CL-MVSNet: Unsupervised Multi-view Stereo with Dual-level Contrastive Learning (Supplemental Materials)

1. Depth Map Filtering and Fusion

After per-view depth estimation, depth maps inferred through CL-MVSNet will be filtered via photometric consistency and geometric consistency before depth fusion. During photometric consistency filtering, pixels with low confidence will be discarded from depth maps according to the confidence maps generated by our CL-MVSNet. In the geometric consistency filtering, pixels with depth estimation inconsistent across neighboring views will be considered outliers and filtered out. For DTU [1] and BlendedMVS [6], we use the same filtering strategy as RC-MVSNet [2]. Specifically, the depth threshold, consistent number, and probability threshold are set to 0.001, 3, and 0.8, respectively. For Tanks&Temples [4], we adopt the dynamic consistency checking strategy proposed in [5]. Finally, pixels in filtered depth maps will be projected to the world coordinate system to produce 3D dense point clouds.

2. More Ablation Studies on DTU

In this section, we will conduct additional ablation studies to provide more comprehensive information about our proposed CL-MVSNet.

2.1. Norms of Photometric Consistency Loss

We first conduct ablation studies to evaluate the effects of different norm-based photometric consistency losses, as shown in Tab. 1.

Table 1. **Ablation study of different norm-based photometric consistency losses on DTU [1].**

Norm	Acc.↓	Comp.↓	Overall↓
$L_{0.25PC}$	0.376	0.298	0.337
$L_{0.5PC}$	**0.375**	**0.283**	**0.329**
$L_{0.75PC}$	0.380	0.292	0.336
L_{1PC}	0.392	0.286	0.339
L_{2PC}	0.391	0.293	0.342

2.2. Number of Views & Input Resolution

As shown in Tab. 2, this study demonstrates that CL-MVSNet can be used with any number of input views and can handle input images of different sizes.

Table 2. **Ablation study of resolution $H \times W$ and number of input views N on DTU [1].**

N	$H \times W$	Acc.↓	Comp.↓	Overall↓
3	1184 ×1600	0.392	0.319	0.356
5	1184 ×1600	**0.375**	0.283	**0.329**
7	1184 ×1600	0.381	**0.279**	0.330
9	1184 ×1600	0.384	0.285	0.335
5	864 ×1152	0.390	0.289	0.340
5	512 ×640	0.448	0.370	0.409

2.3. Occlusion Rate

The pixel-level contrastive sample is constructed with the pixel-level occlusion rate α. To investigate how this hyperparameter influences the performance of the model, we conduct an ablation study with different settings and compare the quantity results, as shown in Tab. 3.

Table 3. **Ablation study of the occlusion rate α for the pixel-level contrastive sample on DTU [1].**

α	Acc.↓	Comp.↓	Overall↓
0	0.378	0.309	0.344
0.1	**0.375**	**0.283**	**0.329**
0.2	0.380	0.291	0.336
0.3	0.398	0.321	0.360

3. Comparison to SOTA End-to-end Unsupervised Method

The SOTA end-to-end unsupervised method RC-MVSNet [2], employs a rendering consistency network to build additional supervisory signals, which may be ineffective at times due to the inherent gap between novel view synthesis and depth estimation. Specifically, their proposed depth rendering consistency loss relies on the sampled point

candidates near the object surface, which are sampled according to the depth inferred from the backbone network. If the backbone network produces incorrect depth estimations, the sampled points will be also unreasonable, and the rendered depth will be wrong in the end. This can result in the depth rendering consistency loss failing to guide the model effectively. In contrast, our method utilizes dual-level contrastive learning to construct more effective supervisory signals, boosting the accuracy, completeness, and overall quality of 3D reconstruction results.

Moreover, during the training phase, our method converges much faster and consumes less memory than RC-MVSNet. RC-MVSNet requires 15 epochs with 11 hours per epoch for training on two NVIDIA Tesla V100s and consumes 14.5 GB memory for each GPU. In comparison, under the same conditions, CL-MVSNet needs 16 epochs with 5 hours per epoch to converge and 12 GB memory for each GPU.

4. Smooth Operation

Similar to Smooth L1, we have applied a smooth operation to avoid the significant gradient change near the zero point:

$L_{0.5}(e) = \begin{cases} ke^2 + b, & e < \beta \\ \|e\|_{\frac{1}{2}}, & e >= \beta \end{cases}$. Hence, our network can converge smoothly.

