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

Aditya Rauniyar*[1], Omar Alama*[1], Silong Yong[1], Katia Sycara[1], and Sebastian Scherer[1]

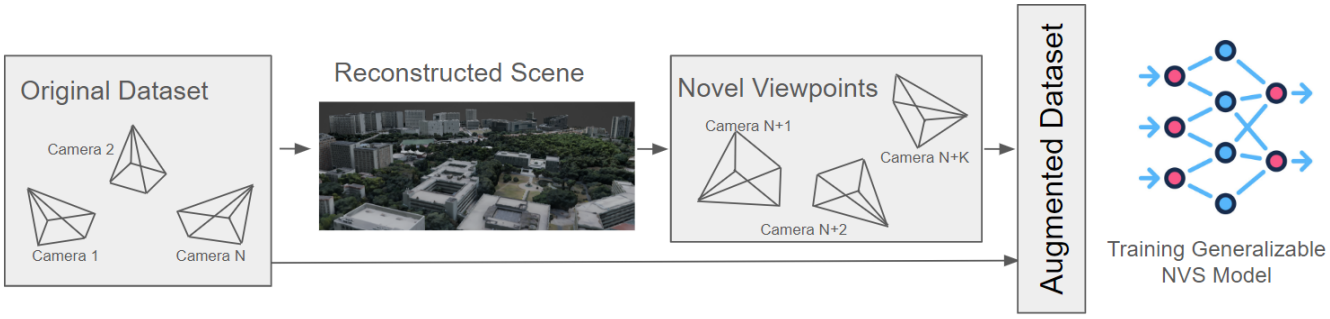


Fig. 1: Aug3D addresses challenges with low-overlap clusters in large-scale outdoor datasets for generalizable novel view synthesis by reconstructing scenes, sampling camera poses to mitigate overlap issues, and combining these samples with the real dataset, resulting in improved performance.

***Abstract*— Recent photorealistic Novel View Synthesis (NVS) advances have increasingly gained attention. However, these approaches remain constrained to small indoor scenes. While optimization-based NVS models have made attempts to address this, generalizable feed-forward methods—offering significant advantages—remain underexplored. In this work, we train PixelNeRF, a feed-forward NVS model, on the large-scale UrbanScene3D dataset. We propose four training strategies to cluster and train on this dataset, highlighting that performance is hindered by limited view overlap. To address this, we introduce Aug3D, an augmentation technique that leverages reconstructed scenes using traditional Structure-from-Motion (SfM). Aug3D generates well-conditioned novel views through grid and semantic sampling to enhance feed-forward NVS model learning. Our experiments reveal that reducing the number of views per cluster from 20 to 10 improves PSNR by 10%, but the performance remains suboptimal. Aug3D further addresses this by combining the newly generated novel views with the original dataset, demonstrating its effectiveness in improving the model's ability to predict novel views. https://aug3Dim.github.io**

I. Introduction

Photorealistic Novel View Synthesis (NVS) plays a vital role in applications requiring immersive experiences, such as AR/VR. As these methods gain popularity, there is a growing need to extend their capabilities to outdoor environments. In this study, we introduce Aug3D, a reconstruction-based augmentation technique designed to adapt existing outdoor datasets for NVS applications.

Generalizable models, exemplified by works like PixelNeRF [38] and Splatter-Image [24], render photorealistic novel views applicable to a wider range of inputs. 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.

Alternatively, we take inspiration from scene-specific NeRF approaches in the research community, such as MegaNeRF [30] and VastGaussian [15], 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. The first arises from how these scenes are typically captured using drones, often employing constant-altitude grid scans over regions of interest [21], [30]. 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 work focuses on object-centric scene captures, 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 [11]), where input images are placed in an object-centric way 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

*The authors share equal contribution.

[1]A. Rauniyar, O. Alama, S. Yong, K.Sycara and S. Scherer are with the Robotics Institute, School of Computer Science at Carnegie Mellon University, Pittsburgh, PA, USA. {arauniya, oalama, silongy, sycara, basti}@andrew.cmu.edu

This work is supported by the Defense Science and Technology Agency Singapore contract DST000EC124000205.

new elements across input images, and (2) feature input images that are closely spaced with controlled variations in view poses (e.g., DTU [11], Shapenet [4] dataset).

