Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sensitivity of Underwater Fish Mask AP to Geometric Instance Splitting

Junjie Qiang

qjj1067@126.com

Fujian University of Technology

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11722722/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Sensitivity of Underwater Fish Mask AP to Geometric Instance Splitting

Junjie Qiang[1*]

[1*]Fujian University of Technology,Fuzhou,China.

Corresponding author(s). E-mail(s): qjj1067@126.com;

Abstract

Instance segmentation scores depend on how reference pixels are grouped into instances. For one non-crowd reference mask partitioned into disjoint parts, ideal predictions equal to the parts have a single-target score ceiling of $\boldsymbol{C(a) = 10 \sum_{t \in T} 1\{t \leq a\}}$, where $\boldsymbol{a}$ is the largest part's area share and $\boldsymbol{T}$ contains the ten overlap thresholds. We measured pooled mask average precision (AP) on 1,050 underwater images and 3,669 source fish masks, preserving their foreground unions and holding predictions fixed. For two-way splits over nominal rates of 1–10%, each additional percentage point split lowered AP by approximately 0.6–1.1 points. Near-1% two-way scenarios yielded estimated losses of 0.56–1.44 points. In the three-way scan, all ideal children of masks with realized $\boldsymbol{a < 0.50}$ fell below all ten thresholds against their source mask. With frozen detections evaluated against the split references, pooled AP fell by about 1.8 points at the nominal 1% rate and approached zero at full splitting. In one two-way full-split condition the checkpoint gap changed sign, with the paired confidence interval spanning zero.

Keywords: Instance segmentation, Underwater imagery, Mask average precision, Annotation granularity, Geometric sensitivity

1 Introduction

Underwater instance segmentation is used to locate and delineate individual objects in marine imagery. WaterMask and related underwater benchmarks support the development and evaluation of underwater segmentation systems using pixel-level instance annotations [1]. In a Common Objects in Context (COCO) evaluation, these annotations determine both the masks against which detections are matched and the number of instances in the recall denominator [2].

A foreground region can be represented as one mask or as multiple masks with the same pixel union. Changing this representation can alter threshold eligibility, the number of available matches, and the pooled precision–recall curve even when the detector outputs are identical. For a prediction equal to one part of a disjoint partition, its intersection over union (IoU) against the full source mask is the part's area share. This identity motivates a controlled study in which a is the largest child's share of its parent mask area and f is the percentage of source fish instances selected for splitting.

Our contributions are as follows.

1. We derive a closed-form threshold-coverage ceiling for ideal part predictions against an isolated non-crowd reference mask, with $a \geq 1/k$ for a partition into k disjoint parts.

2. We quantify pooled Mask AP sensitivity to geometric splitting on 1,050 images and 3,669 source fish instances. The estimated loss over nominal rates of 1–10% is 0.6–1.1 AP points per additional percentage point split, depending on the target area share.
3. We place the absolute AP changes alongside reported underwater method gaps and estimate checkpoint differences using paired group resampling. The observed low-rate ordering persisted, and the sign change at one full-split endpoint remained uncertain.

2 Related work

2.1 Underwater instance segmentation

WaterMask introduced UIIS and a mask head for underwater imagery [1]. USIS-SAM introduced USIS10K for salient underwater instances and adapted the Segment Anything Model (SAM) to this task [3, 4]. UIIS10K expanded the collection of pixel-level underwater annotations and accompanied the UWSAM framework (Li et al., 2025, arXiv preprint 2505.15581v4). DiveSeg adapts DINO representations through underwater alignment and object-prior prompting, while BiPA uses bilevel prompt adaptation [5, 6]. In underwater fish video segmentation, Saleh et al. used self-supervision to reduce reliance on manual annotations during training [7]. We use the UIIS10K fish category and freeze both checkpoints' outputs during evaluation.

