
A APPENDIX

A.1 PRELIMINARIES

We discuss semantic distillation in radiance fields.

Semantic Distillation. Prior work (Zhi et al., 2021; Kobayashi et al., 2022) has explored distilling semantic information from foundation models into NeRFs and GS for open-vocabulary semantic segmentation, object localization, and scene-editing. We limit our discussion to distillation methods that learn a neural field (Kerr et al., 2023), due to their generality to both NeRFs and GS and faster training/inference speeds (Shorinwa et al., 2024a). Neural semantic distillation methods train a semantic field $f : \mathbb{R}^3 \mapsto \mathbb{R}^d$ mapping a 3D point to a d -dimensional semantic embedding. In general, these methods leverage hashgrid encodings (Müller et al., 2022) to learn varying resolutions of semantic detail. The semantic field is trained simultaneously with the radiance field with direct supervision by the ground-truth embeddings from the teacher network, e.g., CLIP/DINO.

A.2 EVALUATION SETUP

We provide additional details on the task and evaluation setup.

Datasets. We consider nine scenes across three benchmark datasets in open-vocabulary semantic localization in radiance fields and a robotics-centric dataset. Specifically, we evaluate the distilled semantic features in the *Ramen*, *Teatime*, and *Waldo_kitchen* scenes in the LERF dataset (Kerr et al., 2023); the *Bed*, *Covered Desk*, and *Table* scenes in the 3D-OVS dataset (Liu et al., 2023); and the *Office*, *Kitchen*, and *Drone* scenes in the robotics dataset (Shorinwa et al., 2025). While the 3D-OVS dataset is limited to primarily small-scale scenes, the LERF and robotics datasets provide broader coverage over both relatively small-scale and large-scale scenes. Moreover, whereas the LERF and 3D-OVS datasets were primarily collected using a smartphone camera, the robotics dataset was collected using cameras onboard a drone and quadruped. We represent each dataset using a unique icon in the subsequent figures.

Implementation. We train a GS and NeRF representation per scene using the camera poses provided with the datasets for 30000 iterations within the Nerfstudio framework (Tancik et al., 2023). To extract visual-geometry semantics, we use the official VGGT checkpoints provided by the authors of (Wang et al., 2025). Likewise, we extract visual-only semantics from DINOv2 (Oquab et al., 2023) and DINOv3 (Siméoni et al., 2025) using their official checkpoints and use the CLIP ResNet model RN50x64 for language semantics. We train the radiance fields on an Nvidia L40 GPU with 48GB VRAM, although we emphasize that in many cases, the GPU memory is not the main compute bottleneck in our implementation. In fact, preliminary experiments were conducted on an Nvidia GeForce RTX 4090 GPU with 24GB VRAM.

Metrics. We evaluate the relative performance between spatially-grounded and visual-only semantics using task-specific metrics. For semantic content analysis, we use the geometric fidelity factor (GFF) metric (see Section 4). For semantic object localization and radiance field inversion, we use the perceptual metrics SSIM, PSNR, and LPIPS (see Section 5), and the SE(3) error with rotation in degrees and translation in meters (see Section 6), respectively. We compute these metrics across 100 novel (unseen) camera poses in most scenes. In the 3D-OVS dataset, we use the training camera poses due to the limited scene coverage.

A.3 ADDITIONAL SEMANTIC CONTENT RESULTS

Below, we include visualizations for the PCA and segmentation experiments, sampling one or two camera views from each scene.

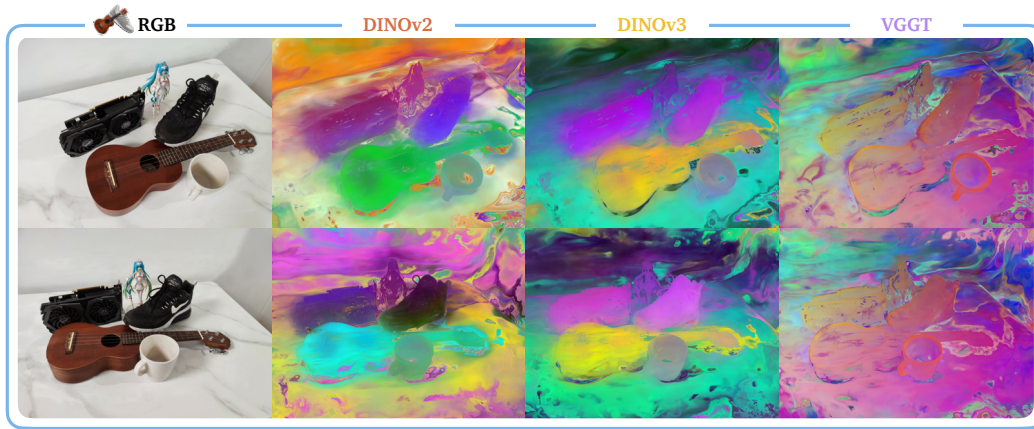


Figure 7: Semantic content of distilled features visualized for GS for the 3D-OVS table scene.

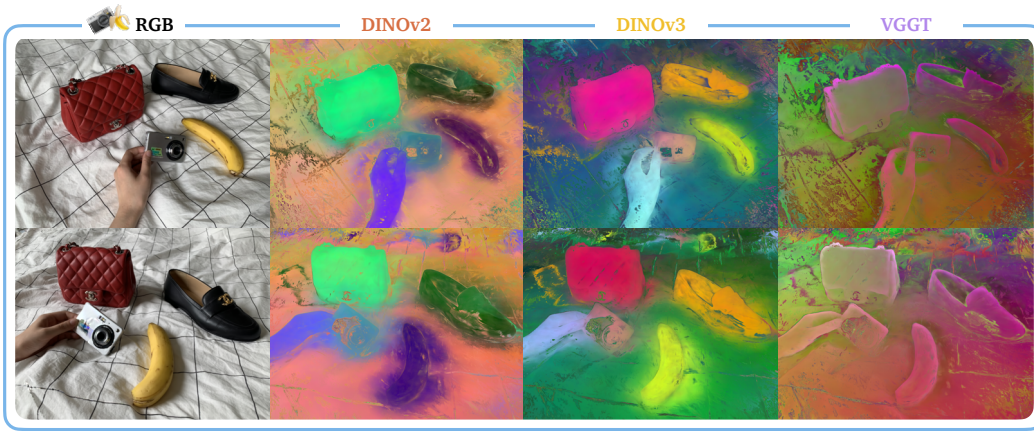


Figure 8: Semantic content of distilled features visualized for GS for the 3D-OVS bed scene.

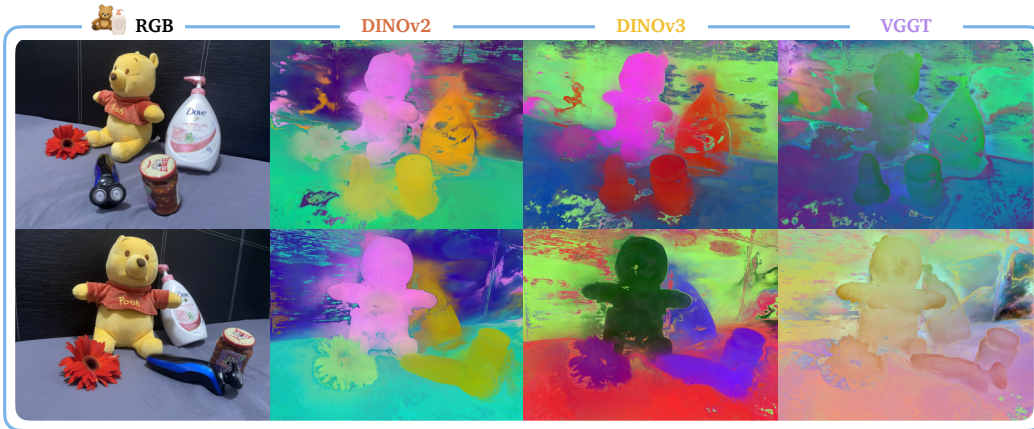


Figure 9: Semantic content of distilled features visualized for GS for the 3D-OVS covered desk scene.

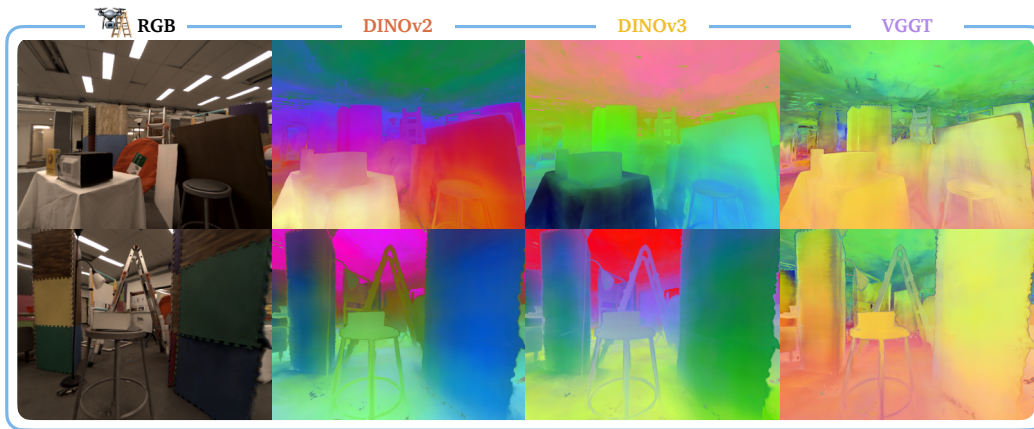


Figure 10: Semantic content of distilled features visualized for GS for the drone scene.

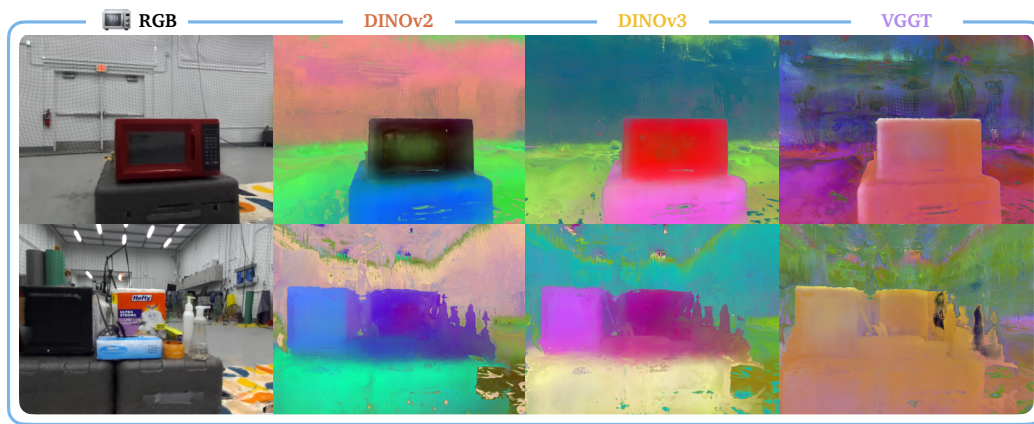


Figure 11: Semantic content of distilled features visualized for GS for the quadruped kitchen scene.

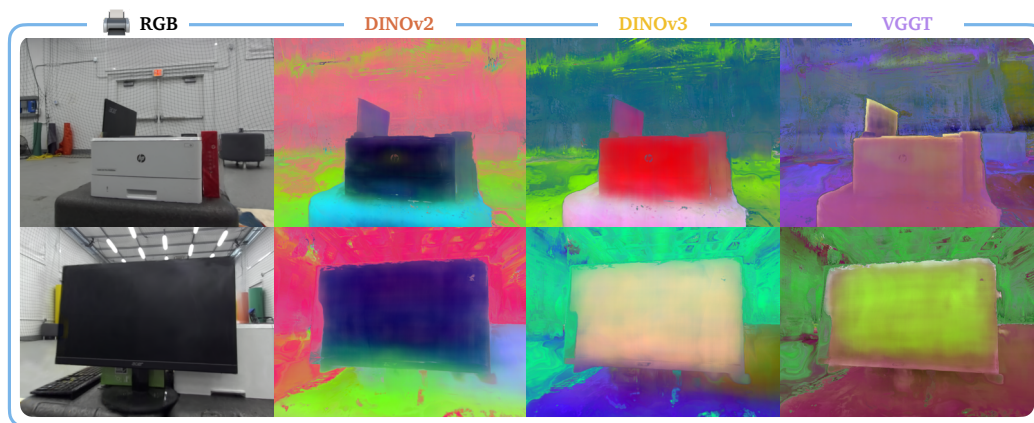


Figure 12: Semantic content of distilled features visualized for GS for the quadruped office scene.

