ObjectGS: Object-aware Scene Reconstruction and Scene Understanding via Gaussian Splatting

Supplementary Material

7. Training Overhead

Table 8 compares training time, FPS, and GPU memory across different instance counts. Even with about 100 instances, overhead remains minimal with efficient parallel rasterizer. Notably, since our one-hot ID encoding is not learnable parameters, it will not significantly increase training overhead. Meanwhile, we can optionally encode only a subset of target instances or leverage category hierarchies, avoiding the waste and inflexibility of fixed-length representations under long-tailed distributions. Therefore, in real applications, our method is both more flexible and scalable.

8. Voting Algorithm

We provide the pseudo code of Algorithms 1 to 3 to clearly demonstrate the proposed voting strategies.

9. More Visualization

We provide more visualization results as shown in Figs. 9 to 14, which includes visualization of OVS segmentation results, panoptic segmentation results, and 3D instance segmentation with point clouds.

Algorithm 1 Object ID Majority Voting

Input:
  Point cloud: $P_{3D} = \{p_1, p_2, \ldots, p_M\}$
  Object ID maps: $L = \{L_1, L_2, \ldots, L_N\}$
  Camera poses: $C = \{C_1, C_2, \ldots, C_N\}$
Initialization:
  labels $= \emptyset$
for each point $p_i \in P_{3D}$ **do**
  for each camera pose $C_j \in C$ **do**
    $x_i =$ Project(p_i, C_j)
    Append $L_j(x_i)$ to labels$[p_i]$
  end for
end for
for each point $p_i \in P_{3D}$ **do**
  if labels$[p_i] \neq \emptyset$ **then**
    frequency(ID) $=$ Counter(labels$[p_i]$)
    ID $= \arg\max$ frequency(ID)
  end if
  Update $p_i = (x_i, y_i, z_i,$ object ID$)$
end for
Output: Updated point cloud P_{3D} with object IDs.

Algorithm 2 Object ID Probability-based Voting

Input:
  Point cloud: $P_{3D} = \{p_1, p_2, \ldots, p_M\}$
  Object ID maps: $L = \{L_1, L_2, \ldots, L_N\}$
  Camera poses: $C = \{C_1, C_2, \ldots, C_N\}$
Initialization:
  labels $= \emptyset$
for each point $p_i \in P_{3D}$ **do**
  for each camera pose $C_j \in C$ **do**
    $x_i =$ Project(p_i, C_j)
    Append $L_j(x_i)$ to labels$[p_i]$
  end for
end for
for each point $p_i \in P_{3D}$ **do**
  if labels$[p_i] \neq \emptyset$ **then**
    frequency(ID) $=$ Counter(labels$[p_i]$)
    ID $=$ Random(Prob $=$ frequency(ID))
  end if
  Update $p_i = (x_i, y_i, z_i,$ object ID$)$
end for
Output: Updated point cloud P_{3D} with object IDs.

Algorithm 3 Object ID Correspondence-based Voting

Input:
  Point cloud: $P_{3D} = \{p_1, p_2, \ldots, p_M\}$
  Object ID maps: $L = \{L_1, L_2, \ldots, L_N\}$
  Correspondences: $C = \{C_1, C_2, \ldots, C_N\}$
Initialization:
  labels $= \emptyset$
for each point $p_i \in P_{3D}$ **do**
  for each correspondence $C_j \in C$ **do**
    $x_i =$ Project(p_i, C_j)
    Append $L_j(x_i)$ to labels$[p_i]$
  end for
end for
for each point $p_i \in P_{3D}$ **do**
  if labels$[p_i] \neq \emptyset$ **then**
    frequency(ID) $=$ Counter(labels$[p_i]$)
    ID $= \arg\max$ frequency(ID)
  end if
  Update $p_i = (x_i, y_i, z_i,$ object ID$)$
end for
Output: Updated point cloud P_{3D} with object IDs.