2.2 Annotation quality and reference uncertainty

LVIS and UVO illustrate how annotation coverage is tied to the benchmark task [8, 9]. Northcutt et al. showed that correcting test labels can change benchmark conclusions [10]. COCONut and COCO-ReM revisited mask annotations and detection benchmark labels, respectively [11, 12]. Kimhi et al. benchmarked spatial annotation noise in segmentation (2025, arXiv preprint 2406.10891v3), and Nordström et al. studied how label noise affects segmentation objectives and threshold choice [13]. Murrugarra-Llerena and Jung found that bounding-box AP can be sensitive to annotation variation for small objects and strict IoU thresholds [14]. Our perturbation changes instance count and grouping while preserving the foreground union. The resulting curves measure sensitivity at specified split rates and area shares.

2.3 AP, boundaries, and scale

TIDE decomposes errors in detection and instance segmentation, and Localization Recall Precision separates aspects of performance that pooled AP can obscure [15, 16]. Dave et al. studied how detection caps can affect category independence and rankings in large-vocabulary evaluation (2022, arXiv preprint 2102.01066v2). The COCO evaluator also uses distinct matching and ignore rules for crowd annotations. Our ceiling concerns a non-crowd target, and the detection cap stays fixed across comparisons. COCO evaluator

Boundary IoU measures contour agreement and underlies Boundary AP [17]. Panoptic Quality separately incorporates segment matching and overlap [18]. Small-object benchmark work motivates care with scale-specific comparisons [19]. We analyze standard fish Mask AP across all object sizes. Mask R-CNN provides the architecture of our frozen checkpoints. Mask Scoring R-CNN calibrates mask quality scores, and DynaMask adapts mask resolution [20–22]. The masks and confidence scores from our checkpoints remain fixed throughout each scan.

3 Methods

Large language model assistants supported manuscript drafting and code preparation, as detailed in Statements and Declarations.

Figure 1 illustrates the intervention using an archived single-image diagnostic from outside the main evaluation cohort. Both reference variants use the same duplicate-normalized parent region.

3.1 Single-target threshold coverage

Let M be a nonempty, non-crowd source mask, and let $M_1, \ldots, M_k$ be nonempty, disjoint submasks whose union equals M. For ideal part predictions $P_j = M_j$, define $a_j = |M_j|/|M|$ and $a = \max_j a_j$. Then

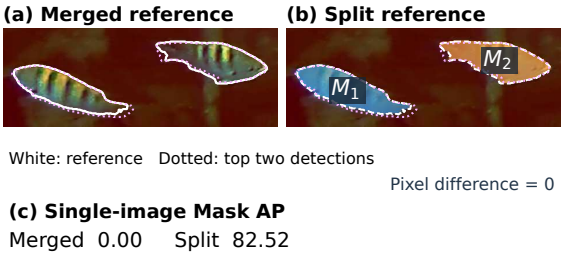

Fig. 1 Reference-grouping intervention on UIIS10K image `test_1465.jpg` (case P10). Panels (a) and (b) share the same crop, frozen detections and scores, and foreground pixel union. The white dashed contour in (b) traces the union of the two child masks. Panel (c) reports single-image fish Mask AP over all eight frozen detections. The dotted contours in (a) and (b) show the two highest-scoring detections. Image credit: UIIS10K (Li et al.)

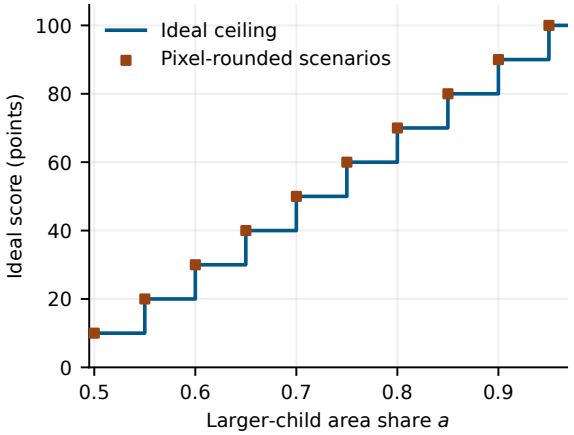

Fig. 2 Ideal single-target threshold coverage. The step function is indexed by the realized larger-child share a. For each target area share in the area scan, a marker shows the median realized a on the horizontal axis and the mean of the per-mask ideal scores on the vertical axis.