Aug3D: To address the challenges, we introduce Aug3D 1, an augmentation camera sampling strategy to adapt large outdoor scene datasets such as UrbanScene3D [16] and Mill-19 [30] 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 data collection via drone flight requires 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, advances in photorealistic scene-specific NVS models such as Mega-NeRF [30], Block-NeRF [25], and VastGaussians [15] 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 [38].
- 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.

II. Related Work

Novel View Synthesis. Novel view synthesis (NVS), tackles the challenge of synthesizing novel RGB views with a set of RGB input views (without necessarily constructing explicit 3D geometry). NVS has seen a rapid growth of interest with the advent recent breakthroughs in learning/neural based methods. These neural methods can be broadly classified into surface or volumetric based approaches[26]. Neural surface approaches reason about the surfaces in the scene either representing them implicitly with zero level set functions [41], [12], with continuous parametric methods [29], [2], or explicitly using meshes [22], or points/surfels [33], [1]. Neural volumetric approaches reason about the volumes occupied by elements in the scene that represent them implicitly [19], [23], as a Neural Radiance Field (NeRF) [17], as a set of volumetric primitives [13], or explicitly as voxel grids [37], [10] or multiplane images [34].

In this work, we focus our evaluation on neural volumetric methods specifically, the NeRF [17] lines of work due to its significant success in high fidelity novel view synthesis and the existence of efforts 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 and 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 [38] conditions NeRF on pixel aligned features recovered by projecting a query point onto feature maps of the input views. IBRNet [32] uses a similar approach but uses transformers. MVSNeRF[6] 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[36] constructs a frustum volume aligned with the target view allowing them to utilize 3D convolutions to predict the volume. Similar recent works [24], [5], [8] have worked on generalizing 3D Gaussian splatting through input image conditioning.

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 [14], meshes [31], or signed distance functions [20]. However, there is an increased interest in using neural volumetric representations for their high-fidelity reconstructions. [30], [25], [40] 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[30] and BirdNeRF[40] focus on bird view reconstruction, while Block-NeRF[25] focuses on street view. BungeeNeRF [35] 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 [21] presents a multi-modal approach of combining lidar information with RGB signals to address exposure differences in outdoor scenes. VastGaussian [15] 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 and are unsuitable for online reconstruction during navigation or data acquisition. 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 [9] and CutMix [39] 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 have been developed. Notably, Mix3D [18] mixes elements/meshes from different synthetic indoor scenes to compose new scenes that are not necessarily semantically reasonable to improve generalizability following the effective techniques of domain randomization[27], [28]. Their work however is done in a limited indoor setting for 3D semantic segmentation. There exists very few works [3], [7] that tackle augmentation for feed-forward NVS, they only augment in 2D image space, severely limiting the variations introduced.

III. Approach

A. Data curation for Generalizable NVS

Large-scale urban scene data are not readily amenable for generalizable NVS as the data covers a huge baseline. For example, urbanscene3d [16] 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; clustering images meaningfully is critical. We test different algorithms as shown in Fig. 2 to achieve that targeting the following criteria: First, it is pivotal to cluster images in the scan that are related to each other (i.e. 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 Fig. 4.

1) Capture Sequence grouping: Using the capture sequence—defined as the order in which images are captured along the camera's trajectory over time—to cluster images is a straightforward but naive approach. Abrupt changes in the camera's trajectory can result in images within the same cluster capturing entirely different parts of the scene, as illustrated in Fig.2.a. Additionally, this method overlooks valuable images from later in the sequence that capture the same scene region but are excluded due to their temporal position.

2) Grid-Based Grouping: In this approach, a grid is overlaid on the ground plane, and cameras are clustered based on proximity to the centers of grid cells, with each cluster containing the K nearest neighbors to a grid cell center. While straightforward, this method has limitations: cameras that are close in Euclidean space may have vastly different viewing frustums, leading to poor clustering results, as shown in Fig.2.b. To address this, we added an angular constraint to ensure that cameras within a cluster are not only spatially close but also oriented in roughly the same direction. Despite this refinement, the approach still struggles to group images that capture the same scene area from different angles.

3) Ray intersection with ground plane: To capture both the Euclidean distance and the viewing distance, we projected the center pixel of each image into the world frame so that it intersects with the ground plane. We then use the distances between the intersection points as our clustering metric. A drawback of this approach is that you need to estimate the distance from the ground plane to each camera. To achieve this, we use Metashape to run SfM on small hand-picked images and calculate the height of the cameras relative to the ground plane. We then use this height to calculate all other camera distances to the ground plane, assuming 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 level.

