Noisy Annotations in Segmentation

Anonymous Author(s)
Affiliation
Address
email

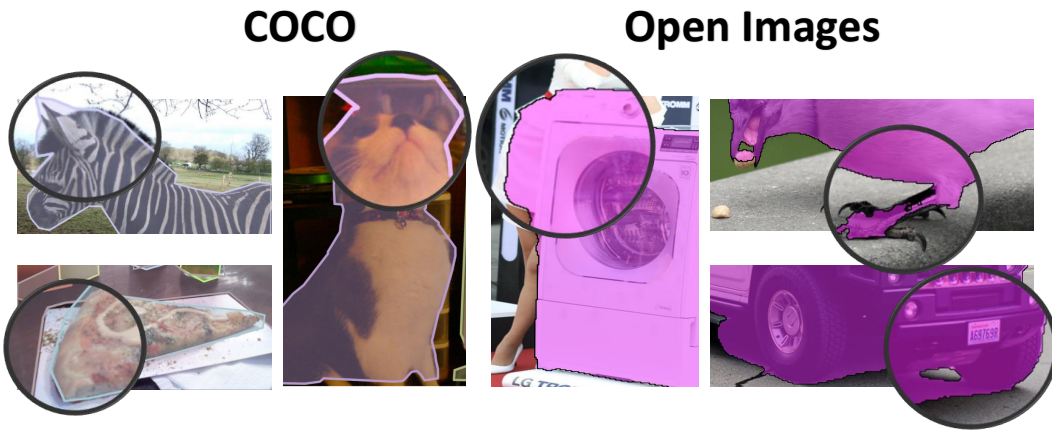


Figure 1: Annotation noise found in both manually labeled data and weakly annotated data. These errors include incomplete or over-extended masks, and ambiguous boundaries.

Abstract

We propose four noise-augmented benchmarks—**COCO-N**, **CityScapes-N**, **VIPER-N** and the weak-annotation track **COCO-WAN**—that provide a unified test-bed for studying annotation noise in instance segmentation. A parametric engine stochastically perturbs mask boundaries, drifts spatial extents, flips categories and omits instances at three severity tiers, producing Monte-Carlo variants of any COCO-style corpus. Evaluating popular segmentation models such as Mask R-CNN, Mask2Former, YOLACT and SAM reveals up to 35 % drops in mask mAP under moderate noise, underscoring the limits of current learning-from-noisy-labels techniques when errors are spatial rather than purely categorical. All proposed **Benchmark-N** suite establishes a reproducible baseline for noise-aware segmentation and motivates future work on robust objectives, data-centric annotation pipelines and noise-adaptive architectures.

1 Introduction

Deep learning–driven instance segmentation underpins safety-critical applications ranging from autonomous driving to medical imaging. Its success hinges on *precise* pixel–level supervision, yet large, rapidly curated datasets inevitably contain erroneous masks. In echocardiography, for example, a modest 5% boundary error around the left-ventricular cavity can swing the ejection-fraction estimate from 45% to 39–50%, potentially tipping a diagnosis from borderline normal to pathological. Such high-stakes scenarios demand segmentation models that remain reliable when labels are imperfect.

Unfortunately, almost all noisy-label benchmarks—focuses on *class* noise for image classification. Spatial distortions, instance omissions and prompt-induced biases that plague instance segmentation

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

are far less explored, and there is no unified test bed for studying them at scale. Without realistic benchmarks, it is unclear how fragile current models are or which learning strategies truly help.

We close this gap with **Benchmark-N**, a suite of four noise-augmented datasets that inject empirically grounded spatial corruptions into both real (**COCO-N**, **CityScapes-N**) and synthetic (**VIPER-N**) data, plus a weak-annotation track (**COCO-WAN**) built with foundation-model prompts. A parametric generator produces controllable *boundary imprecision, spatial drift, category confusion* and *instance omission* at three severity tiers, enabling Monte-Carlo stress tests of any segmentation pipeline. Comprehensive experiments across Mask R-CNN, Mask2Former, YOLACT and SAM reveal sharp performance drops even under mild noise, exposing limitations of current learning-from-noise methods.

Our contributions are:

- A stochastic, task-agnostic noise model that synthesises diverse, realistic annotation errors for instance segmentation.
- Four publicly released benchmarks—**COCO-N**, **CityScapes-N**, **VIPER-N** and **COCO-WAN**—with reproducible "low/mid/high" noise presets.
- An extensive empirical study showing that popular CNN and transformer architectures lose up to $\sim$35% mAP under hard noise, underscoring the need for noise-aware training.

2 Related Work

Noisy-label benchmarks. Classification studies typically flip labels at random or via confusion matrices (CIFAR-N Wei et al. [2022], Clothing1M Xiao et al. [2015]); detection work jitters boxes or drops objects Mao et al. [2021], Ryoo et al. [2023]. In dense prediction, mask opening/closing Lu et al. [2014], Li et al. [2023] and class flips in medical data Nordström et al. [2022] leave object extent mostly intact, missing boundary jaggedness, spatial drift and omissions observed in practice.

Weak or coarse labels. Polygon-level Cityscapes-Coarse and Mapillary Vistas Cordts et al. [2016], click-based OpenImages Kuznetsova et al. [2020], and SAM-generated SA-1B Kirillov et al. [2023] support weak-supervised training but are not designed as robustness tests. Our **COCO-WAN** turns SAM masks—with controlled prompt noise—into such a benchmark.

Learning with noisy labels (LNL). Dense-task LNL adapts classification ideas: Adaptive Early-Learning Correction Liu et al. [2022], spatial Markov refinement Yao et al. [2023], and federated aggregation Wu et al. [2023]. Each uses bespoke or domain-specific corruptions, limiting comparability. Spatial noise thus remains largely un-benchmarked; our datasets provide the first multi-domain, reproducible test bed for boundary-level errors.

3 Annotation-Noise Generator

Accurate segmentation hinges on pixel–level agreement between an image and its ground-truth mask. In practice, annotation pipelines introduce *annotation noise*—any mismatch between the ideal (oracle) mask M^* and the dataset mask M. We first catalogue common error modes, then formalise a stochastic generator that injects them with tunable severity.

3.1 Empirical Taxonomy of Annotation Errors

A manual sweep of COCO Lin et al. [2014], Cityscapes Cordts et al. [2016], OpenImages Kuznetsova et al. [2020] and LVIS Gupta et al. [2019] reveals four recurrent error families (illustrated in Fig. 1):

Boundary Imprecision — coarse or jagged outlines that over- or undershoot the true contour.

Spatial Drift — near-rigid shifts of an entire mask, typically caused by inattentive clicks or snapping heuristics.

Category Confusion — visually similar classes swapped (e.g. *bus*$\nrightarrow$*truck*), reflecting annotator ambiguity or taxonomy overlap.

Instance Omission — thin, occluded or low-contrast objects partially or fully omitted.

Automated polygon simplifiers, box-to-mask converters and prompt-based foundation models can exacerbate these patterns by eroding fine structures or hallucinating plausible yet wrong regions.