$$\mathrm{IoU}(P_j, M) = a_j, \qquad \sum_{j=1}^{k} a_j = 1, \qquad a \geq 1/k.$$

For the isolated source target, a threshold t admits a match exactly when $t \leq a$. COCO allows at most one match to this non-crowd target at each threshold. On the 0–100 scale, the score ceiling for this single-target setting, with the best eligible part ranked first, is

$$C(a) = 10 \sum_{t \in \{0.50, 0.55, \ldots, 0.95\}} \mathbf{1}\{t \leq a\}.$$

For $k = 2$, $a \geq 0.50$, so at least one threshold is reachable. If each part occupies at least 10% of the source mask, the ceiling ranges from 10 to 90. When $a \geq 0.95$, all ten thresholds are reachable and $C(a) = 100$. For $k \geq 3$, threshold coverage is zero when $a < 0.50$. Fig. 2 shows the resulting steps. Pooled AP also depends on detection scores, other masks, competing matches, and recall normalization.

3.2 Cohorts, frozen predictions, and scoring

The geometric screen used 3,702 non-crowd fish masks from 1,060 test images and 1,058 near-duplicate groups. It used the existing sampling frame and excluded groups in the annotation pilot and the earlier cross-group case evaluation. For the evaluation cohort, we selected the largest eligible source fish mask per group, with a minimum of 100 native pixels and the lowest source annotation ID used to resolve area ties. The resulting cohort contained 1,050 groups, one image per group, and 3,669 source fish annotations when all fish on those images were included.

The original Mask R-CNN R50-FPN checkpoint contributed 8,988 cached fish detections. The group-trained R50-FPN checkpoint contributed 8,786 detections from a single local CPU inference pass. The original split randomly assigned groups of byte-identical images, while the group-trained split kept near-duplicate groups intact. Training procedures also differed. We evaluated the frozen detections using fish-only COCO Mask AP pooled over images, IoU thresholds 0.50–0.95 in increments of 0.05, and a maximum of 100 detections per image for the main AP score. Images with no fish detections remained in the evaluation. Mask area for predictions was computed from the encoded masks. COCO evaluator

3.3 Geometric screen and area scan

The primary screen used eight-connected components covering at least 10% of the source mask as seeds and assigned residual mask pixels to the nearest seed. The extended screen additionally checked narrow connections at native scales of one

to five pixels in connected masks. Both rules used source geometry alone.

In the area scan we split one selected target per image, for a fixed split rate $f = 100 \times 1{,}050/3{,}669 \approx 28.6$ percentage points (28.6%). For each target, we sorted its pixels by horizontal or vertical coordinate, used the other coordinate to break ties, and split the ordered pixels at the prescribed area count. The target a took ten values from 0.50 to 0.95 in increments of 0.05. The larger count was rounded upward to a whole pixel. The encoded children were verified to be disjoint and to have the same pixel union as the source.

3.4 Randomized split-rate scan and inference

We randomly permuted all 3,669 source fish instances with a fixed seed and took nested prefixes at nominal split rates of 1%, 2%, 5%, 10%, 25%, 50%, and 100%. We used left-to-right cuts at target larger-child shares of 0.654 and 0.90, rounding to the nearest feasible whole-pixel count. Hereafter, the area settings are labeled by these target values of a. The same selected instances were used at both shares and for both checkpoints. Every selected source mask was replaced by two nonempty, disjoint masks with exactly the same pixel union. All other annotations and all predictions were held fixed. The realized split rate was $f = 100n/3{,}669$ in percentage points, where n is the selected-instance count. At the lowest nonzero setting, 37 instances (1.008%) on 34 images (3.24%) were split.

We selected the nominal 1%, 2%, 5%, and 10% conditions and fitted $\mathrm{AP}(f) = \mathrm{AP}(0) - sf$ using the realized split rates in percentage points. The intercept was fixed at source AP. We obtained pointwise 95% percentile confidence intervals from 1,000 paired near-duplicate-group bootstrap draws. Each draw used the same group weights across models and conditions, recomputed the realized split rate, and refitted the slope. The intervals are conditional on the fixed randomized assignment. For the full-range reference we linearly interpolated between the source AP and the AP at a 100% split rate.