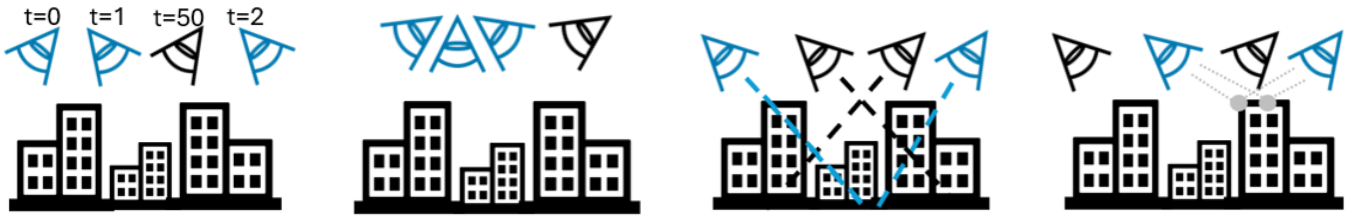
4) SfM shared points: To ensure that images within a cluster view the same structures, we perform a full Structure-from-Motion (SfM) process for each scene and use the number of shared points among different camera views as the metric for clustering. This approach consistently produced the best results, as illustrated in Fig. 4, while effectively avoiding edge cases seen in previous methods. The core idea is that cameras observing the same scene exhibit high correspondence, which we capture by computing a similarity matrix for all images in the scene using SfM. Based on this matrix, we uniformly select cluster centers across the scene and determine the top K views for each cluster according to their similarity scores. Like all clustering methods, this approach requires careful tuning of cluster size to achieve optimal performance.

B. Augmentation

Recognizing the challenges 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. First we reconstruct 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 below.

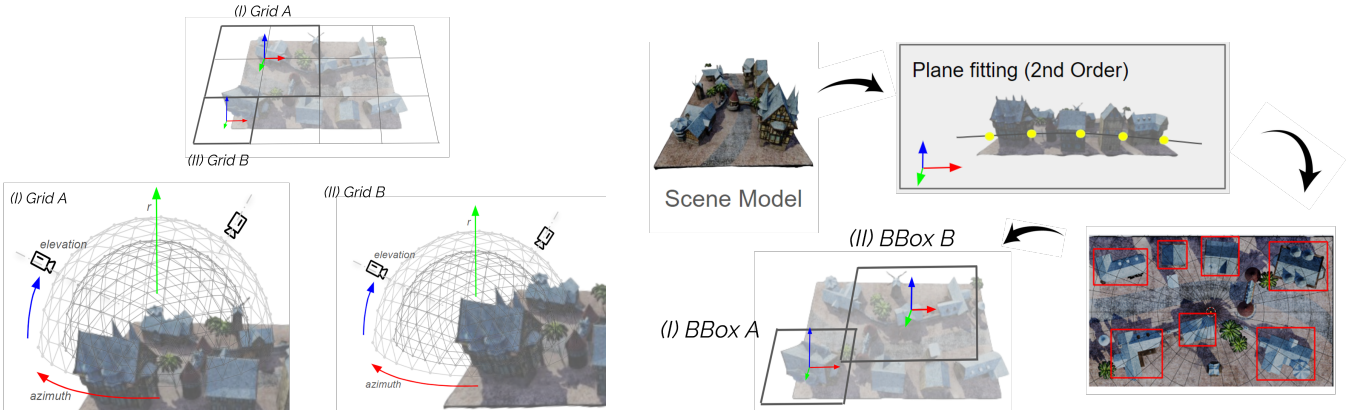
Background on scene sentric dome sampling: A common approach for sampling images from a reconstructed scene uses an Archimedean spiral or dome above the mesh, as shown in Fig.3, commonly applied in models like PixelNeRF [38]. While effective for standard setups, it struggles with large scenes, often resulting in flat, disproportionate reconstructions and reliance on simple homography transformations. To address this, we propose two improved camera sampling strategies for larger scenes.