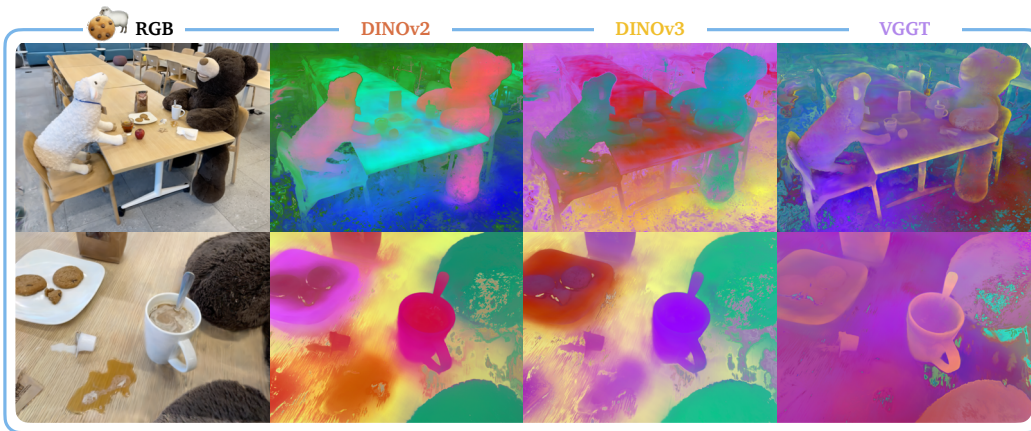


Figure 13: Semantic content of distilled features visualized for GS for the LERF teatime scene.



Figure 14: Semantic content of distilled features visualized for GS for the LERF ramen scene.

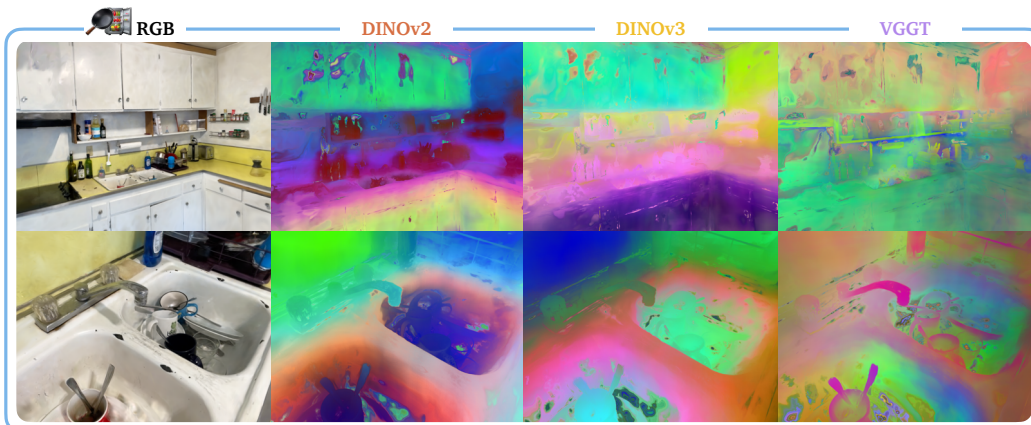


Figure 15: Semantic content of distilled features visualized for GS for the LERF kitchen scene.

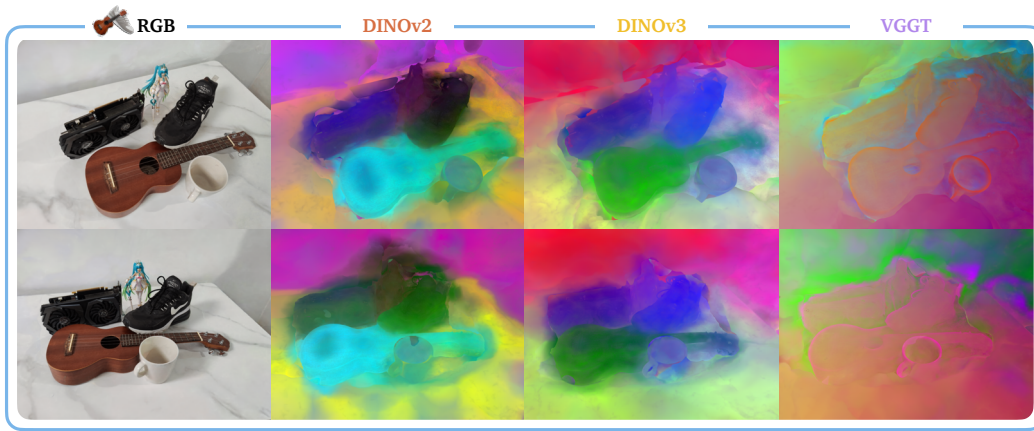


Figure 16: Semantic content of distilled features visualized for NERF for the 3D-OVS table scene.

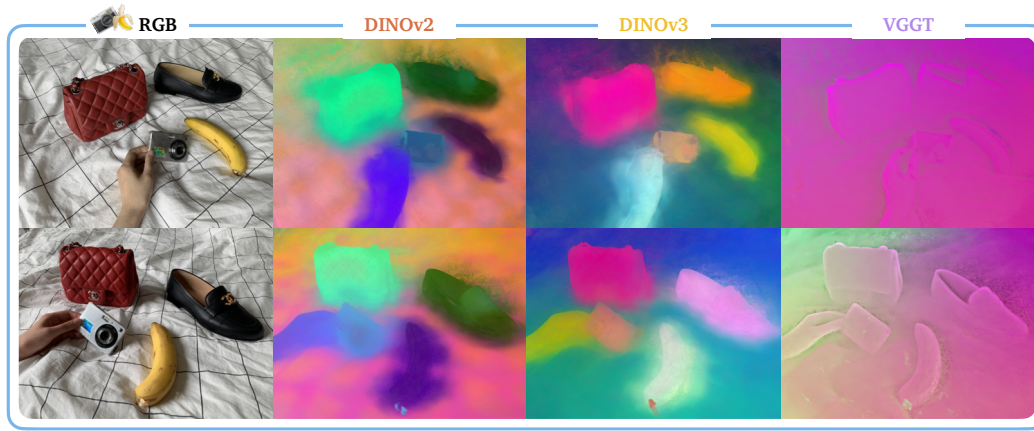


Figure 17: Semantic content of distilled features visualized for NERF for the 3D-OVS bed scene.

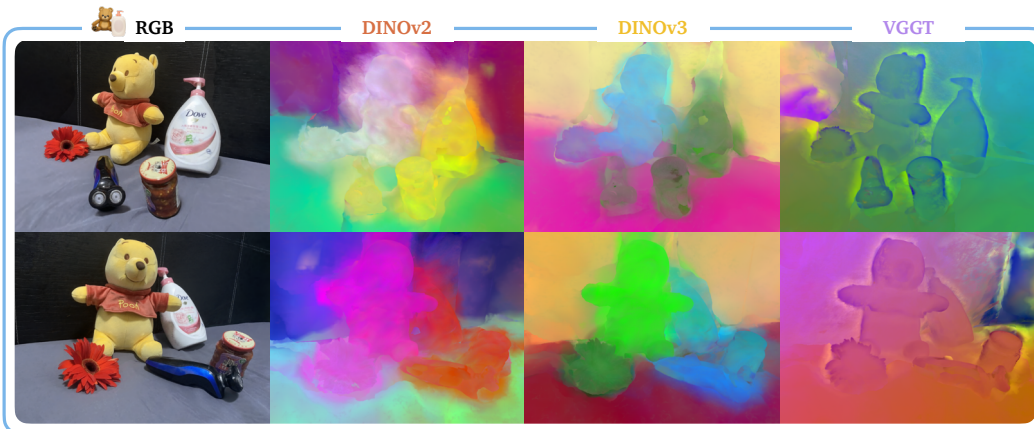


Figure 18: Semantic content of distilled features visualized for NERF for the 3D-OVS covered desk scene.

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

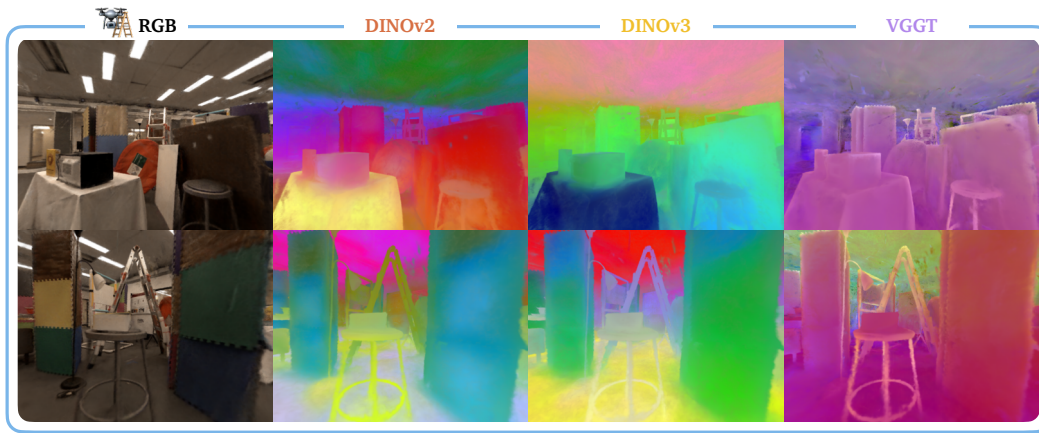


Figure 19: Semantic content of distilled features visualized for NERF for the drone scene.



Figure 20: Semantic content of distilled features visualized for NERF for the quadruped kitchen scene.

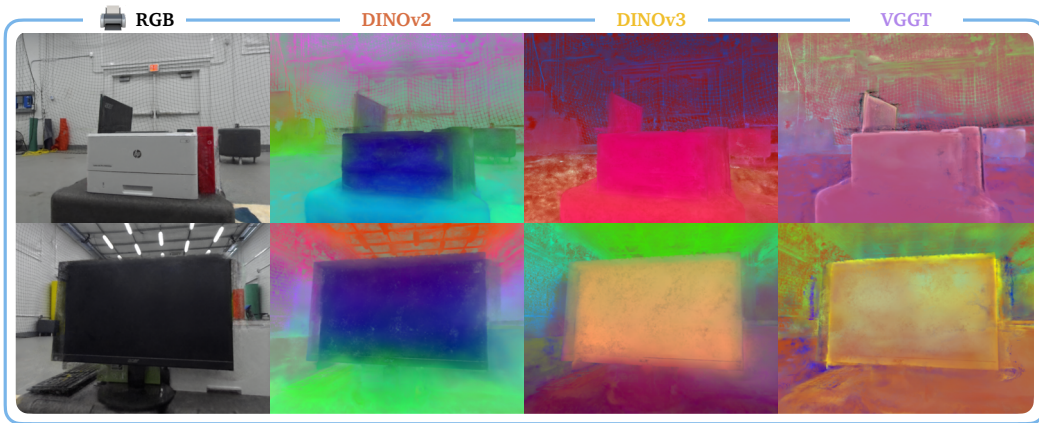


Figure 21: Semantic content of distilled features visualized for NERF for the quadruped office scene.

