contain images that are farther and make the cluster contain views that share lesser tie points. Best PSNR

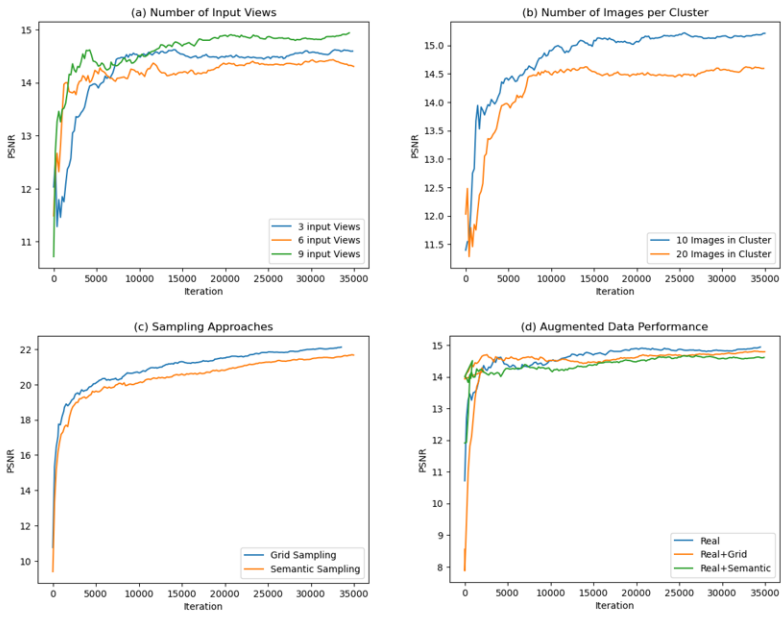

Figure 4. (a) Varying the number of input views on campus real data with 20 images per cluster. (b) Varying the number of

for 10 images per cluster: 22.94. Best PSNR for 20 images per cluster: 20.03.

5.2. Synthetic dataset

Recognizing the challenges of training directly on the real dataset, we explore training on data generated by reconstructing the real scene and sampling well conditioned images for pixel nerf training. In 4.c, we show the results of training using data acquired through the multiscale grid approach vs the semantic plane fitting approach. The graph shows that grid sampling achieves better PSNR on its validation data (Acquired through the same sampling method). Best PSNR for Grid: 29.12. Best PSNR for Semantic: 28.79. For a fairer comparison we further test on the residence scene with a manual sampling approach where we manually draw bounding boxes of buildings in the scene. Here Semantic achieves a best PSNR of 21.8 while grid achieves 21.67. The difference is minute and we conclude that both sampling approaches work reasonablly well. Further qualitative evaluation is available in the appendix.

5.3. Augmented dataset

In this section, we focus on the augmented dataset, where we combine the real dataset with additional semantic information. Specifically, we experiment with two augmentation strategies: (1) Real + Semantic and (2) Real + Grid. We evaluate the performance of PixelNeRF on these augmented datasets and compare the results with those obtained from the original datasets. Our goal is to assess how the inclusion of semantic information or grid-based sampling influences the model's ability to synthesize novel views. Results are shown in 4.d. which show that contrary to our initial hypothesis augmenting through mixing synthetic and real did not improve performance.

6. Conclusion

To conclude, in this work we explore the training of generalizable NVS for large scale urban scenes. We show and explore the challenges of training directly on data acquired. We show that training on object centric, even if reconstructed through SfM, whether grid or semantically sampled improves performance dramatically. We explore the direction of adding the sampled images from the reconstruction to the real data and its affect on performance. We have learned a lot through this project and we believe that augmentation in the generalizable NVS space is an exciting area that is still underexplored, and we are excited to continue in this direction.

Acknowledgment

The authors would like to thank Silong Yong, Yaqi Xie, Simon Stepputtis, Sebastian Scherer, Katia Sycara for their assistance, advice, and feedback. Shubham Tulsiani, and the instructors of 16-825: Learning for 3D Vision (Spring 2024) for their invaluable guidance and support.