Table 8. Training time, FPS, and GPU memory comparison

Scene	#Objects	Training time		FPS		GPU memory	
		GS Grouping	Ours	GS Grouping	Ours	GS Grouping	Ours
bed (3DOVS)	7	94 min	72 min	100	80	∼15G	∼10G
sofa (3DOVS))	24	55 min	31 min	110	90	∼18G	∼12G
1ada (ScanNet++)	63	68 min	69 min	90	50	∼40G	∼35G
3e8b (ScanNet++)	80	71 min	113 min	80	40	∼40G	∼45G
0d2e (ScanNet++)	90	73 min	112 min	80	40	∼40G	∼45G

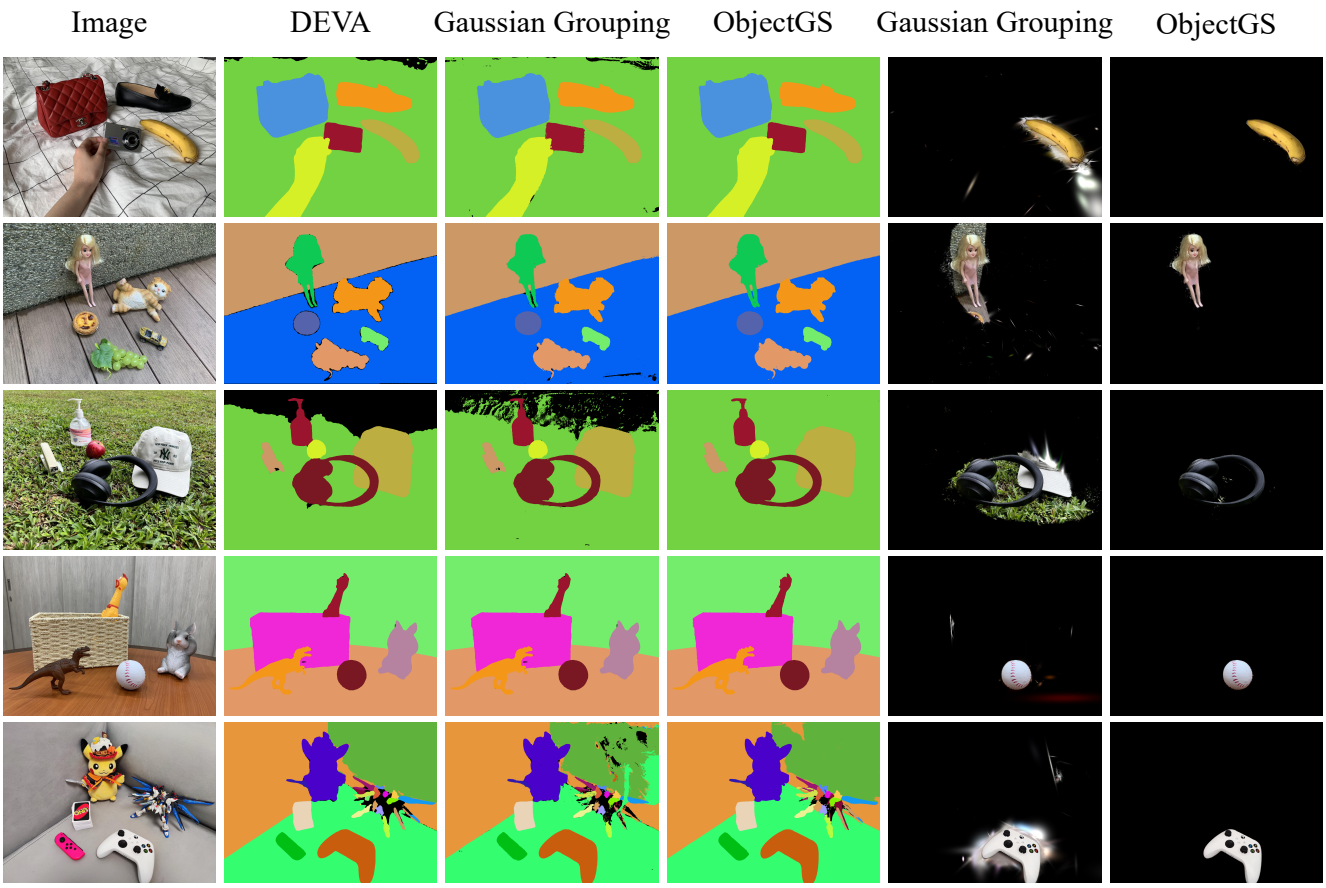

Figure 9. Qualitative comparison of open vocabulary segmentation and 3D object query on the 3DOVS dataset.