3.5 Three-way splits and cut direction

For the three-way scan we used the same nested assignment at all seven nominal split rates. We sorted source pixels left to right and formed three disjoint, nonempty blocks. The largest block used the feasible pixel count closest to the target share ($a = 0.34$, 0.40, or 0.45), subject to a realized share below 0.50 whenever possible. The remaining pixels were divided as evenly as possible. When the below-0.50 constraint was infeasible, we used the smallest feasible largest block. We verified the union pixelwise and checked the ideal child-to-parent IoUs using the encoded masks.

For the direction check we repeated the two-way scan at nominal rates of 1%, 2%, 5%, and 10%, with top-to-bottom coordinate ordering. We held the selected masks, child pixel counts, and predictions fixed. To compare the fitted slopes we subtracted the left-to-right slope from the top-to-bottom slope within each of the same 1,000 group bootstrap draws.

4 Results

4.1 Geometry and fixed-rate area scan

The area scan splits one selected target per image, or 1,050 of 3,669 source instances ($f = 28.6\%$), and varies a. The separate rate scan uses nested random selections and reaches $f = 100\%$ by splitting all 3,669 masks. Its area settings are $a = 0.654$ and $a = 0.90$.

The two geometric screens triggered on 35/3,702 (0.95%) and 49/3,702 (1.32%) masks. Both returned six masks partitioned into at least three parts. The extended screen returned 43 two-part cases, with a ranging from 0.517 to 0.894 and a median of 0.654. These are geometry-based screening statistics.

At fixed $f = 28.6\%$, the original checkpoint's AP increased as a rose from 0.50 to 0.95. It rose from 13.78 to 39.37 for left-to-right cuts and from 13.64 to 39.40 for top-to-bottom cuts. The original checkpoint led in all 20 conditions, by 0.17–0.89 AP points, compared with 1.19 points under the source annotations. This comparison is descriptive and uses a fixed, target-selected cohort of masks.

4.2 Low-rate slopes and departure from linear mixing

Figure 3 shows the split-rate curves and the low-rate fits.

At $a = 0.654$, AP fell by **1.09 points per additional percentage point of instances split** for the original checkpoint (95% CI 1.02–1.17) and **1.07 points** for the group-trained checkpoint (0.99–1.15). At $a = 0.90$, the corresponding slopes were **0.60** (0.57–0.64) and **0.59** (0.56–0.63). Within each checkpoint, the paired slope difference between area settings was 0.49 (0.44–0.53) for the original and 0.48 (0.43–0.52) for the group-trained checkpoint, in AP points per percentage point split.

At $f = 1.008\%$, the original checkpoint lost 1.05 AP points (0.70–1.47) at $a = 0.654$ and 0.64 points (0.42–0.84) at $a = 0.90$. The largest residual across the four low-rate fits was 0.22 AP points. In the full scan, all six interior rates yielded lower AP than endpoint interpolation. Around $f = 50\%$, the deficit reached 5.75–12.65 points across the four curves. For the original checkpoint at $a = 0.654$, the full-range average decline of 0.46 AP points per percentage point is less than half the low-rate slope of 1.09. The declines per percentage point were smaller over the higher-rate intervals. This difference in slopes motivates using the local fit for near-1% scenarios. Global precision–recall accumulation and the changing recall denominator can both contribute to nonlinear behavior.

4.3 Checkpoint gap and endpoint uncertainty

The source-annotation gap was +1.194 AP points (95% CI 0.638–1.733), with the gap defined as original minus group-trained AP. Table 1 shows an initial fluctuation followed by narrower gaps at higher split rates. The paired low-rate slope difference $\Delta s = s_{\text{original}} - s_{\text{group-trained}}$ was **+0.019** AP points per percentage point at $a = 0.654$ (95% CI −0.002 to +0.048) and **+0.010** at $a = 0.90$ (−0.005 to +0.028). Both intervals include zero, leaving the systematic low-rate change in the gap unresolved.