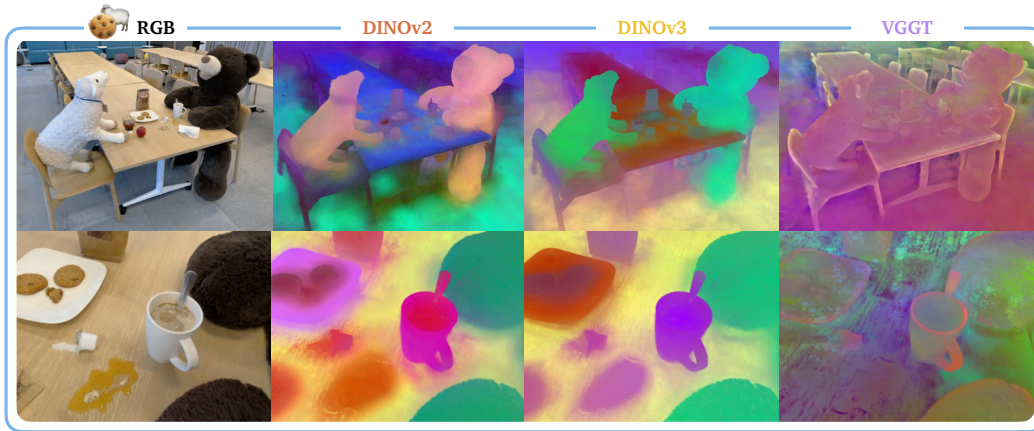


Figure 22: Semantic content of distilled features visualized for NERF for the LERF teatime scene.

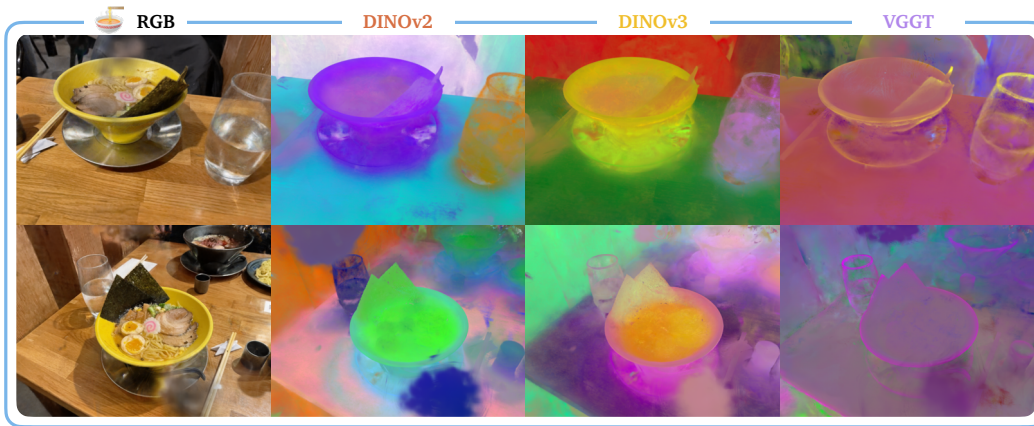


Figure 23: Semantic content of distilled features visualized for NERF for the LERF ramen scene.

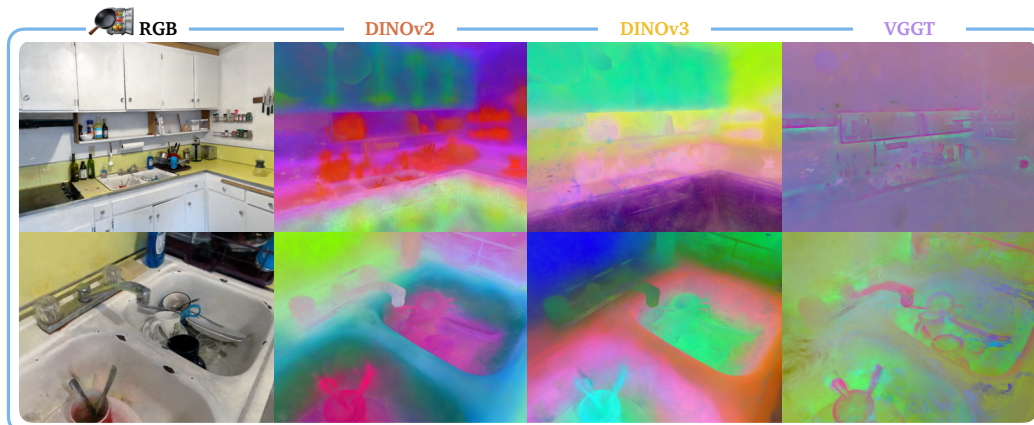


Figure 24: Semantic content of distilled features visualized for NERF for the LERF kitchen scene.

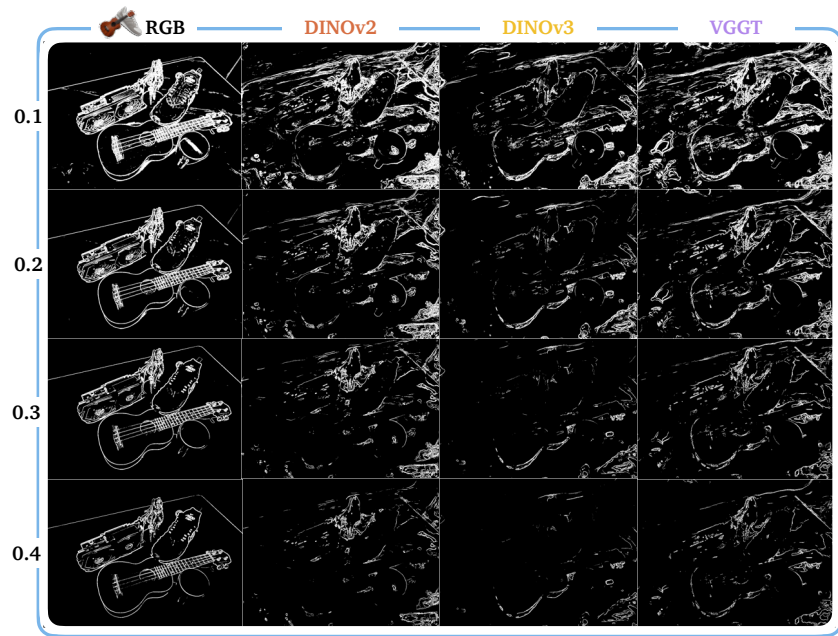


Figure 25: GFF of visual-geometry and visual-only features visualized for GS for the 3D-OVS table scene.

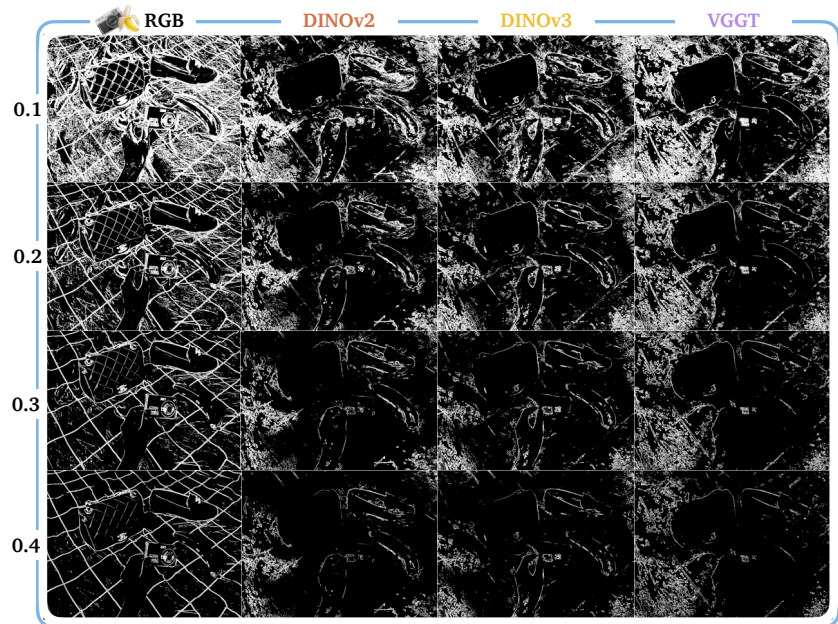


Figure 26: GFF of visual-geometry and visual-only features visualized for GS for the 3D-OVS bed scene.

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

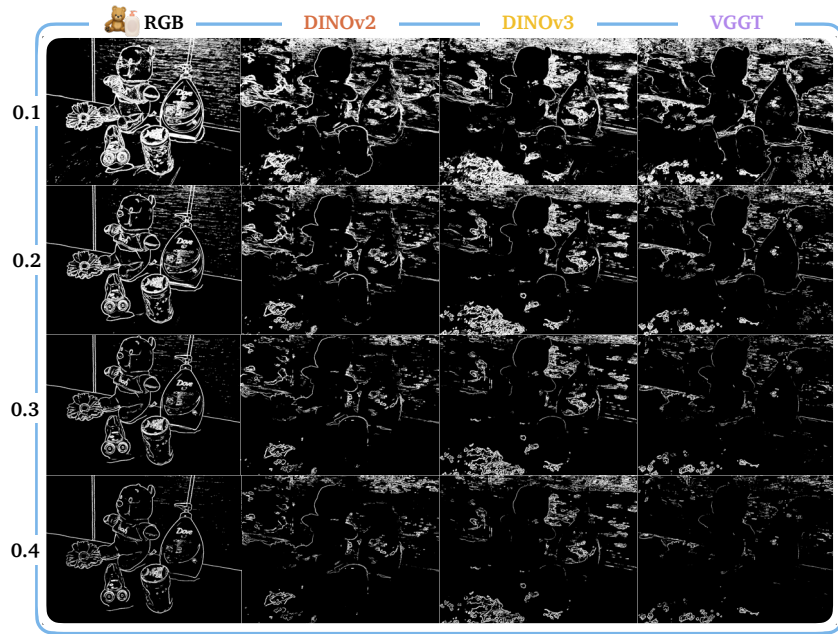


Figure 27: GFF of visual-geometry and visual-only features visualized for GS for the 3D-OVS desk scene.

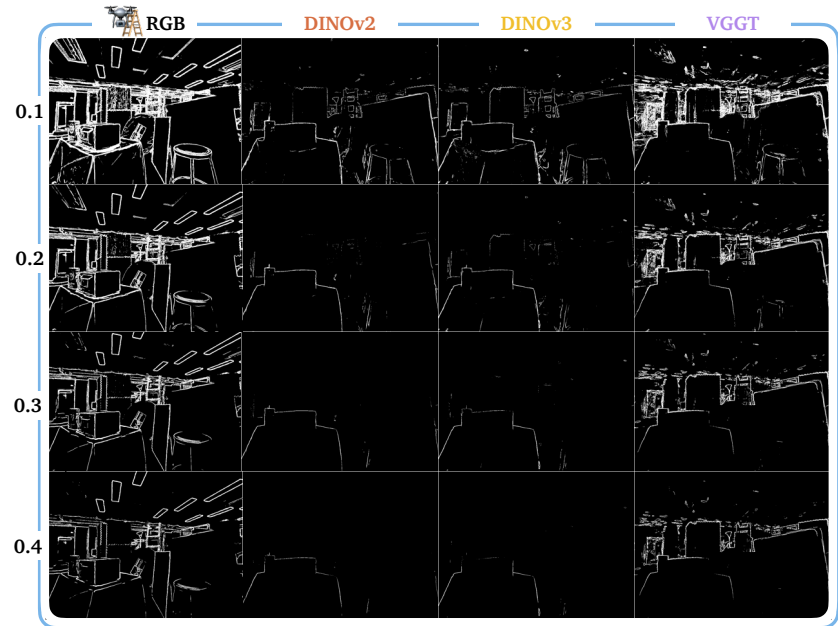


Figure 28: GFF of visual-geometry and visual-only features visualized for GS for the drone scene.

