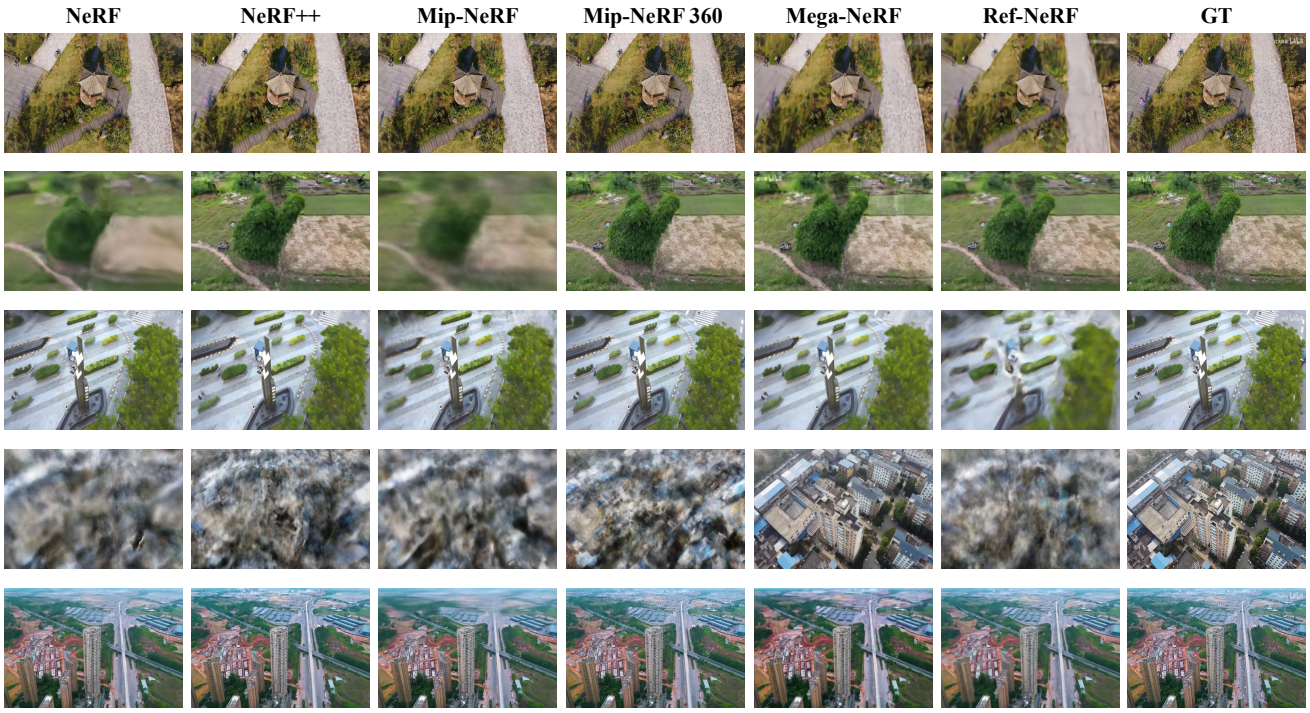
A Large-Scale Outdoor Multi-modal Dataset and Benchmark for Novel View Synthesis and Implicit Scene Reconstruction : Appendix

In this supplementary material, we provide the appendix section and a supplemental video to better understand our dataset and benchmarks. This appendix involves more qualitative and quantitative results (cf. Sec. 1 and Sec. 2), experiments on the Mill-19 dataset(cf. Sec. 3), details of our dataset generation method (cf. Sec. 4), and dataset analysis (cf. Sec. 5). The supplemental video contains a brief introduction to our dataset, some examples in detail, and more comprehensive synthesis results in surrounding views or progressive views.