The original checkpoint retained its lead in all eight low-rate comparisons, with pointwise paired intervals above zero. At $a = 0.654$, the interval already included zero at the nominal 50% split rate. At the 100% split rate, both checkpoints scored below 7 AP (6.56 and 6.65). The point estimate changed sign, and the paired interval spanning zero leaves the ranking uncertain at this endpoint. At $a = 0.90$, the original checkpoint retained a positive endpoint gap.

4.4 Three-way splitting and the zero-coverage case

At each target share, 3,668 of 3,669 source masks had realized $a < 0.50$. For each of these masks, no ideal child prediction reached any of the ten IoU thresholds against its parent, giving $C(a) = 0$. The remaining mask contained four pixels. Three nonempty children necessarily have sizes $(2, 1, 1)$, so $a = 0.50$ and $C(a) = 10$. This mask was retained and entered the intervention only at $f = 100\%$. All 11,007 partitions across the three shares preserved the source pixel union exactly and had no child overlap.

A prediction equal to the unsplit parent has IoU with each child equal to that child's area share. With $a < 0.50$, all three IoUs fall below the minimum threshold. An exact parent prediction can reach the minimum threshold for the largest child in a two-way split because its share is at least 0.50. Archived detections can also match individual children.

Table 2 shows that pooled AP fell by about 1.8 points at the nominal 1% rate. At $f = 100\%$, pooled AP was 0.0052–0.0103 across both checkpoints and all three target shares. This pooled result reflects both detection matching and the tripling of the recall denominator from 3,669 to 11,007 fish masks. The single-target ceiling $C(a) = 0$ for $a < 0.50$ follows from the area-share identity independently of pooling. These are controlled low-overlap scenarios. Their assigned split rates do not estimate the prevalence of three-way merges. Online Resource 1 gives the curves for each target share and the mask-level checks.

4.5 Sensitivity to cut direction

At $a = 0.654$, top-to-bottom slopes were 1.078 (95% CI 1.008–1.159) for the original checkpoint and 1.050 (0.979–1.134) for the group-trained checkpoint. At $a = 0.90$, they were 0.592 (0.560–0.632) and 0.582 (0.546–0.620), respectively. The