3.2 Parametric Noise Model

Let (M, c) denote a binary instance mask and its class label. We inject noise by composing five independent perturbations; each perturbation is sampled *i.i.d.* per instance, so every invocation yields a corrupted dataset.

Approximation: Simplify the polygon via Douglas–Peucker with tolerance $\varepsilon \sim \mathcal{N}^{+}(\mu_{\text{approx}}, \sigma_{\text{approx}})$.

Localization: Displace each vertex by $(\Delta x, \Delta y)$ where $\Delta x, \Delta y \sim \mathcal{N}(\mu_{\text{loc}}, \sigma_{\text{loc}})$ with random signs.

Scale: With equal probability, dilate or erode M by a square kernel of size $K \sim \max\{1, \lfloor \mathcal{N}(\mu_{\text{scale}}, \sigma_{\text{scale}}) \rfloor\}$. *Class Confusion:* With probability p_{cls}, replace c by a sibling inside the same super-category, following empirical confusion matrices Northcutt et al. [2021].

Deletion: With probability p_{del}, drop the instance altogether, mimicking missed objects.

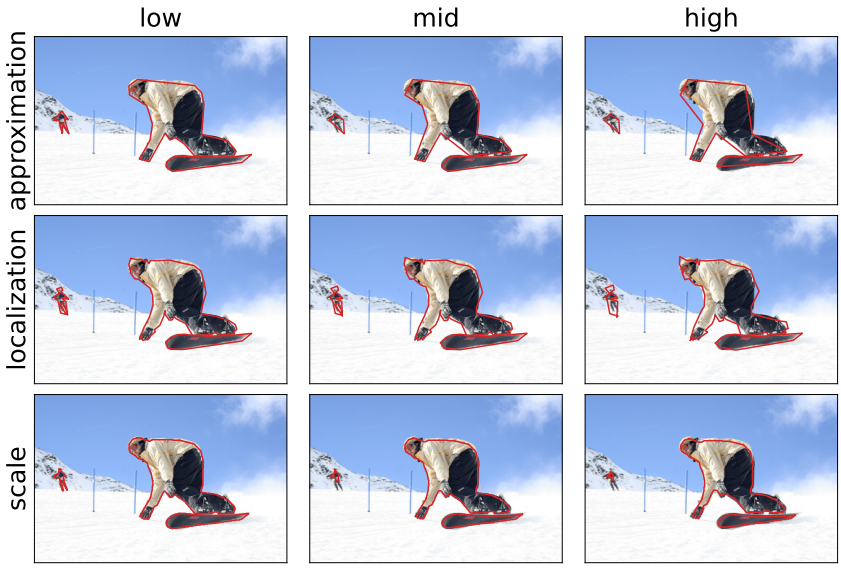


Figure 2: Illustrating the effects of the spatial noise with varying intensity.

3.3 Severity Presets and Reproducibility

Our open-source tool Benchmark-N suite[1] applies the above process to any COCO-style dataset. Three presets—*Low*, *Mid*, *High*—scale $(\mu, \sigma, p_{\text{cls}}, p_{\text{del}})$ as detailed in Table 1. Because the generator is purely stochastic, one can draw multiple corrupted variants, enabling Monte-Carlo robustness studies instead of a single "clean vs. noisy" split.

This formulation cleanly decouples the *empirically grounded taxonomy* (Sec. 3.1) from the *synthetic noise engine* (Sec. 3.2), providing a rigorous basis for analysing segmentation robustness under realistic annotation imperfections.

All variables are sampled *i.i.d.* across instances, yielding a truly stochastic benchmark—unlike previous works that commit to a single "clean vs. noisy" split Nordström et al. [2022], Lad and Mueller [2023], Yao et al. [2023], Liu et al. [2022]. Three presets (*low/mid/high*) correspond to increasing (μ, σ) pairs (Table 1, Appx.). Our public tool Benchmark-N applies these transformations with a single command, enabling reproducible stress-tests of segmentation pipelines.

[1] https://anonymous.4open.science/r/noisy_labels-0C70/README.md

Intensity	Low	Medium	High
$(\mu_{\text{approx}}, \sigma_{\text{approx}})$	$(5, 2.5)$	$(10, 2.5)$	$(15, 10)$
$(\mu_{\text{local}}, \sigma_{\text{local}})$	$(2, 0.5)$	$(3, 0.5)$	$(4, 2)$
$(\mu_{\text{scale}}, \sigma_{\text{scale}})$	$(3, 1)$	$(5, 1)$	$(7, 4)$
p_{class}	0.05	0.05	0.05
p_{delete}	0.05	0.05	0.05

Table 1: Noise parameters used to produce the noisy annotations that compose **Benchmark-N**.

4 Benchmark

4.1 Synthetic Dataset: VIPER

In order to validate our noise model under perfectly labeled conditions, we turn to the VIPER dataset Richter et al. [2017], which is derived from the GTA V game engine. VIPER provides high-fidelity, pixel-accurate annotations for every object and region in the scene, making it a "clean" baseline for testing the pure effect of annotation noise.

Figure 3: Examples from **VIPER-N** benchmark. Top row shows the clean annotations, second row the low noise regime, third present the midum annotation noise and last row the high annotation noise.

Because VIPER's segmentation maps are automatically rendered in a synthetic environment, the ground-truth annotations exhibit none of the spatial inaccuracies common in human-labeled datasets. This allows us to inject our prescribed noise types in a fully controlled way, without mixing in any preexisting labeling errors.

Experimental Results We train and evaluate the popular Mask R-CNN on VIPER-N and compare to the noise-free VIPER baseline. Figure 3 illustrates qualitative examples of clean vs. noisy labels, and Table 2 quantifies performance drops by model and noise level. Notably, even low-level spatial distortions can reduce precision significantly, confirming the sensitivity of modern architectures to subtle label corruptions.

Table 2: mAP on VIPER-N at four noise severities (higher is better)

Noise	All	S	M	L
Clean	15.8	6.0	44.3	60.6
Low	13.8	4.7	38.0	57.3
Mid	12.3	4.0	31.4	55.2
High	10.7	2.6	29.0	53.6

VIPER-N thus provides a controlled, synthetic test bed that highlights each model's vulnerabilities to annotation noise when all else—lighting, context, labeling scale—is held constant.

4.2 COCO-N & CityScapes-N

Finally, the noise integrated with the same noise strategies into widely used real-world datasets, producing **COCO-N** and **CityScapes-N**. Unlike VIPER, these datasets already contain minor human labeling errors, meaning our injected noise adds a further layer of realism. Below are the key steps and summary results.

We apply the exact same noise operations (§3.2) to each instance in COCO Lin et al. [2014] and Cityscapes Cordts et al. [2016] train splits. In line with VIPER-N, we create three tiers of severity (low, mid, high) by increasing the morphological kernel size, polygon simplification tolerance, and class confusion probabilities. Figure 7a illustrate the performance degradation on those, as well as LVIS Gupta et al. [2019] dataset, more details in the supp. materials.

