ISSN: 2664-9640

Industrial Part Detection based on CAD Model Matching and Appearance–Silhouette Score Fusion

Minglei Zhu[1,*], Junjie Shen[1], and Junjie Ma[2]

[1] School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

[2] VEM China Co., Ltd., Wuxi 214100, China

*Corresponding Author

Abstract

Part detectors based on computer-aided design (CAD) models must rank candidate regions by confidence. To assess the complementary value of CAD silhouette information and its effect on proposal ranking and industrial part detection, this study combines a library-wide mean silhouette score with appearance confidence. Experiments use Segment Anything Model (SAM) proposals and DINOv2 appearance features on XYZIBD RGB images. An internally pre-specified development rule selected appearance scoring because of its higher coverage AUC (0.645 versus 0.620), making the conditional detector identical to the appearance baseline. On the held-out set, pre-specified unit-weight silhouette addition increased the class-agnostic average precision (CAAP) point estimate by 1.28 percentage points from a 1.25% baseline, with a 95% difference interval of −0.20 to 2.68. Adding SAM-predicted IoU yielded a difference of −0.33 points, with an interval of −1.22 to 0.13. Instance coverage was 17.3% at IoU ≥ 0.5, defining the range of instances recoverable through rescoring. Post hoc rescoring of saved CNOS proposals yielded CAAP differences of +2.91 and −0.01 points for the study and native CNOS scorers, respectively. Classwise AP increased slightly for the study scorer and decreased for the native scorer. These findings support evaluating appearance and silhouette score fusion within complete detection pipelines while accounting for instance coverage, scoring configuration, and classwise AP.

Keywords

Industrial Part Detection; CAD Model Matching; Confidence Score Fusion; Silhouette Similarity; Unseen-Object Detection.

1. Introduction

Computer-aided design (CAD) models provide geometric references for recognizing previously unseen parts without collecting annotated training photographs for each identity. Classical gradient templates and chamfer matching exploit object shape to recognize textureless targets [1, 2], while modern CAD-based detectors generate image proposals and match their visual features against rendered CAD views [3, 4]. In industrial bins, detectors must identify the corresponding CAD model and rank numerous candidate regions by confidence. Confidence construction therefore connects model matching with the selection of final detections.

Existing methods combine several sources of evidence to construct detection confidence. CNOS aggregates template similarities [3], SAM6D incorporates identity-specific geometric consistency [4], and MUSE uses proposal objectness [5]. This study examines a library-wide mean silhouette score, defined as the average similarity between a proposal mask and normalized silhouettes across the CAD library. The score summarizes shape compatibility across all candidate identities and provides a proposal-level assessment independent of

ISSN: 2664-9640

appearance-based identity selection. Adding it to appearance confidence provides a direct way to examine the complementary role of CAD silhouette information in proposal ranking.

Preliminary experiments on TLESS [6] and YCB-Video [7] motivated the investigation of mean silhouette scoring. These experiments used different datasets and appearance-aggregation settings, highlighting the need to evaluate score fusion under a fixed experimental protocol. The present study therefore examines the contribution of the mean silhouette score on XYZIBD using pre-specified scoring comparisons and an object-disjoint development and held-out partition.

To assess the complementary value of silhouette information and examine the relationship between development-set ranking and held-out detection performance, we conduct experiments on XYZIBD RGB images. The pipeline uses SAM to generate proposals, DINOv2 to extract appearance features, and rendered CAD templates for identity matching. All scoring rules preserve proposal boxes and predicted identities while changing confidence ranking and subsequent detection selection. The experiment uses five development layouts and ten held-out layouts containing different object identities, enabling assessment under this partition of objects and layouts.

The scoring comparisons and development decision rule were recorded in an internal analysis plan before the XYZIBD images were downloaded or processed. The decision rule compared the standalone coverage AUCs of silhouette and appearance scores and selected appearance scoring, making the primary conditional-versus-appearance contrast zero by construction. Two additional pre-specified comparisons add the silhouette score or SAM's predicted mask-quality score to appearance confidence with unit weight. CAAP serves as the primary metric, while classwise AP provides a secondary assessment that incorporates predicted identity. Standalone ranking and contribution to a combined score address different questions [8], and this distinction guides the analysis.

Post hoc diagnostics further quantify proposal coverage and variation across layouts. Held-out instance coverage was 17.3% at IoU ≥ 0.5, defining the instances recoverable by rescoring the fixed proposal boxes. Post hoc rescoring of stronger saved CNOS proposals recorded different CAAP responses under the study and native CNOS scorers, accompanied by a small increase and a decrease in classwise AP, respectively. Together, the pre-specified scoring comparisons and post hoc diagnostics connect proposal coverage, scoring configuration, and final detection metrics, providing experimental evidence for assessing appearance and silhouette score fusion in this industrial setting.

2. Related Work

2.1. CAD-Based Detection and Confidence Fusion

CAD-based part detection commonly combines proposal generation with rendered-template matching. CNOS matches SAM masks against rendered CAD views using DINOv2 descriptors and averages the five highest similarities for each object identity [3]. NIDS-Net combines Grounding DINO boxes, SAM masks, and adapted foreground DINOv2 features [9]. Grounding DINO provides text-guided open-set object detection [10]. These approaches incorporate both proposal quality and appearance matching into the detection pipeline.

Confidence fusion further incorporates proposal-quality information into detection scoring. MUSE multiplies absolute and relative template similarities by transformed Grounding DINO objectness [5]. NOCTIS combines box and mask confidence with appearance matching [11]. These methods construct detection-ranking scores from both matching evidence and information about the proposals themselves.

ISSN: 2664-9640

Geometry and object-set information also contribute to confidence construction. SAM6D combines semantic, local-appearance, and projected-box-overlap scores, weighting the geometric term by a feature-derived visible ratio [4]. DepthNOCTIS introduces depth-based geometric consistency [12]. The CLON preprint determines cue weights from the object set before inference and uses language-guided proposals [13]. These approaches illustrate different ways of incorporating object models, geometric relationships, and object-set information into scoring.