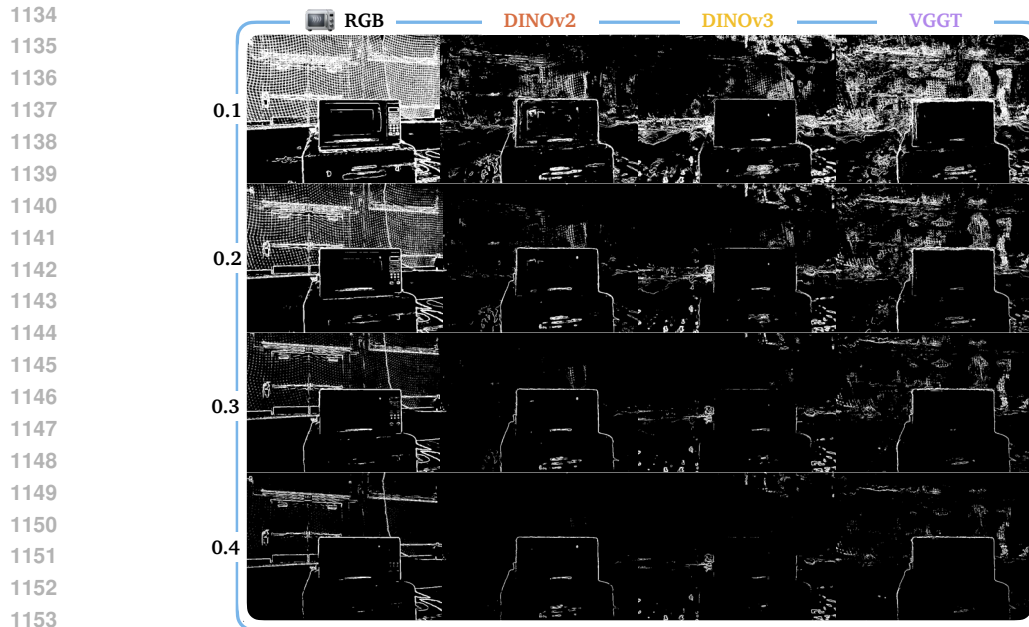


Figure 29: GFF of visual-geometry and visual-only features visualized for GS for the quadruped kitchen scene.

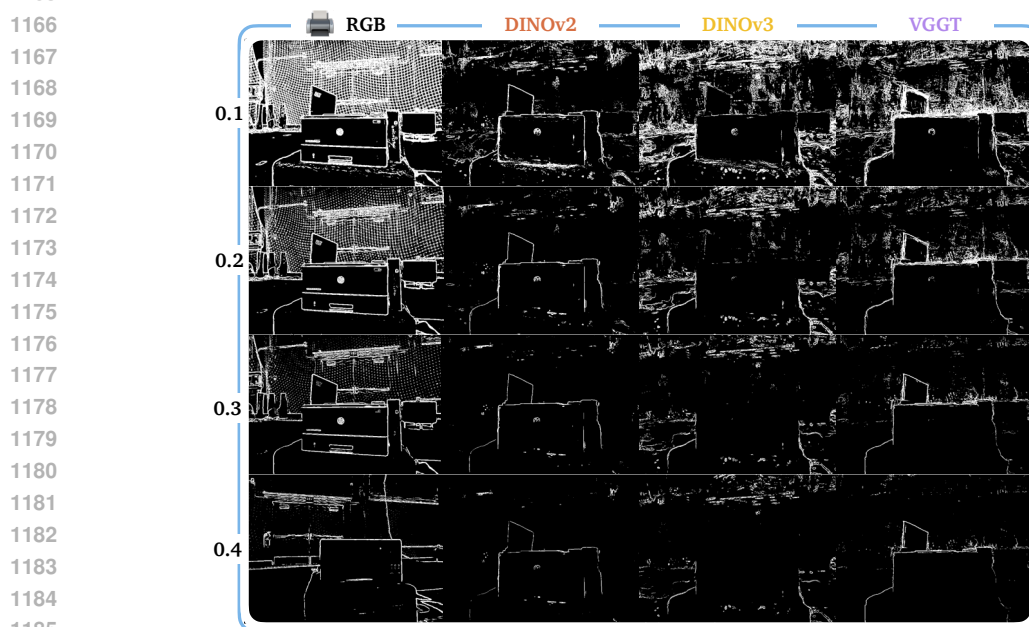


Figure 30: GFF of visual-geometry and visual-only features visualized for GS for the quadruped office scene.

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

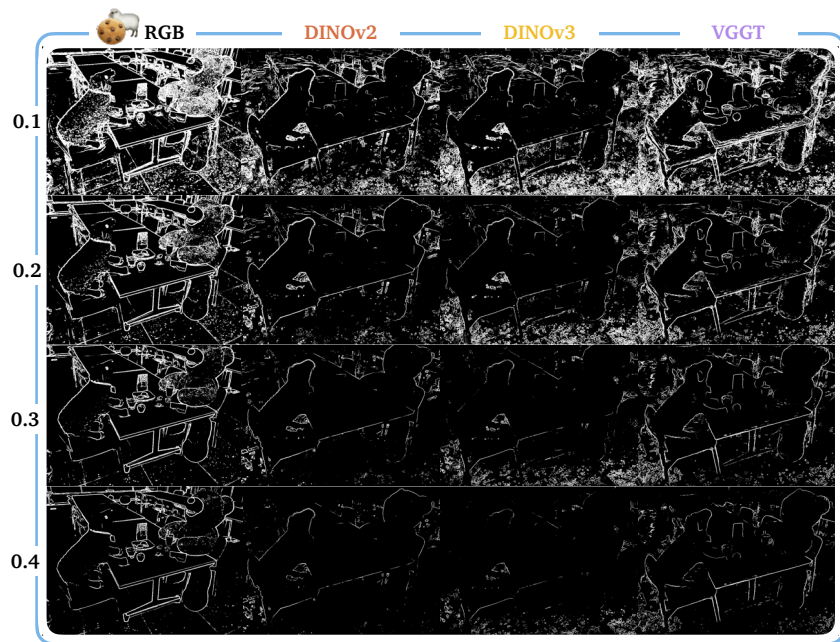


Figure 31: GFF of visual-geometry and visual-only features visualized for GS for the LERF teatime scene.

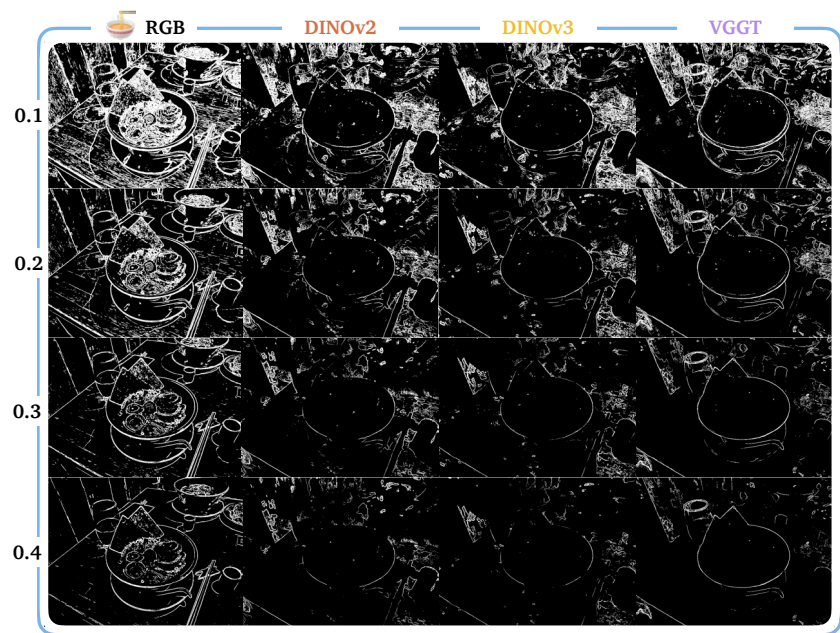


Figure 32: GFF of visual-geometry and visual-only features visualized for GS for the LERF ramen scene.

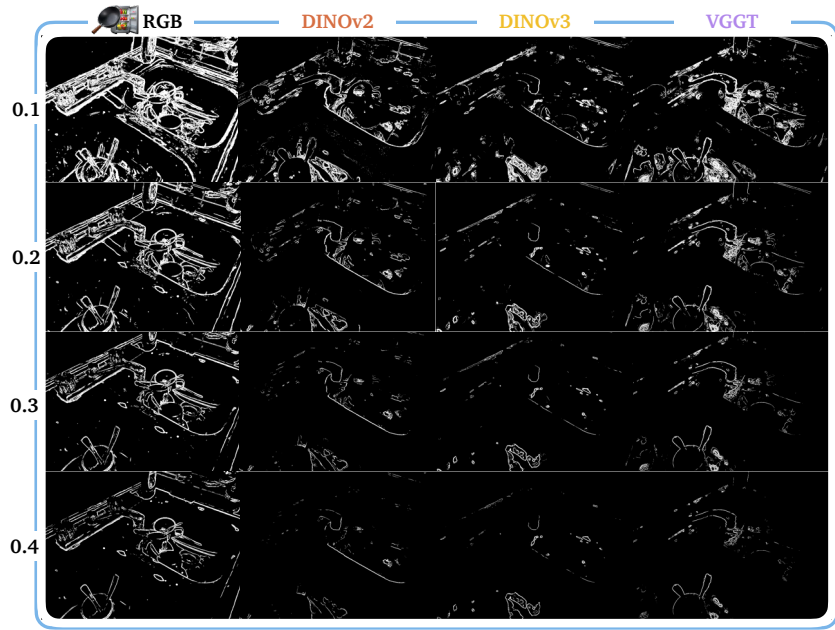


Figure 33: GFF of visual-geometry and visual-only features visualized for GS for the LERF kitchen scene.

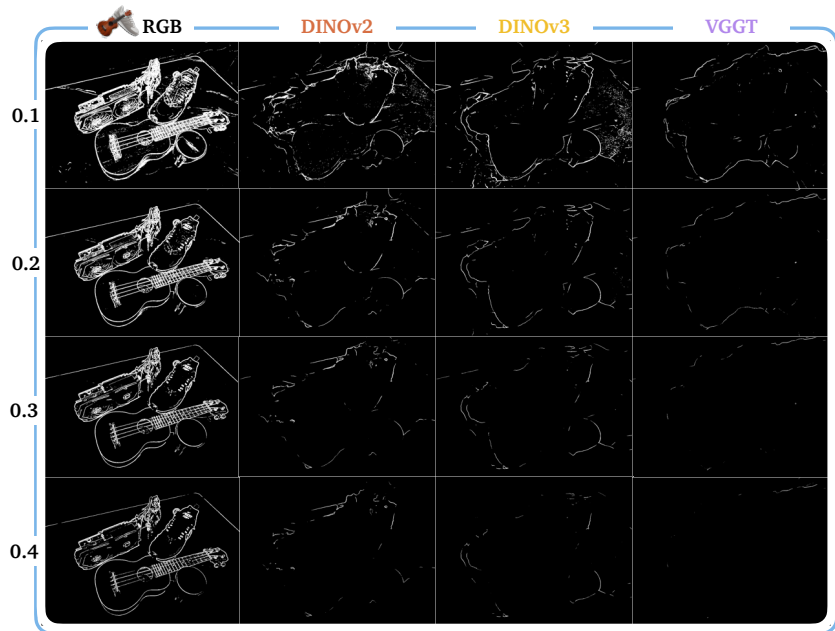


Figure 34: GFF of visual-geometry and visual-only features visualized for NERF for the 3D-OVS table scene.

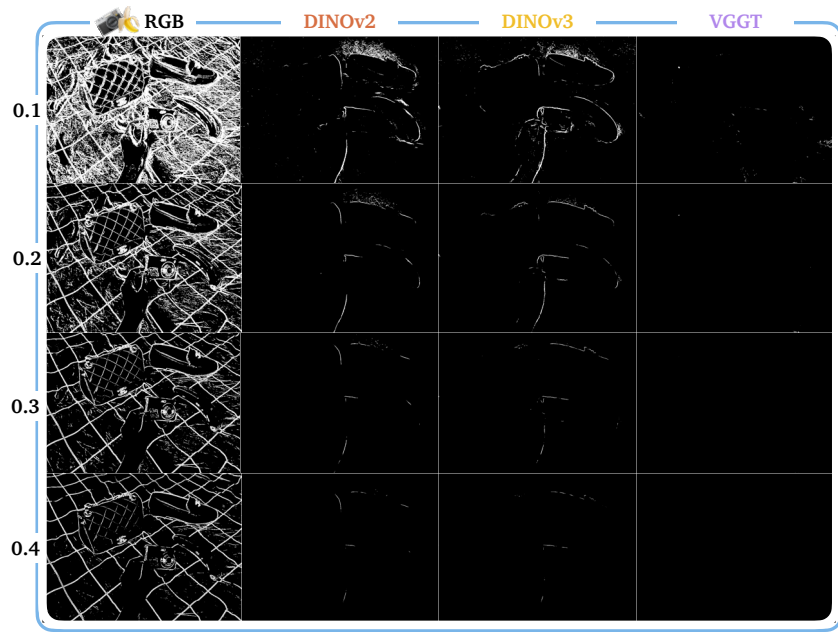


Figure 35: GFF of visual-geometry and visual-only features visualized for NERF for the 3D-OVS bed scene.