Results Across Popular Models. Table 3 shows how varius models Mask R-CNN (R-50/R-101), Mask2Former (R-50/Swin), YOLACT fare on both **COCO-N** and **CityScapes-N** for all three noise tiers, as well as HTC Chen et al. [2019b] and SOLO Wang et al. [2020] for **COCO-N** Across the board, we see a notable dip in both standard mAP as well as boundary-focused metrics Cheng et al. [2021a] in supp materials. For **COCO-WAN**, we report fewer architectures, as full report will be released upon acceptance. Interestingly, transformer-based architectures (e.g., Swin in Mask2Former) appear slightly more robust to misaligned boundaries, but no model is immune to severe disruptions.

To assess the effect of label noise, we evaluate the performance of various instance segmentation models using our newly developed benchmark. We apply the various levels of noise, presenting **COCO-N** and **CityScapes-N**, providing insights into their robustness and adaptability. For more details about the models and datasets refer to the implantation details in the supplementary materials. Table 3 present the findings Mask-RCNN (M-RCNN) He et al. [2017], YOLACT Bolya et al. [2019], SOLO Wang et al. [2020], HTC Chen et al. [2019a] and Mask2Former (M2F) Cheng et al. [2021b]. **Clean** denote the performance of a model on the original annotations, where **Easy**, **Mid** and **Hard** correspond to the definition in Table 1. The reported numbers in the table represent mask mean average precision (AP) and boundary mask mean average precision (AP^{b}), respectively. More experiments involving LVIS dataset Gupta et al. [2019] and learning with noisy labels in sup. materials. All models trained and evaluated by standard training procedure[2]. We obtained additional experiments include cardiac unltrasound data in Appendix A, more evaluation metrics, models and datasets in Appendix E, and most notably, evaluate both **zero-shot** and **fine-tune** SAM Kirillov et al. [2023] on our proposed benchmark in Appendix F.

Our experiments demonstrate that label corruption leads to a degradation in model performance. Specifically, Mask R-CNN with a ResNet50 backbone retains approximately 80.6%, 71.7%, and 64.4% of its performance under Easy, Medium, and Hard noise conditions, respectively, on the **COCO-N** benchmark. The same model exhibits a more dramatic performance drop on the **CityScapes-N** benchmark, managing to retain only 72.8%, 60.9%, and only 45% under the corresponding noise levels. This trend is consistent across all tested models, suggesting that the impact is more crucial when less data is available, but might be easier to mitigate when using more data, even with the same portion of label noise.

This study demonstrates that all models are affected by labeling bias and exhibit diminished performance to varying extents, highlighting differing sensitivities to label noise. Notably, transformers

[2] openmmlab/mmdetection/model_zoo

Table 3: Evaluation Results of Instance Segmentation Models under Different Benchmarks, reporting mAP.

Dataset	Model	Backbone	Clean	Easy	Mid	Hard
COCO-N	M-RCNN	R-50	34.6	27.9	24.8	22.3
	YOLACT		28.5	26.4	23.3	20.8
	SOLO		35.9	25.2	17.1	12.4
	HTC		34.1	-	28.4	25.5
	M2F		42.9	33.5	30.1	26.7
	M-RCNN	R-101	36.2	28.8	31.8	23.7
	M2F	Swin-S	46.1	39.6	37.9	33.6
CityScapes-N	M-RCNN	R-50	36.1	26.4	22.0	16.3
	YOLACT		19.3	19.1	17.1	13.6
	M-RCNN	R-101	37.0	33.7	30.7	27.0
COCO-WAN	M-RCNN	R-50	34.6	32.8	24.4	21.6
	Cascade M-RCNN		35.9	26.8	25.7	24.2
	M2F		42.9	39.2	31.9	26.2
	M2F	Swin-s	46.1	42.9	34.4	28.4

display greater resilience, retaining 73% on the Hard benchmark, effectively mitigating the adverse effects of noisy labels compared to the convolution counterpart. This observation underscores the potential of using transformer-based architectures in scenarios where robustness to label noise is crucial. Our findings offer preliminary guidance for selecting or designing robust instance segmentation models in practical applications where encountering label noise is inevitable.

Implications. Given their critical role as mainstream benchmarks, **COCO-N** and **CityScapes-N** offer a practical measure of model reliability under imperfect labels. This can guide future research in developing noise-aware training strategies, data-cleaning pipelines, or architectures that gracefully handle label distortion. Our publicly released tool ensures that anyone can replicate these noisy benchmarks, tune the noise parameters, or adapt them to new datasets.

4.3 COCO-WAN (Weakly ANnotated)

Modern annotation pipelines commonly employ Vision Foundation Models (VFMs) Zhang et al. [2025] to reduce the dependence on fully manual labeling. While VFMs trained on large-scale data can produce high-quality masks, they often introduce systematic biases, since they overlook fine details. Due to the extend of tasks this models solves, for a specific context, they require some prompt that provides a task-specific context, as illustrated in fig. 4a. Specifically, we examine Segment Anything Model (SAM) Kirillov et al. [2023], prompting the model with either bounding-box, points, partial masks or text queries, incorporating noises based on the model and queries biases.

(a) prompt-based VFM. Points, boxes, and text guide the mask decoder.

(b) Point Prompt

(c) Box Prompt

Figure 4: prompt-based VFM (left) and example SAM masks using different prompts (middle and right).

We have put into test three kinds of weak annotations as prompts, **Points**- one point per instance in the middle of the object mask. **Boxes**- the bounding box from the annotations, and **Text**- we

Table 4: COCO-WAN prompt quality (mAP, b-mAP Cheng et al. [2021a]) into Grounded-SAM.

Prompt	Type	Clean		Noisy	
		AP	AP^B	AP	AP^B
Original labels		34.6	20.6	-	-
Point	$\sim \mathcal{U}$	24.4	15.7	21.6	13.7
Box	$+\mathcal{N}(0, 2)$	32.8	19.7	25.3	15.9
Text	{cls}	22.0	14.1	-	-

Table 5: Mask2Former fine-tune vs. Gounded-SAM Ren et al. [2024] text prompted noise. Reporting AP_m and AP_b as mask and box mAP respectively.

Model	Clean		Noisy	
	AP_m	AP_b	AP_m	AP_b
R-50	42.9	45.7	26.2	24.1
R-101	43.4	46.1	26.7	25.1
Swin-S	46.1	49.3	28.4	31.4
Swin-B	48.2	51.5	30.3	33.2

fed the class label from the annotations into Grounded-DINO, and used the boxes output as a box query, similar to Grounded-SAM Ren et al. [2024]. We examine the effect of noise on the prompts in Table 4, incorporate noise into the points, by randomly sample one point from the mask, and to the boxes by adding Gaussian noise ($\mathcal{N}(0, 2)$) into one of the box corners.

In Table 5, we examine how a transformer backbone (Swin-S Liu et al. [2021]) impacts the Mask2Former Cheng et al. [2021b] model's robustness to noise. This noise degrades the models (on both mask and bounding box) by approximately 48% in R-50 and 37.3% in Swin-S. Although still notably affected by noise, this trend aligns with the results on **COCO-N** and **CityScapes-N** as reflected from Table 3.

Figure 4 illustrates how different prompt types can lead to varying degrees of segmentation noise, as for the given example bounding box captures the background instead of the actuall object, while a point is sufficient to produce high quality mask.