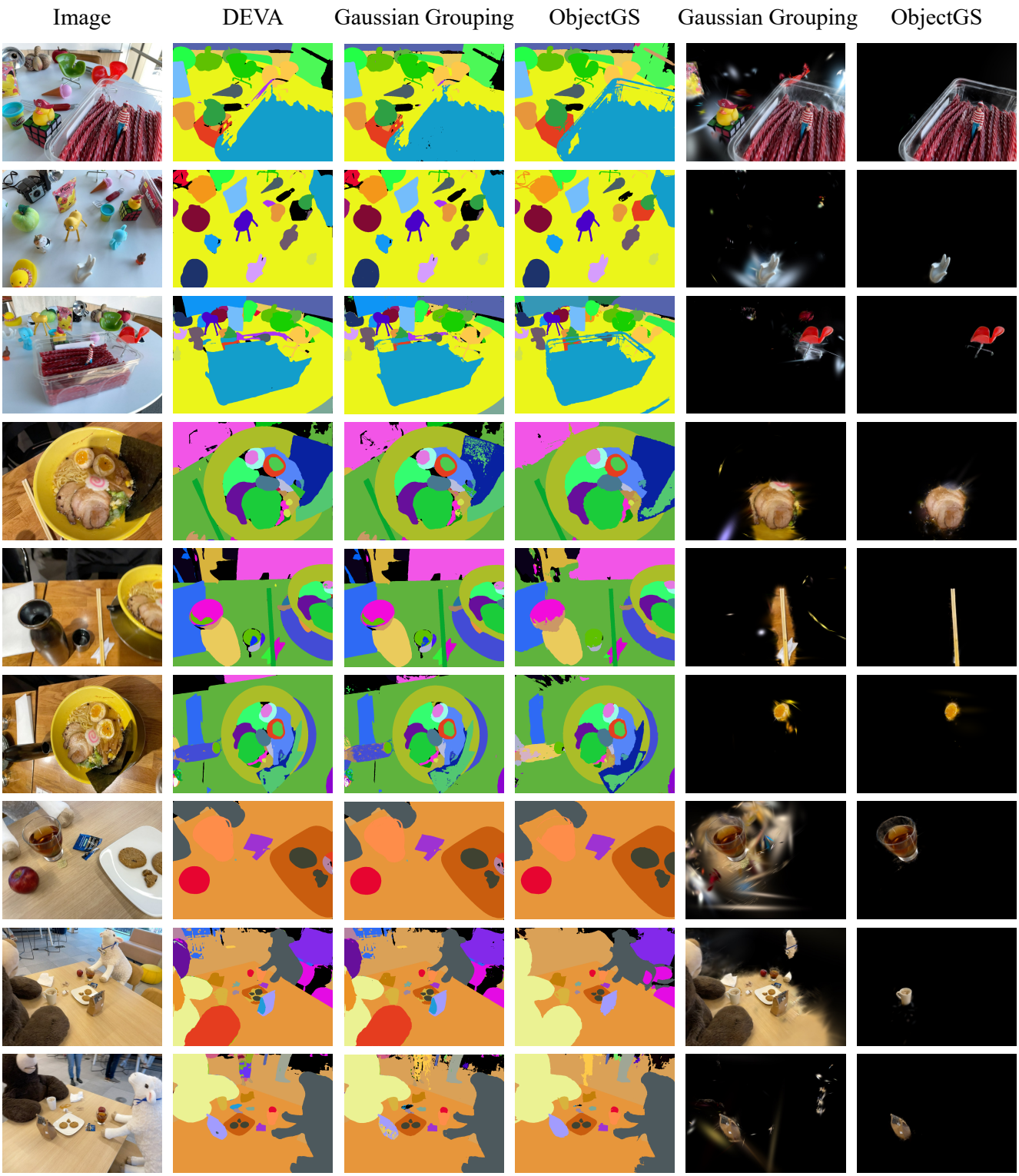

Figure 10. Qualitative comparison of open vocabulary segmentation and 3D object query on the LERF-Mask dataset.