The library-wide mean silhouette score studied here aggregates normalized shape similarities across the entire CAD library, providing a proposal-level score independent of predicted identity. Its conceptual role is a CAD-informed objectness prior [14]. With proposal boxes and predicted identities fixed, we add this score to appearance confidence with unit weight and evaluate its effect on detection ranking and selection. Silhouette similarity, SAM-predicted mask quality, and appearance confidence represent different information. We compare them as ranking scores while maintaining the conceptual distinction between confidence ranking and probability calibration [15, 16].

2.2. Proposal Coverage and Detection Evaluation

Work on object proposals shows how recall and localization constrain detection performance [17]. BOP's recent reports identify unseen-object detection as a pipeline bottleneck [18, 19], and the XYZIBD paper reports difficulties in industrial scenes [20]. BioDet addresses background artifacts through preprocessing and proposal filtering [21]. Our instance-coverage measurements quantify this established problem under a recorded RGB configuration; its causes require separate controlled measurements.

AUC and AP summarize different aspects of ranking [22]. Coverage AUC gives equal weight to covered–uncovered proposal pairs and can label several proposals for one instance as covered. We use the COCO evaluation framework [23]; its detection implementation permits each eligible non-crowd annotation to match once at a given IoU threshold and averages AP over multiple thresholds [24]. Detection scores also affect suppression and truncation. Pooled AUC additionally compares proposals across layouts, so it can differ from within-layout discrimination. These definitions guide the comparison of score ranking and detection performance.

3. Methods

3.1. Dataset and Experimental Protocol

We use RealSense RGB images from the publicly released validation split of XYZIBD, which contains industrial parts densely arranged in bins [20]. The experiment includes 15 physical layouts, each comprising 50 images at a resolution of 1280 × 720 pixels, for a total of 750 images. RGB input continues the image-only pipeline used in the preliminary studies. The BOP Toolkit's default XYZIBD configuration uses XYZ grayscale images at a different resolution [25]. Our results correspond to a local RGB evaluation protocol on the public validation split.

The images are partitioned by layout into development and held-out sets. The five layouts whose identifiers are multiples of 15 form the 250-image development set, and the remaining ten layouts form the 500-image held-out set. Each layout contains one CAD identity, and its layout identifier is also its BOP scene identifier. Table 1 lists the layout-to-model mapping. The two subsets contain disjoint object identities, while all 15 CAD models remain matching candidates in both subsets.

ISSN: 2664-9640

Table 1. Layout and CAD model assignments for the development and held-out sets

Group	Layout identifiers	CAD model identifiers, same order
Development	0,15,30,45,60	16,13,5,9,11
Held out	5,10,20,25,35,40,50,55,65,70	17,2,8,1,12,4,10,15,6,14

Evaluation uses the amodal bounding boxes provided by bbox_obj, which include the projected extent of occluded object regions. Boxes are clipped to image boundaries, annotations with positive width and height are retained, and box area is calculated as width × height. All instances satisfying these conditions are included, irrespective of visibility. BOP's unseen-object 2D detection task also uses amodal boxes [18]; its official conversion derives boxes from full-object masks and marks low-visibility instances as ignored [25]. These annotation-processing and inclusion rules define the local evaluation protocol used in this study.

The internal analysis plan was finalized before the XYZIBD images were downloaded or processed. It specified the partition, fusion coefficient, similarity aggregation, identity assignment, proposal settings, detection selection, and primary CAAP comparison. Development predictions were generated before the corresponding annotations were read, and the development decision was recorded before held-out inference. Fixed silhouette addition, fixed SAM addition, and paired layout resampling were also pre-specified. Post hoc analyses conducted after inspecting the results included instance-coverage diagnostics, development-set bootstrap resampling, layout-omission analyses, score summaries, secondary classwise AP intervals, and CNOS analyses.

3.2. Detection Pipeline and Confidence Scoring

Figure 1 illustrates the CAD-based part detection and scoring pipeline. SAM first generates candidate masks and their bounding boxes from the input image. Appearance features of each proposal are then compared with rendered CAD views to determine object identity and appearance confidence. In parallel, a silhouette descriptor measures similarity between the proposal mask and library silhouettes, providing a shape score independent of identity selection. All scoring rules preserve proposal boxes and predicted identities while adjusting proposal ordering through score fusion.