1. More Qualitative Results

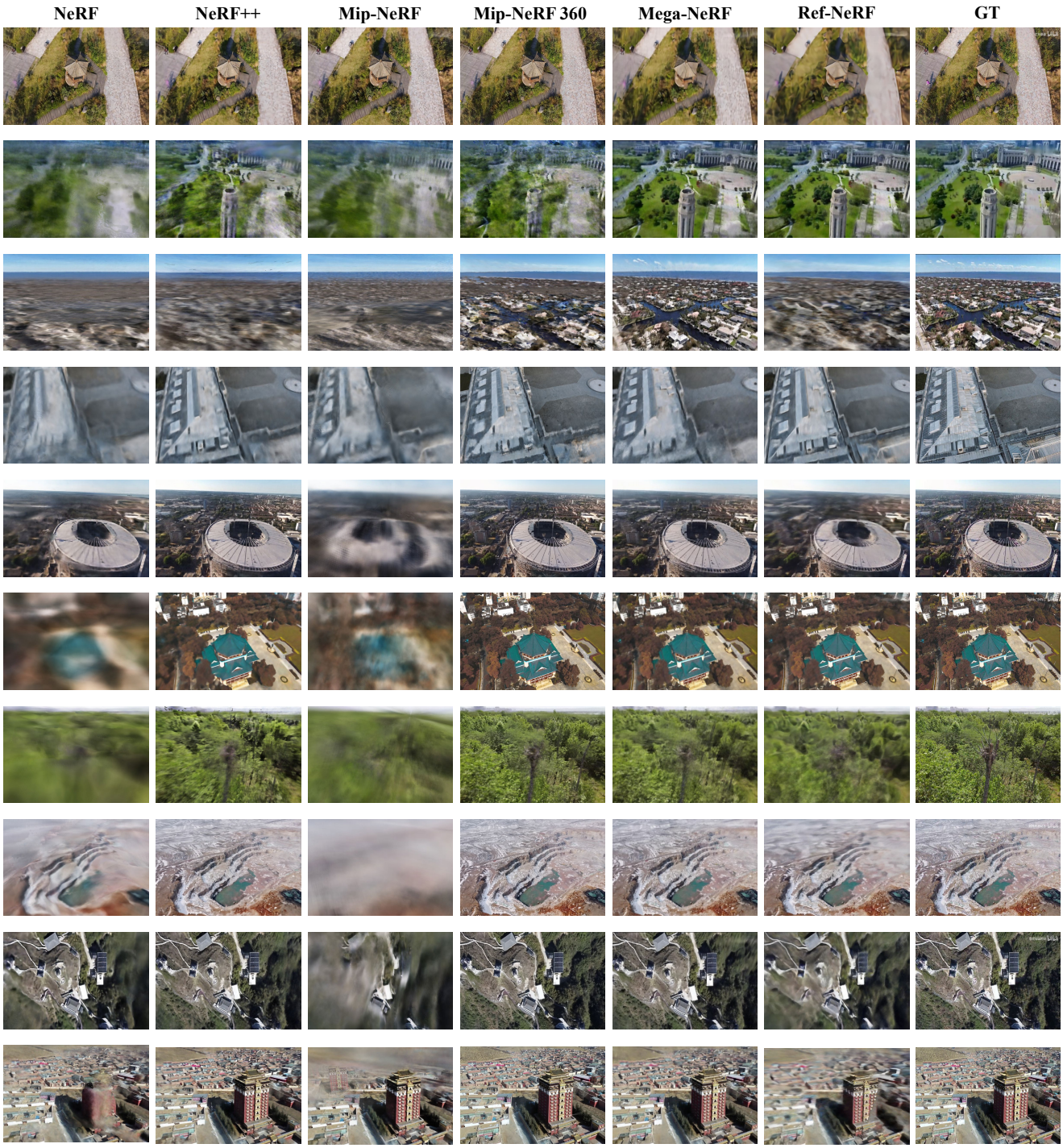
1.1. Novel View Synthesis

In our main manuscript, we can only provide the visualization results of five scenes due to the length limitation. In this section, more qualitative results are presented to demonstrate the novel view synthesis ability of each method (cf. Fig.1).



Part 1 / 2

Figure 1. More qualitative visualization results for novel view synthesis (zoom-in for the best of views) on our OMMO dataset.



Part 2 / 2

Figure 1. More qualitative visualization results for novel view synthesis (zoom-in for the best of views) on our OMMO dataset.

1.2. Scene Representation

To further demonstrate that our OMMO dataset can well support surface or scene reconstruction tasks including NeRF-based methods, we visualize more shape results by various representations (*cf*. Fig. 2). Among them, plenoctree, mesh, and dense points are provided by Mega-NeRF [6], InstantNGP [2], and Colmap [3, 4], respectively.