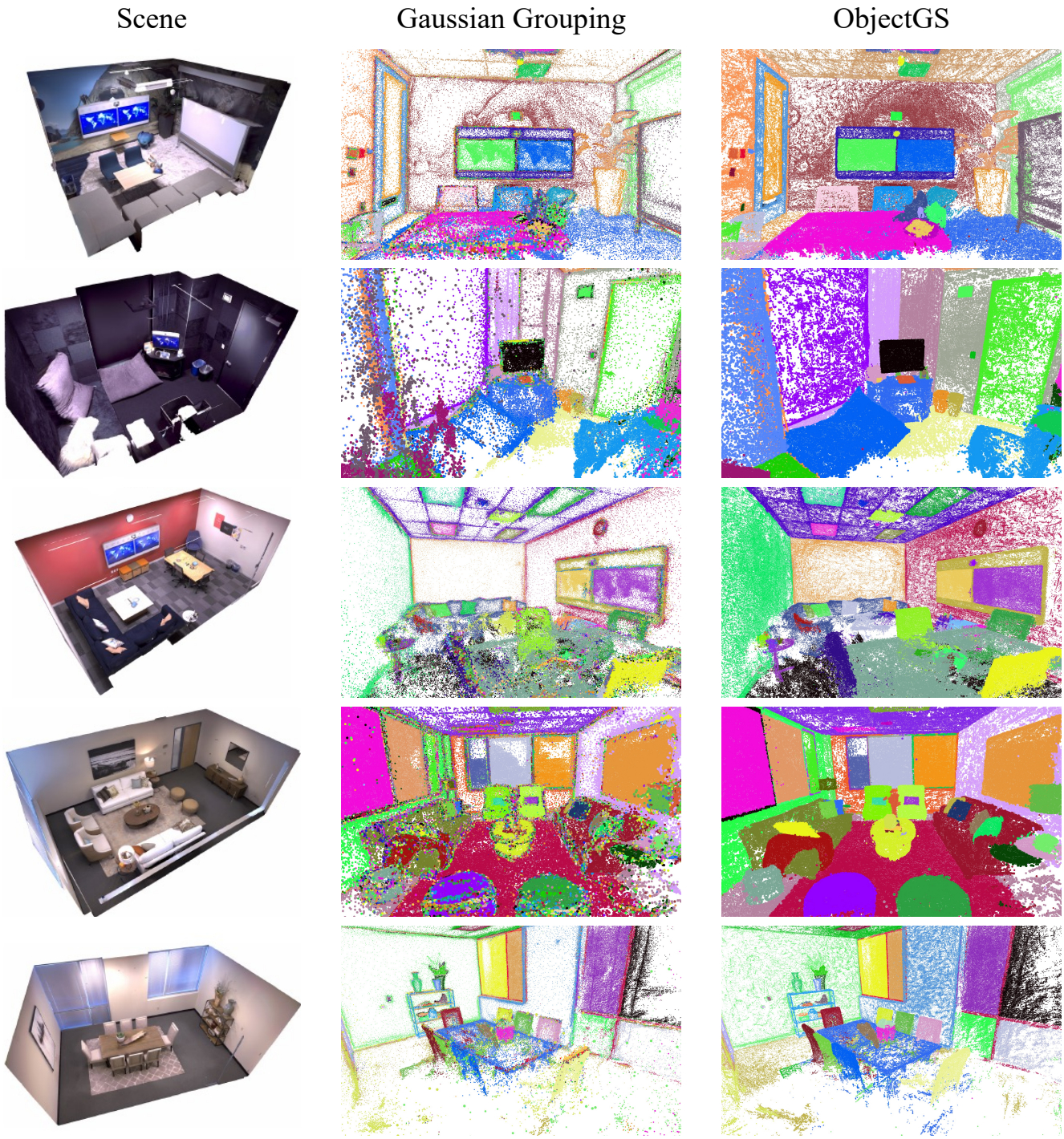

Figure 11. Qualitative comparison of 3D panoptic segmentation on the Replica dataset.