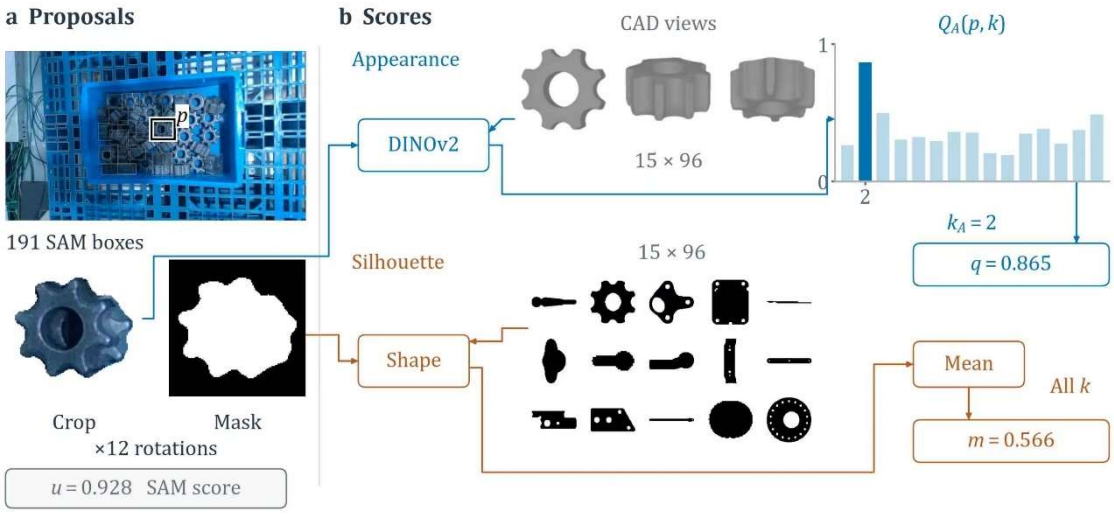


Figure 1. CAD-based part detection and confidence scoring

ISSN: 2664-9640

Proposals are generated using SAM ViT-B [27] with a 32 × 32 prompt grid, a predicted-IoU threshold of 0.88, a stability threshold of 0.95, and zero crop layers. Masks are retained when their area is at least 20 pixels and at most 75% of the image area. These settings are inherited from the preliminary pipeline and remain fixed throughout the industrial experiments.

Appearance features are extracted using a frozen DINOv2 ViT-S/14 encoder [28]. Each CAD model is centered and scaled to a maximum vertex radius of one, then rendered orthographically against a white background from 96 Fibonacci-sphere directions. Each proposal crop and its corresponding mask undergo 12 joint rotations at 30° intervals for subsequent appearance and silhouette similarity computation.

Let p, r, k and v denote a proposal, rotation, CAD model, and rendered view, respectively. Appearance and silhouette cosine similarities are denoted by $A_{p,r,k,v}$ and $S_{p,r,k,v}$. Appearance scoring first aggregates over rotations and then over views. For each model and view, the maximum appearance similarity across rotations is

$$a_{p,k,v} = \max_r A_{p,r,k,v} \tag{1}$$

The five highest-scoring views for that model form the set $V_{p,k}^{(5)}$, and the model-level appearance score is

$$Q_{\mathrm{A}}(p,k) = \frac{1}{5} \sum_{v \in V_{p,k}^{(5)}} a_{p,k,v} \tag{2}$$

The predicted identity and appearance confidence are defined as

$$k_{\mathrm{A}}(p) = \arg\max_k Q_{\mathrm{A}}(p,k) \tag{3}$$

$$q(p) = Q_{\mathrm{A}}\left(p, k_{\mathrm{A}}(p)\right) \tag{4}$$

The silhouette descriptor concatenates 64 × 64 occupancy and signed-distance grids. Signed distances are divided by 64 and clipped to [−0.5, 0.5] before vector normalization. The mean silhouette score averages silhouette similarities over all rotations, CAD models, and rendered views:

$$m(p) = \frac{1}{12 \cdot 1440} \sum_r \sum_k \sum_v S_{p,r,k,v} \tag{5}$$

Here, 1440 corresponds to 96 rendered views for each of the 15 CAD models. The score $m(p)$ summarizes normalized shape compatibility across the entire library and has the same value for every candidate identity of a given proposal. SAM's raw predicted mask IoU is denoted by $u(p)$ and used in the subsequent scoring comparisons.

3.3. Development Decision Rule and Detection Selection

A proposal is defined as covered when its bounding box achieves an intersection over union (IoU) of at least 0.5 with any eligible annotation in the same image. Instance coverage is the proportion of annotated instances covered by at least one proposal. Each annotation

ISSN: 2664-9640

contributes once to this proportion, and coverage is determined independently for each annotation rather than through one-to-one matching. Counts are based on image annotations, so the same physical part appearing in multiple frames contributes an instance in each frame. Coverage at IoU 0.75 follows the same definition with the higher threshold. Before non-maximum suppression (NMS), coverage AUC evaluates how well a score ranks covered proposals above uncovered proposals, with tied score pairs contributing 0.5.

The pre-specified decision rule compares the standalone coverage AUCs of the mean silhouette score $m(p)$ and appearance confidence $q(p)$, pooling all development proposals. The decision variable d equals 1 when the AUC of $m(p)$ exceeds that of $q(p)$, and 0 otherwise, with ties assigned to appearance-only scoring. The decision requires both AUC values to be defined. The resulting conditional confidence score is

$$C(p) = q(p) + d \cdot m(p) \tag{6}$$

Two additional pre-specified comparisons use fixed score addition: $q(p) + m(p)$ for silhouette addition and $q(p) + u(p)$ for SAM predicted-IoU addition. Both assign a unit coefficient to the added score and are evaluated independently of the development decision. All scoring rules preserve the appearance-selected identity $k_A(p)$ and the bounding box of each proposal.

For each scoring rule, classwise NMS uses an IoU threshold of 0.5 to suppress overlapping boxes assigned to the same identity. The detector then retains the 100 highest-scoring detections per image across all identities, or all detections when fewer than 100 remain, without a confidence threshold. Consequently, the scoring rule affects both suppression order and the final set of retained detections. Class-agnostic metrics are computed from these selected detections without an additional NMS step.

3.4. Metrics and Statistical Analysis

Detection performance was evaluated using COCO bounding-box average precision (AP), computed with pycocotools 2.0.11 at IoU thresholds from 0.50 to 0.95 in increments of 0.05 [24]. Classwise AP averages performance across annotated object identities, whereas class-agnostic AP (CAAP) evaluates localization and confidence ranking with all identities treated as a single category. CAAP was pre-specified as the primary metric to assess identity-independent confidence scoring under fixed identity assignments. Classwise AP served as the secondary metric of CAD-model-based detection performance. Class-agnostic average recall with a maximum of 100 detections per image (CAAR100) was evaluated over the same IoU thresholds.