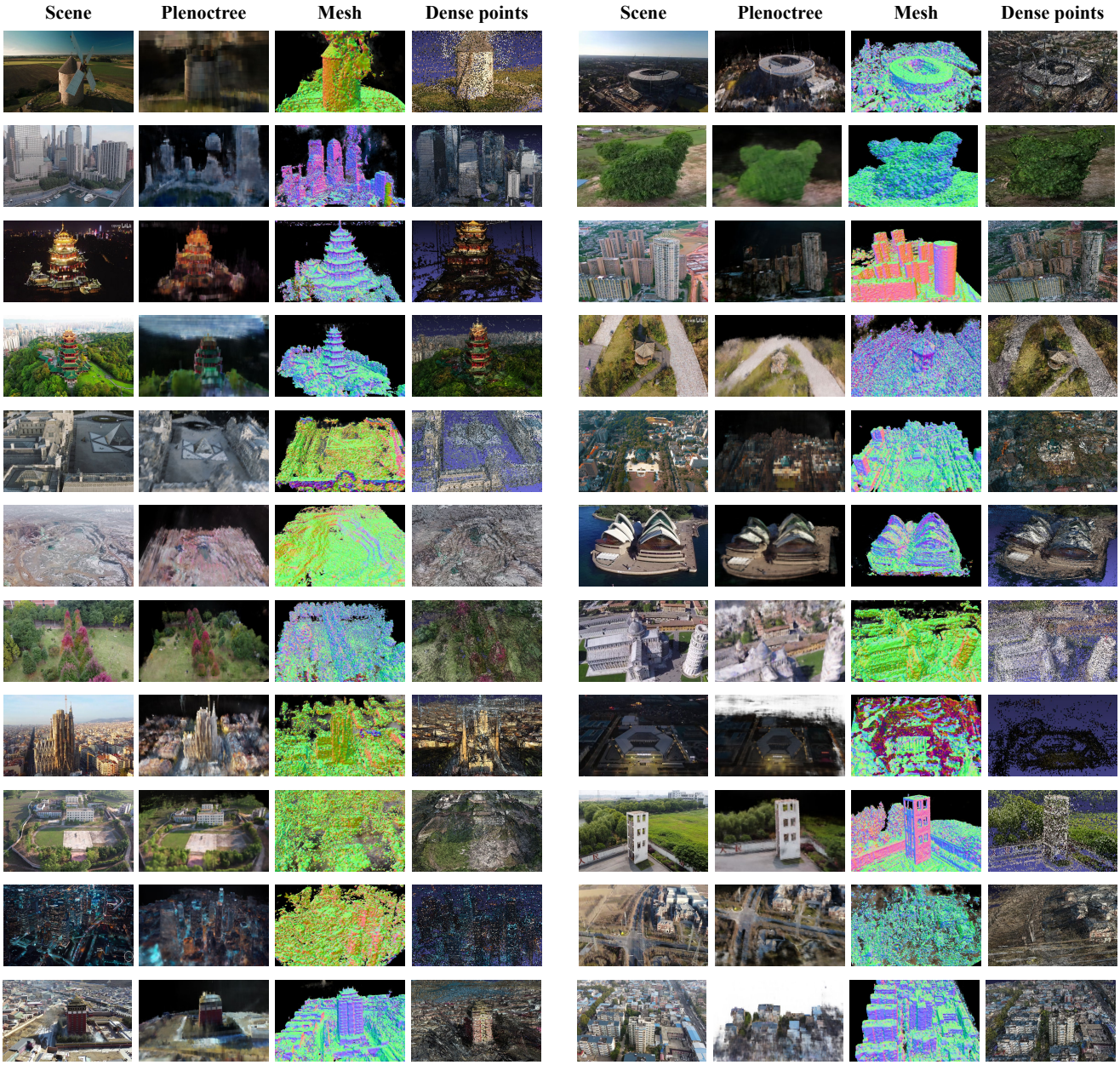


Figure 2. More qualitative visualization results for various scene representations (zoom-in for the best of views) through the state-of-the-art methods on the OMMO dataset.

2. More Quantitative Results

Multi-modal NeRF Synthesis. Our manuscript has shown that even without a well-designed module for injecting text information, the performance of both NeRF [1] and CoCo-INR [7] methods have improved. Since the textual prompts contain more global features about rich geometry or appearance information, which are shared by different views in the scene to guarantee the network to synthesis view-consistency results. We hope to inspire more image-text multi-modal NeRF methods to synthesize photo-realistic results and decent geometry by exploring effective ways to make full use of textual prompts. The benchmark on each scene and the sub-benchmarks on different scene types are shown in Tab. 1 and Tab. 2.

Table 1. Benchmark for multi-modal NeRF synthesis. We present the performance of text-assisted novel view synthesis based on existing methods on our OMMO dataset. ↑ means the higher, the better.