References

[1] Agisoft LLC: Agisoft Metashape (2021), https://

www.agisoft.com/ 5

[2] Aliev, K.A., Sevastopolsky, A., Kolos, M., Ulyanov, D., Lempitsky, V.: Neural point-based graphics. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 696–712. Springer (2020) 2

[3] Bhattad, A., Dundar, A., Liu, G., Tao, A., Catanzaro, B.: View generalization for single image textured 3d models. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6077–6086 (2021), https://api.semanticscholar.org/CorpusID:235417325 2

[4] Bortolon, M., Del Bue, A., Poiesi, F.: Vm-nerf: tackling sparsity in nerf with view morphing. In: International Conference on Image Analysis and Processing. pp. 63–74. Springer (2023) 3

[5] Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository 1

[6] Charatan, D., Li, S., Tagliasacchi, A., Sitzmann, V.: pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction (Dec 2023), http://arxiv.org/abs/2312.12337, arXiv:2312.12337 [cs] 2

[7] Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14104–14113. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01386, https://ieeexplore.ieee.org/document/9711430/ 2

[8] Chen, T., Wang, P., Fan, Z., Wang, Z.: Aug-nerf: Training stronger neural radiance fields with triple-level physically-grounded augmentations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15191–15202 (2022) 3

[9] Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. arXiv preprint arXiv:2403.14627 (2024) 2

[10] DeVries, T., Taylor, G.W.: Improved Regularization of Convolutional Neural Networks with Cutout (Nov 2017), http://arxiv.org/abs/1708.04552, arXiv:1708.04552 [cs] 3

[11] Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5501–5510 (2022) 2

[12] Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanaes, H.: Large Scale Multi-view Stereopsis Evaluation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 406–413. IEEE, Columbus, OH, USA (Jun 2014). https://doi.org/10.1109/CVPR.2014.59, https://ieeexplore.ieee.org/document/6909453 1

[13] Kellnhofer, P., Jebe, L., Jones, A., Spicer, R.P., Pulli, K., Wetzstein, G.: Neural lumigraph rendering. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4285–4295 (2021), https://api.semanticscholar.org/CorpusID:232307471 2

[14] Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics **42**(4), 1–14 (Aug 2023). https://doi.org/10.1145/3592433, https://dl.acm.org/doi/10.1145/3592433 2

[15] Lan, Z., Yew, Z.J., Lee, G.H.: Robust Point Cloud Based Reconstruction of Large-Scale Outdoor Scenes. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9682–9690. IEEE, Long Beach, CA, USA (Jun 2019). https://doi.org/10.1109/CVPR.2019.00992, https://ieeexplore.ieee.org/document/8953959/ 3

[16] Lin, J., Li, Z., Tang, X., Liu, J., Liu, S., Liu, J., Lu, Y., Wu, X., Xu, S., Yan, Y., Yang, W.: VastGaussian: Vast 3D Gaussians for Large Scene Reconstruction 1, 2, 3

[17] Lin, L., Liu, Y., Hu, Y., Yan, X., Xie, K., Huang, H.: Capturing, Reconstructing, and Simulating: the UrbanScene3D Dataset (Jul 2022), http://arxiv.org/abs/2107.04286, arXiv:2107.04286 [cs] 2, 3, 5

[18] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis 2

[19] Nekrasov, A., Schult, J., Litany, O., Leibe, B., Engelmann, F.: Mix3D: Out-of-Context Data Augmentation for 3D Scenes. In: 2021 International Conference on 3D Vision (3DV). pp. 116–125. IEEE, London, United Kingdom (Dec 2021). https://doi.org/10.1109/3DV53792.2021.00022, https://ieeexplore.ieee.org/document/9665916/ 3

[20] Niemeyer, M., Mescheder, L.M., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3501–3512 (2019), https://api.semanticscholar.org/CorpusID:209376368 2