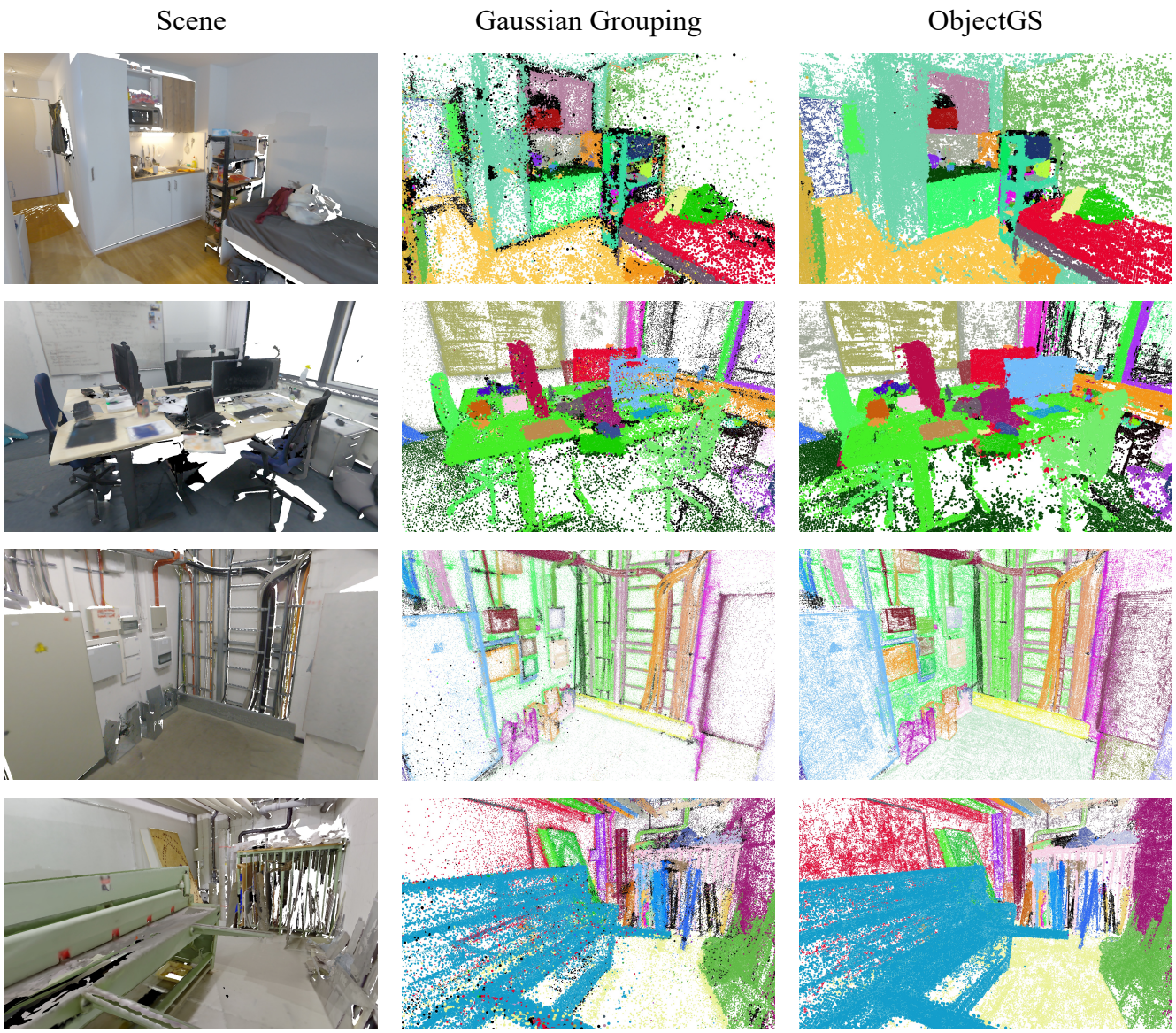

Figure 12. Qualitative comparison of 3D panoptic segmentation on the Scannet++ dataset.