Qualitatively, SAM generally captures coarse object boundaries well, but Figure 5 shows how color and texture biases may cause missing or conflated parts, particularly in challenging scenes (e.g. without noticeable approximation errors). For instance, certain darker regions or closely colored objects can be merged or overlooked, signaling a lack of task oriented context. As a practical example, the middle image pair shows the pants and face of the standing person are not included in the mask due to the stark contrast in color from the light shirt. On the right image, we observe annotations with shape (stove-top) and instances of conflating potential objects (stove and cabinet) due to color biases. More qualitative results show in the supp. materials. In Figure 6 we see yet another example for auto-annotations excel in masks fidelity and even finding missing annotations, such as the portrait in the top-left pair, however, it commonly struggle with crowded annotations, as demonstrated in the bottom image, where the text was crowd orange and the mask include mostly the basket. this reflect the need to explore open vocabulary VFM that may overcome this annotation obstacle.

Figure 5: **Annotation quality** comparing COCO labels (left) and COCO-WAN labels using box queries (right)

This emphasizes the importance of developing more robust annotation strategies—both in prompt design and in subsequent label refinement—when relying on VFMs for real-world segmentation tasks.

4.4 Qualitatively Analysis

To evaluate how each noise independently affects model performance, we conduct an ablation using Mask R-CNN He et al. [2017] (ResNet-50 backbone) trained on the whole COCO with only one noise type active at various severity levels. Table 7 summarizes the quantitative impact on standard

Figure 6: Compared annotations between COCO (left) and text prompt weak annotations (right).

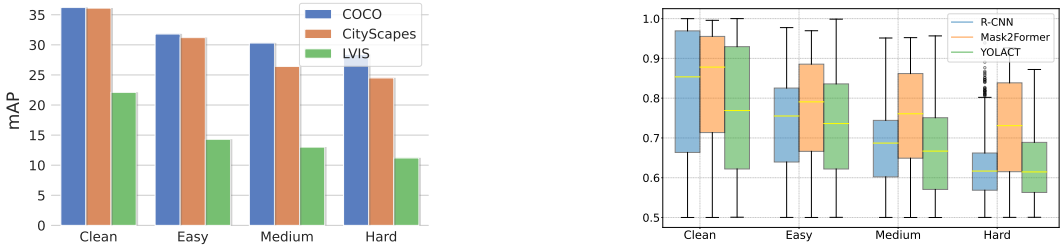
metrics like mAP and boundary-level mAP (B-mAP) Cheng et al. [2021a]. Figure 11 visualizes performance declines for increasing noise severity.

Scale Noise (especially erosions) severely affects boundary fidelity, leading to the largest drop in performance, yet easy to fix by a pre-process morphological counter operation that bring the masks close to clean (e.g., opening or closing), thus, we chose to scale at random.

Localization and *Approximation Noise* subtly degrade object outlines, though moderate levels of displacement do not drastically lower global mAP.

Class Confusion chiefly impacts recognition accuracy; the reduced classification confidence leads to a measurable mAP drop, but less so on boundary metrics.

Deletion yields fewer total annotations, skews training and causes a performance loss.



(a) Mask-RCNN performance on COCO, CityScapes and LVIS across three noise levels.

(b) Confidence scores (threshold > 0.5) of Mask-RCNN, Mask2Former and YOLACT under increasing noise.

Figure 7: Effect of annotation noise on segmentation quality (left) and prediction confidence (right).

Our experiments indicate that various architectures and backbones exhibit notable sensitivity to label noise, affecting both mask quality and prediction confidence. As shown in Figure 7b, higher noise levels correlate with reduced confidence scores, underscoring the vulnerability of model predictions to annotation accuracy. This effect is further illustrated in Figure 10, where increased noise leads to misclassification, causing the model to generate multiple conflicting predictions for a single instance.

5 Limitations

One limitation is that **Benchmark-N** suit targets four dominant error families (boundary imprecision, spatial drift, category confusion, instance omission). It does not yet cover multi-instance merge/split mistakes, or temporal label noise in videos. Future iterations should extend the taxonomy and validate it with larger human studies.

COCO-WAN perturbs point and box prompts with zero-mean Gaussian noise. Other real-world biases—e.g. inconsistent text queries across annotators—are not modeled, and could alter the observed failure modes of SAM or Grounded-SAM.

This work *measures* robustness; it does not propose a noise-aware training algorithm. Consequently, conclusions about "limitations of current LNL techniques" are empirical, not prescriptive.

Finally, because the benchmark re-uses publicly available images, we do not study privacy leakage or disparate performance across demographic groups.

6 Discussion

Our experiments demonstrate that label noise—whether from imprecise human annotations, automated tools, or weak prompts—can substantially degrade the performance of instance segmentation models. We introduced both synthetic and weakly annotated benchmarks that systematically capture real-world noise patterns, ranging from boundary misalignments to class confusion and missing instances. Even moderate levels of noise can erode confidence in model predictions and lead to notable mAP reductions, highlighting the sensitivity of current architectures to spatial inaccuracies.

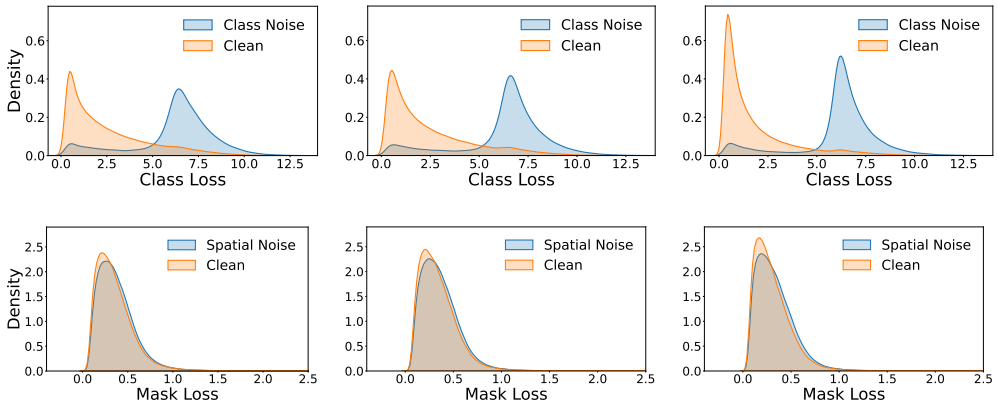
In particular, our results show that (1) models trained on large datasets like COCO and Cityscapes are far from robust under moderate noise, exhibiting over 10% drops in mask mAP, (2) scale noise severely mislead boundary-based metrics, and (3) while prompt-based foundation models reduce labeling effort, they also introduce new biases, and themselves are not fully immune to noisy prompts. These outcomes underscore the gap between current label-noise handling strategies—mostly devised for image classification—and the complexities of segmentation tasks, where spatial quality is paramount.

6.1 Confidence and Loss Analysis

Our study reveals that various architectures and backbones exhibit sensitivity to noise, impacting not only mask quality but also confidence in instance identification. As illustrated in Figure 7b, increased label noise correlates with diminished confidence in model predictions, underscoring the vulnerability of different model architectures to labeling accuracy.

