MUSE: Model-based Uncertainty-aware Similarity Estimation for zero-shot 2D Object Detection and Segmentation

Supplementary Material

CONTENT:

A. Vision Foundation Model Configurations

In this section, we provide detailed configurations of the vision foundation models employed in our framework. Table 1 summarizes the specific hyperparameter setups for each model, including Grounding DINO and SAM2. These configurations ensure fair comparison and reproducibility of our experiments. The top_k parameter of Grounding DINO refers to the top-k boxes, which is sorted by box confidence.

Backbone	Hyperparameter	Setting
(a) Grounding DINO		
SWIN-B	caption(prompt)	objects
	box_threshold	0.1
	top_k	50
(b) SAM2 (w Grounding DINO)		
Hiera-L	mask_threshold	0.0
	max_hole_area	0.0
	max_sprinkle_area	0.0

Table 1. Backbone and hyperparameter settings for vision foundation models.

B. Ablation on Hyperparameters α, β, γ

In this section, we conduct additional ablation studies, including sensitivity analysis on three hyperparameters—α, β, and γ. We systematically vary each parameter while keeping the others fixed, and report their effect on both detection and segmentation performance.

Here, α controls the balance between class-level and appearance embeddings, β adjusts the weighting between absolute and relative similarity scores, and γ rescales the objectness prior to compensate for uncertainty. This section analyzes how each hyperparameter contributes to the overall performance and provides guidelines for robust configuration.

Based on these results, we set the default configuration of our framework as $\alpha = 0.5$, $\beta = 0.8$, and $\gamma = 0.1$. These values consistently yielded the best trade-off between detection accuracy, segmentation quality, and robustness across datasets, and are therefore fixed as the default settings in all main experiments.

B.1. Ablation Study of α

Table 2 reports the sensitivity analysis of α, which controls the balance between class-level and appearance embeddings. We observe that the performance gradually improves as α increases from 0.0 to 0.5, indicating that combining both global (class) and local (patch) information is beneficial. In particular, $\alpha = 0.5$ achieves the best overall performance across all datasets, showing that an equal weighting between the two embedding types provides the most robust trade-off. When α exceeds 0.5, the performance saturates or slightly decreases, suggesting that overemphasizing either global or local features alone is suboptimal.

These results validate our design choice of integrating class-level and appearance embeddings through a balanced weighting scheme, which captures both semantic consistency and geometric details for more accurate matching.

B.2. Ablation Study of β

Table 3 summarizes the effect of the hyperparameter β, which adjusts the balance between the absolute similarity score S_{abs} and the relative similarity score S_{rel}. We observe that $\beta = 0.8$ achieves the best overall performance across detection and segmentation tasks. This indicates that giving higher weight to the absolute similarity is generally beneficial, as it directly measures the alignment between proposals and templates. At the same time, assigning a smaller weight to the relative similarity ($1 - \beta = 0.2$) is crucial in difficult scenarios. In such cases, relative similarity plays a decisive role in distinguishing between visually similar hard negatives by leveraging class-wise normalization.

Therefore, the combination of a dominant absolute score with a complementary relative score provides the most effective trade-off, yielding robust performance while mitigating errors from hard negative proposals.

B.3. Ablation Study of γ

Table 4 presents the ablation study on γ, which rescales the objectness prior $P(O|p)$ to compensate for uncertainty in