Scene ID	Scene Types	Camera Tracks	Lighting Conditions	**NeRF** [1] w/o Prompts			**NeRF** [1] w/ Prompts			**CoCo-INR** [7] w/o Prompts			**CoCo-INR** [7] w/ Prompts		
				PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
1	Buildings	Irregular	Day	16.93	0.369	0.744	16.89	0.366	0.729	14.31	0.432	0.788	14.81	0.431	0.785
2	Small area	Circles	Day	15.31	0.442	0.694	15.61	0.465	0.711	16.04	0.597	0.632	16.25	0.597	0.626
3	Citys	Lines	Day	14.38	0.278	0.556	14.42	0.277	0.573	15.59	0.485	0.616	16.62	0.509	0.585
4	Buildings	Circle	Night	25.39	0.859	0.431	24.94	0.851	0.425	21.61	0.876	0.480	21.87	0.879	0.481
5	Small area	Circles	Day	22.26	0.670	0.531	21.31	0.652	0.564	18.16	0.657	0.597	20.08	0.675	0.573
6	Natural scenes	Circles	Day	24.09	0.679	0.504	23.78	0.655	0.535	19.65	0.630	0.576	19.39	0.627	0.578
7	Buildings	Lines	Day	5.36	0.166	0.747	6.25	0.183	0.697	16.53	0.628	0.679	15.38	0.567	0.654
8	Citys	Circle	Day	21.14	0.496	0.594	21.55	0.510	0.571	16.94	0.413	0.687	16.57	0.407	0.704
9	Citys	Lines	Day	14.92	0.344	0.744	15.02	0.345	0.749	13.70	0.340	0.773	13.68	0.340	0.765
10	Citys	Irregular	Day	22.26	0.550	0.626	22.44	0.551	0.624	18.81	0.536	0.694	18.62	0.535	0.693
11	Buildings	Circles	Night	22.36	0.816	0.420	22.58	0.820	0.412	17.08	0.746	0.494	17.35	0.747	0.491
12	Small area	Circles	Day	22.41	0.594	0.533	22.80	0.608	0.512	17.87	0.475	0.658	17.81	0.473	0.659
13	Buildings	Lines	Day	22.27	0.592	0.608	23.12	0.619	0.576	16.55	0.532	0.698	17.02	0.542	0.671
14	Small area	Lines	Day	19.85	0.554	0.569	20.73	0.591	0.534	15.44	0.485	0.663	15.19	0.482	0.665
15	Small area	Circles	Day	20.35	0.527	0.552	20.70	0.549	0.533	16.37	0.407	0.702	16.45	0.411	0.689
16	Natural scenes	Circles	Day	17.86	0.397	0.631	17.53	0.362	0.647	15.37	0.384	0.633	15.24	0.376	0.640
17	Natural scenes	Circles	Day	22.02	0.571	0.610	22.23	0.575	0.596	20.52	0.575	0.619	19.38	0.527	0.648
18	Small area	Lines	Day	26.06	0.754	0.428	26.48	0.770	0.402	17.31	0.527	0.658	17.35	0.532	0.664
19	Small area	Circles	Day	14.20	0.399	0.726	14.19	0.397	0.720	15.41	0.388	0.701	15.82	0.413	0.694
20	Citys	Circles	Day	22.84	0.613	0.499	23.30	0.636	0.465	18.28	0.434	0.676	18.09	0.431	0.685
21	Natural scenes	Circles	Day	22.59	0.514	0.532	22.99	0.541	0.508	17.08	0.358	0.744	17.28	0.359	0.720
22	Buildings	Lines	Day	16.53	0.466	0.733	20.404	0.539	0.598	14.86	0.408	0.759	14.73	0.406	0.772
23	Natural scenes	Lines	Day	18.99	0.405	0.669	19.09	0.405	0.671	17.57	0.335	0.673	17.43	0.332	0.701
24	Natural scenes	Lines	Day	19.32	0.386	0.696	18.52	0.379	0.708	18.63	0.347	0.765	18.27	0.341	0.814
25	Natural scenes	Lines	Day	24.72	0.550	0.528	25.24	0.576	0.496	20.15	0.434	0.717	20.29	0.434	0.711
26	Buildings	Irregular	Day	8.56	0.242	0.564	8.56	0.242	0.564	9.19	0.336	0.924	9.23	0.341	0.913
27	Citys	Irregular	Day	4.54	0.006	0.705	4.91	0.249	0.818	16.19	0.443	0.699	16.07	0.443	0.687
28	Small area	Circles	Day	24.48	0.660	0.479	24.32	0.630	0.493	20.12	0.536	0.643	21.13	0.595	0.621
29	Buildings	Circle	Day	22.98	0.608	0.540	23.58	0.631	0.516	16.57	0.439	0.733	17.93	0.453	0.716
30	Natural scenes	Irregular	Day	20.23	0.522	0.605	21.02	0.559	0.569	12.36	0.431	0.760	15.40	0.450	0.719
31	Citys	Circles	Night	18.97	0.365	0.645	19.09	0.371	0.634	17.88	0.465	0.685	17.57	0.459	0.704
32	Citys	Irregular	Day	17.99	0.582	0.621	18.00	0.582	0.628	17.01	0.623	0.588	16.94	0.622	0.590
33	Citys	Irregular	Day	5.79	0.007	0.745	5.79	0.007	0.744	15.20	0.436	0.761	14.68	0.431	0.770
Mean	-	-	-	**18.72**	**0.484**	**0.600**	**19.01**	**0.500**	**0.592**	**16.80**	**0.489**	**0.681**	**16.97**	**0.490**	**0.678**

Table 2. More sub-benchmarks for multi-modal NeRF synthesis. We divide our dataset into subsets based on different scene types, camera trajectories, and lighting conditions, and provide sub-benchmarks under different settings. ↑ means the higher, the better.