The primary comparison was the CAAP difference between conditional scoring and appearance-only scoring. When $d = 0$, the two rules produce identical predictions, yielding a difference of zero by construction. The two additional pre-specified comparisons evaluated fixed silhouette addition and fixed SAM addition against appearance-only scoring. For each comparison, a 95% percentile interval was estimated from 1,000 paired bootstrap resamples of the 10 held-out layouts. Each resample drew 10 complete layouts with replacement, using the same sampled layouts and multiplicities for both scoring rules. Pooled COCO AP was recomputed for each resample, and the interval bounds were obtained from the 2.5th and 97.5th percentiles of the resulting differences. Differences were calculated before rounding.

For identity accuracy, each covered proposal was assigned a reference identity from the annotation with the highest IoU. Ties were resolved using the stored annotation order. The predicted identity was then compared with this reference identity.

For the spatial diagnostic, the annotation envelope was defined as the smallest axis-aligned rectangle enclosing all eligible annotation boxes in an image. Proposal-box centers strictly outside this envelope were counted, with points on the boundary treated as inside. This

ISSN: 2664-9640

measure characterizes the spatial relationship between proposals and annotated objects; it is a geometric diagnostic rather than a semantic classification of background pixels.

3.5. CNOS Comparison and Post Hoc Rescoring

A post hoc comparison with CNOS used the dataset authors' adapted implementation [29] on the same 500 held-out images under the same local annotation and evaluation protocol. This implementation uses SAM ViT-H, DINOv2 ViT-L/14 and 42 physically based rendered (PBR) views per CAD model. Its native configuration applies a score threshold of 0.15 and retains the 100 highest-scoring candidates before classwise NMS at an IoU threshold of 0.25. Size filtering retains bounding boxes and masks with areas greater than 0.0025 and 0.0003 of the image area, respectively. CNOS therefore limits the candidate count before NMS, whereas the study configuration limits the detection count after NMS.

The complete-system comparison jointly varies proposal generation, appearance features, reference views and detection selection. A separate rescoring analysis examined silhouette addition on a fixed proposal set by reusing all 25,172 saved CNOS proposals that passed size filtering. Both analyses were designed after inspection of the held-out results and are reported separately from the pre-specified comparisons. Full query scores were recomputed using the saved proposals and existing reference templates.

Two paired rescoring comparisons evaluated appearance-only scoring against fixed silhouette addition. The first used the study's DINOv2 ViT-S/14 appearance scorer with 12 query rotations and its detection-selection procedure. The second used native CNOS DINOv2 ViT-L/14 appearance scores, PBR reference views and native detection selection. Within each comparison, the two scoring rules retained the same proposal boxes and appearance-selected identities. The fused rule added the study's mean silhouette score, computed from orthographic reference views, with a coefficient of 1. Each scoring rule applied its corresponding configuration's threshold, detection limit and NMS procedure using its own scores.

The recomputed native appearance predictions exactly reproduced the bounding boxes, identities and scores of the CNOS run reported in Table 3 for all 500 images. The saved proposal data contained no SAM predicted-IoU values, so the paired rescoring comparisons evaluated silhouette addition only. Mean silhouette scores restricted to the appearance-selected CAD model were also examined as descriptive controls.

4. Results

4.1. Development Decision

The development set yielded 46,483 proposals, of which 873 were covered at IoU ≥ 0.5. Appearance confidence q achieved a standalone coverage AUC of 0.645, compared with 0.620 for the mean silhouette score m. The pre-specified decision rule therefore set $d = 0$ and retained appearance-only scoring. Consequently, conditional scoring produced the same detection predictions as appearance-only scoring, and the primary CAAP difference was zero by construction.

Post hoc analysis of the combined scores yielded development coverage AUCs of 0.664 for $q + m$ and 0.554 for $q + u$. Although m had a lower standalone AUC than q, their sum achieved a higher AUC than q alone. This result demonstrates that standalone score ranking and combined-score ranking provide different assessments in this development set. The pre-specified decision was retained, and the detection performance of the fixed additions was evaluated separately on the held-out set.

ISSN: 2664-9640

4.2. Pre-specified Held-out Comparisons

Table 2 summarizes the held-out performance of the pre-specified scoring rules. Appearance-only scoring achieved a CAAP of 1.25%. Relative to this baseline, fixed silhouette addition yielded a CAAP difference of +1.28 percentage points (pp), with a 95% interval of [−0.20, 2.68], whereas fixed SAM addition yielded −0.33 pp [−1.22, 0.13] (Figure 2). Both intervals included zero. Fixed silhouette addition increased the CAAP point estimate, while classwise AP, coverage AUC and CAAR100 decreased. The standalone coverage AUC of the mean silhouette score was 0.571, compared with 0.581 for appearance confidence.

Table 2. Detection performance of the pre-specified scoring rules on the held-out set

Scoring rule	Classwise AP	CAAP	Coverage AUC	CAAR100
Appearance only	4.55	1.25	0.581	7.21
Fixed silhouette addition	4.44	2.54	0.564	6.93
Fixed SAM addition	3.60	0.92	0.478	5.98

All scoring rules retained 100 detections per image. Before NMS, appearance matching assigned the correct identity to 1,949 of the 2,428 covered proposals, giving an identity accuracy of 80.3%. This proposal-level accuracy was identical across scoring rules because identity assignments were fixed. CAAP pooled detections across all 15 candidate identities, including five without held-out annotations, whereas classwise AP averaged performance over the 10 annotated identities. Thus, high-scoring detections from any candidate identity contributed to the class-agnostic ranking. Differences between CAAP and classwise AP reflect category pooling, score ranking and detection selection under fixed proposal-level identity assignments.

Post hoc analysis of classwise AP yielded differences of −0.11 pp [−0.80, 0.50] for fixed silhouette addition and −0.95 pp [−1.52, −0.31] for fixed SAM addition, relative to appearance-only scoring. The silhouette-addition interval spanned zero, whereas the SAM-addition interval lay entirely below zero.