α	AP(core)	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V
0.0	0.508 / 0.502	0.513 / 0.476	0.488 / 0.475	0.473 / 0.448	0.333 / 0.433	0.469 / 0.378	0.600 / 0.619	0.683 / 0.688
0.1	0.518 / 0.512	0.515 / 0.479	0.494 / 0.482	0.506 / 0.480	0.335 / 0.434	0.483 / 0.388	0.610 / 0.628	0.686 / 0.692
0.2	0.525 / 0.518	0.514 / 0.479	0.499 / 0.485	0.536 / 0.509	0.335 / 0.435	0.488 / 0.391	0.615 / 0.634	0.685 / 0.691
0.3	0.529 / 0.522	0.513 / 0.479	0.500 / 0.483	0.563 / 0.537	0.336 / 0.435	0.492 / 0.394	0.614 / 0.633	0.685 / 0.692
0.4	0.532 / 0.525	0.513 / 0.478	0.497 / 0.482	0.586 / 0.557	0.336 / 0.433	0.494 / 0.395	0.616 / 0.635	0.685 / 0.692
0.5	**0.533 / 0.525**	**0.511 / 0.477**	**0.494 / 0.478**	**0.601 / 0.573**	**0.337 / 0.433**	**0.490 / 0.391**	**0.616 / 0.635**	**0.684 / 0.690**
0.6	0.532 / 0.525	0.508 / 0.474	0.490 / 0.474	0.610 / 0.583	0.336 / 0.432	0.484 / 0.385	0.616 / 0.616	0.683 / 0.689
0.7	0.530 / 0.523	0.504 / 0.472	0.486 / 0.470	0.615 / 0.590	0.335 / 0.430	0.477 / 0.380	0.616 / 0.633	0.680 / 0.686
0.8	0.528 / 0.520	0.502 / 0.469	0.482 / 0.467	0.618 / 0.593	0.335 / 0.428	0.469 / 0.374	0.613 / 0.630	0.676 / 0.682
0.9	0.525 / 0.518	0.498 / 0.466	0.479 / 0.463	0.614 / 0.593	0.334 / 0.426	0.464 / 0.369	0.612 / 0.630	0.672 / 0.679
1.0	0.521 / 0.514	0.494 / 0.463	0.474 / 0.459	0.612 / 0.593	0.332 / 0.422	0.458 / 0.365	0.609 / 0.626	0.666 / 0.673

Table 2. Ablation study on α. Each entry reports Detection (top) / Segmentation (bottom). $\alpha = 0.5$ achieves the best overall trade-off across datasets.

β	AP(core)	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V
0.0	0.484 / 0.477	0.487 / 0.449	0.472 / 0.463	0.379 / 0.348	0.300 / 0.405	0.489 / 0.386	0.594 / 0.615	0.667 / 0.673
0.1	0.497 / 0.490	0.500 / 0.464	0.474 / 0.465	0.435 / 0.405	0.313 / 0.416	0.489 / 0.387	0.597 / 0.617	0.671 / 0.677
0.2	0.505 / 0.498	0.506 / 0.470	0.477 / 0.467	0.467 / 0.440	0.321 / 0.424	0.490 / 0.387	0.601 / 0.621	0.673 / 0.680
0.3	0.513 / 0.506	0.509 / 0.474	0.480 / 0.469	0.500 / 0.473	0.328 / 0.430	0.491 / 0.389	0.605 / 0.625	0.675 / 0.682
0.4	0.518 / 0.511	0.510 / 0.475	0.484 / 0.472	0.529 / 0.496	0.332 / 0.432	0.490 / 0.389	0.607 / 0.628	0.676 / 0.684
0.5	0.524 / 0.516	0.511 / 0.476	0.488 / 0.474	0.557 / 0.525	0.334 / 0.433	0.491 / 0.390	0.609 / 0.630	0.679 / 0.686
0.6	0.528 / 0.520	0.511 / 0.476	0.491 / 0.477	0.578 / 0.547	0.334 / 0.432	0.492 / 0.391	0.612 / 0.632	0.681 / 0.688
0.7	0.532 / 0.524	0.511 / 0.477	0.494 / 0.479	0.591 / 0.560	0.336 / 0.433	0.491 / 0.391	0.615 / 0.635	0.683 / 0.690
0.8	**0.533 / 0.525**	**0.511 / 0.477**	**0.494 / 0.478**	**0.601 / 0.573**	**0.337 / 0.433**	**0.490 / 0.391**	**0.616 / 0.635**	**0.684 / 0.690**
0.9	0.533 / 0.525	0.511 / 0.477	0.492 / 0.475	0.608 / 0.581	0.337 / 0.432	0.484 / 0.388	0.617 / 0.634	0.683 / 0.689
1.0	0.529 / 0.521	0.510 / 0.476	0.482 / 0.465	0.609 / 0.585	0.336 / 0.431	0.475 / 0.379	0.613 / 0.630	0.677 / 0.681