Scene ID	Sub-benchmark	**NeRF** [1] w/o Prompts			**NeRF** [1] w/ Prompts			**CoCo-INR** [7] w/o Prompts			**CoCo-INR** [7] w/ Prompts		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
1,4,7,8,11,13,22,26	**Buildings**	17.32	0.501	0.605	18.04	0.516	0.572	15.88	0.546	0.689	15.87	0.540	0.684
2,5,12,14,15,18,19,28,29	**Small areas**	20.88	0.579	0.561	21.08	0.588	0.554	17.03	0.501	0.665	17.56	0.515	0.656
3,8,9,10,20,27,31,32,33	**Cities**	15.87	0.360	0.637	16.06	0.392	0.645	16.62	0.464	0.687	16.54	0.464	0.687
6,16,17,21,23,24,25,30	**Natural scenes**	21.23	0.503	0.597	21.30	0.507	0.591	17.67	0.437	0.686	17.84	0.431	0.691
2,4,5,6,8,11,12,15,16,17,19,20,21,28,31	**Circles**	21.08	0.573	0.559	21.13	0.575	0.555	17.89	0.529	0.635	18.02	0.532	0.634
3,7,9,13,14,18,22,23,24,25	**Lines**	18.24	0.450	0.628	18.93	0.468	0.600	16.63	0.452	0.700	16.60	0.449	0.700
1,10,26,27,29,30,32,33	**Irregular**	14.91	0.361	0.644	15.15	0.398	0.649	14.96	0.460	0.743	15.46	0.463	0.734
ALL-{4, 11,31}	**Day**	18.37	0.465	0.610	18.69	0.482	0.602	16.59	0.468	0.694	16.77	0.469	0.690
4, 11,31	**Night**	22.24	0.680	0.499	22.20	0.681	0.490	18.86	0.696	0.553	18.93	0.695	0.559

3. Experiments on the Mill-19 Dataset

In addition, we surveyed real outdoor large-scale scene dataset from the same type of drone perspective. Among them, Mega-NeRF's Mill-19 [6] meets the requirements. It contains two scenes: building and rubble, which contain 1940 and 1678 images respectively, as well as 30 billion and 26 billion pixels/rays. However, due to the close-up view settings in Mill-19, there is less overlap between different viewpoints. According to the author's report, a single image contains only 0.062 and 0.050 of the scene. Therefore, it is difficult to apply this dataset on most NeRF models which requires more overlap. Considering the need for comparison, we uniformly sample the perspective of the original data of the above scenes.After adjusting the number of views, it is difficult to obtain good results on other NeRF models or even converge. The visualization results on MipNeRF-360 with better convergence are still blurry . In contrast, OMMO dataset is more friendly and universal to most NeRF models and can perform more comprehensive and standard comparisons.

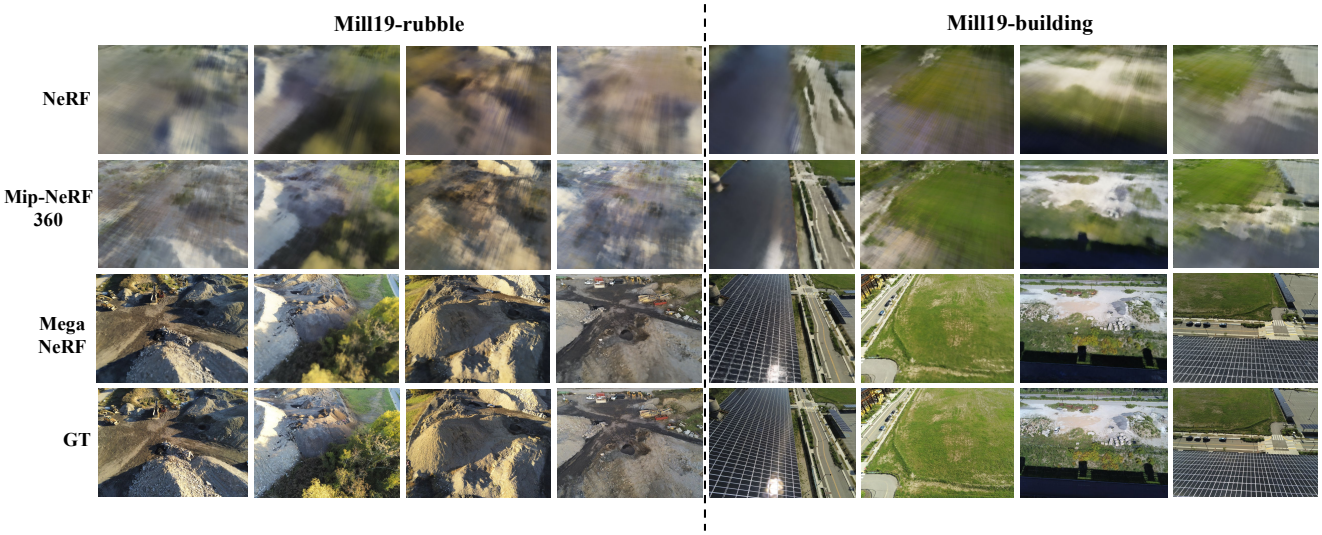


Figure 3. The visualization results of the mill19 dataset on several representative models, respectively, are the nerf model with poor convergence , the result of mip-nerf 360 with better convergence after uniformly sampling the perspective and the performance result of 25 sub-modules of mega-nerf under full perspective.