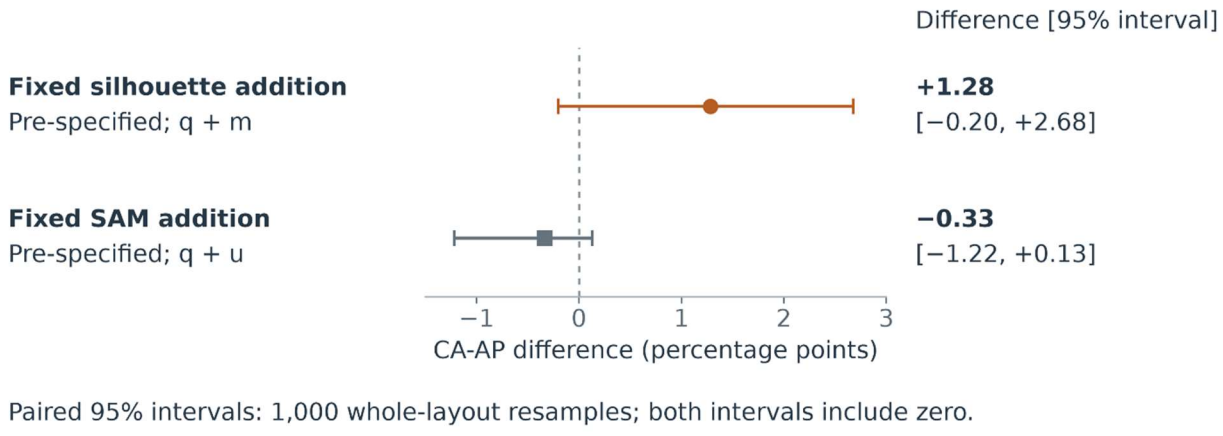
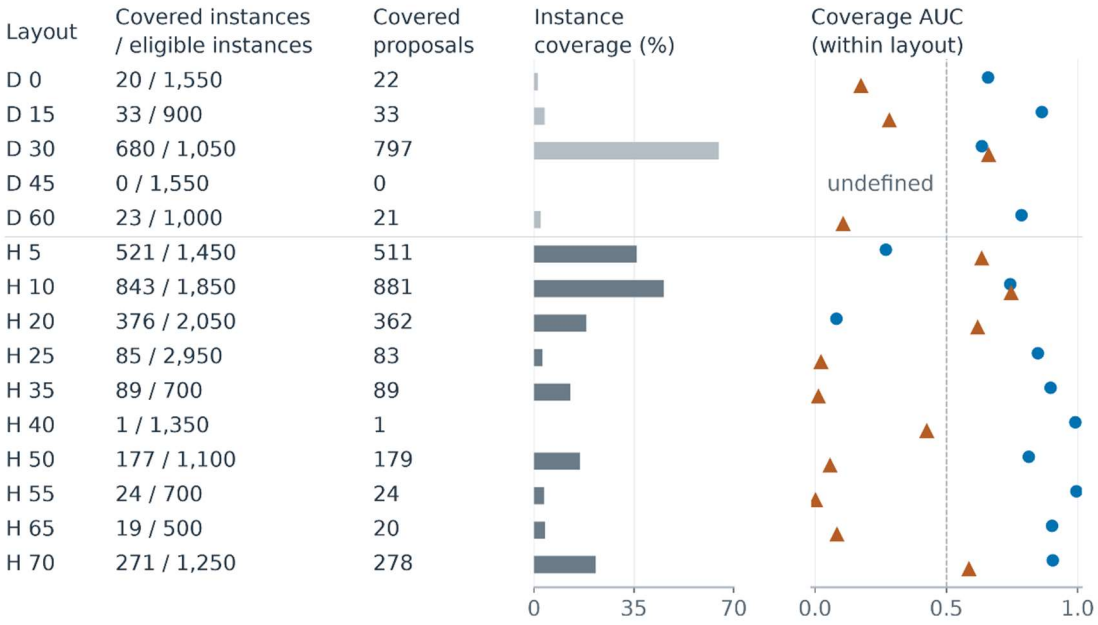


Figure 2. Held-out CAAP differences (pp) from appearance-only scoring with 95% intervals

4.3. Post Hoc Instance Coverage and Layout Heterogeneity

Post hoc analysis showed substantial variation in instance coverage across layouts (Figure 3). At IoU $\geq$ 0.5, proposals covered 756 of 6,050 development instances (12.5%) and 2,406 of 13,900 held-out instances (17.3%). Eight layouts had coverage below 4%, while the remaining seven exceeded 10%. All layouts were retained in the analysis.

ISSN: 2664-9640



D: development; H: held-out. Each layout has 50 images.
AUC: blue circles = appearance q; orange triangles = mean silhouette score m.
Layout 45: no covered proposal. Layout 40: one covered proposal; descriptive only.

Figure 3. Instance coverage and within-layout coverage AUC at IoU ≥ 0.5

Covered development observations were concentrated in layout 30, which accounted for 89.9% of covered instances and 91.3% of covered proposals. Within this layout, the mean silhouette score m achieved a higher coverage AUC than appearance confidence q (0.661 versus 0.634). In contrast, the AUCs of m in layouts 0, 15 and 60 were 0.175, 0.283 and 0.108, respectively. Pooled AUC incorporates both within-layout and cross-layout comparisons.

A post hoc bootstrap analysis assessed how development-layout composition affected the pooled AUC comparison. Each of 1,000 resamples drew five complete development layouts with replacement. The pooled AUC of m exceeded that of q in 0/349 resamples omitting layout 30, 18/403 including it once, 134/191 including it twice and 57/57 including it at least three times. After excluding layout 45, the pooled AUC difference between m and q remained negative at −0.027.

