ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes

Supplementary Material

A. Details of Data Collection

Our hardware setup is shown in Fig. 1. We aim to capture large spaces as a single scene, rather than splitting them into separate rooms in order to provide more context as well as increase the complexity for downstream tasks. Captures from the three sensors are performed as close together in time as possible to avoid inconsistencies in lighting between the different modalities.

Figure 1: Our hardware setup consists of a Faro Focus Premium laser scanner, Sony Alpha 7 IV DSLR camera, and iPhone 13 Pro with a LiDAR sensor.

A.1. Laser Scan

We use the 1/4 resolution and 2x quality setting for the Faro scanner, which takes about 2 minutes for each scan. We estimate the normals of the point cloud as the cross product of neighboring scan points in the corresponding 2D scan grid, and voxelize the points to a $1mm$ resolution.

Point clouds from different scans are first merged at a $1mm$ resolution. The resulting point cloud is chunked into overlapping cubes of side $0.5m$. Poisson reconstruction [5] is applied on each chunk with a depth of 9 (grid size of 2^9) which gives a grid cell size of $0.5m/512 < 1mm$. The chunk overlap is set to 150 grid cells. The resulting chunk meshes are first trimmed to match the original point cloud, and then clipped by 75 grid cells. The vertices of these chunk meshes are finally joined with a $1mm$ threshold to form a single mesh. Quadric edge collapse is then applied [4] to the full resolution mesh, to obtain smaller meshes with 12.5%, 5% and 1.5% of the original number of faces.

A.2. DSLR Images

Our DSLR capture settings are optimized for training and evaluating of novel view synthesis methods. Specifically, we fix the white balance and exposure time to have consistent lighting throughout the scene. We use a wide field-of-view fisheye lens which provides larger overlap between images, and empirically improves both camera pose registration and novel view synthesis. Exposure time is set to 1/100s to avoid flickering effects from indoor lights and minimize motion blur.

A.3. iPhone

In contrast to the DSLR capture, our iPhone recordings are performed in the default automatic mode of the iPhone, making novel view synthesis more challenging. We show the setting comparison between DSLR and iPhone in Tab. 1.

We show the comparison of the point cloud generated from the iPhone depth map and 3D geometry obtained from the laser scanner in Fig. 2. The 3D geometry from the scanner is much cleaner and preserves finer details.

Settings	DSLR	iPhone
Auto white balance	✗	✓
Auto focus	✗	✓
Auto exposure	✗	✓
Field of view (deg)	180	71

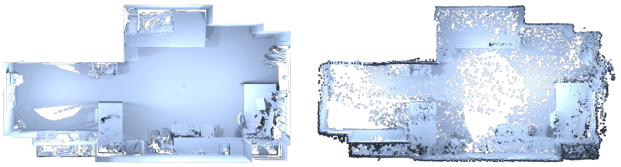
Table 1: Comparison of capture settings between DSLR and iPhone.

B. Benchmark

The ScanNet++ dataset will be made public, along with an online benchmark for the following tasks.

B.1. Novel View Synthesis

Given a set of training images of a scene and unseen camera poses, a method must synthesize the views at the

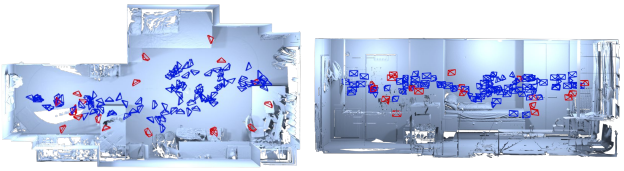


(a) Mesh from laser scan (b) Point cloud from iPhone LiDAR

Figure 2: Comparison of 3D reconstructions from the laser scan and low-resolution point cloud from the iPhone. Depth images from iPhone LiDAR are noisy and low resolution.

unseen poses. The unseen poses are captured at challenging viewpoints independently of the training trajectory (e.g., Fig. 3). Evaluation metrics include PSNR, LPIPS and SSIM.