Table 3. Ablation study on β. Each entry reports Detection (top) / Segmentation (bottom). $\beta = 0.8$ achieves the best trade-off by emphasizing absolute similarity while still retaining relative similarity to handle hard negatives.

γ	AP(core)	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V
0.1	**0.533 / 0.525**	**0.511 / 0.477**	**0.494 / 0.478**	**0.601 / 0.573**	**0.337 / 0.433**	**0.490 / 0.391**	**0.616 / 0.635**	**0.684 / 0.690**
0.2	0.526 / 0.516	0.514 / 0.478	0.497 / 0.480	0.530 / 0.494	0.334 / 0.434	0.505 / 0.402	0.616 / 0.634	0.687 / 0.693
0.3	0.518 / 0.508	0.512 / 0.476	0.494 / 0.477	0.477 / 0.444	0.328 / 0.430	0.510 / 0.406	0.615 / 0.631	0.688 / 0.694
0.4	0.510 / 0.500	0.509 / 0.473	0.491 / 0.473	0.438 / 0.405	0.324 / 0.426	0.509 / 0.404	0.609 / 0.626	0.688 / 0.693
0.5	0.503 / 0.493	0.507 / 0.470	0.487 / 0.469	0.411 / 0.377	0.320 / 0.422	0.508 / 0.403	0.603 / 0.619	0.687 / 0.692

Table 4. Ablation study on γ. Each entry reports Detection (top) / Segmentation (bottom). $\gamma = 0.1$ achieves the best performance, particularly improving results on TUD-L where the raw objectness prior is underestimated due to the generic prompt "items".

proposal confidence. Since Grounding DINO is prompted with a very general text query, such as *"items"*, the raw objectness scores are often underestimated, requiring an appropriate rescaling. We observe that $\gamma = 0.1$ yields the best overall performance across datasets. In particular, the TUD-L dataset shows a notable gain at $\gamma = 0.1$, indicating that a small degree of scaling effectively boosts weak proposals without over-amplifying false positives. As γ increases beyond 0.1, performance gradually degrades due to excessive scaling, which introduces more false detections. Therefore, we set $\gamma = 0.1$ as the default in our framework, as it provides the most stable trade-off between compensating for underestimated objectness scores and suppressing false positives.

C. Additional Qualitative Results

In this section, we provide additional qualitative results on the BOP-Classic-Core, BOP-H3, and BOP-Industrial datasets to further demonstrate the effectiveness of our framework. These examples complement the quantitative evaluations and illustrate the robustness of our approach across diverse scenarios.

Figure 1. Qualitative results of LM-O dataset.

Figure 2. Qualitative Results of T-LESS dataset.

Figure 3. Qualitative results of TUD-L dataset.