Score ranking also varied across held-out layouts. Layouts 5 and 20 accounted for 37.3% of covered held-out instances, but their appearance AUCs were 0.269 and 0.080, respectively. These values indicate that uncovered proposals tended to rank above covered proposals. Silhouette AUC ranged from 0.003 to 0.083 in layouts 25, 35, 50, 55 and 65, compared with 0.587–0.747 in layouts 5, 10, 20 and 70. Median annotation visibility was 0.970 in development and 0.899 in the held-out set, with all instances exceeding 0.1. The same pre-specified decision rule was retained across layouts.

Development layout 45 had zero instance coverage. Its annotation boxes had median width and height of 34 and 233 pixels, compared with 45 and 50 pixels for proposal boxes; the maximum proposal-to-annotation IoU was 0.065. Figure 4 illustrates this geometric mismatch alongside a covered example. In layout 45, frame 0, all 174 proposal centers lay outside the annotation envelope, and none of the 31 instances was covered. In held-out layout 10, frame 19, 159 of 191 proposal centers lay outside the envelope, while 26 of 37 instances were covered. The displayed examples were selected post hoc. In each proposal panel, four boxes were selected by SAM predicted IoU and four additional boxes by annotation overlap, solely for visualization.

ISSN: 2664-9640

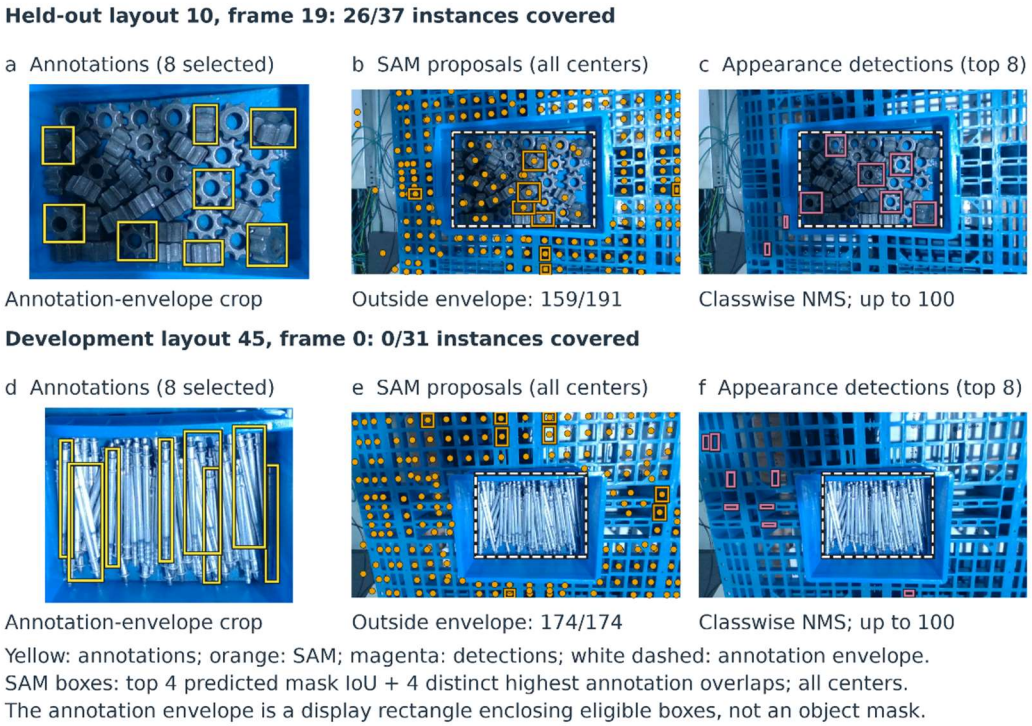


Figure 4. Annotations, SAM proposals and appearance-based detections in two example frames. Adapted from XYZIBD [20, 26], CC BY-NC-SA 4.0

SAM's predicted mask-quality score u had standalone coverage AUCs of 0.229 in development and 0.204 in the held-out set. This score estimates mask quality, while coverage labels are defined by proposal-box overlap with annotations. Raw predicted-IoU values, including values above 1, were retained as model scores. Held-out means and population standard deviations were 0.627 (0.058) for q, 0.541 (0.029) for m and 0.947 (0.030) for u. The standard deviation of m was approximately half that of q, showing that unit coefficients preserved different levels of score variability. These statistics describe the score distributions; the pre-specified fusion coefficients remained unchanged.

4.4. Post Hoc CNOS Comparison and Rescoring

The dataset-adapted CNOS configuration achieved higher instance coverage and detection performance than the study configuration on the same 500 held-out images (Table 3). Its 25,172 size-filtered proposals covered 31.0% of annotated instances at IoU ≥ 0.5 and 19.0% at IoU ≥ 0.75. After detection selection, CNOS produced 23,066 detections, with classwise AP of 12.77%, CAAP of 12.69% and CAAR100 of 16.84%. A separate CPU re-evaluation reproduced the AP and AR values. These results compare complete configurations that differ jointly in proposal generation, appearance features, reference views and detection selection.

Table 3. Comparison of the study and CNOS configurations on the held-out set

Configuration	IoU 0.5	IoU 0.75	CAP	CAAP	CAAR
Study	17.3	10.9	4.55	1.25	7.21
adapted CNOS:	31.0	19.0	12.77	12.69	16.84

Rescoring the same saved CNOS proposals yielded different responses to silhouette addition under the two appearance settings (Table 4). With the study's DINOv2 ViT-S/14 scorer, silhouette addition increased CAAP from 9.56% to 12.47%, corresponding to a difference of

ISSN: 2664-9640

+2.91 percentage points (pp) with a 95% interval of [0.77, 4.48]. Classwise AP increased from 13.29% to 13.47%, and CAAR100 changed from 17.33% to 17.35%.