1) Multiscale Grid Sampling: A straightforward approach involves dividing the scene into cells at varying grid scales, as shown in Fig.3a. Using multiple scales helps prevent the model from overfitting to a single scale. Virtual domes are then placed over each cell, and cameras are uniformly sampled within a limited azimuth and elevation range. To



(a) Capture sequence grouping (b) Grid-based grouping (c) Ray intersection grouping (d) SfM shared points

Fig. 2: Scene clustering methods for training GNVS models. Colored cameras represent cameras within the same cluster. (a), (b) and (c) show edge cases where these methods would cluster wrong images into a scene.



(a) Multiscale Grid Sampling (b) Semantic Sampling

Fig. 3: Two types of augmentation to reduce low overlap among outdoor scene datasets: (a) Multiscale Grid Sampling and (b) Semantic Sampling. The left figure shows dynamic camera placements for varying grid scales, and the right figure illustrates focused sampling around urban regions.

avoid manually fine-tuning grid scales for each scene, we dynamically adjust them based on the scene's height-to-width/length ratios. This ensures finer grids for large scenes and coarser grids for smaller ones, as illustrated in Fig.3a.

2) Semantic Building Sampling: This approach focuses on underrepresented areas, such as urban regions, which are often overshadowed by forest-dominated samples. Unlike the multiscale grid method that uniformly samples the scene, this method uses semantic camera sampling to identify urban areas as regions of interest and concentrates camera samples around them. As illustrated in Fig.3b, this strategy reduces forest overrepresentation and improves dataset diversity by prioritizing urban scenes.

Plane fitting: We simplify perform building detection using a geometric approach: fitting a plane to the Kth percentile of points (sorted by Z height) in the scene point cloud via least-squares. This plane slices the point cloud, rendering a top-down orthographic view, which is converted into binary masks and then bounding boxes. These bounding boxes initialize dome placements for targeted camera sampling.

To enable multiscale novel view synthesis (NVS), we extend this by combining bounding boxes. For each detected box, we merge it with 1 to M nearest boxes, creating clusters that represent individual buildings and multi-building regions, ensuring comprehensive and scalable scene coverage as showing in Fig.3b.

IV. Experimental Setup

Dataset: For our experimental analysis, we focus exclusively on the Campus scene from the UrbanScene3D [16] dataset. This scene spans an area of $1.3 \times 10^6 m^2$ and includes 178 objects, providing diverse urban structures for evaluation.

Metric: In evaluating our model, we will apply a combination of quantitative metrics and qualitative assessments. Quantitatively, we will utilize the Peak Signal-to-Noise Ratio (PSNR) to measure the fidelity of the reconstructions against the corrupting noise. Our approach will include visual inspections to assess the realistic rendering of the scenes.

Comparison: Our analysis involves comparing the performance of PixelNeRF on the real dataset with its performance on an augmented dataset that combines real and synthetic data. To achieve this, we first evaluate PixelNeRF's performance on the real dataset alone, ensuring that the data is curated effectively. We identify the best-performing approach using the proposed four preprocessing methods described in Section III-A. Once this baseline is established, we integrate augmentations generated through Aug3D, employing the two augmentation strategies detailed in Section III-B, and compare the results to assess the impact of augmentation.