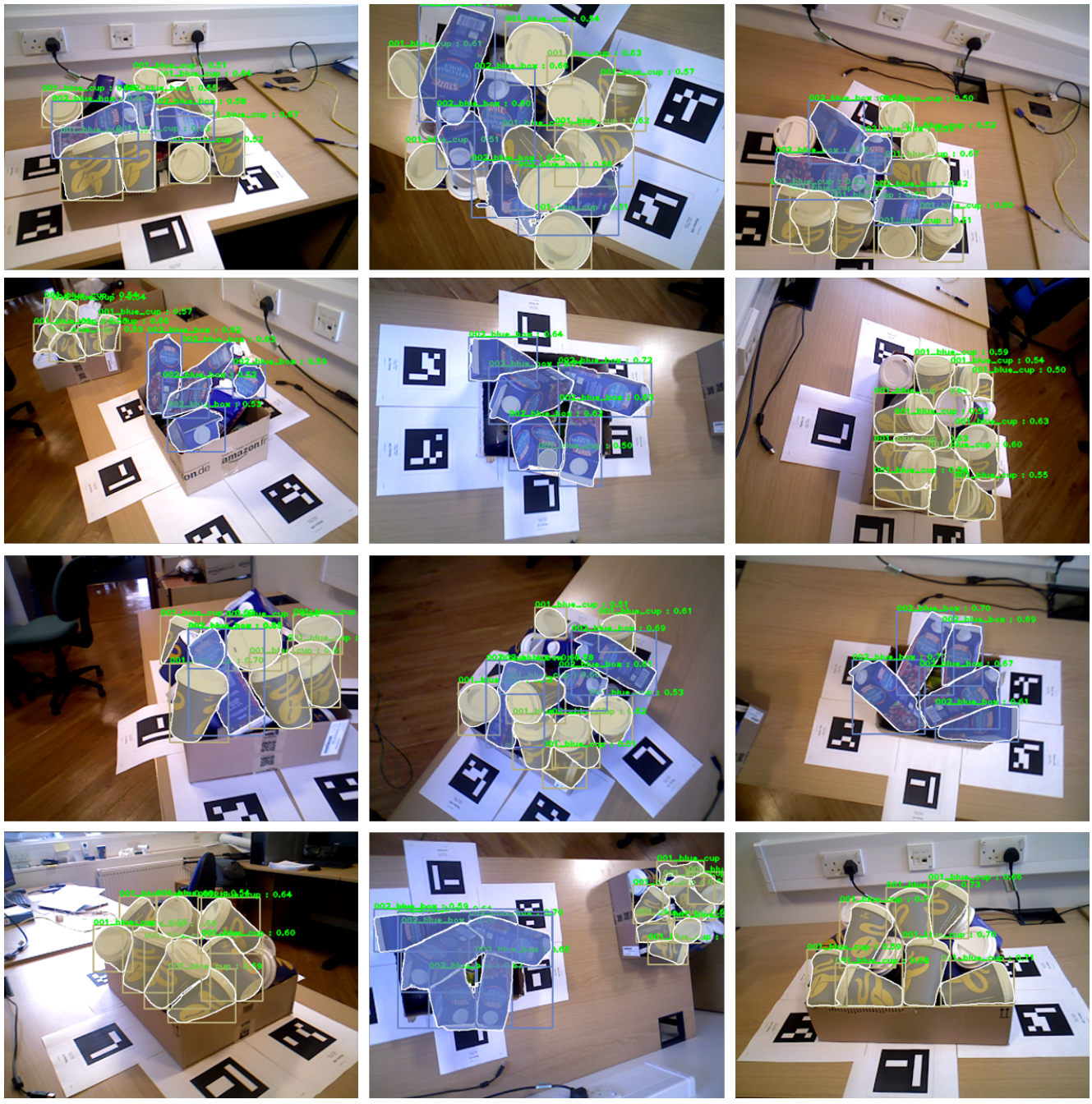

Figure 4. Qualitative results of IC-BIN dataset.