To benchmark methods trained on iPhone data, we evaluate against DSLR images as ground truth. This setting is more challenging since the methods are expected to produce high-quality outputs based on commodity-level input.



(a) Top-down view (b) Side view

Figure 3: The unseen poses (red) of our DSLR capture for evaluation are challenging since they are very different from the training trajectory (blue) in terms of translation and orientation.

B.2. 3D Semantic Understanding

Given colored meshes of scenes and posed RGB images, we evaluate predictions of 3D semantic segmentation methods on vertices against the ground-truth vertex labels; this is similar to the semantic benchmark on ScanNet [1]. Evaluation metrics include per-class intersection-over-union (IoU) and mean IoU. Similarly, 3D instance segmentation methods are evaluated on ground truth instance masks as well as semantic labels, and evaluation metrics include AP25, AP50 and AP similar to the ScanNet benchmark.

Our benchmark will include more than 100 frequent object classes for both tasks, and evaluation is performed against the multilabeled ground truth – which is not possible in any prior 3D semantic scene understanding benchmark. Hence, we will allow submissions to provide more than one prediction per vertex.

Following ScanNet [1], we maintain a hidden test set and build an online public evaluation website. This website will provide for entries to the latest state-of-the-art methods to facilitate comparisons for new submissions. Importantly, the test set for novel view synthesis tasks does not overlap with the one for semantic understanding tasks, so that the input meshes of the latter cannot be used to guide the former.

C. 3D Semantic Understanding

Semantic Annotation Semantic annotation is performed using a web interface that allows the annotator to apply a free-text label to every mesh segment. Annotation of one scene takes about 1 hour on average, after which a verification pass is done by another annotator to fix incorrect labels. A set of guidelines with reference images is provided to the annotators to perform consistent annotation over similar classes, such as shelf, cabinet, cupboard, wardrobe, bookshelf. Importantly, the guidelines describe common cases for multilabel annotation such as "jacket on a chair", "bedsheet on a bed" and so on. The distribution of the most frequent multilabeled classes is shown in Fig. 4.

Qualitative Results Further qualitative results of semantic and instance segmentation baselines are shown in Fig. 6.

D. Novel View Synthesis on iPhone Data

After training on iPhone data, we apply color correction based on optimal transport between color distributions [2, 3] on the output in order to compare with the DSLR ground truth. The visual comparisons are shown in Fig. 5.

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[2] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014. 2

[3] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. 2

[4] Michael Garland and Paul S Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216, 1997. 1

[5] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 1

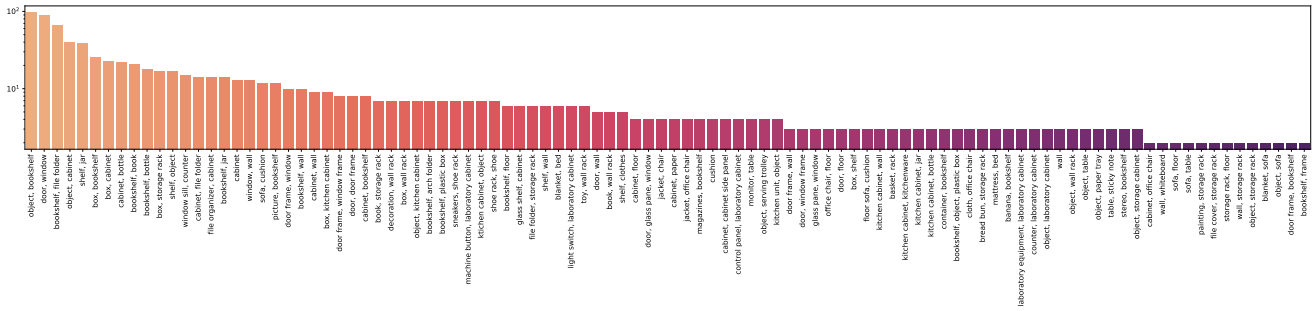


Figure 4: Top 100 pairs of semantic classes that are multi-labeled together on a log scale.