Table 4. Silhouette-score addition on the saved CNOS proposals

Appearance setting	Score	CAP	CAAP	CAAR	CAAP difference
Study	q	13.29	9.56	17.33	Reference
Study	$q + m$	13.47	12.47	17.35	+2.91 [0.77,4.48]
Native	q	12.77	12.69	16.84	Reference
Native	$q + m$	11.97	12.69	16.96	−0.01 [−1.01,0.79]

With native CNOS scoring, the CAAP difference was −0.01 pp, with a 95% interval of [−1.01, 0.79]; both CAAP values rounded to 12.69%. Classwise AP decreased from 12.77% to 11.97%, while CAAR100 changed from 16.84% to 16.96%. The study-scorer interval lay entirely above zero, whereas the native-scorer interval spanned zero. These post hoc results characterize the observed layouts under the respective appearance and detection-selection settings.

On the saved proposal set, the library-wide mean silhouette score had a coverage AUC of 0.628, compared with 0.804 for the study's appearance confidence and 0.941 for native CNOS confidence. Restricting silhouette averaging to the CAD model selected by the corresponding appearance scorer yielded coverage AUCs of 0.772 for study-selected identities and 0.698 for CNOS-selected identities. Both exceeded the library-wide silhouette AUC. These identity-specific scores were evaluated as descriptive ranking controls; the fused-detector comparisons in Table 4 used the library-wide mean silhouette score with the fixed unit coefficient.

5. Discussion

5.1. Score Fusion and Detection Performance

The pre-specified experiment evaluated whether a development decision based on standalone coverage AUC could guide silhouette-score addition. The mean silhouette score achieved a lower development AUC than appearance confidence, setting $d = 0$. Conditional scoring consequently reproduced the appearance-only predictions, and the primary CAAP difference was zero by construction. The fixed-addition comparisons directly evaluated fusion on the held-out set: silhouette addition yielded a CAAP difference of +1.28 percentage points (pp), with a 95% interval of [−0.20, 2.68], while SAM addition yielded −0.33 pp [−1.22, 0.13]. Both intervals spanned zero.

The metrics showed different responses to silhouette addition. The CAAP point estimate increased, while classwise AP, coverage AUC and CAAR100 decreased. Coverage AUC measures the ordering of covered and uncovered proposals, including multiple proposals associated with the same instance. Detection AP applies one-to-one matching at each IoU threshold and evaluates performance across multiple thresholds [22, 24]. Scores also affect NMS and the selection of the highest-ranked detections. These differences support evaluating confidence fusion through several complementary metrics, with classwise AP assessing performance when predicted identities are taken into account.

The development results further distinguish standalone score quality from the performance of a score combination [8]. Although $m$had a lower standalone AUC than q, the combined score $q + m$ achieved a higher development AUC, at 0.664 versus 0.645. Thus, the standalone screening rule rejected a combination with better development ranking in this experiment. The

ISSN: 2664-9640

result illustrates the importance of evaluating the combination itself. Its contribution to detection performance is then assessed through the complete selection and evaluation pipeline on held-out data.

5.2. Proposal Coverage and the CNOS Comparison

Proposal coverage imposed a substantial constraint on the study configuration. At IoU ≥ 0.5, its proposals covered 17.3% of held-out instances, leaving 82.7% without a qualifying box. Rescoring preserves this coverage limit because proposal locations remain fixed. Coverage and score ordering also varied across layouts, with some relatively well-covered layouts exhibiting appearance AUCs below 0.5. These observations identify proposal availability and confidence ranking as separate challenges within the evaluated pipeline, consistent with previous work on proposal recall and industrial detection [17–21].

The adapted CNOS configuration achieved higher complete-system performance on the same images. Its instance coverage was 31.0% at IoU ≥ 0.5, compared with 17.3% for the study configuration, and its CAAP was 12.69%, compared with 1.25%. Classwise AP likewise increased from 4.55% to 12.77%. This comparison establishes the performance gap between the configurations while jointly varying proposal generation, appearance features, reference views and detection selection.

Rescoring the same saved CNOS proposals revealed different responses to silhouette addition under the two scoring settings. With the study scorer, CAAP increased from 9.56% to 12.47%, corresponding to +2.91 pp [0.77, 4.48], while classwise AP increased from 13.29% to 13.47%. With native CNOS scoring, the CAAP difference was −0.01 pp [−1.01, 0.79], and classwise AP decreased from 12.77% to 11.97%. The study scorer with silhouette addition therefore remained slightly below native CNOS in CAAP while achieving higher classwise AP. These results characterize configuration-specific responses to fusion on a common proposal set.

5.3. Limitations and Future Work

The experimental scope comprises RGB images from one dataset, one CAD library and a fixed SAM ViT-B/DINOv2 ViT-S/14 configuration. The partition holds out object identities together with layouts, and covered development observations are concentrated in one layout. The bootstrap intervals describe variation across the 10 observed held-out layouts. The experimental plan was recorded internally before image download and processing, while the coverage diagnostics and CNOS analyses were designed after inspection of the results. Independent evaluation on additional layouts and object identities is therefore the next step in assessing the reproducibility of these findings.

Coverage was measured using clipped amodal bounding boxes, and the spatial diagnostics describe geometric relationships between proposals and annotations. These measurements establish the observed coverage constraint; identifying its causes requires controlled comparisons of proposal generation, object scale, rendering and image preprocessing. The mean silhouette score also removes absolute size and uses a fixed unit coefficient. Its role is that of a shape-based ranking score, and the fusion procedure operates on raw score scales rather than calibrated probabilities.

Future experiments should evaluate score combinations directly within the complete detector using a pre-specified development procedure followed by independent held-out evaluation. Identity-specific silhouette scoring provides one candidate for such an experiment: restricting averaging to the appearance-selected CAD model increased standalone coverage AUC from 0.628 to 0.772 with study-selected identities and to 0.698 with CNOS-selected identities. This observation motivates testing its contribution to fused detection and connects with the use of identity-specific geometric evidence in SAM6D [4] and DepthNOCTIS [12]. Controlled

ISSN: 2664-9640

comparisons should examine proposal quality, appearance scoring and silhouette aggregation separately, while reporting instance coverage, classwise AP and CAAP together.