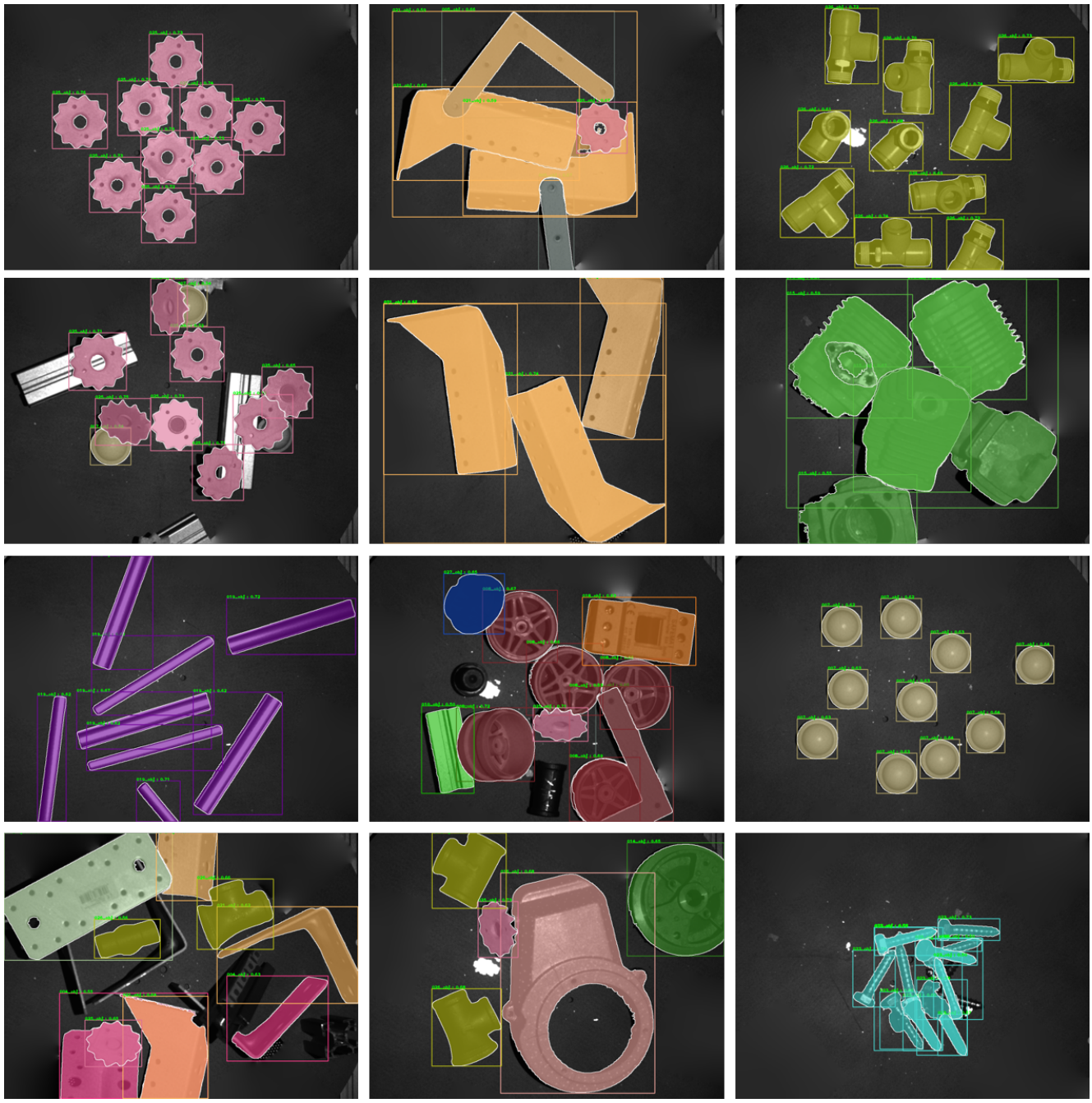

Figure 5. Qualitative results of ITODD dataset.