This reduction in confidence is further evidenced in Figure 10, where increased label noise results in poorer mask quality and reduced confidence in the classification head.

We examine the model's ability to distinguish noisy from clean annotations. Figure 8 shows two experiments: in the first, 40% of instances contain class noise; in the second, 40% have medium-level spatial noise. Under class noise, the model's classification losses form two roughly distinct Gaussian distributions, suggesting partial separation of clean and noisy samples. By contrast, when spatial noise is introduced, the losses remain intermixed throughout training. This highlights the challenge of boundary-level label errors for methods relying on loss-based separation. Further experimental details and additional results on learning with noisy labels appear in the supplementary.



(a) Epoch 2 (b) Epoch 6 (c) Epoch 10

Figure 8: Class and Mask Loss Distribution of Mask-RCNN (R50) trained on COCO easy benchmark at different epochs during training.

References

Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation, 2019. 5, 17

Kai Chen, Wanli Ouyang, Chen Change Loy, Dahua Lin, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, and Jianping Shi. Hybrid task cascade for instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019a. doi: 10.1109/cvpr.2019.00511. URL http://dx.doi.org/10.1109/CVPR.2019.00511. 5

Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation, 2019b. 5

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019c. 18

Bowen Cheng, Ross Girshick, Piotr Dollar, Alexander C. Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021a. doi: 10.1109/cvpr46437.2021.01508. URL http://dx.doi.org/10.1109/CVPR46437.2021.01508. 5, 7, 8, 14

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv*, 2021b. 5, 7, 17

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding, 2016. 2, 5

Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. 2, 5

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 18

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2017. 5, 7, 17

Moshe Kimhi, Shai Kimhi, Evgenii Zheltonozhskii, Or Litany, and Chaim Baskin. Semi-supervised semantic segmentation via marginal contextual information. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=i5yKW1pmjW. 18

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2, 5, 6, 19

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 2

Vedang Lad and Jonas Mueller. Estimating label quality and errors in semantic segmentation data via any model, 2023. 3

Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Østvik, Frédéric Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D'hooge, Lasse Løvstakken, and Olivier Bernard. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Transactions on Medical Imaging*, 38:2198–2210, 2019. 13

Peixia Li, Pulak Purkait, Ajanthan Thalaiyasingam, Majid Abdolshah, Ravi Garg, Hisham Husain, Chenchen Xu, Stephen Gould, Wanli Ouyang, and Anton van den Hengel. Semi-supervised semantic segmentation under label noise via diverse learning groups. In *ICCV 2023*, 2023. URL https://www.amazon.science/publications/semi-supervised-semantic-segmentation-under-label-noise-via-diverse-learning-groups. 2

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. 2, 5

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2016. 18

Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2596–2606. IEEE, June 2022. doi: 10.1109/cvpr52688.2022.00263. URL http://dx.doi.org/10.1109/CVPR52688.2022.00263. 2, 3

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 7, 18

Zhiwu Lu, Zhenyong Fu, Tao Xiang, Peng Han, Liwei Wang, and Xin Gao. Learning from weak and noisy labels for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:486–500, 2014. 2

Jiafeng Mao, Qing Yu, Yoko Yamakata, and Kiyoharu Aizawa. Noisy annotation refinement for object detection, 2021. 2

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf. 15

Marcus Nordström, Henrik Hult, Jonas Söderberg, and Fredrik Löfman. On image segmentation with noisy labels: Characterization and volume properties of the optimal solutions to accuracy and dice, 2022. 2, 3

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, April 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12125. URL http://dx.doi.org/10.1613/jair.1.12125. 3

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 7

Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, page 2232–2241. IEEE, October 2017. doi: 10.1109/iccv.2017.243. URL http://dx.doi.org/10.1109/ICCV.2017.243. 4, 18

Kwangrok Ryoo, Yeonsik Jo, Seungjun Lee, Mira Kim, Ahra Jo, Seung Hwan Kim, Seungryong Kim, and Soonyoung Lee. Universal noise annotation: Unveiling the impact of noisy annotation on object detection, 2023. 2

Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. *SOLO: Segmenting Objects by Locations*, page 649–665. Springer International Publishing, 2020. ISBN 9783030585235. doi: 10.1007/978-3-030-58523-5_38. URL http://dx.doi.org/10.1007/978-3-030-58523-5_38. 5

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2019. doi: 10.1109/iccv.2019.00041. URL http://dx.doi.org/10.1109/ICCV.2019.00041. 18

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm. 2

Nannan Wu, Zhaobin Sun, Zengqiang Yan, and Li Yu. Feda3i: Annotation quality-aware aggregation for federated medical image segmentation against heterogeneous annotation noise, 2023. 2

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 2, 15

Jiachen Yao, Yikai Zhang, Songzhu Zheng, Mayank Goswami, Prateek Prasanna, and Chao Chen. Learning to segment from noisy annotations: A spatial correction approach, 2023. 2, 3

Yixin Zhang, Shen Zhao, Hanxue Gu, and Maciej A Mazurowski. How to efficiently annotate images for best-performing deep learning-based segmentation models: An empirical study with weak and noisy annotations and segment anything model. *Journal of Imaging Informatics in Medicine*, pages 1–13, 2025. 6

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2020. 18

A Ejection Fraction Analysis in the CAMUS Dataset

The CAMUS dataset Leclerc et al. [2019] provides 2D echocardiographic images along with high-quality, expert-annotated labels of the left ventricle (LV). A critical clinical metric in these annotations is the left ventricle's *ejection fraction* (EF), defined as:

$$\mathrm{EF} = \frac{\mathrm{EDV} - \mathrm{ESV}}{\mathrm{EDV}} \times 100\%, \tag{1}$$

where EDV is the end-diastolic volume (i.e., the LV volume at its most dilated state) and ESV is the end-systolic volume (the LV volume at maximal contraction). EF offers a succinct quantification of cardiac pump efficiency: a healthy range is typically considered to be above 50%, while borderline or reduced EF can indicate impaired cardiac function.

Clinical Implications and Risks. Misestimations of the LV boundary—especially at the end-diastolic or end-systolic frames—can propagate into disproportionate errors in volume computations. Even small annotation noise around the boundary may shift the EF from borderline-normal (e.g., 45%) to a clearly abnormal ($\approx 39\%$) or misleadingly normal ($\approx 50\%$) reading. Such inaccuracies pose a risk for misdiagnosis or delayed therapeutic intervention, since EF underlies critical clinical decisions, including the prescription of certain medications, lifestyle interventions, or further diagnostic procedures.

Noise-Induced Errors. Figure 9 (to be added) illustrates how a noisy annotation around the LV boundary at end-diastole can lead to an overestimation or underestimation of EDV. When combined with an equally skewed ESV, the net EF deviation can be clinically significant. We examine morphological dilation of the ESV boundary, along with moderate localization noise in both EDV and ESV, using the "low" noise setup described in the main text.