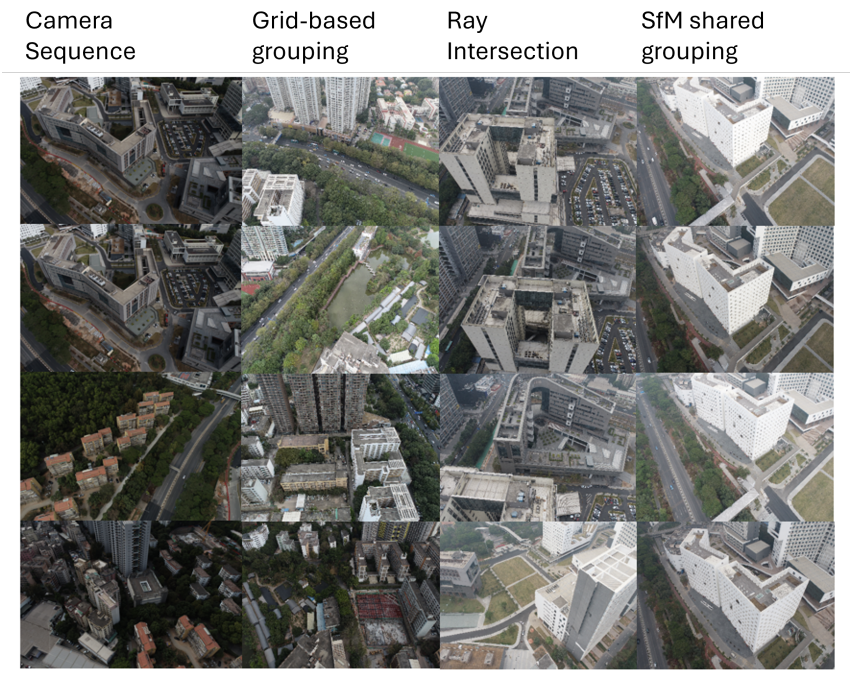


Fig. 4: Qualitative comparison of clustering methods for aerial image grouping. Each column represents a method: (a) Camera Sequence groups images with overlap in scenes 1 and 2, but misses 3 and 4. (b) Grid-Based grouping overlaps scenes 1 and 3, missing others. (c) Ray Intersection captures overlap in scenes 1, 2, and partly 3, but not 4. (d) SfM Shared grouping achieves high overlap across all scenes, demonstrating superior performance.

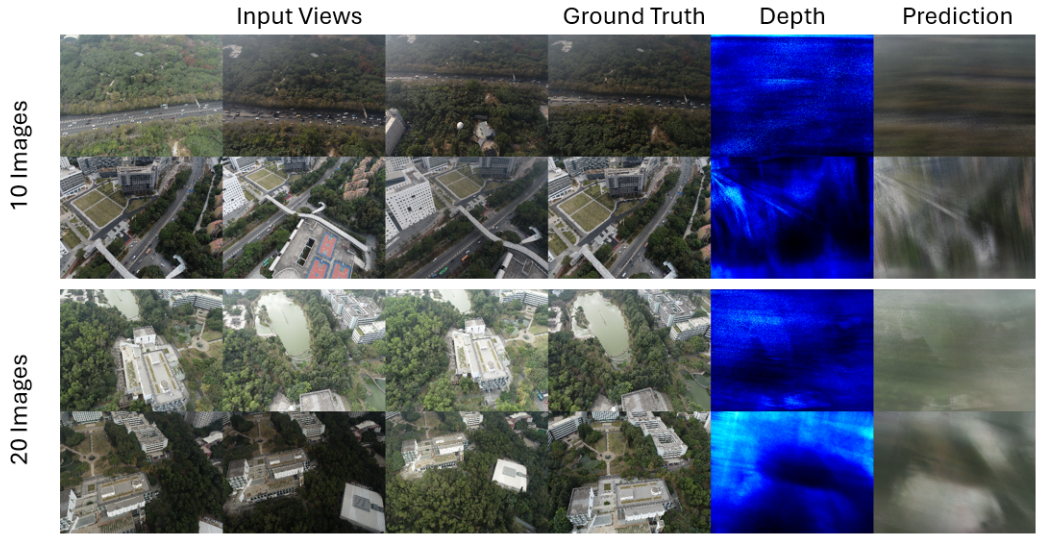


Fig. 5: Qualitative comparison of models trained with 10 images per cluster versus 20 images per cluster using the SfM shared points method on a campus scene. Reducing the cluster size from 20 to 10 demonstrates marginally improved visual quality. The first row represents fine-grained predictions, while the second row shows coarse-grained predictions. Columns depict Input Views, Ground Truth, Depth, and Predictions.

Compute Setup: PixelNeRF [38] is run with 256 hidden layers and fixed encoder weights, adhering to its default configuration to meet low computational requirements. We utilize two 32GB Tesla V100 GPUs to evaluate the real Campus dataset. Grid-based augmentation, combined with the real dataset, is processed on a single 24GB NVIDIA RTX 3090 Ti. All other experiments are conducted using a 10GB NVIDIA RTX 3080, ensuring consistency across setups where applicable.