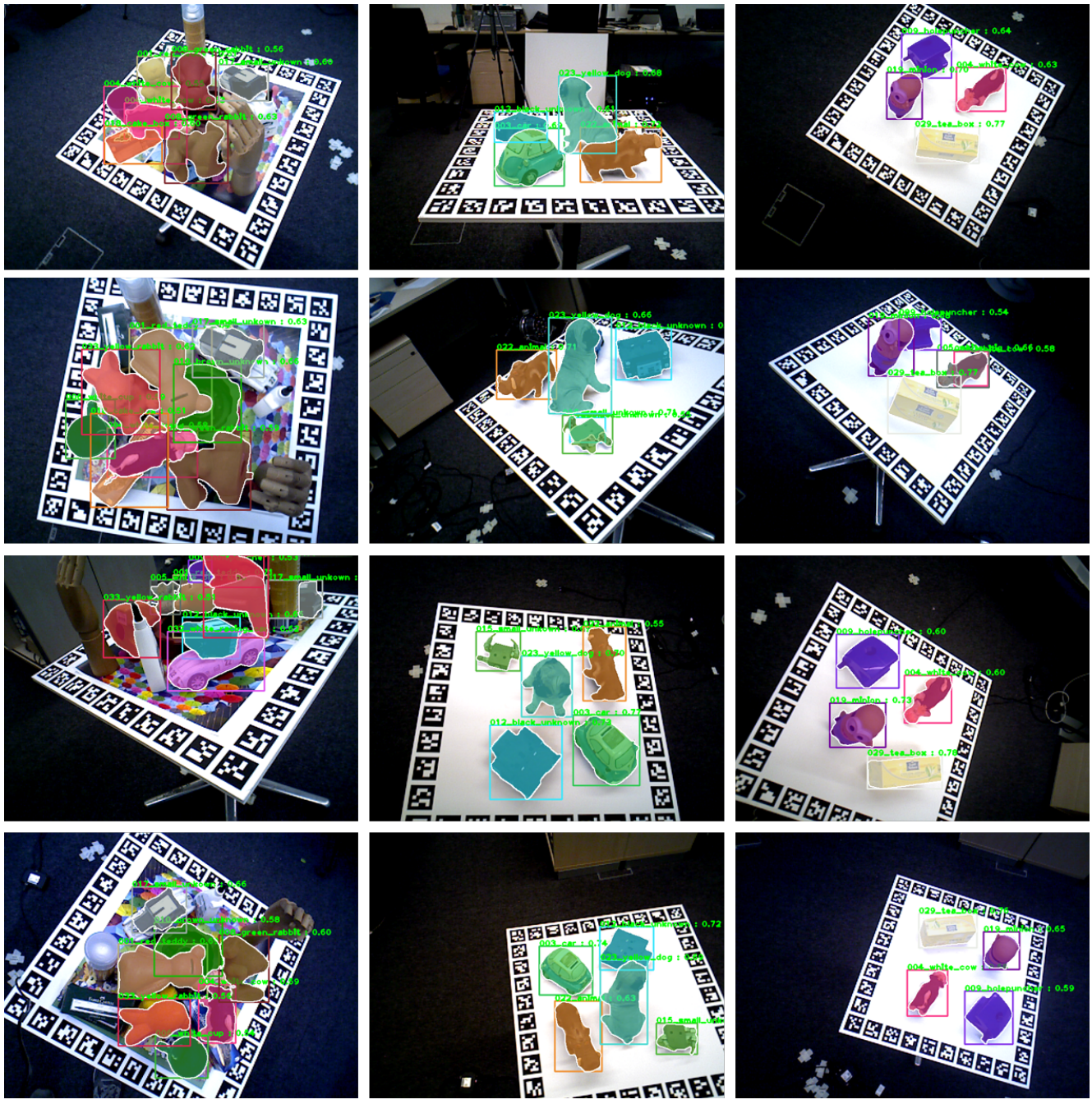

Figure 6. Qualitative results of HB dataset.

Figure 7. Qualitative results of YCB-V dataset.

Figure 8. Qualitative results of HANDAL dataset.

Figure 9. Qualitative results of HOPEv2 dataset.

Figure 10. Qualitative results of HOT3D dataset.