5. Limitation and Future Work

Our model has addressed the limitations of indistinguishable regions and view-dependent effects, but the accurate depth estimation in object edge areas remains a challenge. It is worth noting that this is a common problem in unsupervised MVS methods. To mitigate this issue, we adopt an edge-aware depth smoothness loss proposed in [3], which is based on the assumption that the gradient maps of the input reference image and the inferred depth map should be similar. However, this simple assumption may be invalid in many cases. For instance, there may be significant color gradient changes within the same object. In the near future, we will explore a more effective approach to address this problem.

6. More Visualization Results

We visualize the reconstructed 3D point clouds from DTU [1] evaluation set, TanksTemples [4] set, and BlendedMVS [6] respectively in Fig. 1, Fig. 2 and Fig. 3. And it is important to note that our model has been trained solely on the DTU [1] training set without any fine-tuning. Our CL-MVSNet shows its robustness and generalizability on various scenes.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016.

[2] Di Chang, Aljaž Božič, Tong Zhang, Qingsong Yan, Yingcong Chen, Sabine Süsstrunk, and Matthias Nießner. Rc-mvsnet: unsupervised multi-view stereo with neural rendering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 665–680. Springer, 2022.

[3] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv preprint arXiv:1905.02706*, 2019.

[4] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[5] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, pages 674–689. Springer, 2020.

[6] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.