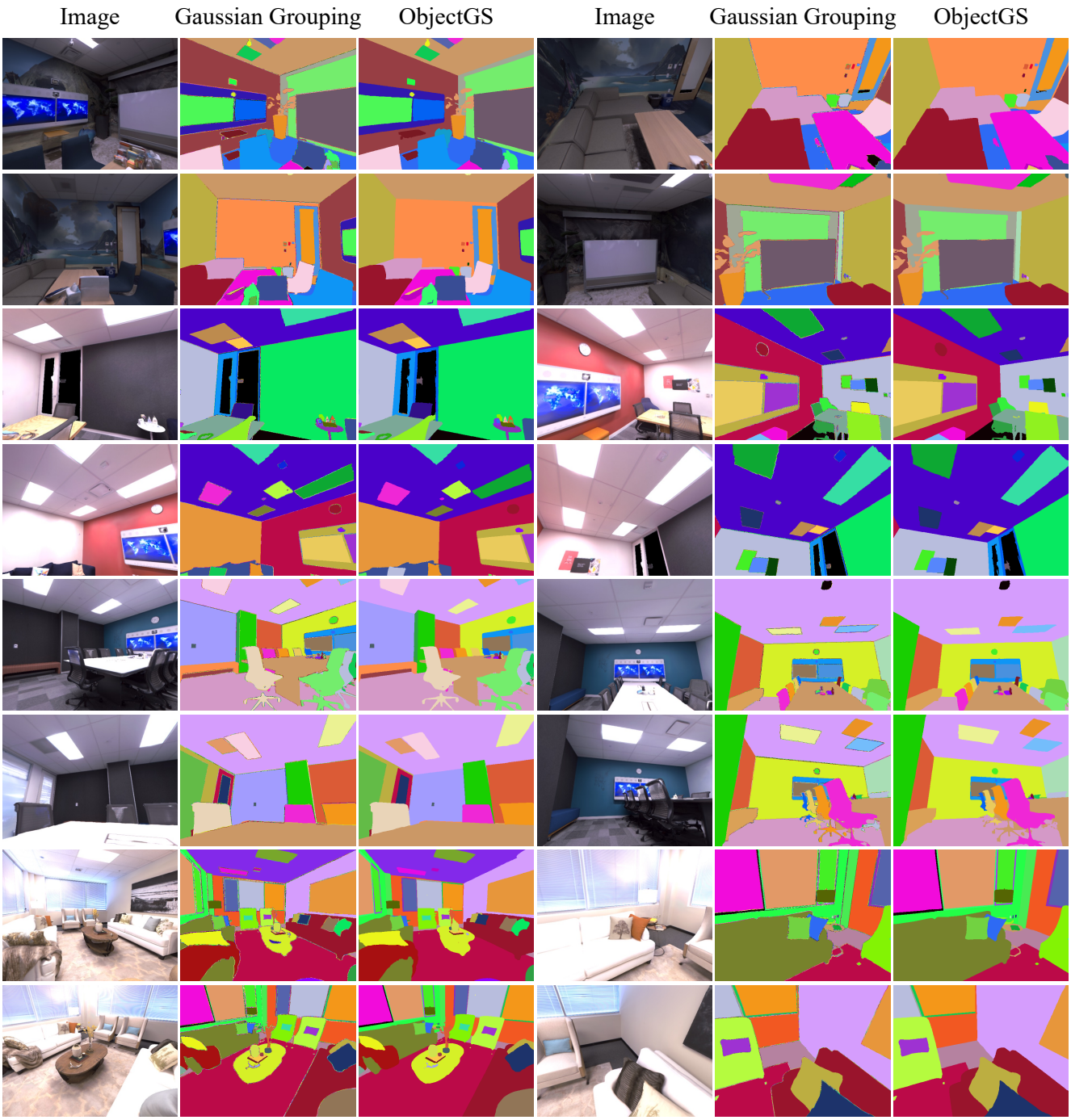

Figure 13. Qualitative comparison of 2D panoptic segmentation on the Replica dataset.