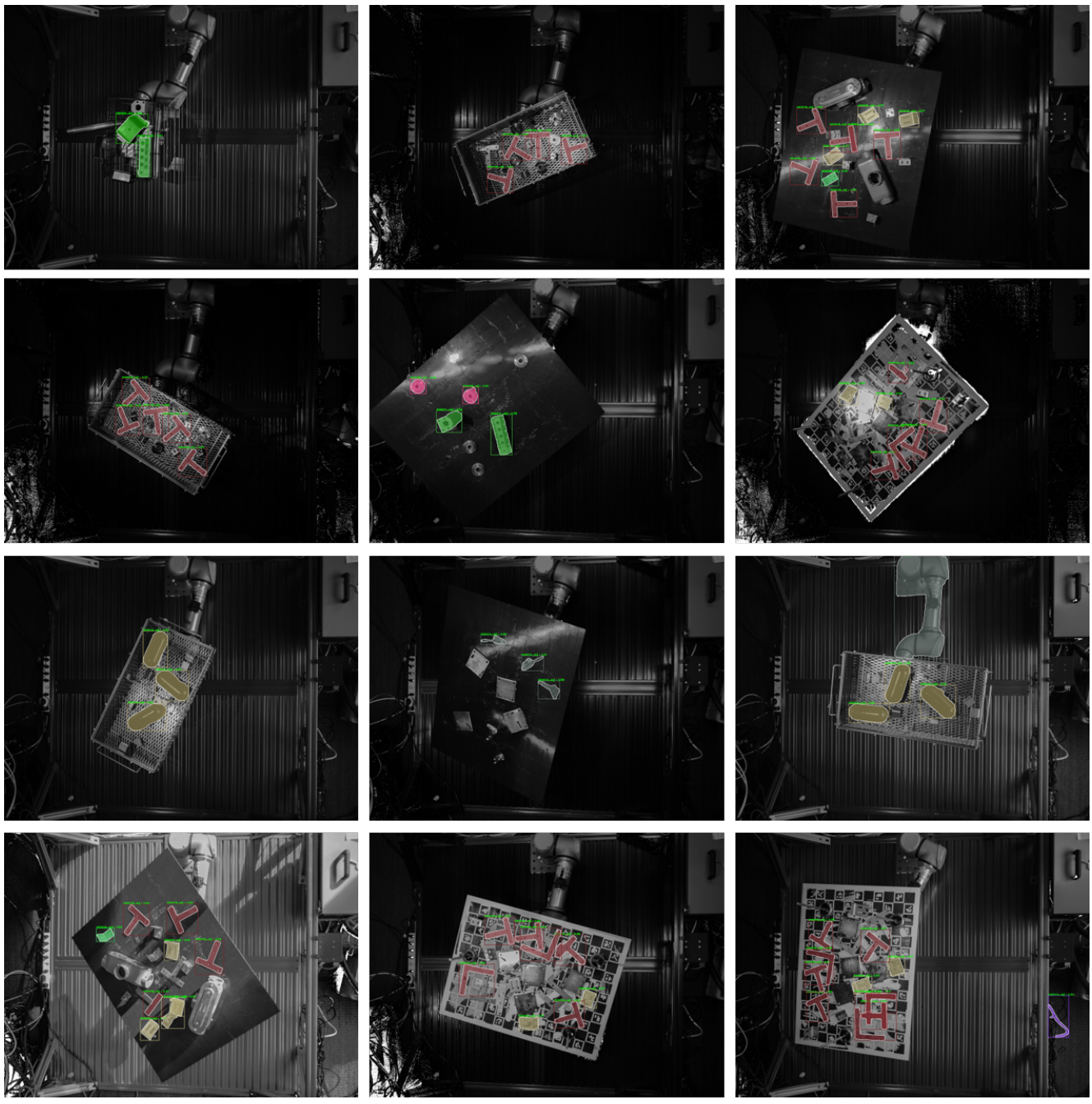

Figure 11. Qualitative results of IPD dataset.