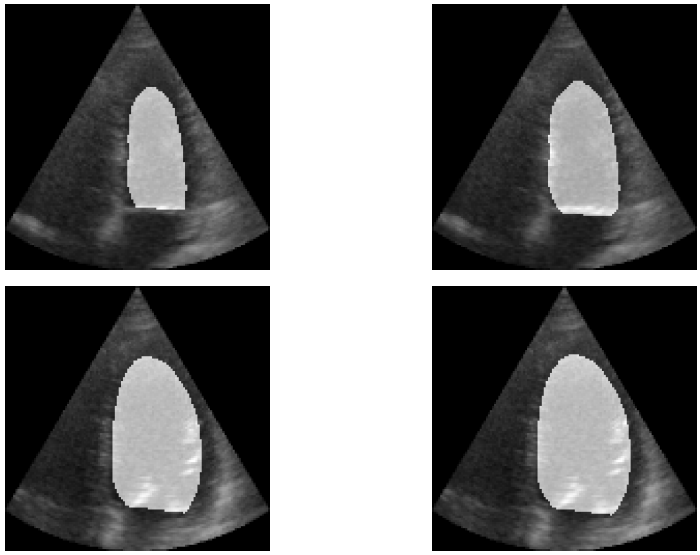


Figure 9: Example of ESV (top) and EDV (bottom) from the CAMUS dataset (left) and their noisy counterparts (right). Even modest boundary distortions can shift EF calculations significantly.

Evaluation Under Noisy Labels. We trained a simple convolution-based U-Net model, as described in Leclerc et al. [2019], on both **clean** and **noisy** CAMUS annotations, and compared the results in Table 6. Evaluation metrics are **Dice Index** for segmentation overlap of the left ventricle (LV) at end-systolic (ES) and end-diastolic (ED) frames,**EF Error** as mean absolute error compared to 2D compute of EF values from the labels in percentage points (p.p), as well as **HD** (Hausdorff Distance) for boundary alignment.

As Table 6 indicates, the model trained on noisy labels tends to yield worse Dice overlap and a higher EF error than when trained on clean labels, underscoring the sensitivity of medical diagnostics to

Table 6: Comparing UNET results on clean vs noisy CAMUS data.

Training	**Dice (%)**		**EF Error**	**HD (mm)**
Data	ES	ED	(p.p.)	(ED frame)
Clean	86.9	91.1	2.1	6.3
Noisy	82.1	87.5	4.5	11.25

Table 7: Ablate the performance evaluation of Mask R-CNN with Spatial label noise across all data on **COCO-N**.

Severity	Low		Medium		High	
Metric	mAP	B-mAP	mAP	B-mAP	mAP	B-mAP
Clean	34.6	20.6	34.6	20.6	34.6	20.6
Dilation	32.8	18.5	29.1	14.2	26.4	10.3
Erosion	29	15.7	22	9.5	17.4	5.3
Opening	34.6	20.7	34.7	20.7	34.6	20.6
Random Scale	34.1	20.4	32.4	18.5	30.8	17.1
Shifting	28.2	15.4	26.6	14.0	21.1	8.6
Localization	34.4	20.4	34.2	20.1	33.5	19.4
Approximation	34.7	20.8	32.5	18.8	30	16.3

annotation precision. Crucially, this discrepancy demonstrates that even modest boundary errors can propagate into clinically important EF ranges, highlighting the urgency of robust noise-handling strategies in echocardiographic segmentation tasks.

B Additional Experiments

Figure 11 compare the mAP and boundary mAP of original vs. noisy annotations. The top row illustrates the morphological operations used for scale-based spatial distortion, while the bottom row shows the specific noise types we apply in our benchmark.

Table 8: Evaluation results of instance segmentation models (Boundary mAPCheng et al. [2021a]) under various noise levels.

Dataset	**Model**	Clean	Easy	Medium	Hard
COCO-N	M-RCNN (R50)	20.6	18.9	17.5	16.3
	M-RCNN (R101)	22.2	20.4	19.0	17.4
	M2F (R50)	30.0	28.6	26.7	23.8
	M2F (Swin-S)	32.6	30.9	29.3	26.2
	YOLACT (R50)	15.7	14.4	13.5	12.4
Cityscapes-N	M-RCNN (R50)	33.4	28.4	24.7	22.8
	M-RCNN (R101)	34.3	30.7	29.0	25.4
	YOLACT (R50)	16.5	16.5	14.5	13.3

To further validate our noise design choices and their impact, we obtained additional experiments. As presented in Table 9, we evaluated the traditional symmetric and asymmetric class noise on instance segmentation using MASK-RCNN with two different backbones to assess the resulting performance

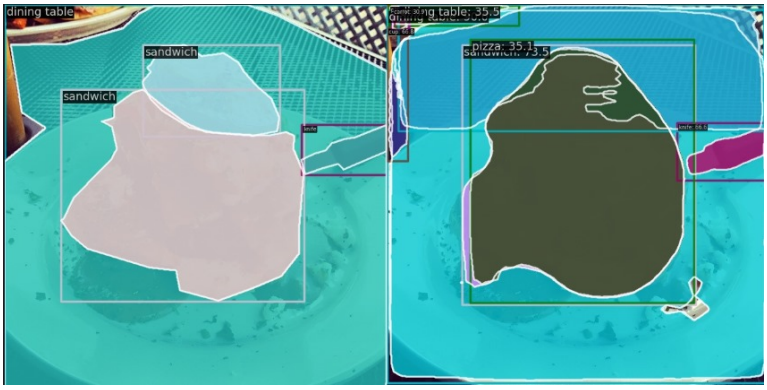
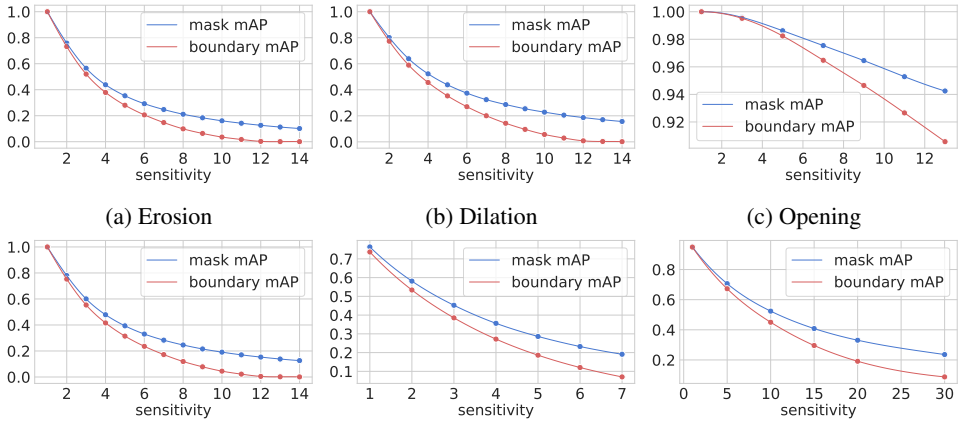


Figure 10: Visual results of Mask-RCNN using the **COCO-N** easy benchmark. Since the model is uncertain it observe different objects (pizza and sandwich in the bottom image) fooling the NMS operation.



(a) Erosion (b) Dilation (c) Opening

(d) Scale (e) Localization (f) Approximation

Figure 11: The mAP and boundary-mAP metrics between real annotations from COCO dataset and their **COCO-N** annotations counterpart.

degradation. "Sym $p\%$" refers to symmetric class confusion with probability p, while "Asym $p\%$" denotes mislabeling concentrated in a smaller set of classes Natarajan et al. [2013], Xiao et al. [2015].