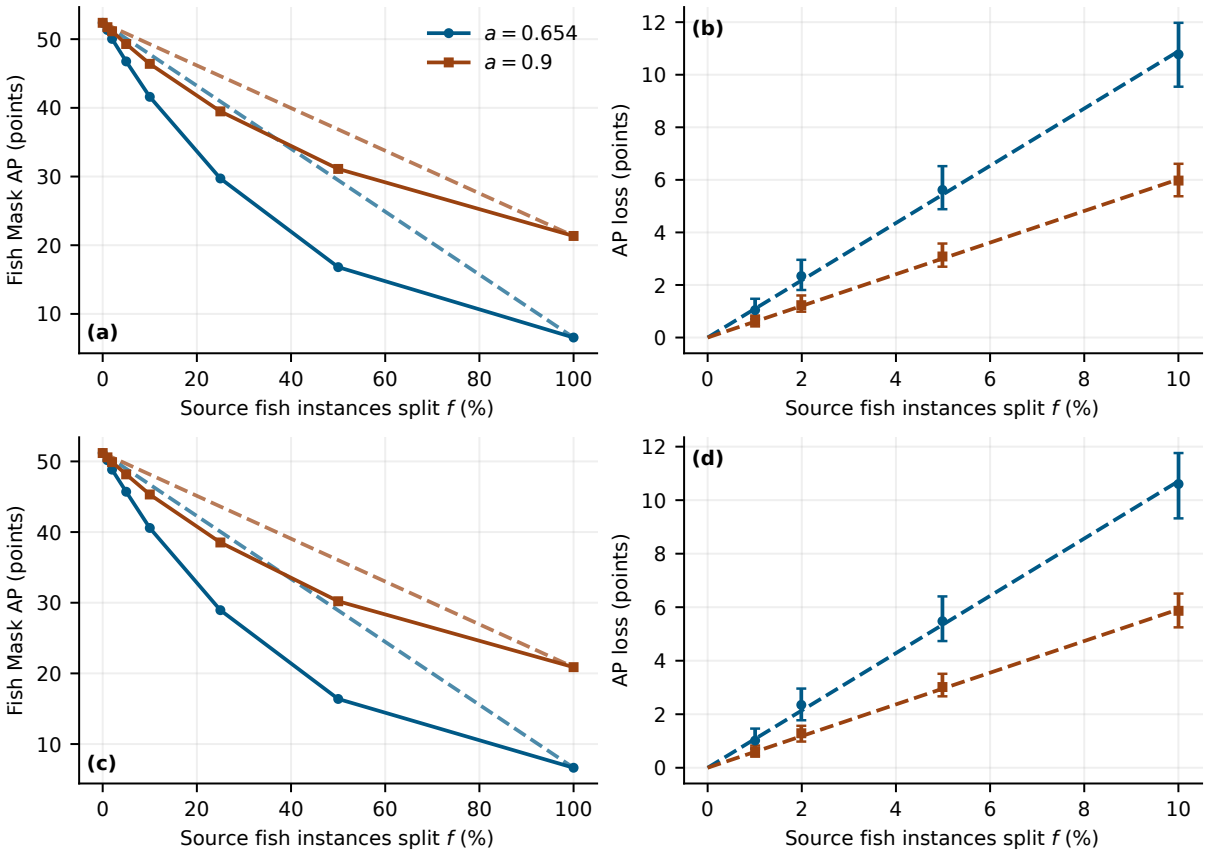

Fig. 3 Fish Mask AP as a function of the instance split rate f among 3,669 source fish instances. Panels (a) and (b) show the original checkpoint, and (c) and (d) show the group-trained checkpoint. Panels (a) and (c) show pooled AP across the full range, with dashed endpoint interpolations. Panels (b) and (d) show AP loss at low rates, with dashed local fits and pointwise 95% paired bootstrap intervals. Circles and squares mark target area shares of $a = 0.654$ and $a = 0.90$, respectively

Table 1 Original minus group-trained fish Mask AP. Parentheses give pointwise 95% paired bootstrap confidence intervals. The displayed split rate uses the full source-instance denominator.

Realized f	Instances split	$a = 0.654$	$a = 0.90$
0.000%	0	+1.194 (0.638, 1.733)	+1.194 (0.638, 1.733)
1.008%	37	+1.165 (0.661, 1.722)	+1.168 (0.640, 1.721)
1.990%	73	+1.203 (0.623, 1.720)	+1.252 (0.607, 1.706)
4.988%	183	+1.054 (0.530, 1.586)	+1.127 (0.603, 1.657)
10.003%	367	+1.023 (0.456, 1.478)	+1.087 (0.580, 1.592)
24.993%	917	+0.772 (0.317, 1.172)	+0.960 (0.528, 1.394)
50.014%	1,835	+0.429 (−0.005, 0.844)	+0.899 (0.491, 1.197)
100.000%	3,669	−0.084 (−0.218, 0.185)	+0.467 (0.202, 0.744)

paired top-minus-left slope differences at $a = 0.654$ were −0.0118 (−0.0325 to +0.0081) and −0.0210 (−0.0319 to +0.0045). At $a = 0.90$, they were −0.0097 (−0.0214 to +0.0032) and −0.0095 (−0.0248 to −0.0002). All slopes and differences are in AP points per percentage point split. The four point estimates changed by at most 0.022 in these units. The last pointwise interval lies slightly below zero. Online Resource 1 provides the per-rate scores and paired draws.