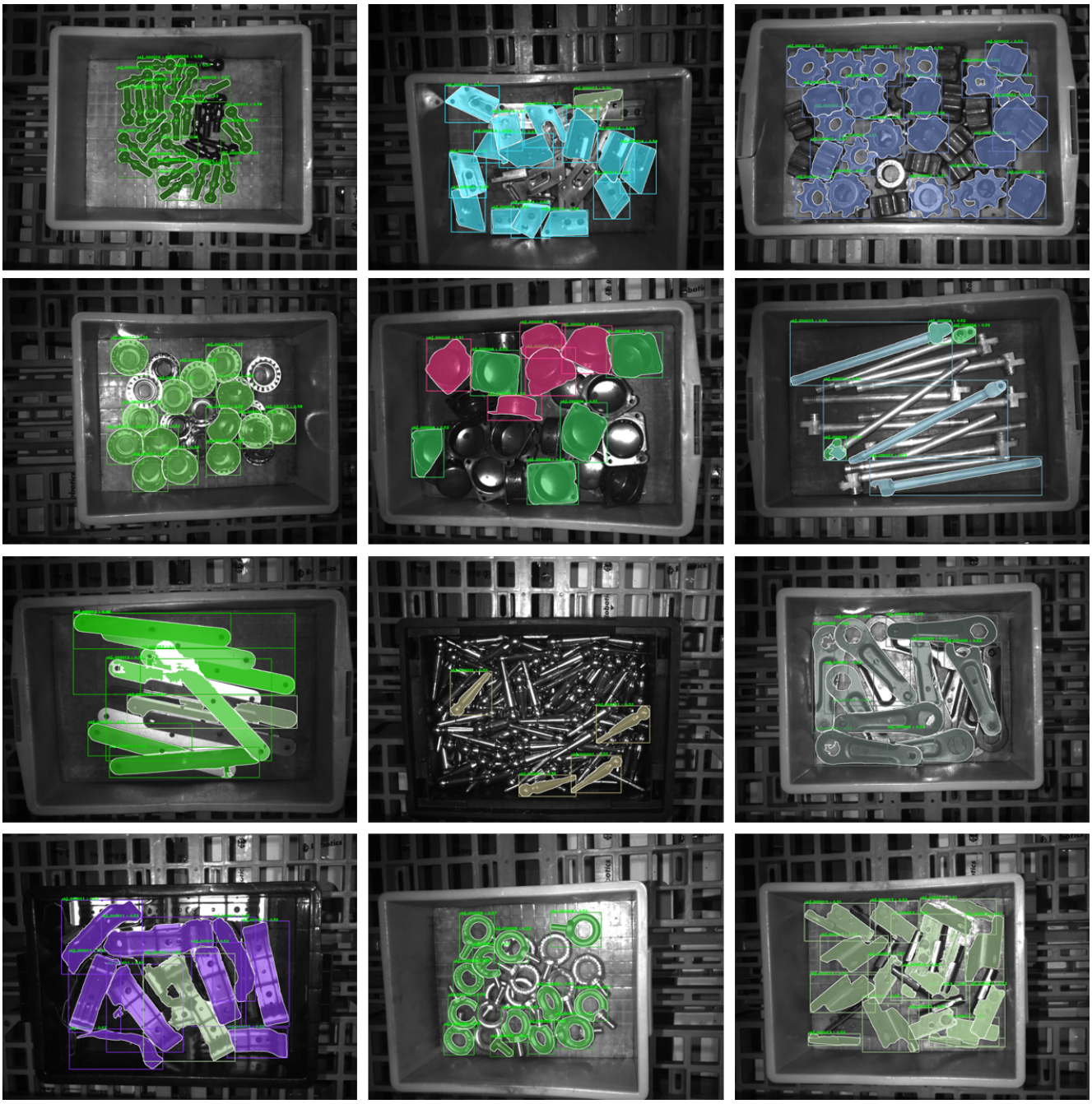

Figure 12. Qualitative results of XYZIBD dataset.