V. Results

Evaluating Data Curation Methods: Experiments on the Campus scene from UrbanScene3D [21] demonstrate that *SfM shared grouping* out of the methods mentioned in Section III-A achieves the best performance for Generalizable Novel View Synthesis using PixelNeRF [38]. Using input images set to 3, a cluster size of 20, and Peak Signal-to-Noise Ratio (PSNR, higher the better) as the evaluation metric, *SfM shared grouping* attains the highest PSNR of 20.03 and an average PSNR of 14.6, outperforming *Camera sequence grouping* and *Grid-based grouping*, with PSNR values of 9.7 and 12.2, respectively as shown in Table I. Qualitative

TABLE I: Performance of Different Clustering Methods

Method	Best PSNR↑	Low PSNR↑	Avg. PSNR↑
Sequence grouping	9.7	0.0	3.5
Grid-Based grouping	12.2	0.0	4.6
Ray intersection	13.6	0.0	9.9
SfM shared grouping	**20.03**	**10.9**	**14.6**

results in Fig. 4 confirm that *SfM shared grouping* provides better visual correspondence and hence leads to stable training performance. Reducing the cluster size from 20 to 10 further improves PSNR to 22.94, additionally highlighting the importance of high overlap within input clusters for reconstruction fidelity, also shown with qualitative results in Fig. 5.

Baseline Performance: Table II details the results for the real and augmented datasets. For the real dataset, using a cluster size of 20 images, we observe a slight decline in PSNR as the number of input views increases. Specifically, the PSNR decreases from 20.03 for 3 input views to 19.59 for 9 input views. This trend suggests that while additional views provide more information, they may also introduce noise or redundancy that hinders GNVS performance.

Aug3D + Real vs Real dataset: The synthetic dataset, reconstructed using *Grid Sampling* and *Semantic Plane Fitting*, achieves PSNR values of 29.12 and 28.79, respectively, with 3 input images, a cluster size of 20. Augmenting the real dataset with these under the same parameters yields the best PSNR of 21.80 for the *Semantic* approach, slightly surpassing *Grid Sampling* at 21.67. These results validate the effectiveness of the Aug3D dataset in enhancing GNVS performance.

VI. Discussion

This work demonstrates the potential of feed-forward Generalizable Novel View Synthesis (GNVS) models like PixelNeRF for large-scale outdoor scenes, exemplified by the UrbanScene3D dataset. To address the need for a dataset curation pipeline, we proposed four clustering strategies, identifying *SfM shared grouping* as the most effective. Reducing the cluster size further improved performance, highlighting the critical role of high-view overlap. Additionally, our Aug3D augmentation method, which generates synthetic views through *Grid Sampling* and *Semantic Plane Fitting*, boosted GNVS performance when integrated with real data. Despite these advances, challenges remain, including mitigating noise from additional input views and ensuring scalability to diverse datasets and models, pointing to future directions in adaptive clustering and semantic-driven 3D augmentation.

TABLE II: Results for various datasets

Dataset	Configuration	Best PSNR
Real Dataset (Baseline)	Input views 3	20.03
	Input views 6	19.95
	Input views 9	19.59
Synthetic Dataset (ours)	Grid Sampling	29.12
	Semantic Plane Fitting	28.79
Aug3D (ours + baseline)	Grid	21.67
	Semantic	21.80

References

[1] 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] 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

[3] 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)

[4] 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

[5] 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]

[6] Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo. In: 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/

[7] 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)

[8] 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)

[9] 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]

[10] 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)

[11] 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

[12] 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

[13] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[14] 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/

[15] 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. In: CVPR (2024)

[16] 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]

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

[18] 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/

[19] 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

[20] 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]

[21] 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/

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

[23] 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

[24] 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]

[25] 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]

[26] 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

[27] 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)

[28] 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)

[29] 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

[30] 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/

[31] 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/

[32] 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)

[33] 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

[34] 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

[35] 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)

[36] 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)

[37] 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

[38] Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: CVPR (2021)

[39] 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/

[40] 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)

[41] 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)

APPENDIX

A. OTHER SEMANTIC SAMPLING METHOD

In addition to the geometric plane fitting approach, we experimented with a second semantic sampling method using the Segment Anything Model (SAM) to detect buildings