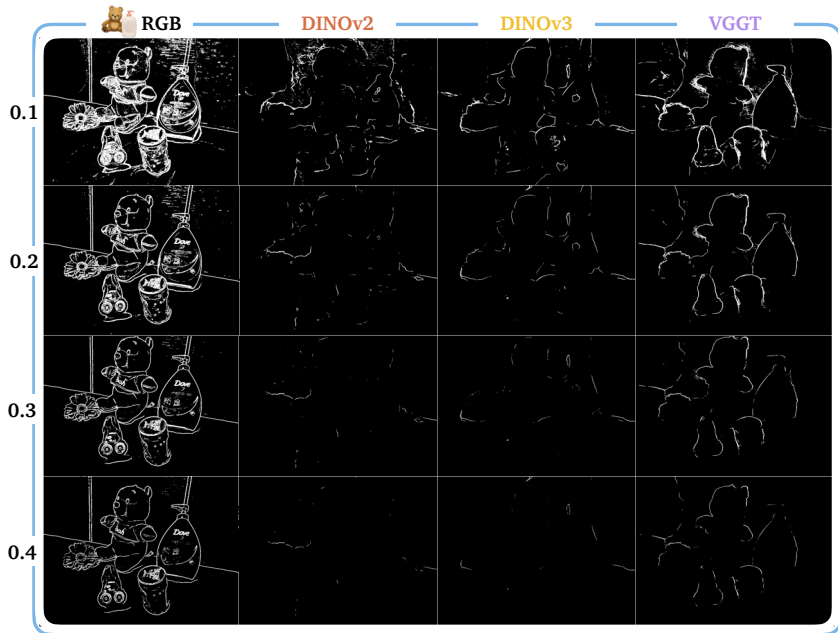


Figure 36: GFF of visual-geometry and visual-only features visualized for NERF for the 3D-OVS desk scene.

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403

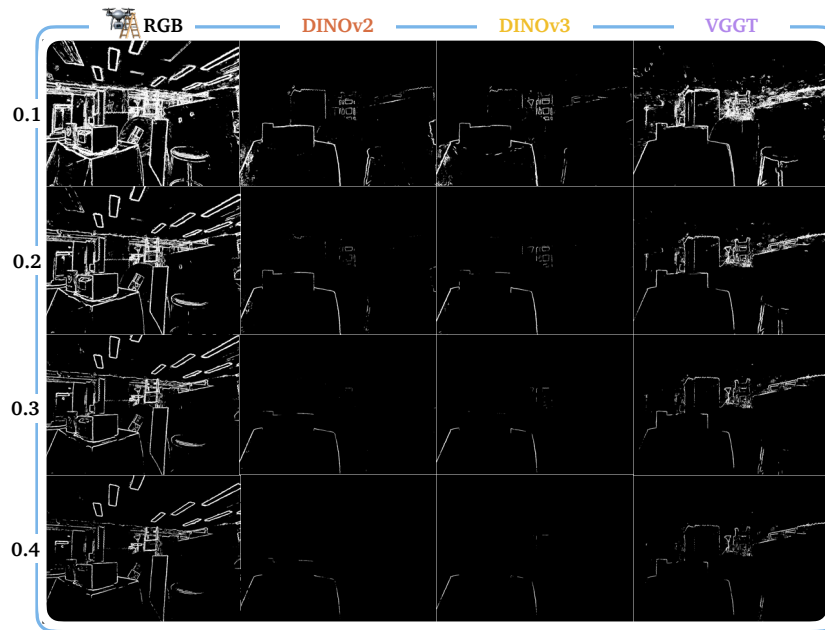


Figure 37: GFF of visual-geometry and visual-only features visualized for NERF for the drone scene.

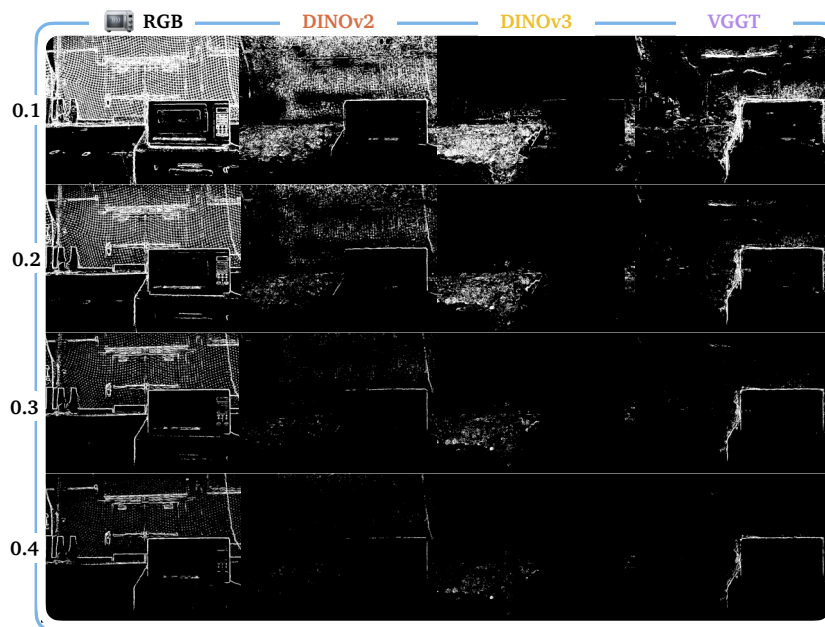


Figure 38: GFF of visual-geometry and visual-only features visualized for NERF for the quadruped kitchen scene.

1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

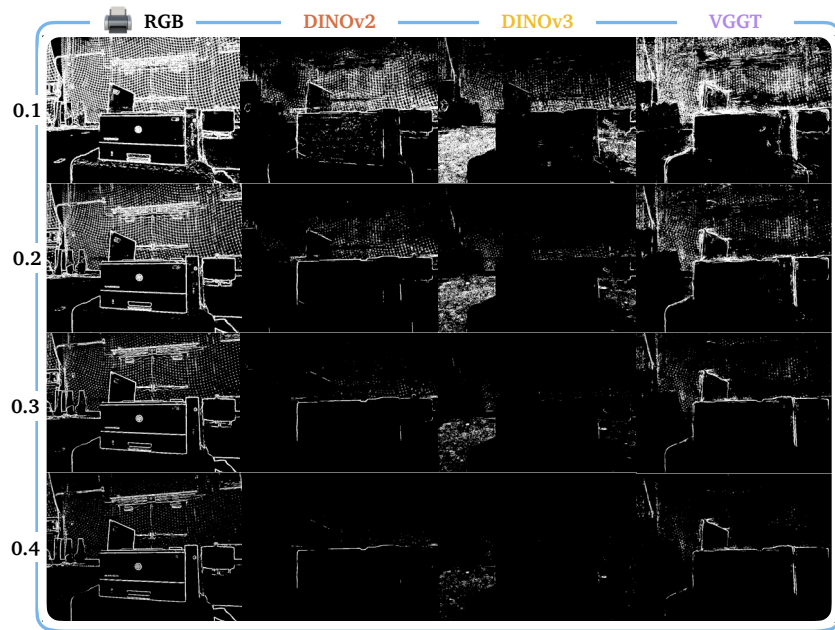


Figure 39: GFF of visual-geometry and visual-only features visualized for NERF for the quadruped office scene.

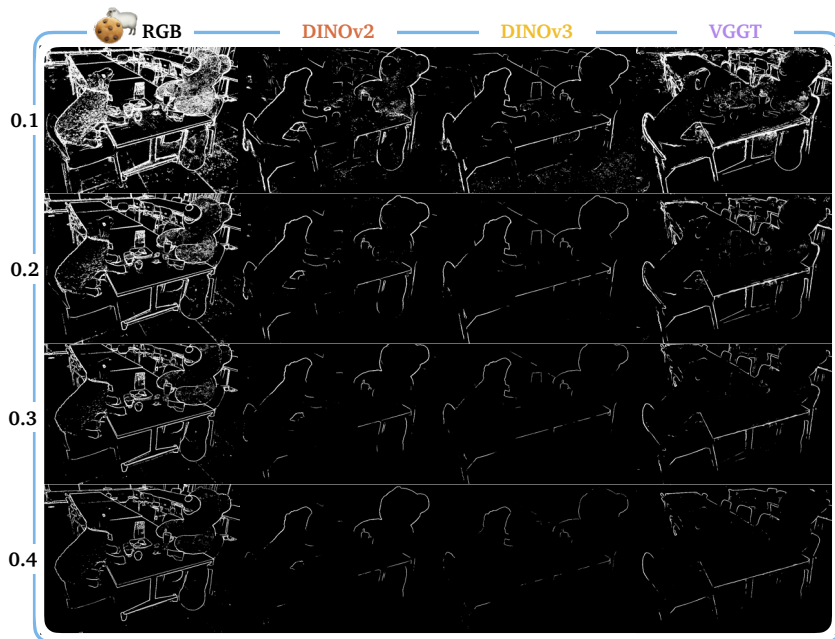


Figure 40: GFF of visual-geometry and visual-only features visualized for NERF for the LERF teatime scene.

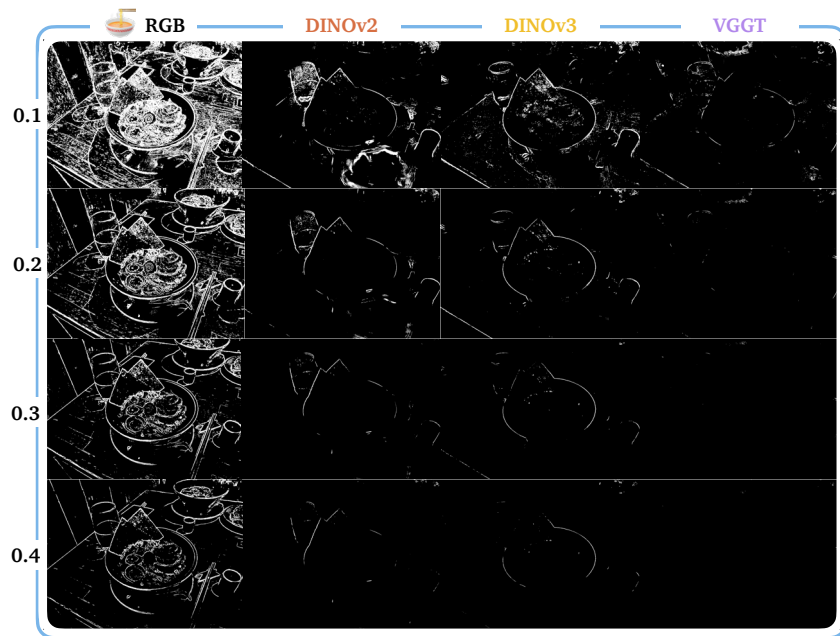


Figure 41: GFF of visual-geometry and visual-only features visualized for NERF for the LERF ramen scene.

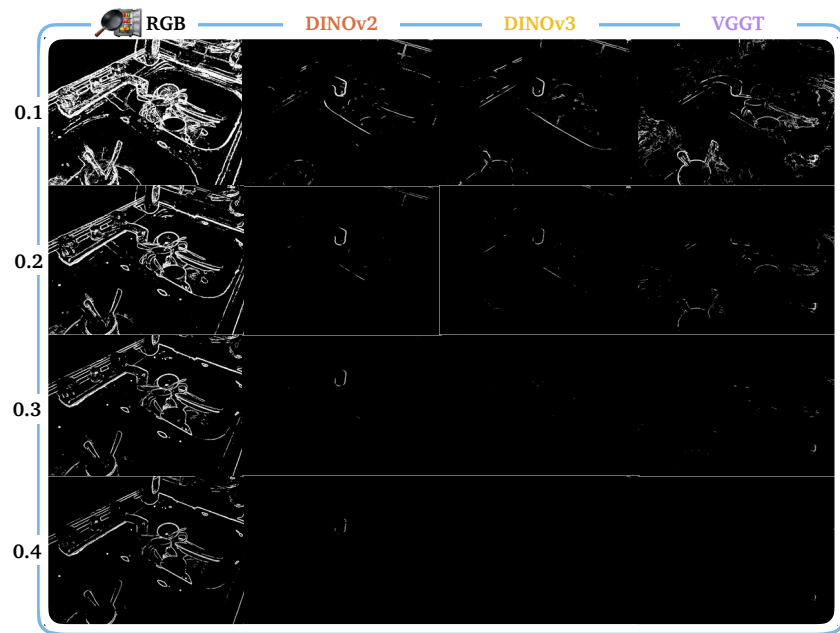


Figure 42: GFF of visual-geometry and visual-only features visualized for NERF for the LERF kitchen scene.