4. Method Details

We show some dropped frames or scenes during dataset generation to better understand our selection and review standard in Fig. 4. At auto assessment stage, the image quality assessment model [5] is employed to remove frames with blur, artifacts, ghosting and incorrect colors caused by overexposure or optical effects. In this way, about 64% of the frames remained, but there are still some low-quality frames with blur, subtitles, abnormal brightness or transparency caused by fading in or out at the beginning or end of the video. So during the manual quality review process, volunteers and experts will work together to remove these frames. After scene calibration and reconstruction, some scenes will fail, such as with insufficient overlap and textures, or forwardly moving camera motion. These fail-to-calibrate scenes cannot meet the requirements of NeRF-based methods, which need to be removed at the manual scene review stage.

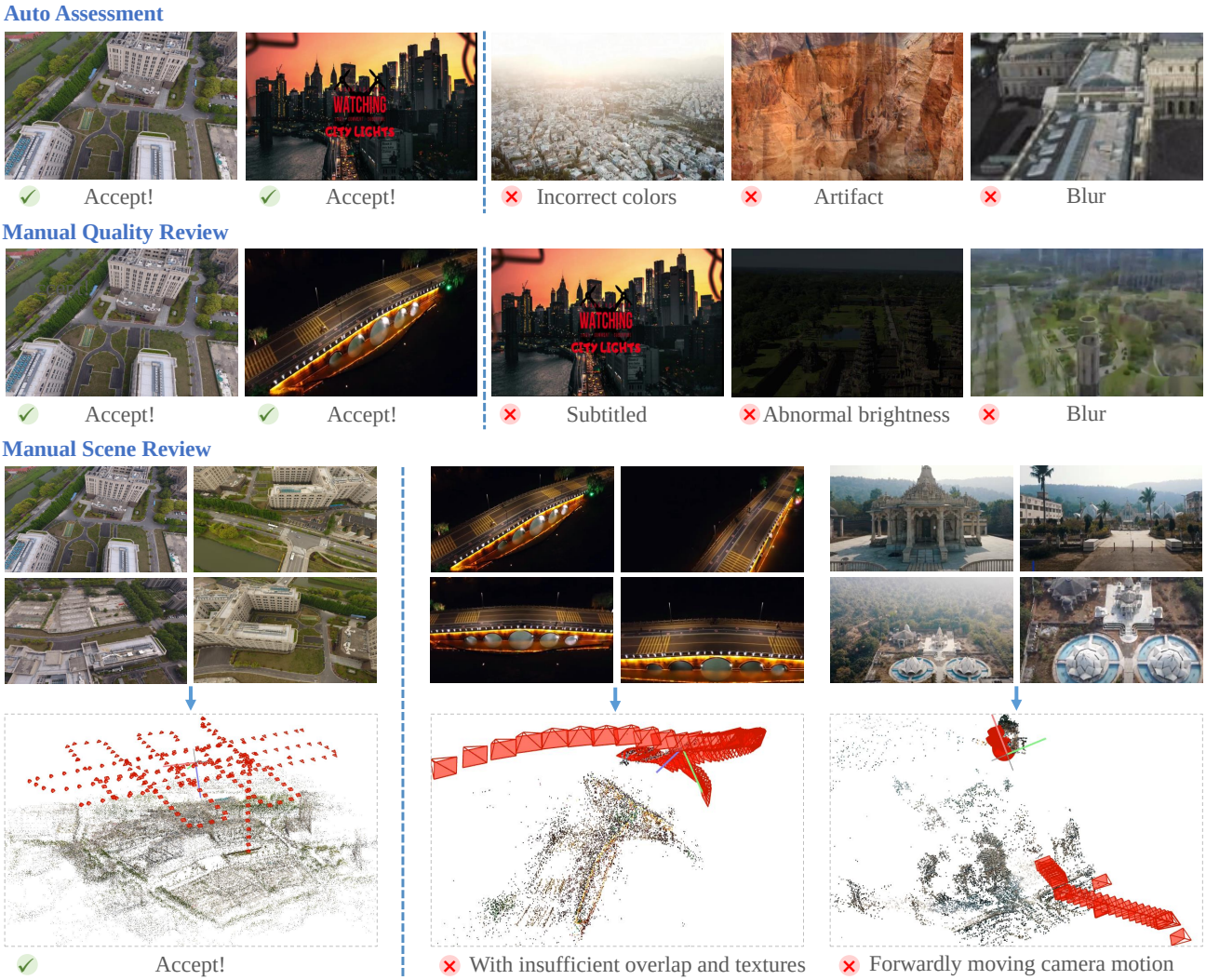


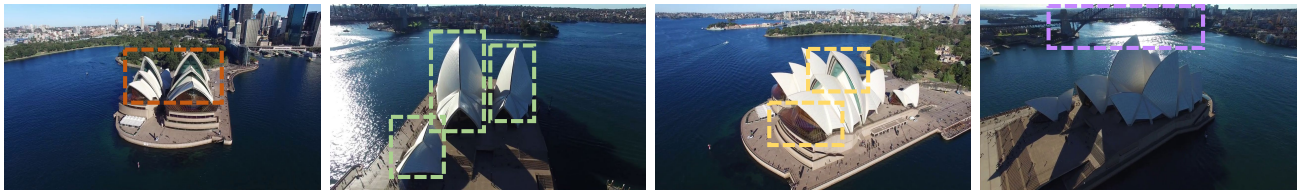
Figure 4. Some examples of dropped frames or scenes at auto assessment, manual quality review, and manual scene review stages. Meanwhile, we also show the number of images and scenes before and after the review at each stage.