Table 9: Class noise ablation reporting mAP^{box} and mAP^{mask}

Models/Labels	**Clean**	**Sym 20%**	**Sym 50%**	**Sym 80%**	**Asym 40%**
M-RCNN (R50)	38/34.6	35.5/31.9	32.2/29.2	22.5/20.2	34.6/31.4
M-RCNN (R101)	40.1/36.2	37.5/33.6	34.5/31	25.2/22.7	36.8/33.2

Next, we examined the affects of label noise and the additional impact of spatial noise on mask quality, as shown in Table 10. We assessed the quality of all masks through the foreground-background segmentation task of a trained model. The results indicate that the mask quality deteriorates more significantly when spatial noise is incorporated along with traditional class noise.

In addition to evaluating the benchmark itself, we extended our analysis to include the impact on object detection performance. Specifically, we examined the Boundary $- mAP$ and mAP^{box} scores, as presented in Tables 8 and11 respectively. This tables highlights the detrimental effects of spatial label noise on the boundaries of the masks, as well as bounding box quality, in addition to the previously discussed impacts on mask quality. By analyzing the mAP^{box}, we aim to demonstrate the broader implications of our noise design choices, showing that spatial noise not only affects segmentation masks but also significantly degrades the performance of object detection tasks. This

Table 10: Foreground-background segmentation results under class and spatial noise. The symbol "+" indicates an added spatial corruption using M-RCNN(R50).

Foregound noise	**bbox**	**segm**	**boundry**
clean	42	35.8	22.4
20 %	40.7	34.9	21.7
20 % + Easy	40.4	34.2	21.2
30% + Medium	39.6	32.7	19.9
40% + Hard	38.7	30.6	18.3
50 %	38.7	32.7	20.7

comprehensive evaluation underscores the robustness of our benchmark in assessing the performance degradation across different aspects of instance segmentation and object detection.

Table 11: Evaluation Results of Instance Segmentation Models under Different Benchmarks reporting AP^{box}.

Dataset	**Model**	Clean	Easy	Medium	Hard
COCO-N	M-RCNN (R50)	38	35.4	34.3	33.4
	M-RCNN (R101)	40.1	37.4	36.5	35.2
	M2F (R50)	45.7	42.2	43.7	44.7
	M2F (Swin-S)	49.3	47.9	47.1	45.7
	YOLACT (R50)	30.8	29.2	28.2	27.7
Cityscapes-N	M-RCNN (R50)	41.5	35.7	32.8	31.2
	M-RCNN (R101)	39.8	32.8	29.6	26.8
COCO-WAN	M-RCNN (R50)	36.3	34.1	25.5	22.4

Finally, we present results on the long-tailed segmentation dataset LVIS, as shown in Table 12. The findings reveal a significant impact, with a 50% reduction in boundary IoU under the hard benchmark conditions. This provides evidence of an exacerbated effect in long-tailed scenarios, highlighting the increased challenges posed by our noise design in datasets with imbalanced class distributions.

Table 12: Performance on LVIS-N (Mask R-CNN R50-FPN). We report mAP / Boundary mAP under various noise levels.

Dataset	Clean		Easy		Medium		Hard	
	AP	AP^B	AP	AP^B	AP	AP^B	AP	AP^B
LVIS-N	22.8	22.1	15.5	14.3	17.7	13	13.3	11.2

C Additional Noise Visualizations

Figure 12 presents additional samples from our benchmark under different intensities of spatial label noise. Each row highlights a specific set of distortions—such as boundary approximations or morphological operations—applied to one or more instances. As the noise severity increases from left to right, the object contours become visibly degraded, illustrating the range of realistic annotation errors our benchmark can simulate.

Figure 15 provides a qualitative comparison of Mask R-CNN and Mask2Former under varying noise levels. The top row shows both models' predictions on a clean COCO image: each accurately delineates the car and surrounding objects with sharp, well-aligned masks. In the middle row, Mask R-CNN is challenged by the Easy, Medium, and Hard variants of COCO-N—its masks become

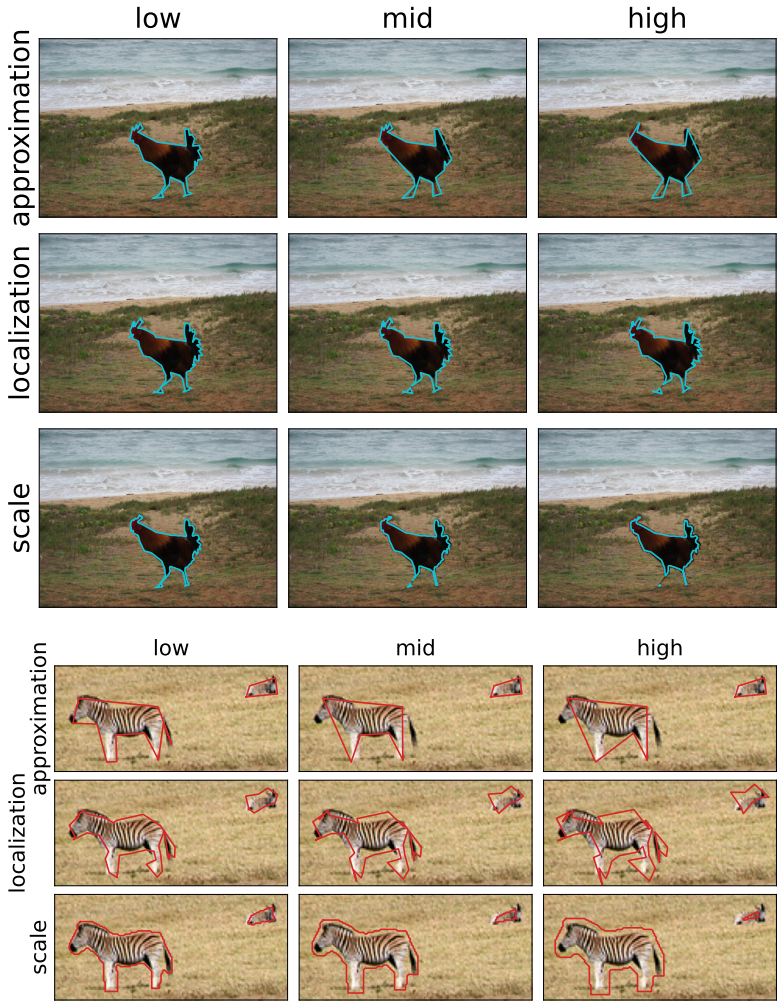


Figure 12: Additional illustrating the effects of the spatial noises on one or two instances with various scales, similar to fig. 2

progressively fragmented, with missing segments and increasingly jagged boundaries. The bottom row reveals that Mask2Former, while initially more robust, also suffers under stronger noise: its Easy predictions remain close to the clean baseline, but Medium and Hard noise lead to boundary bleed and partial omissions. Overall, this figure illustrates that spatial annotation errors systematically degrade mask quality in both CNN- and transformer-based architectures, with severity correlating with noise intensity.