A.4 ADDITIONAL SEMANTIC LOCALIZATION RESULTS

Below, we include semantic localization visualizations, sampling one or two camera views from each scene.

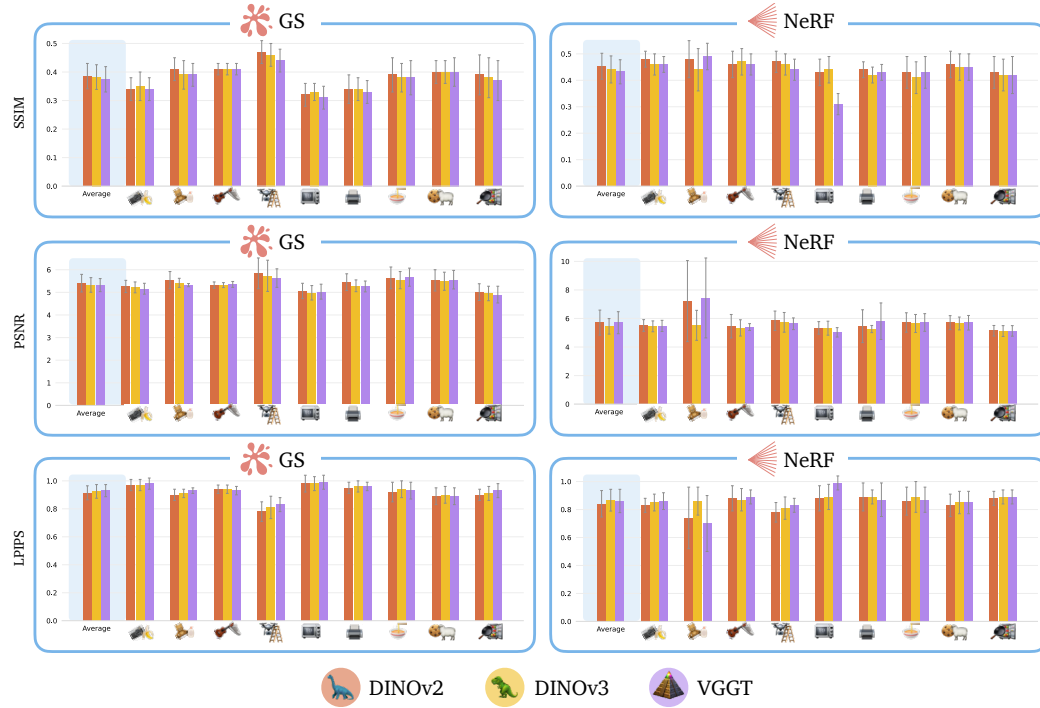


Figure 43: Quantitative results for the three metrics (SSIM, PSNR, LPIPS) of semantic object localization across each scene.

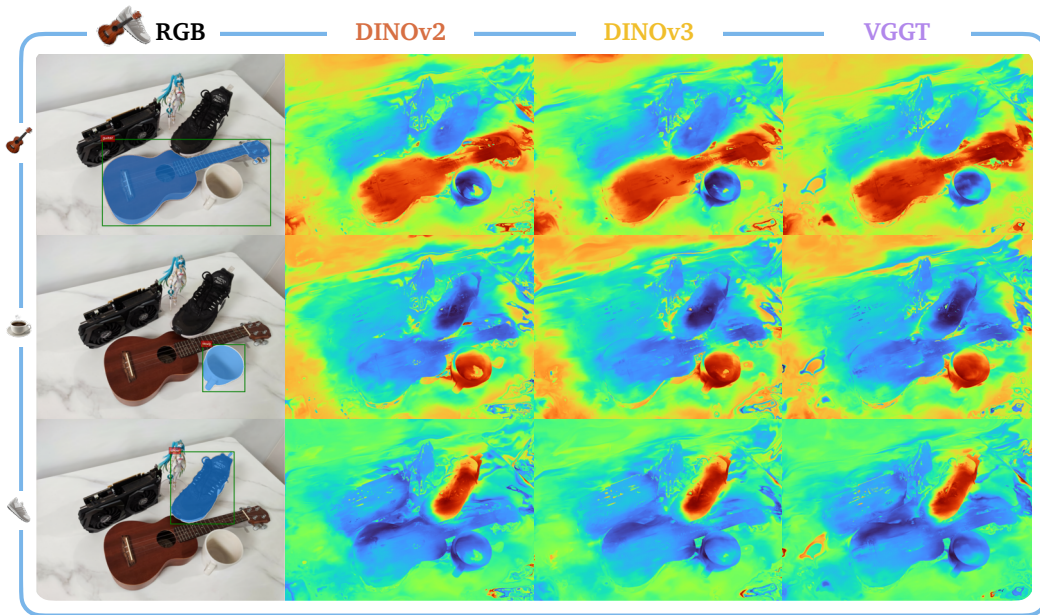


Figure 44: Semantic object localization visualized for GS for the 3D-OVS table scene.

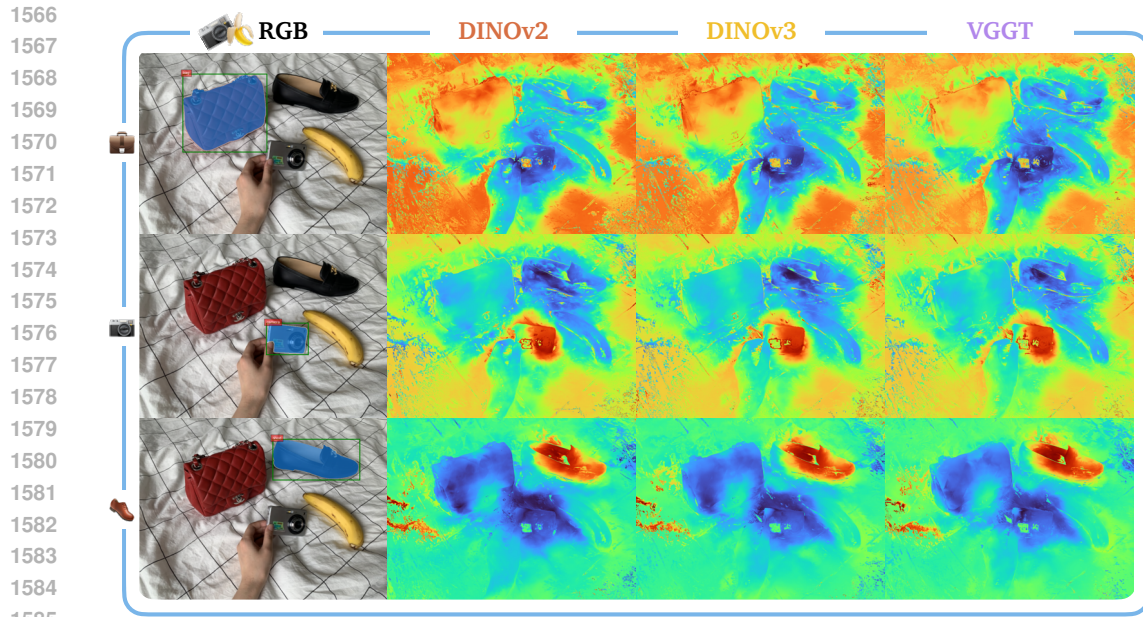


Figure 45: Semantic object localization visualized for GS for the 3D-OVS bed scene.

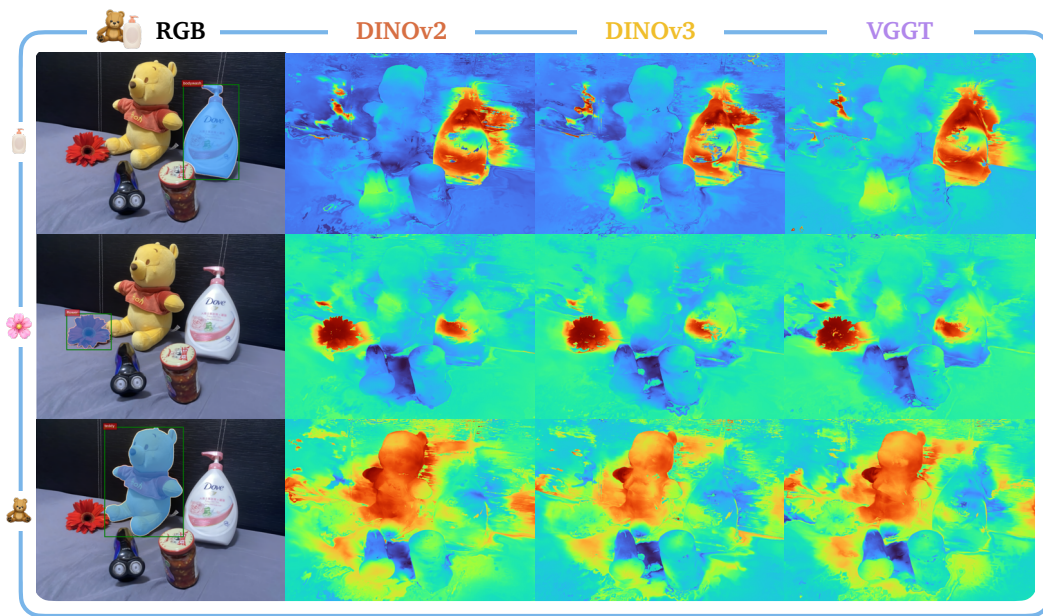


Figure 46: Semantic object localization visualized for GS for the 3D-OVS covered desk scene.

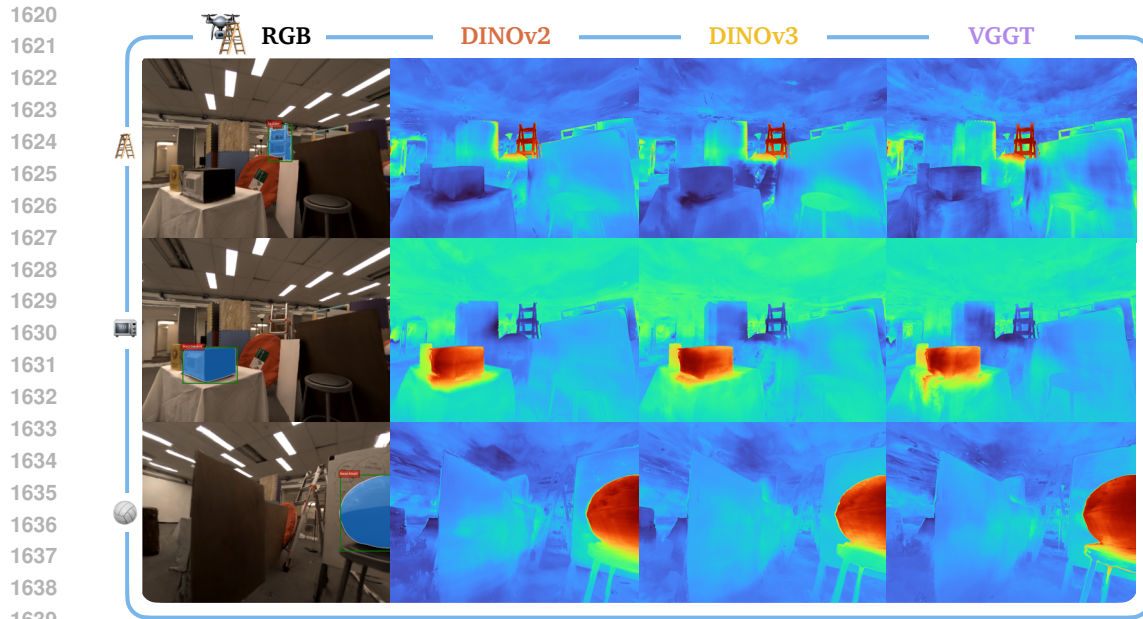


Figure 47: Semantic object localization visualized for GS for the drone scene.