5. Dataset Anaysis

5.1. Textual Prompts

We show an example of scene prompt annotations from our OMMO dataset in Fig. 5. Our prompts annotation comprehensively describes every detail of the scene center and its surrounding environment in many short sentences.

Views



Prompt annotation

- The white building is captured by a circular camera track.
- The building is located on a peninsula surrounded by water on three sides.
- The shape of the building is three shell-shaped sub-buildings, two of which are juxtaposed with larger shells and another one is smaller.
- The two larger sub-buildings are composed of four pointed shells in a cascade.
- The smaller sub-building consists of two back-to-back shells.
- The glass between each layer of shells is yellow or green.
- There are many people around the building.
- There is a white carport in front of the building, where some cars are parked;
- Behind the building is a round island with many trees planted on it.
- There is a bridge across the river on one side of the building.
- There are dense buildings on both sides of the bridge.

Figure 5. **Textual Prompt.** An example of annotations. Several phrases and their corresponding patches are highlighted in the same color.

5.2. Word Statistic

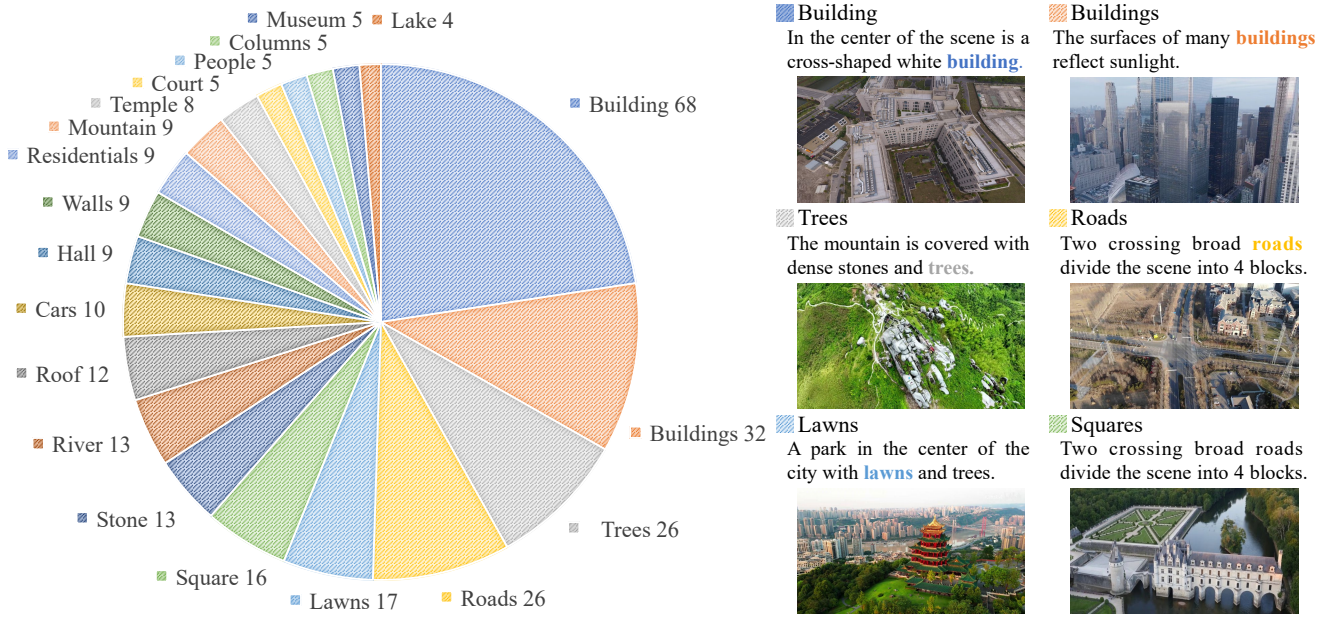


Figure 6. **Word statistic.** Only include nouns that appear more than 4 times in our OMMO dataset.

We report the word statistic for all scene prompts annotations (only including nouns that appear more than 4 times) and some word-scene examples (cf. Fig. 6). It can be seen that our data distribution is comprehensive and reasonable, including building, buildings (architectural complex), trees, roads, lawn and rivers, etc. Meanwhile, the number of keywords can roughly reflect the distribution of different scenes, such as natural scene: urban scene (building, small area, city) is about 1:3.