[21] Oleynikova, H., Millane, A., Taylor, Z., Galceran, E., Nieto, J., Siegwart, R.: Signed Distance Fields: A Natural Representation for Both Mapping and Planning p. 6 p. (2016). https://doi.org/10.3929/ETHZ-A-010820134, http://hdl.handle.net/20.500.11850/128029, artwork Size: 6 p. Medium: application/pdf Publisher: [object Object] 3

[22] Rematas, K., Liu, A., Srinivasan, P., Barron, J., Tagliasacchi, A., Funkhouser, T., Ferrari, V.: Urban Radiance Fields. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12922–12932. IEEE, New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01259, https://ieeexplore.ieee.org/document/9879805/ 1, 3

[23] Riegler, G., Koltun, V.: Free view synthesis. In: European Conference on Computer Vision (2020), https://api.semanticscholar.org/CorpusID:221112229 2

[24] Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhöfer, M.: Deepvoxels: Learning persistent 3d feature embeddings. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2432–2441 (2018), https://api.semanticscholar.org/CorpusID:54444417 2

[25] Szymanowicz, S., Rupprecht, C., Vedaldi, A.: Splatter Image: Ultra-Fast Single-View 3D Reconstruction (Dec 2023), http://arxiv.org/abs/2312.13150, arXiv:2312.13150 [cs] 1, 2

[26] Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-NeRF: Scalable Large Scene Neural View Synthesis (Feb 2022), http://arxiv.org/abs/2202.05263, arXiv:2202.05263 [cs] 1, 2, 3

[27] Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Xu, Z., Simon, T., Nießner, M., Tretschk, E., Liu, L., Mildenhall, B., Srinivasan, P., Pandey, R., Orts-Escolano, S., Fanello, S., Guo, M.G., Wetzstein, G., y Zhu, J., Theobalt, C., Agrawala, M., Goldman, D.B., Zollhöfer, M.: Advances in neural rendering. Computer Graphics Forum **41** (2021), https://api.semanticscholar.org/CorpusID:236162433 2

[28] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 23–30. IEEE (2017) 3

[29] Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S.: Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 969–977 (2018) 3

[30] Tulsiani, S., Kulkarni, N., Gupta, A.K.: Implicit mesh reconstruction from unannotated image collections. ArXiv **abs/2007.08504** (2020), https://api.semanticscholar.org/CorpusID:220546413 2

[31] Turki, H., Ramanan, D., Satyanarayanan, M.: Mega-NeRF: Scalable Construction of Large-Scale NeRFs for Virtual Fly- Throughs. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12912–12921. IEEE, New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01258, https://ieeexplore.ieee.org/document/9878491/ 1, 2, 3, 5

[32] Valentin, J.P., Sengupta, S., Warrell, J., Shahrokni, A., Torr, P.H.: Mesh Based Semantic Modelling for Indoor and Outdoor Scenes. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2067–2074. IEEE, Portland, OR, USA (Jun 2013). https://doi.org/10.1109/CVPR.2013.269, http://ieeexplore.ieee.org/document/6619113/ 3

[33] Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R.,

Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021) 2

[34] Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 7465–7475 (2019), https://api.semanticscholar.org/CorpusID:209405397 2

[35] Wizadwongsa, S., Phongthawee, P., Yenphraphai, J., Suwajanakorn, S.: Nex: Real-time view synthesis with neural basis expansion. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8530–8539 (2021), https://api.semanticscholar.org/CorpusID:232168851 2

[36] Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D.: Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In: European conference on computer vision. pp. 106–122. Springer (2022) 3

[37] Xu, H., Chen, A., Chen, Y., Sakaridis, C., Zhang, Y., Pollefeys, M., Geiger, A., Yu, F.: Murf: Multi-baseline radiance fields. arXiv preprint arXiv:2312.04565 (2023) 2

[38] Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 5732–5741 (2021), https://api.semanticscholar.org/CorpusID:232352425 2

[39] Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural Radiance Fields From One or Few Images 1, 2, 4, 5

[40] Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6022–6031. IEEE, Seoul, Korea (South) (Oct 2019). https://doi.org/10.1109/ICCV.2019.00612, https://ieeexplore.ieee.org/document/9008296/ 3

[41] Zhang, H., Xue, Y., Liao, M., Lao, Y.: Birdnerf: Fast neural reconstruction of large-scale scenes from aerial imagery. arXiv preprint arXiv:2402.04554 (2024) 3

[42] Zhang, J., Yang, G., Tulsiani, S., Ramanan, D.: Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. Advances in Neural Information Processing Systems **34**, 29835–29847 (2021) 2

A.