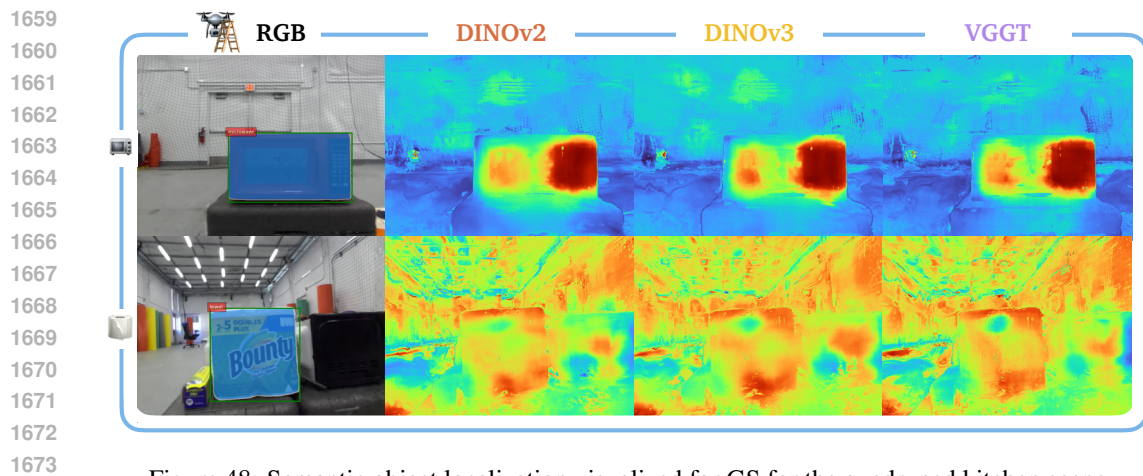


Figure 48: Semantic object localization visualized for GS for the quadruped kitchen scene.

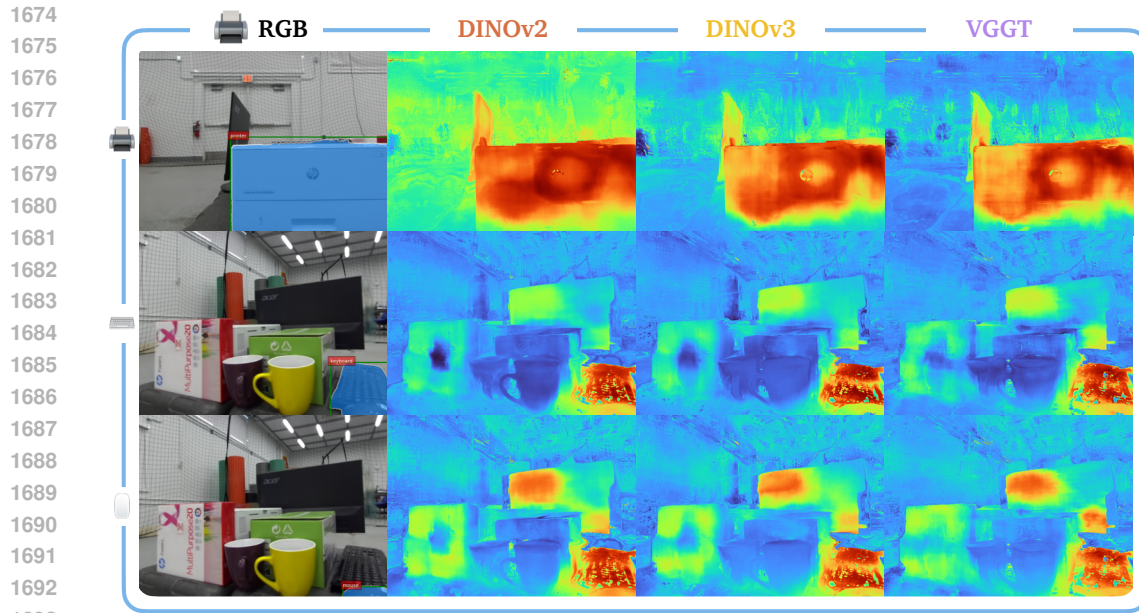


Figure 49: Semantic object localization visualized for GS for the 3D-OVS bed scene.

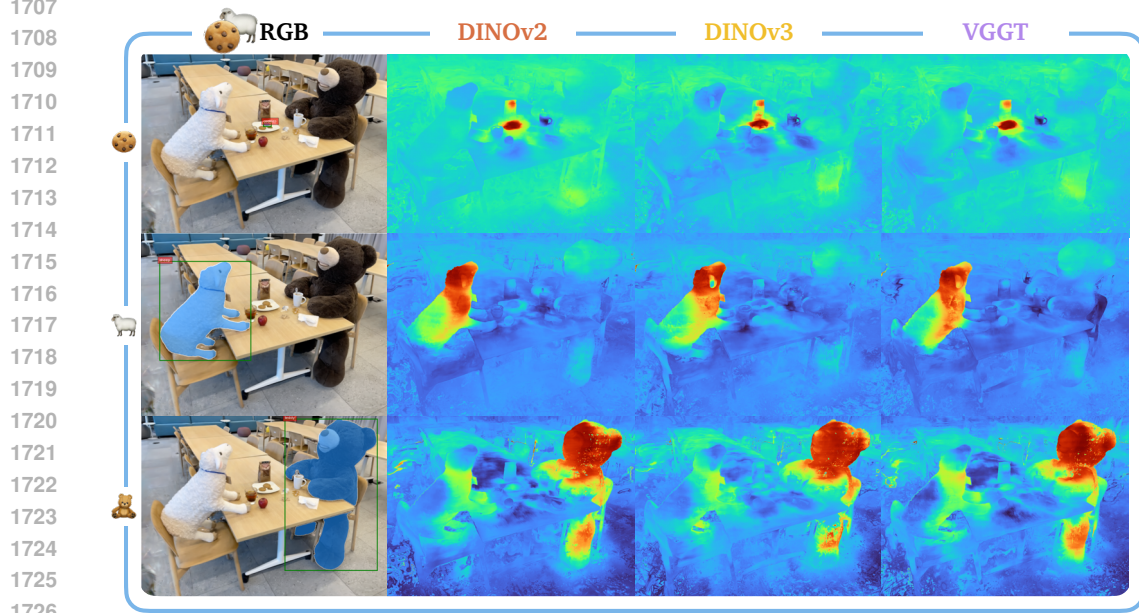


Figure 50: Semantic object localization visualized for GS for the LERF teatime scene.

1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781

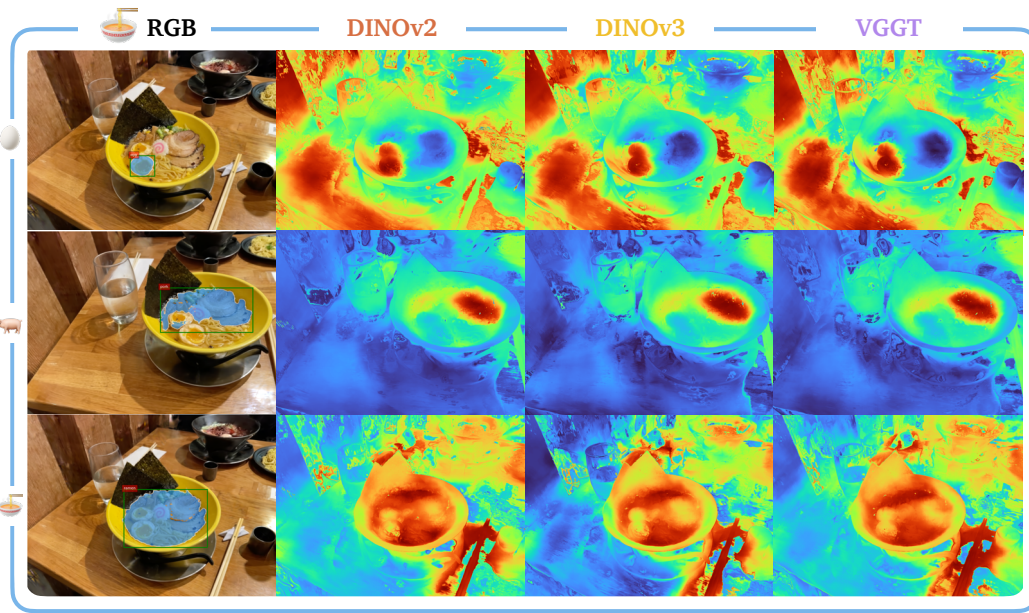


Figure 51: Semantic object localization visualized for GS for the LERF ramen scene.

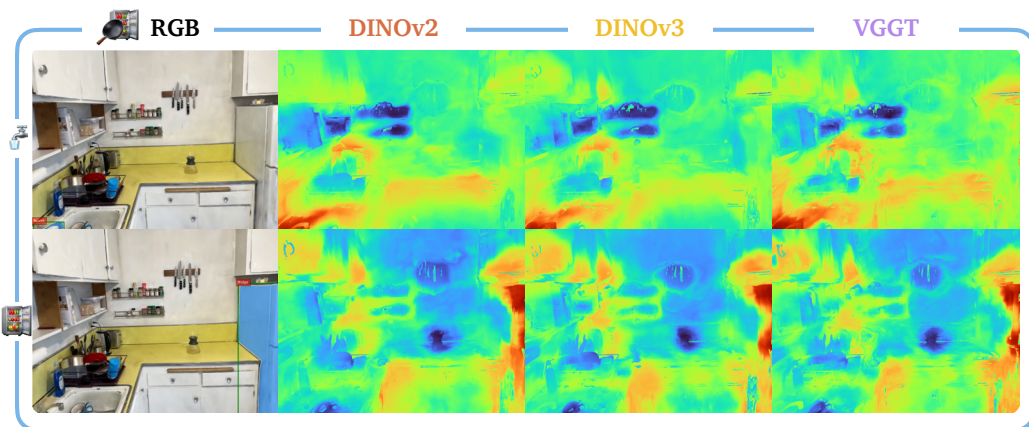


Figure 52: Semantic object localization visualized for GS for the LERF kitchen scene.

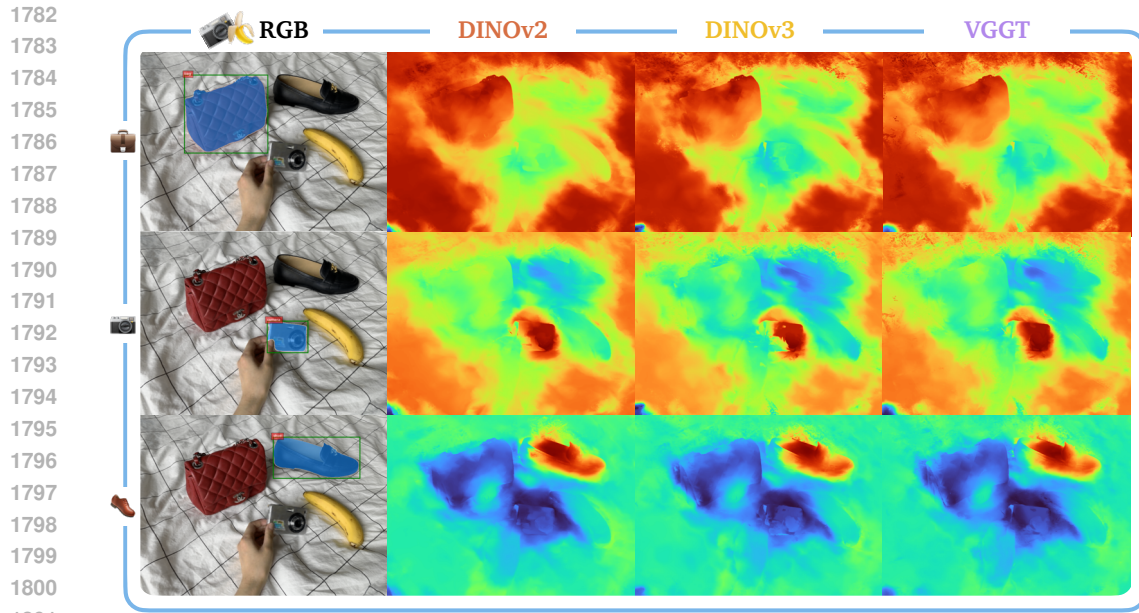


Figure 53: Semantic object localization visualized for NERF for the 3D-OVS table scene.