Table 2 Fish Mask AP under three-way splitting. Entries span the three target area shares. Bootstrap confidence intervals were not computed for this scan

Split rate	Original	Group-trained
0%	52.38	51.18
1.008%	50.56–50.57	49.35–49.36
1.990%	48.54–48.55	47.32–47.33
4.988%	43.32–43.40	42.26–42.27
10.003%	35.67–35.68	34.78–34.79
24.993%	20.44–20.46	19.75–19.76
50.014%	7.18–7.20	6.89–6.91
100%	0.0052–0.0091	0.0059–0.0103

4.6 Distribution of image-level changes

Online Resource 1 shows empirical cumulative distributions of single-image Mask AP changes for the two-way scan. All images with unchanged annotations had zero change. Across the four checkpoint and area-share combinations, AP decreased on 32–33 of the 34 affected images at the nominal 1% rate and on 234–236 of 239 at 10%. The four largest individual losses accounted for 41.3–44.1% of the summed negative image-level changes at 1%. Restoring the source annotations for those images recovered 9.1–20.8% of the pooled AP loss. Pooled AP ranks detections across images and recomputes the precision–recall curve, so this recovery fraction can differ from the share of summed image-level losses. At 10%, restoring the 24 images with the largest individual losses recovered 7.7–10.0%. Pooled COCO Mask AP remains the primary metric.

5 Discussion and limitations

For two-way splits ($k = 2$), using the rounded geometric screening frequencies of 0.95% and 1.32% as hypothetical values of f gives an AP loss envelope of **0.56–1.44 points**. It summarizes eight fitted point estimates across two checkpoints, two area settings, and two rates, without weighting by the empirical area-share distribution. The 0.95% scenario lies slightly below the lowest measured rate. Both screens partitioned six of 3,702 masks (0.16%) into three or more parts. Most flagged masks yielded two parts, including 43 of 49 under the extended screen. The three-way splits produced larger losses at the same positive rates under the tested area settings (Table 2). The cut-direction check yielded similar two-way slopes, and the image-level diagnostic showed losses on most affected images. Table III of the UWSAM preprint reports UIIS10K multiclass mask AP of 35.8 for Mask R-CNN, 37.4 for WaterMask, and 38.7 for UWSAM-Student, giving successive gaps of 1.6 and 1.3 points. These gaps provide an external magnitude reference for the two-way fish-only estimates. The 1.19-point source gap is a descriptive checkpoint comparison. At low two-way rates, the paired slope-difference intervals span zero and the observed ordering is stable. UWSAM Table III

The intervention holds detector predictions and foreground pixel unions fixed while varying geometric subdivision. The study did not establish the biological identities of the child masks or the true prevalence of merged individuals. Consequently, the estimated losses apply to the specified geometric scenarios. The primary comparison covers two checkpoints with a shared architecture and different training procedures. On a systematic 50-image audit, the original checkpoint's CPU rerun and historical GPU cache gave a maximum absolute Mask AP difference of 0.00000 points across the source and 20 area-scan conditions. The audit supports reproducibility for those images and conditions.

A public grayscale YOLO11n-seg checkpoint obtained 3.07 fish AP in the domain-transfer comparison. Its 49.31-point gap from the original checkpoint places that comparison outside the close-score setting of the primary analysis. Model card

6 Conclusion

With fixed predictions and preserved foreground pixel unions, two-way splitting reduced Mask AP by approximately 0.6–1.1 points per additional percentage point split at nominal rates of 1–10%. For the tested three-way partitions with realized $a < 0.50$, ideal child-to-parent matching yielded $C(a) = 0$. Frozen detector AP against the split references fell by about 1.8 points at the nominal 1% rate and approached zero at full splitting. The low-rate two-way checkpoint ordering remained consistent, while one two-way full-split endpoint left the ranking uncertain.

Acknowledgements

Not applicable.

Statements and Declarations

Funding. No funding was received for this work.

Competing interests. The author declares no competing interests.

Author contributions. Junjie Qiang is the sole author and takes responsibility for the study and manuscript.

Use of generative AI. Large language model assistants were used to help draft parts of the manuscript, review the methodology, check numerical claims, and prepare analysis and plotting code. The evaluation used archived detector predictions and reference partitions generated by deterministic pixel-coordinate rules. Reported scores were calculated by the evaluation scripts. Figures were rendered from archived data, including the original image crop and mask overlays in Figure 1. The author takes full responsibility for verifying the AI-assisted material and for the manuscript's content.

Ethics approval and consent. Not applicable. The study reanalyzed existing benchmark annotations and frozen detector predictions.

Data and code availability. The original annotations are distributed by the UIIS10K project. The replication archive (Online Resource 1) contains the cohort manifest, frozen predictions, evaluation scripts, and supporting analysis outputs.