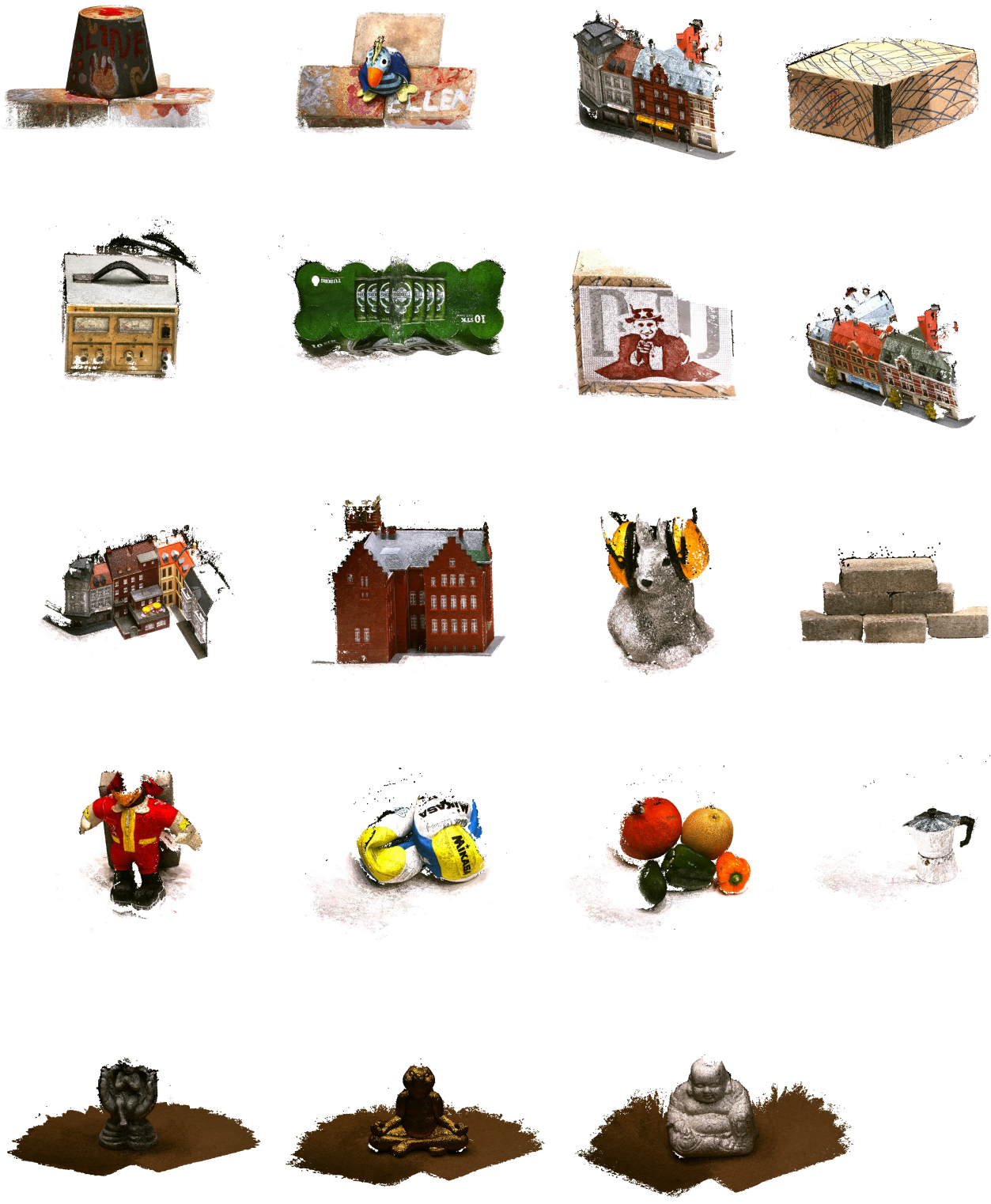

Figure 1. **Reconstruction results on DTU [1].**

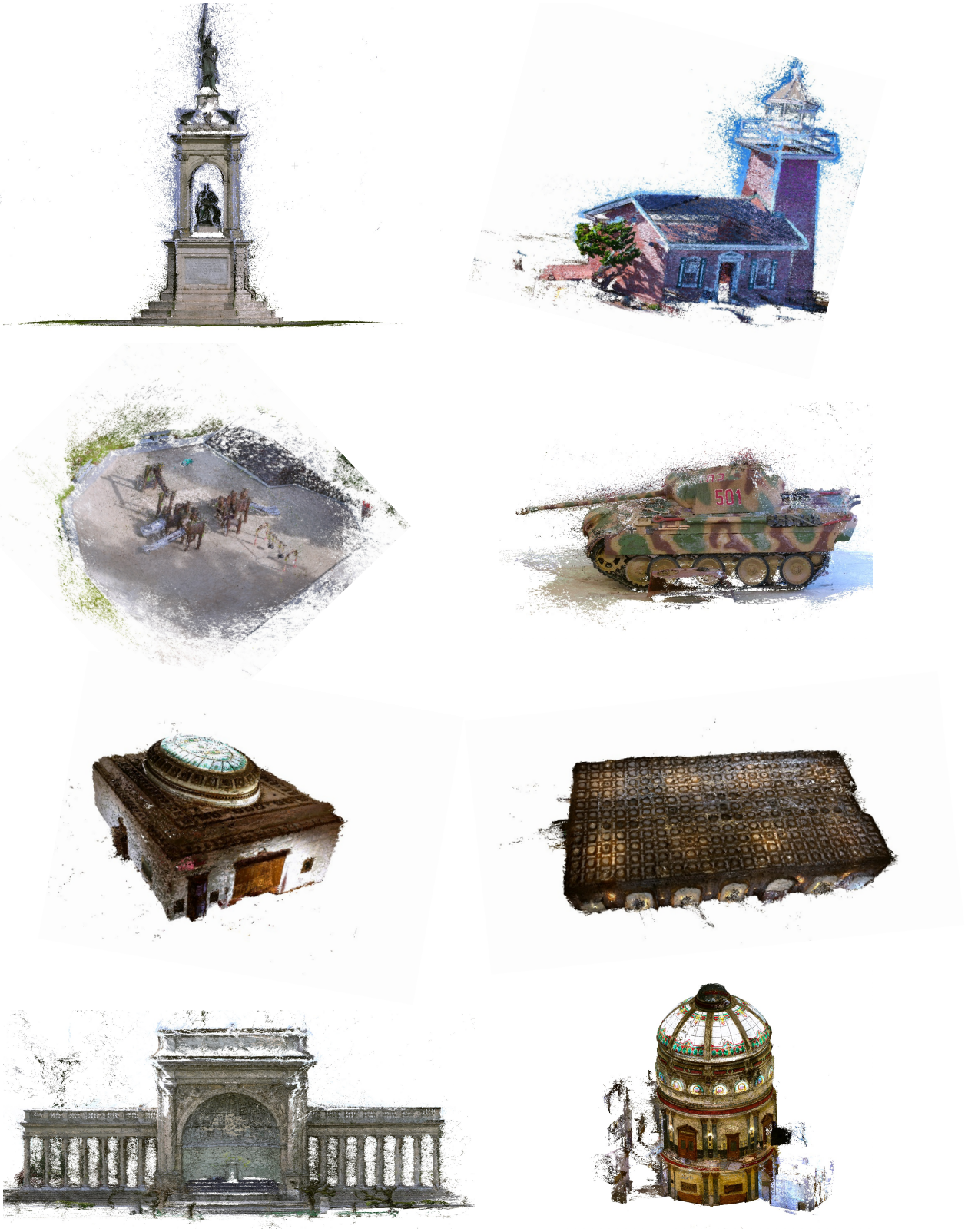