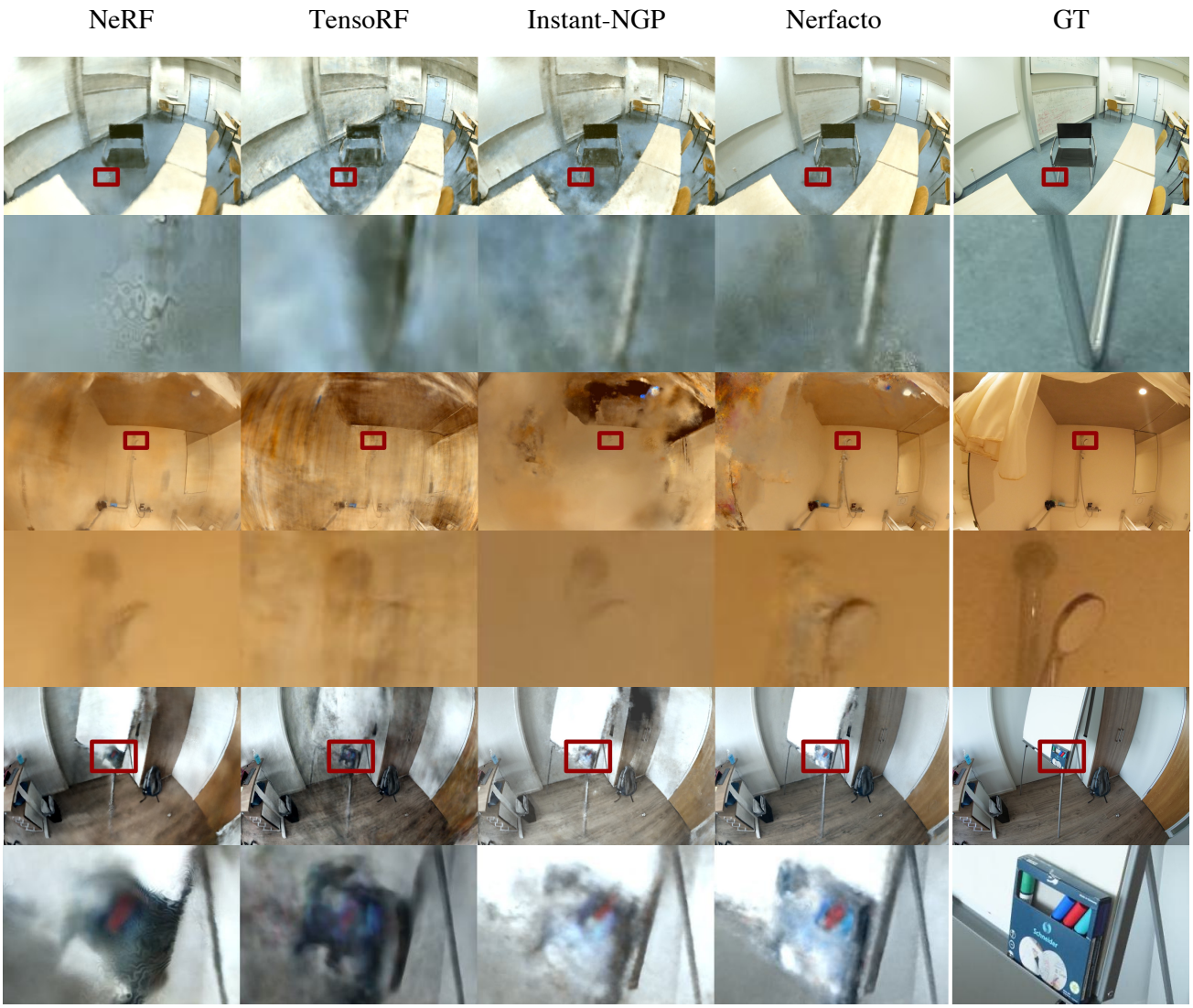
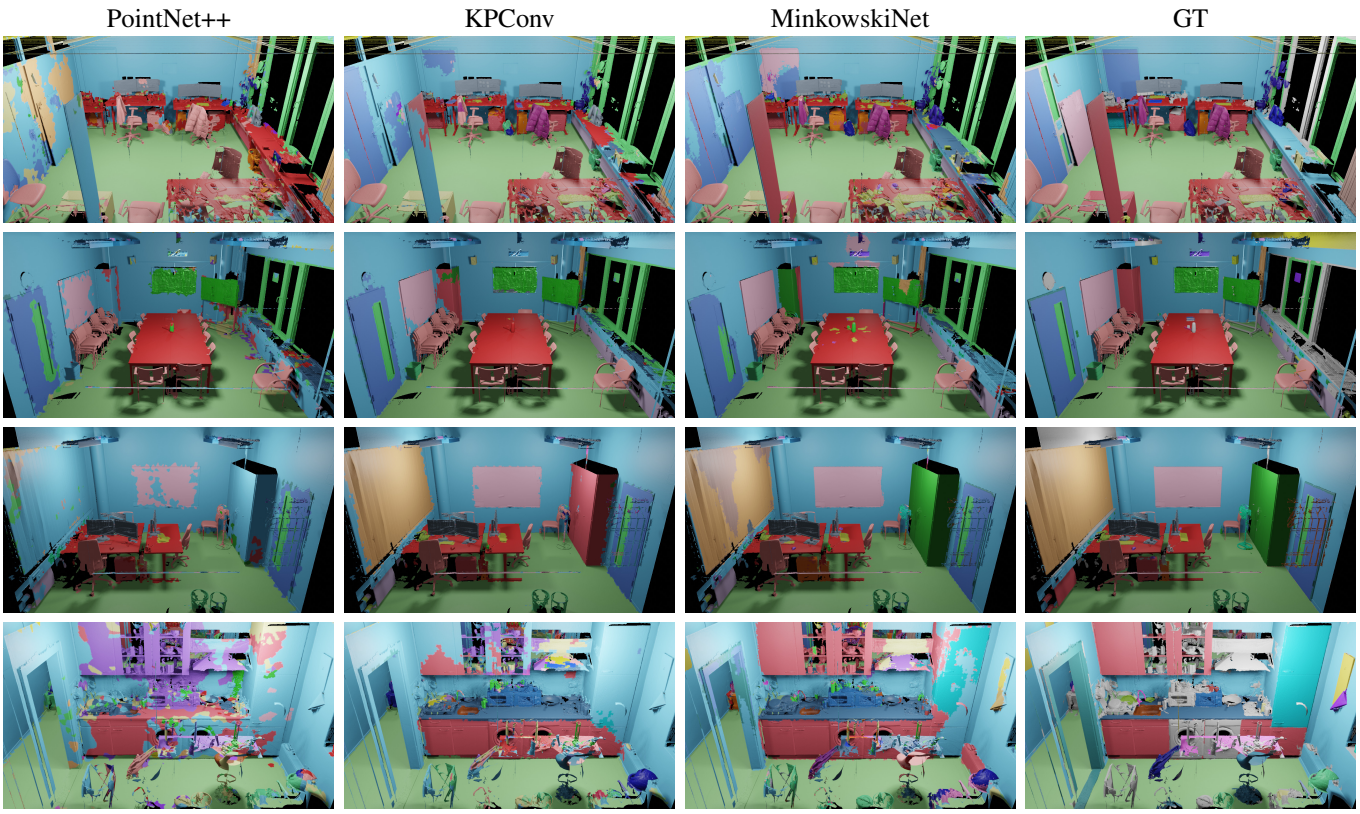
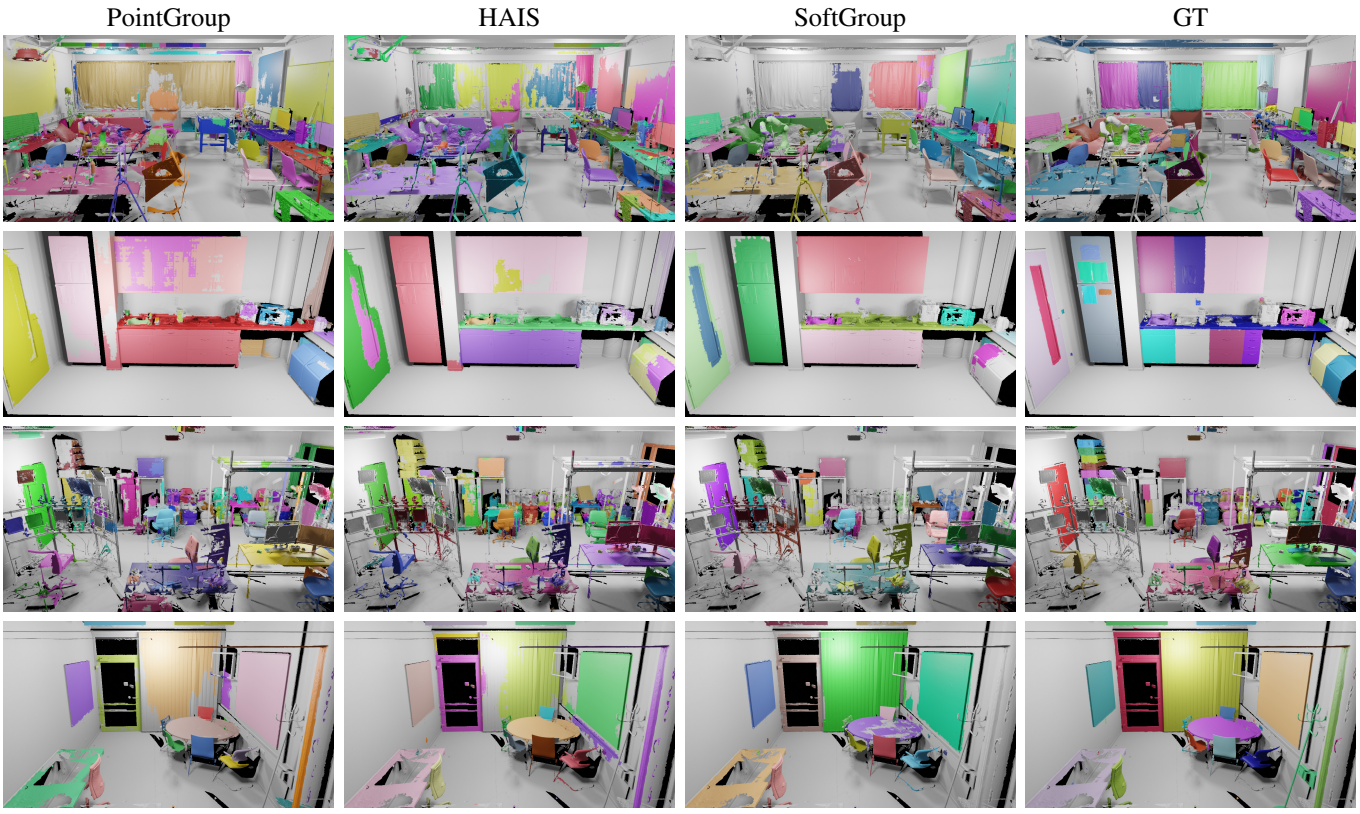


Figure 5: Comparison of different novel view synthesis methods on iPhone data using the DSLR ground truth. Compared to DSLR results shown in the main paper, NVS on commodity iPhone data generates more artifacts and blurry results.



(a) 3D semantic segmentation baselines.



(b) 3D instance segmentation baselines.

Figure 6: Qualitative results of 3D semantic and instance segmentation methods on the validation set of ScanNet++, showing diverse and cluttered scenes. All methods have notable room for improvement on small and ambiguous objects.