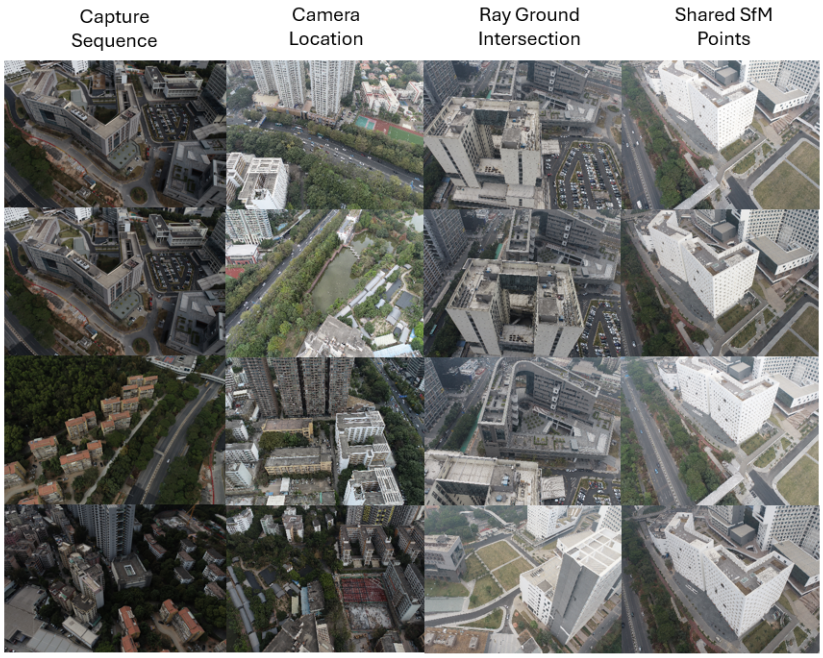

Figure 5. Results of different clustering algorithms. Images in a column belong to the same cluster

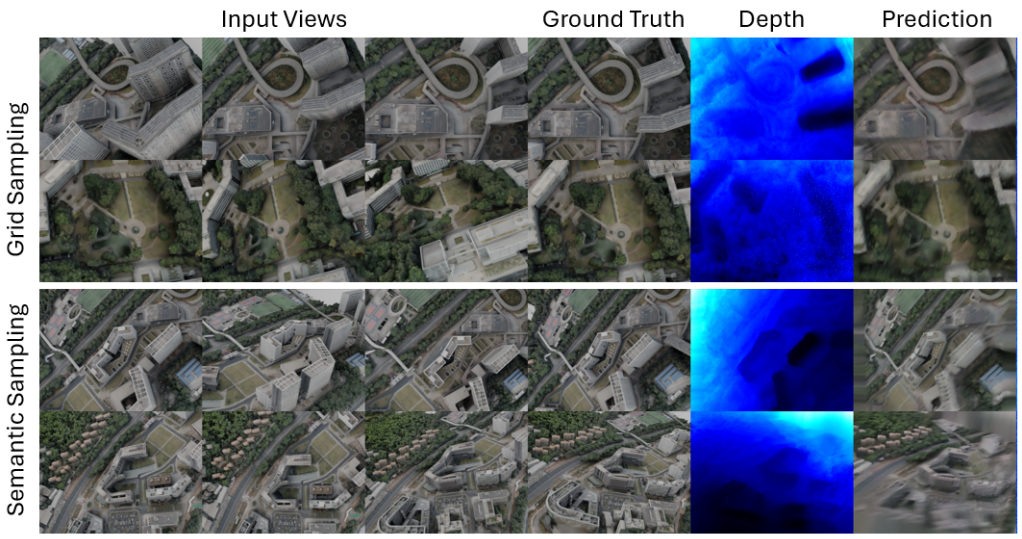

Figure 6. Qualitative results of training only on campus grid sampled vs campus semantically sampled data