Figure 2. **Reconstruction results on Tanks&Templs [4].**

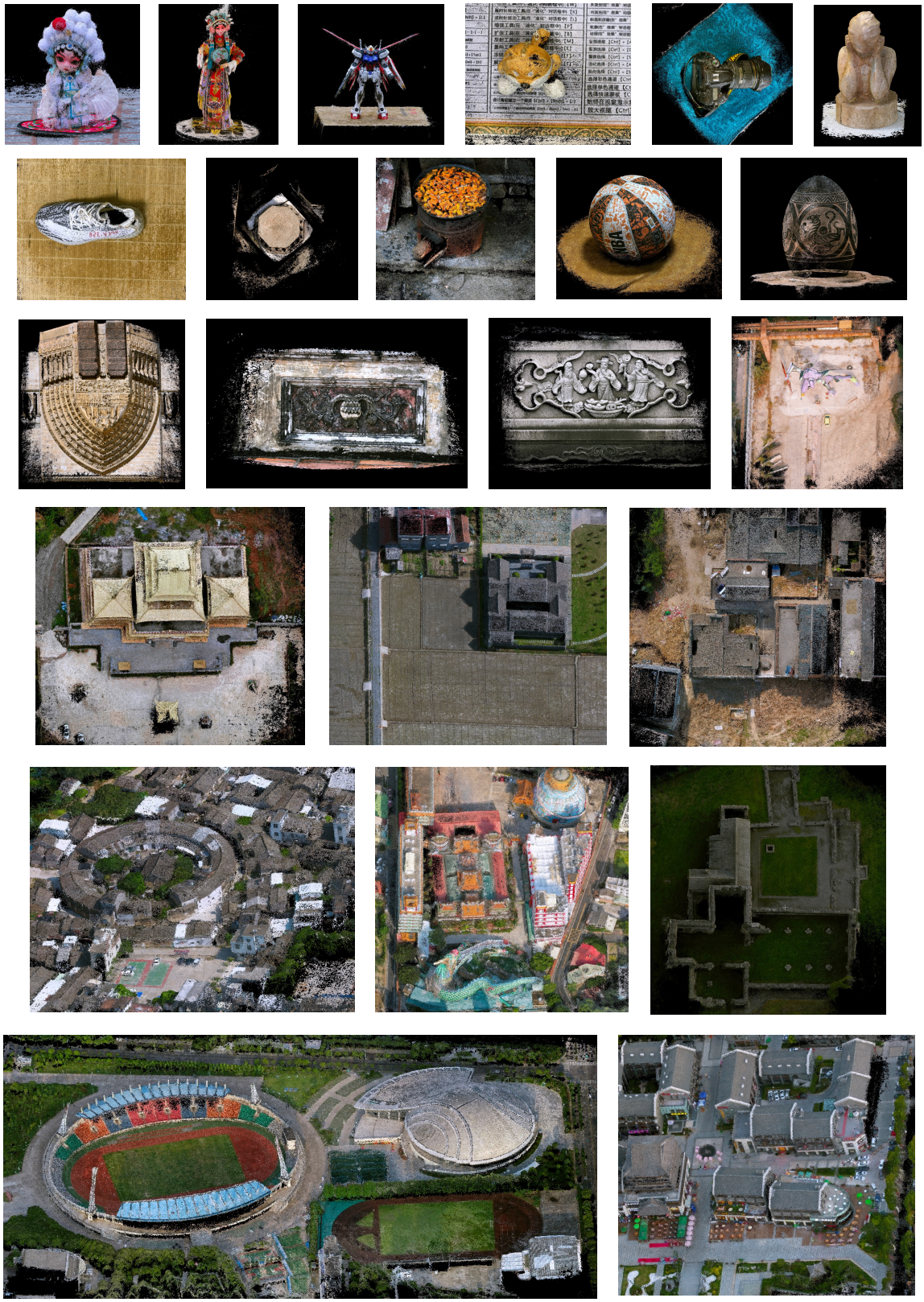

Figure 3. **Reconstruction results on on BlendedMVS [6].**