6. Conclusion

This study evaluated the addition of a library-wide mean silhouette score to appearance confidence for CAD-model-based part detection on XYZIBD. The pre-specified development rule retained appearance-only scoring, making the primary CAAP difference zero by construction. On the held-out set, fixed silhouette addition yielded a CAAP difference of +1.28 percentage points (pp), with a 95% interval of [−0.20, 2.68], while fixed SAM addition yielded −0.33 pp [−1.22, 0.13]. Both intervals spanned zero. Instance coverage of 17.3% at IoU ≥ 0.5 imposed a substantial limit on detection through rescoring.

Post hoc analysis of the saved CNOS proposals showed different responses to silhouette addition: the CAAP difference was +2.91 pp [0.77, 4.48] for the study appearance scorer and −0.01 pp [−1.01, 0.79] for native CNOS scoring. These results characterize the contribution of silhouette fusion under the evaluated configurations and support assessing score combinations within the complete detector, alongside proposal coverage and classwise detection performance. Future evaluation should use new layouts and object identities under a pre-specified protocol.

References

[1] Hinterstoisser, S., et al. (2012). Gradient response maps for real-time detection of texture-less objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(5), 876–888.

[2] Borgefors, G. (1988). Hierarchical chamfer matching: A parametric edge matching algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 10(6), 849–865.

[3] Nguyen, V. N., Groueix, T., Ponimatkin, G., Lepetit, V., & Hodaň, T. (2023). CNOS: A strong baseline for CAD-based novel object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 2126–2132). Paris, France.

[4] Lin, J., Liu, L., Lu, D., & Jia, K. (2024). SAM-6D: Segment anything model meets zero-shot 6D object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 27906–27916). Seattle, USA.

[5] Cho, S., Park, S., & Oh, I. (2026). MUSE: Model-based uncertainty-aware similarity estimation for zero-shot 2D object detection and segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 6090–6100). Tucson, USA.

[6] Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., & Zabulis, X. (2017). T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision* (pp. 880–888). Santa Rosa, USA.

[7] Xiang, Y., Schmidt, T., Narayanan, V., & Fox, D. (2018). PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *Robotics: Science and Systems XIV*. Pittsburgh, USA.

[8] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.

[9] Lu, Y., P, J. J., Guo, Y., Ruozzi, N., & Xiang, Y. (2025). Adapting pre-trained vision models for novel instance detection and segmentation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 13341–13348). Hangzhou, China.

[10] Liu, S., et al. (2024). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision* (Vol. 15105, pp. 38–55). Lecture Notes in Computer Science. Milan, Italy.

[11] Gandyra, M., Santonicola, A., & Beetz, M. (2025). NOCTIS: Novel object cyclic threshold based instance segmentation. arXiv preprint arXiv:2507.01463. https://arxiv.org/abs/2507.01463

ISSN: 2664-9640

[12] Gandyra, M., Santonicola, A., & Beetz, M. (2025). Depth-NOCTIS: Depth-based novel object cyclic threshold based instance segmentation. In *British Machine Vision Conference DIFA Workshop*. Sheffield, UK.

[13] Ji, S., Kwon, Y., & Kim, H.-S. (2026). CLON: Cue-calibrated linguistic object onboarding for zero-shot 6D pose front-ends. arXiv preprint arXiv:2609.04784v1. https://arxiv.org/abs/2609.04784v1

[14] Alexe, B., Deselaers, T., & Ferrari, V. (2012). Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11), 2189–2202.

[15] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). Proceedings of Machine Learning Research. Sydney, Australia.

[16] Küppers, F., Kronenberger, J., Shantia, A., & Haselhoff, A. (2020). Multivariate confidence calibration for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1322–1330). Virtual Conference.

[17] Hosang, J., Benenson, R., Dollár, P., & Schiele, B. (2016). What makes for effective detection proposals? IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(4), 814–830.

[18] Hodaň, T., et al. (2024). BOP Challenge 2023 on detection, segmentation and pose estimation of seen and unseen rigid objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 5610–5619). Seattle, USA.

[19] Nguyen, V. N., et al. (2025). BOP Challenge 2024 on model-based and model-free 6D object pose estimation. arXiv preprint arXiv:2504.02812v4. https://arxiv.org/abs/2504.02812v4

[20] Huang, J., Hu, J., Yu, P. K. T., Ilic, S., Sundermeyer, M., & Busam, B. (2026). XYZ-IBD: Benchmarking robust 6D object pose estimation under real-world industrial complexity. arXiv preprint arXiv:2506.00599v3. https://arxiv.org/abs/2506.00599v3

[21] Hu, J., Xu, H., Huang, J., Yu, P. K. T., Ilic, S., & Busam, B. (2025). BioDet: Boosting industrial object detection with image preprocessing strategies. arXiv preprint arXiv:2510.21000. https://arxiv.org/abs/2510.21000

[22] Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). Pittsburgh, USA.

[23] Lin, T.-Y., et al. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision* (Vol. 8693, pp. 740–755). Lecture Notes in Computer Science. Zurich, Switzerland.

[24] COCO Dataset. (n.d.). pycocotools. https://github.com/cocodataset/cocoapi/blob/master/PythonAPI/pycocotools

[25] Hodaň, T. (n.d.). bop_toolkit. https://github.com/thodan/bop_toolkit

[26] BOP Benchmark. (n.d.). xyzibd dataset. https://huggingface.co/datasets/bop-benchmark/xyzibd

[27] Kirillov, A., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3992–4003). Paris, France.

[28] Oquab, M., et al. (2024). DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research.

[29] Hj, D. (n.d.). Baselines-for-Industrial-Bin-Picking-BOP2025. https://github.com/demianhj/Baselines-for-Industrial-Bin-Picking-BOP2025