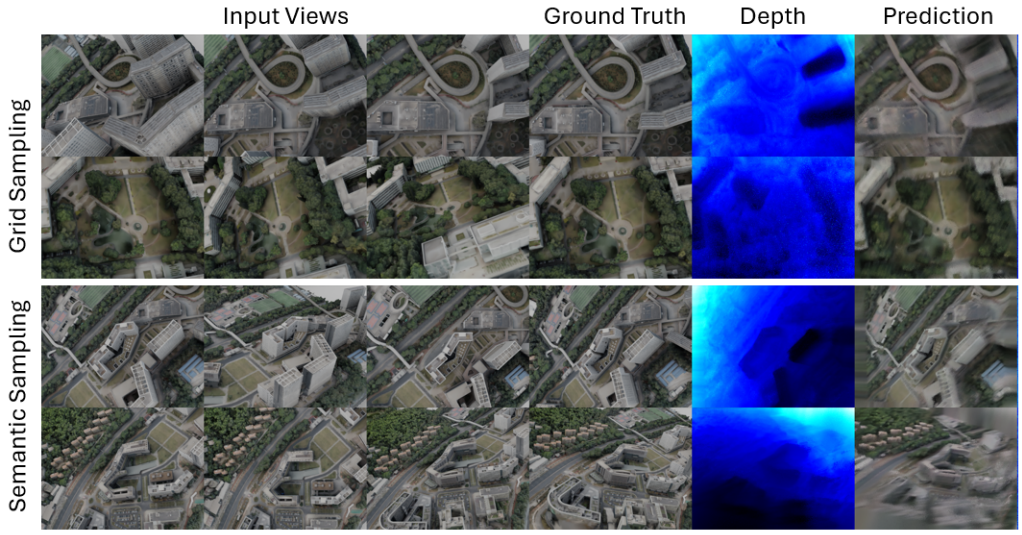


Fig. 6: Qualitative comparison of PixelNeRF trained exclusively on synthetic datasets generated using grid sampling versus semantic sampling methods on the UrbanScene3D *Campus scene*. The first row represents fine-grained predictions, while the second row shows coarse-grained predictions.

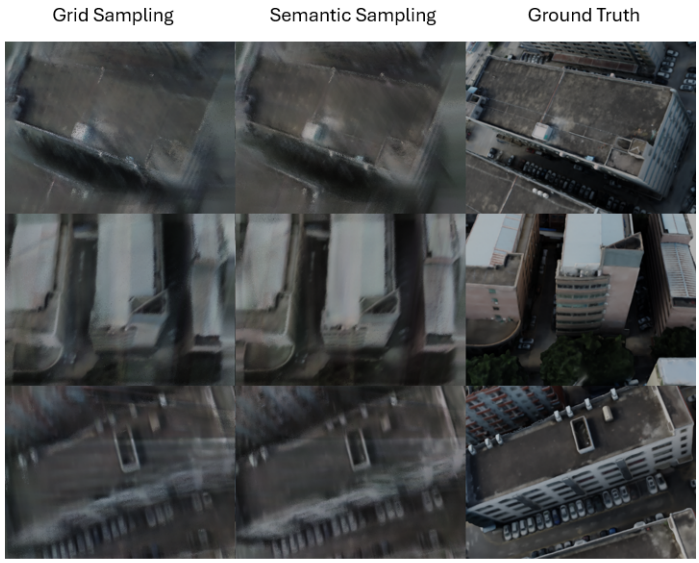


Fig. 7: Qualitative comparison of PixelNeRF trained exclusively on synthetic datasets generated using grid sampling versus semantic sampling methods on the UrbanScene3D *residence scene*, showcasing coarse predictions.

from a top-down view. While SAM showed promise, it was found to be highly sensitive to shadows, resulting in inconsistencies in detecting building structures. Comparatively, the geometric plane fitting method yielded more reliable and accurate results, further emphasizing its suitability for generating semantically meaningful views in diverse lighting conditions.

B. Additional Qualitative Results

To further illustrate the effectiveness of the proposed Aug3D augmentation strategies, we provide qualitative comparisons of the reconstructed scenes using *Grid Sampling* and *Semantic Plane Fitting*.

Figure 6 showcases the qualitative results on the UrbanScene3D Campus scene with 3 input images, comparing models trained exclusively on synthetic datasets generated via grid sampling versus semantic sampling methods. Notably, the semantic sampling approach demonstrates improved reconstruction fidelity, with sharper edges and more accurate structural details, particularly in regions with complex geometries.

In Figure 7, we extend this analysis to the Residence scene, evaluating the same augmentation techniques. Similar trends are observed, with semantic sampling outperforming

grid sampling in preserving finer scene details and mitigating artifacts. The results underline the potential of semantic-driven augmentation to enhance the diversity and quality of synthetic datasets, thereby benefiting GNVS training.

These qualitative evaluations, along with our experiments, reinforce the quantitative findings presented in Section V, validating the advantages of integrating semantic-driven synthetic views into the GNVS pipeline. Future work can explore enhancing SAM's robustness to lighting variations or combining its capabilities with geometric methods for more versatile augmentation strategies.