Online Resource 1. Replication archive for the geometric instance-splitting analyses, including the cohort and intervention manifests, frozen predictions, evaluation code, and paired bootstrap outputs.

References

[1] Lian, S., Li, H., Cong, R., Li, S., Zhang, W., Kwong, S.: WaterMask: Instance segmentation for underwater imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023). https://doi.org/10.1109/ICCV51070.2023.00126

[2] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Computer Vision – ECCV 2014, pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48

[3] Lian, S., Zhang, Z., Li, H., Li, W., Yang, L.T., Kwong, S., Cong, R.: Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. In: Proceedings of the 41st International Conference on Machine Learning, vol. 235, pp. 29545–29559 (2024). Online proceedings

[4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023). https://doi.org/10.1109/ICCV51070.2023.00371

[5] Chen, Z., Zhang, C., Fang, H., Cong, R.: Empowering DINO representations for underwater instance segmentation via aligner and prompter. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, pp. 3201–3209 (2026). https://doi.org/10.1609/aaai.v40i5.37314

[6] Ma, L., Zheng, H., Mao, Y., Liu, J., Xu, C., Xue, X., Wang, Y., He, X., Wang, W.: BiPA: Bilevel prompt adaptation for underwater instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10731–10740 (2026). Online proceedings

[7] Saleh, A., Sheaves, M., Jerry, D., Rahimi Azghadi, M.: Overcoming annotation bottlenecks in underwater fish segmentation: a robust self-supervised learning approach. Signal, Image and Video Processing **19**, 270 (2025) https://doi.org/10.1007/s11760-025-03860-y

[8] Gupta, A., Dollár, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019). https://doi.org/10.1109/CVPR.2019.00550

[9] Wang, W., Feiszli, M., Wang, H., Tran, D.: Unidentified video objects: A benchmark for dense, open-world segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021). https://doi.org/10.1109/ICCV48922.2021.01060

[10] Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. In: NeurIPS Datasets and Benchmarks Track (2021). Online proceedings

[11] Deng, X., Yu, Q., Wang, P., Shen, X., Chen, L.-C.: COCONut: Modernizing COCO segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21863–21873 (2024). Online proceedings

[12] Singh, S., Yadav, A., Jain, J., Shi, H., Johnson, J., Desai, K.: Benchmarking object detectors with COCO: A new path forward. In: Computer Vision – ECCV 2024, pp. 279–295 (2025). https://doi.org/10.1007/978-3-031-72784-9_16

[13] Nordström, M., Maki, A., Hult, H.: The impact label noise and choice of threshold has on cross-entropy and soft-dice in image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20820–20829 (2025). Online proceedings

[14] Murrugarra-Llerena, J., Jung, C.R.: Noise-aware evaluation of object detectors. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2025). Online proceedings

[15] Bolya, D., Foley, S., Hays, J., Hoffman, J.: TIDE: A general toolbox for identifying object detection errors. In: Computer Vision – ECCV 2020 (2020). https://doi.org/10.1007/978-3-030-58580-8_33

[16] Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E.: One metric to measure them all: Localisation recall precision (LRP) for evaluating visual detection tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9446–9463 (2022) https://doi.org/10.1109/TPAMI.2021.3130188

[17] Cheng, B., Girshick, R., Dollár, P., Berg, A.C., Kirillov, A.: Boundary IoU: Improving object-centric image segmentation evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021). https://doi.org/10.1109/CVPR46437.2021.01508

[18] Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019). https://doi.org/10.1109/CVPR.2019.00963

[19] Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J.: Towards large-scale small object detection: Survey and benchmarks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) https://doi.org/10.1109/TPAMI.2023.3290594

[20] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (2017). https://doi.org/10.1109/ICCV.2017.322

[21] Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring R-CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019). https://doi.org/10.1109/CVPR.2019.00657

[22] Li, R., He, C., Li, S., Zhang, Y., Zhang, L.: DynaMask: Dynamic mask selection for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023). https://doi.org/10.1109/CVPR52729.2023.01085

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.zip