5.3. User Instrictions

Our OMMO dataset structure list is shown in Fig. 7. The first-level directory contains the scene list and sub-folders for each scene. Each scene-folder contains scene prompts, the original video, the training and validation split file, and sub-folders for images, camera matrices and frame textual prompts.

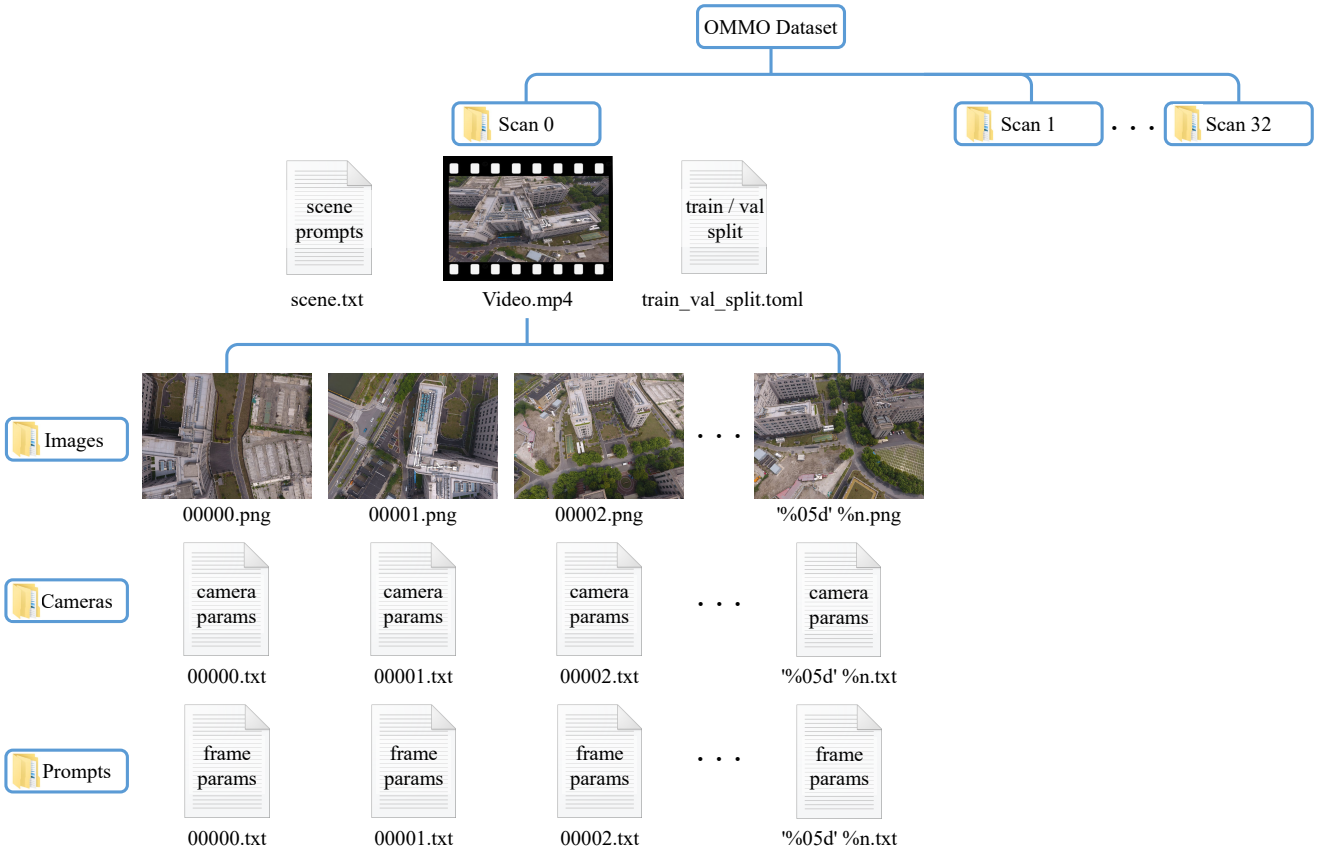


Figure 7. The structure of our OMMO dataset.

6. Ethical Issue

Human privacy is well preserved in OMMO dataset. Since our dataset is captured by drones, pedestrians cannot be identified in most cases because they are small enough (cf. the first row in Fig. 8). For some zoomed-in shots where pedestrians can be identified, we blur their shapes and textures (cf. the second row in Fig. 8). Since identifiable pedestrians occupy few pixels, so it does not harm the view consistency and a robust method will not be severely affected as shown in above benchmarks.

Figure 8. **Privacy Protection.** Identifiable pedestrians will be blurred.

References

[1] 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 *ECCV*, 2020. 4

[2] 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. 3

[3] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3

[4] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016. 3

[5] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018. 6

[6] 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*, pages 12922–12931, 2022. 3, 5

[7] Fukun Yin, Wen Liu, Zilong Huang, Pei Cheng, Tao Chen, and Gang YU. Coordinates are not lonely–codebook prior helps implicit neural 3d representations. *arXiv preprint arXiv:2210.11170*, 2022. 4