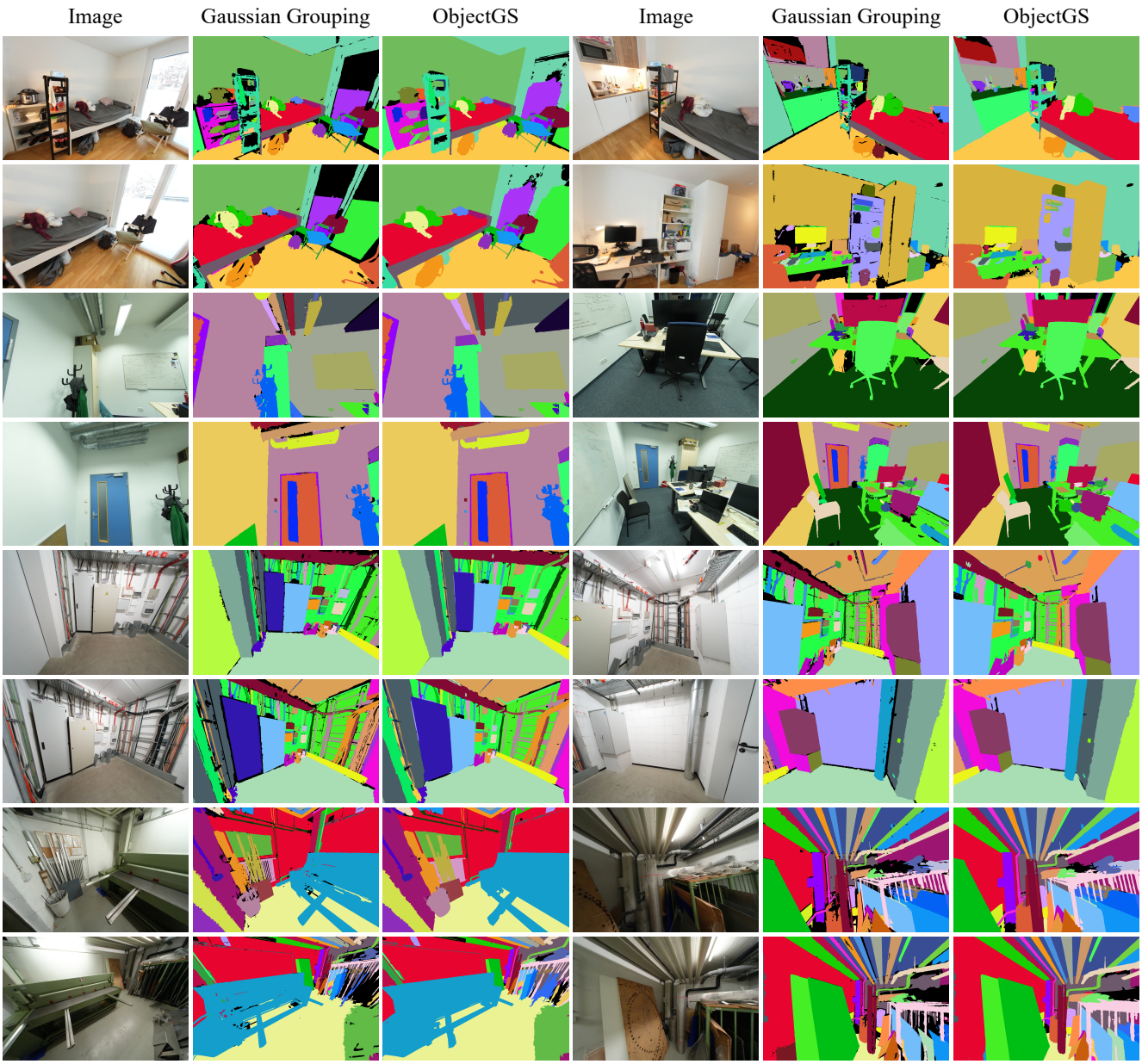

Figure 14. Qualitative comparison of 2D panoptic segmentation on the Scannet++ dataset.