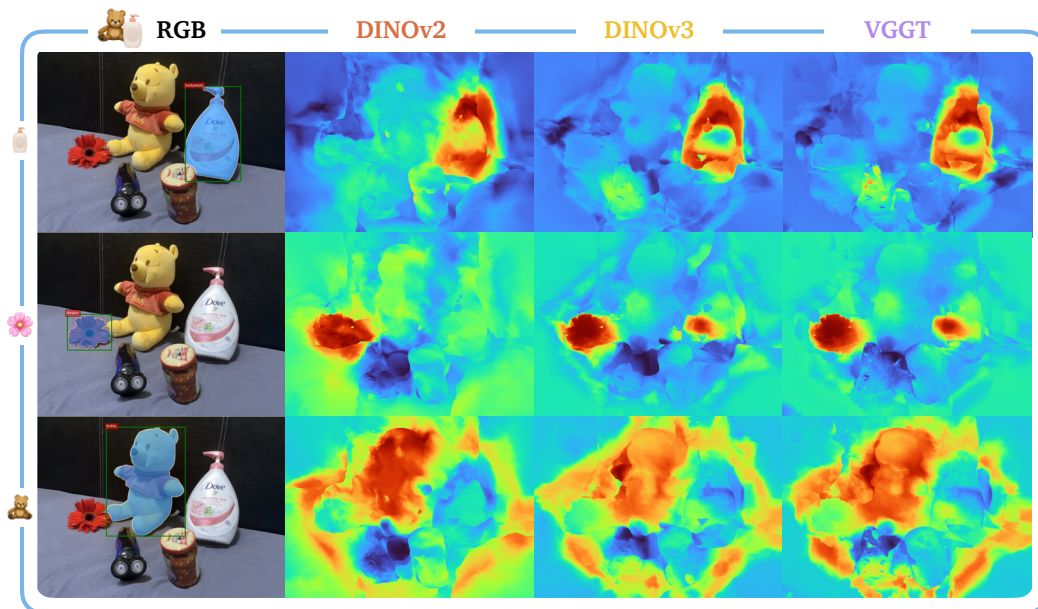


Figure 54: Semantic object localization visualized for NERF for the 3D-OVS bed scene.

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

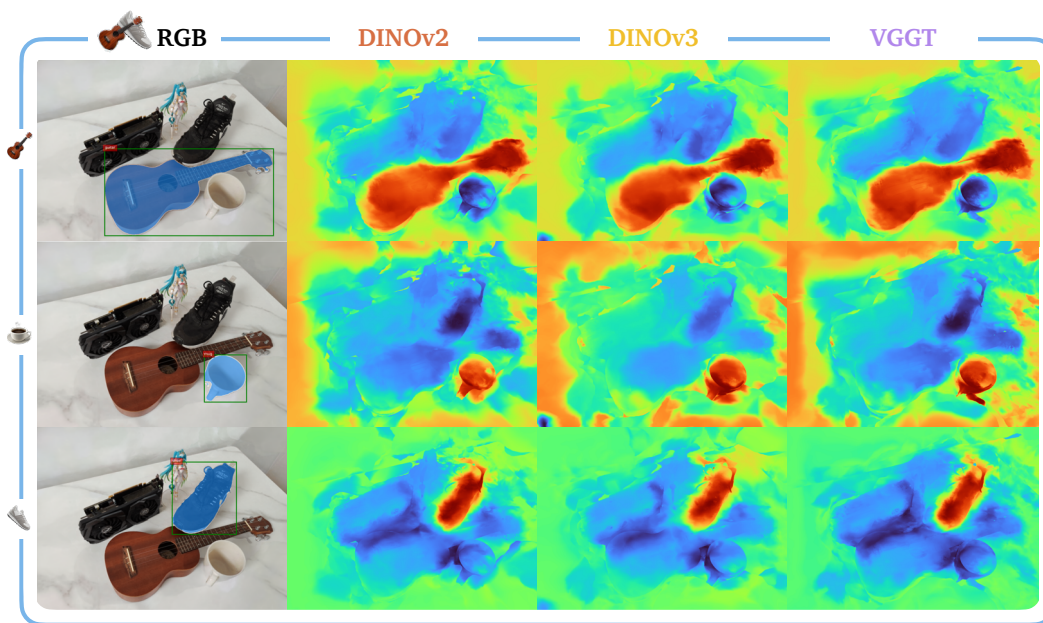


Figure 55: Semantic object localization visualized for NERF for the 3D-OVS covered desk scene.

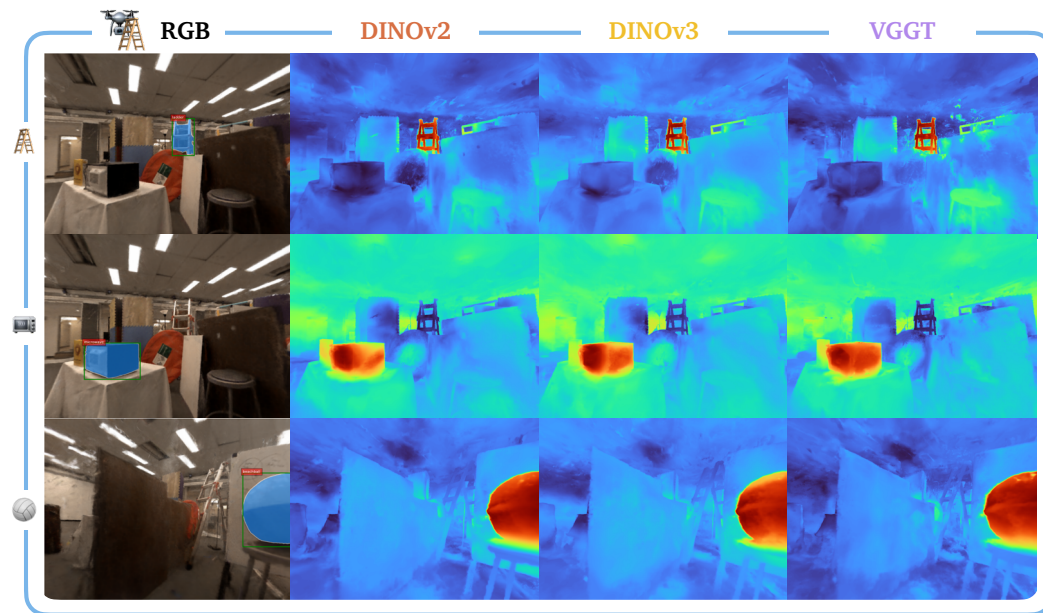


Figure 56: Semantic object localization visualized for NERF for the drone scene.

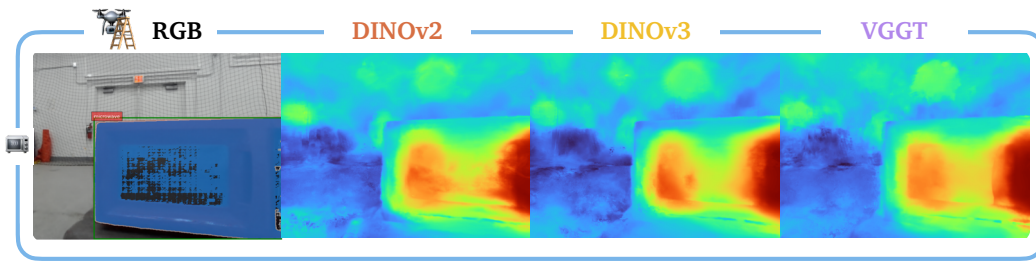


Figure 57: Semantic object localization visualized for NERF for the quadruped kitchen scene.

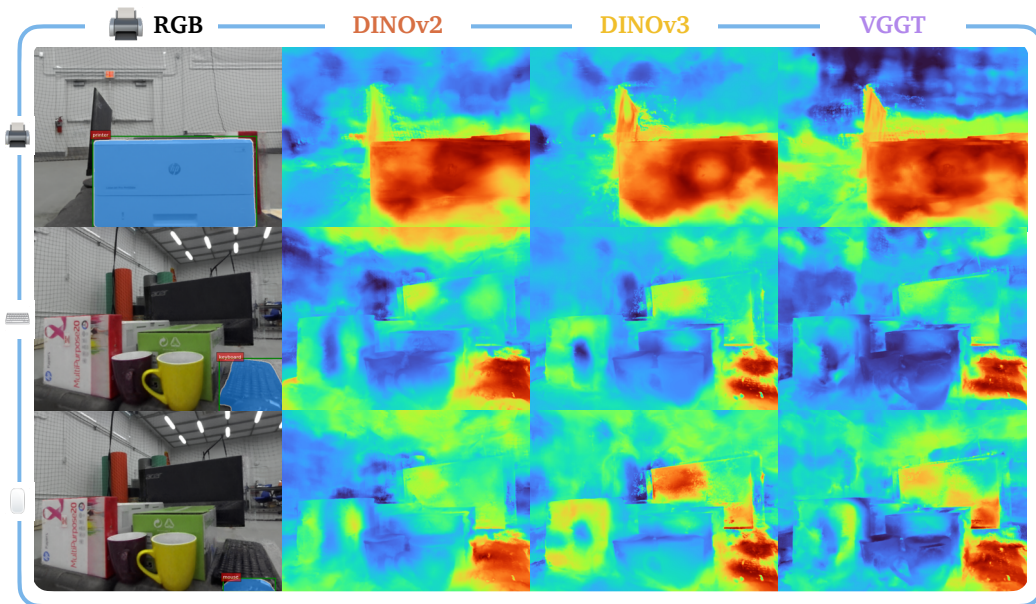


Figure 58: Semantic object localization visualized for NERF for the 3D-OVS bed scene.

1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

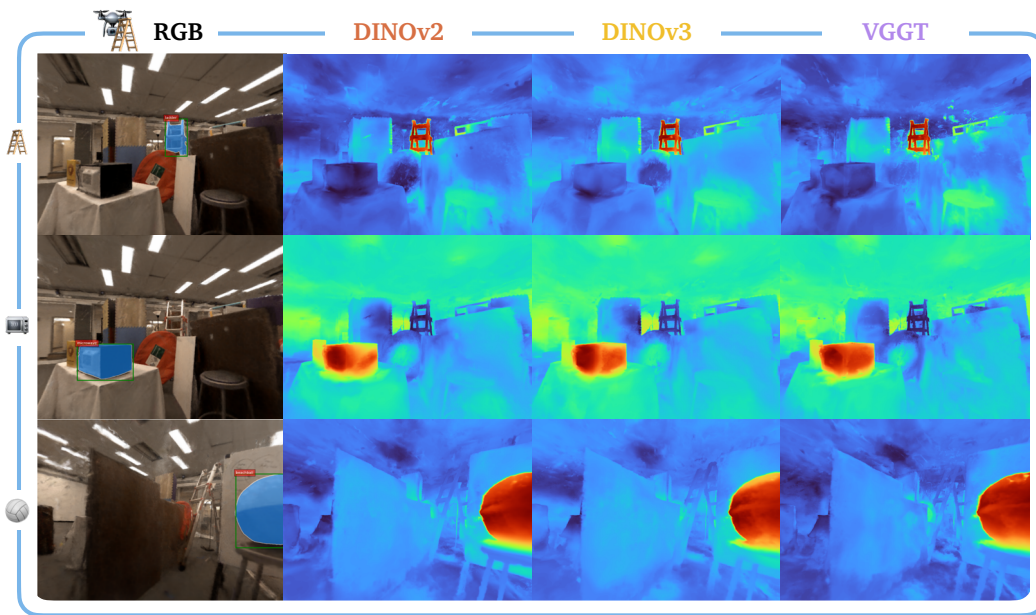


Figure 59: Semantic object localization visualized for NERF for the LERF teatime scene.