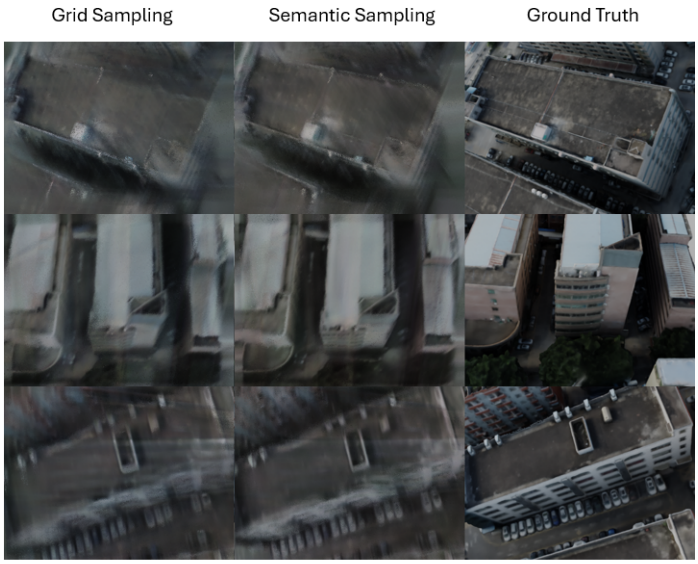

Figure 7. Qualitative results on the residence scene of training only on campus grid sampled vs campus semantically sampled data

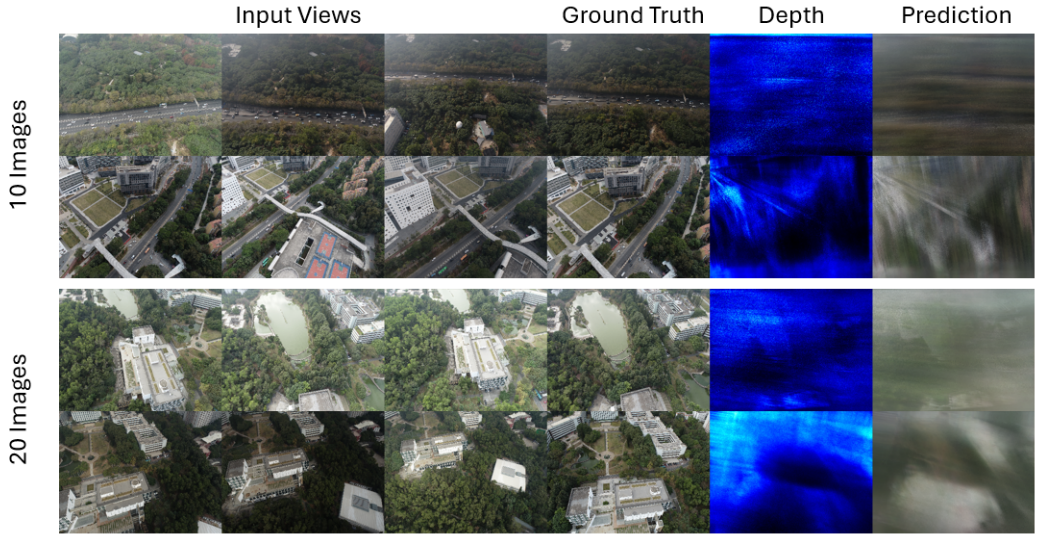

Figure 8. Qualitative results on the campus scene of models trained using 20 images per cluster vs 10 images per cluster using the SfM shared points method