D Implementation Details

This section elaborates on the architectures, datasets, noise definitions, and the levels of asymmetric noise used in our experiments. We also detail the noise intensity applied in the benchmark, along with the hardware configurations and convergence times.

D.1 Architectures

We explore the effects of label noise on various instance segmentation models, encompassing multi-stage (Mask R-CNN He et al. [2017]), single-stage (YOLACT Bolya et al. [2019]), and query-based (Mask2Former Cheng et al. [2021b]) architectures. To achieve a comprehensive analysis, we experimented with different feature extractors, we used convolutional backbones such as ResNet-

50 He et al. [2015] for all models and ResNet-101 for Mask R-CNN, alongside a transformer-based backbone (Swin-B Liu et al. [2021]) for Mask2Former. For the integration of multi-scale features, Feature Pyramid Networks (FPN) Lin et al. [2016] were employed across all models except Mask2Former, which utilizes Multi-Scale Deformable Attention (MSDeformAttn) Zhu et al. [2020], as multi-scale feature representation. All models and configurations implementations from MMDetection Chen et al. [2019c].

D.2 Datasets

COCO dataset for training and evaluating algorithms that segment individual objects within a scene. It contains about 330,000 images, annotated with over 1.5 million instances masked from 80 categories that are also part of 12 super-categories.

Cityscapes dataset is designed for training and evaluating algorithms in urban scene understanding, particularly for segmentation tasks. It comprises a collection of images captured in 50 different cities, featuring 5,000 annotated images with 19 classes for evaluation, covering a range of urban object categories such as vehicles, pedestrians, and buildings.

VIPER VIPER Richter et al. [2017] is a synthetic dataset generated from the GTA V game engine. It provides per-pixel annotations for a broad range of 31 categories in photorealistic urban scenes, making it ideal for benchmarking under controlled conditions. Because VIPER annotations are automatically rendered (rather than hand-labeled), they are virtually free from human annotation errors, allowing precise evaluation of how injected label noise affects segmentation performance.

LVIS dataset is based on COCO images and curated to provide a comprehensive benchmark for instance segmentation, emphasizing rare object categories. It contains over 2 million high-quality instance annotations across 1,203 categories, making it one of the largest and most diverse datasets for instance segmentation. The LVIS dataset is particularly noted for its long-tail distribution of object categories, which poses significant challenges for segmentation algorithms and help us to asses the abilities of segmentation algorithms to deal with label noise in this scenario.

.

D.3 Hardware details

MS-COCO based experiments (include both COCO and LVIS) and VIPER conducted on local machine with 4 Nvidia RTX A6000 or 4 Nvidia RTX 3090, ranging from 20 hours (Mask-RCNN with R50) to 7 days (Mask2Former with SWIN transformer beckbone), training for 12 epochs for all models except YOLACT that trained for 50 epochs. Cityscapes experiments conducted on local machine with one instance of Nvidia RTX 3090, training for 12 epocs for about 12 hours. All experiments use the default configs from MMDetection Chen et al. [2019c].

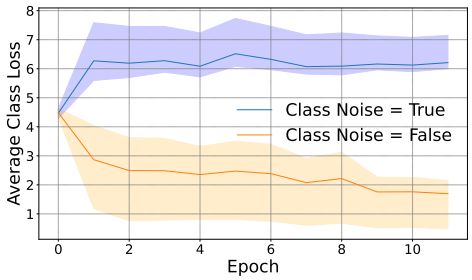
E Learning with Noisy Labels

As described in the paper, class noise is separable, allowing one to derive noisy instances from clean ones (refer to Figure 13a). However, dealing with mask losses is more challenging. The loss of noisy instances consists predominantly of correctly labeled pixels, with only a few noisy ones (refer to Figure 13b). Furthermore, since most spatial noise occurs at the boundaries, these areas are where the model exhibits the least confidence Kimhi et al. [2024]. This complexity makes it impossible to distinguish between pixel-level noisy and clean data, posing a significant challenge in developing a spatial noise solution to learn from noisy labels.

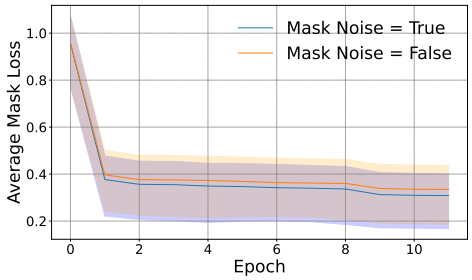
Due to these difficulties, we compared a class noise method to handle noisy labels. Table 13 presents the results on the COCO-N benchmark, comparing standard Cross-Entropy with Symmetric Cross-Entropy Wang et al. [2019]. While there is a marginal improvement, the method still faces challenges as the noise level increases.

Table 13: Evaluation Results of Instance Segmentation with different losses learning with noisy labels trained on COCO-N dataset (mAP / Boundary mAP).

Loss	Clean		Easy		Mid		Hard	
	AP	AP^B	AP	AP^B	AP	AP^B	AP	AP^B
CE	34.6	20.6	31.8	18.9	30.3	17.5	28.4	16.3
SCE	32.5	19.5	32.1	18.9	30.8	17.8	28.8	16.4



(a) Class Loss Separation. Average of the class loss of the Coco dataset, with the 25% and 75% quantiles as margins - per epoch of training.



(b) Mask Loss Separation. Average of the mask loss of the Coco dataset, with the 25% and 75% quantiles as margins - per epoch of training.

F SAM Finetune with label noise

Since for weakly supervision annotations we heavly relay on SAM Kirillov et al. [2023], we exemine how noise in prompt effect the model itself in two setups, *zero-shot*, that corespond to the quality of the masks produced by sam, and *fine-tuning*, as a popular paradigm of using SAM for a downstream application. For the zero-shot, we prompt SAM with the grounded bounding boxes of the validation set of COCO as well as noisy boxes with the **COCO-N hard** type of noise on the validation annotations. For fine-tuned, we exemine fine tuning with both clean and noisy **COCO-N hard** annotations masks. Table 14, shows both mIoU and F1 scores of the masks produced by SAM, showing that the quality of masks can be increased when fine-tuned, compared with zero shot training with high quality prompts. Fine-tuning with noisy annotations however, is less sever, when prompting with cerfully designed prompts, compared to noisy prompts. Our findings suggest that the quality of prompts are fur more important then the qua

Table 14: Evaluation of prompt Instance segmentation on SAM

Annotations	Clean		COCO-N Hard	
Method	IoU	$F1$	IoU	$F1$
Zero-shot	79.78	87.49	67.99	63.30
Fine-tune	79.91	78.6	77.47	76.18

G Biases of Self-annotating Datasets

More visual results of the weakly supervised annotations created by SAM are presented in Figure 14. A significant number of annotations were curated by this process (top row), reducing label noise, particularly in cases where the original annotations suffered from approximation noise. In other instances, where an object is surrounded by similar colors or illumination conditions, the annotations become noisier around the boundaries, exhibiting weak localization noise (middle row).

The specific context of the dataset annotations can influence what the user is looking for. We observed cases where there is ambiguity in the definition of certain objects, such as stove-tops (bottom row).

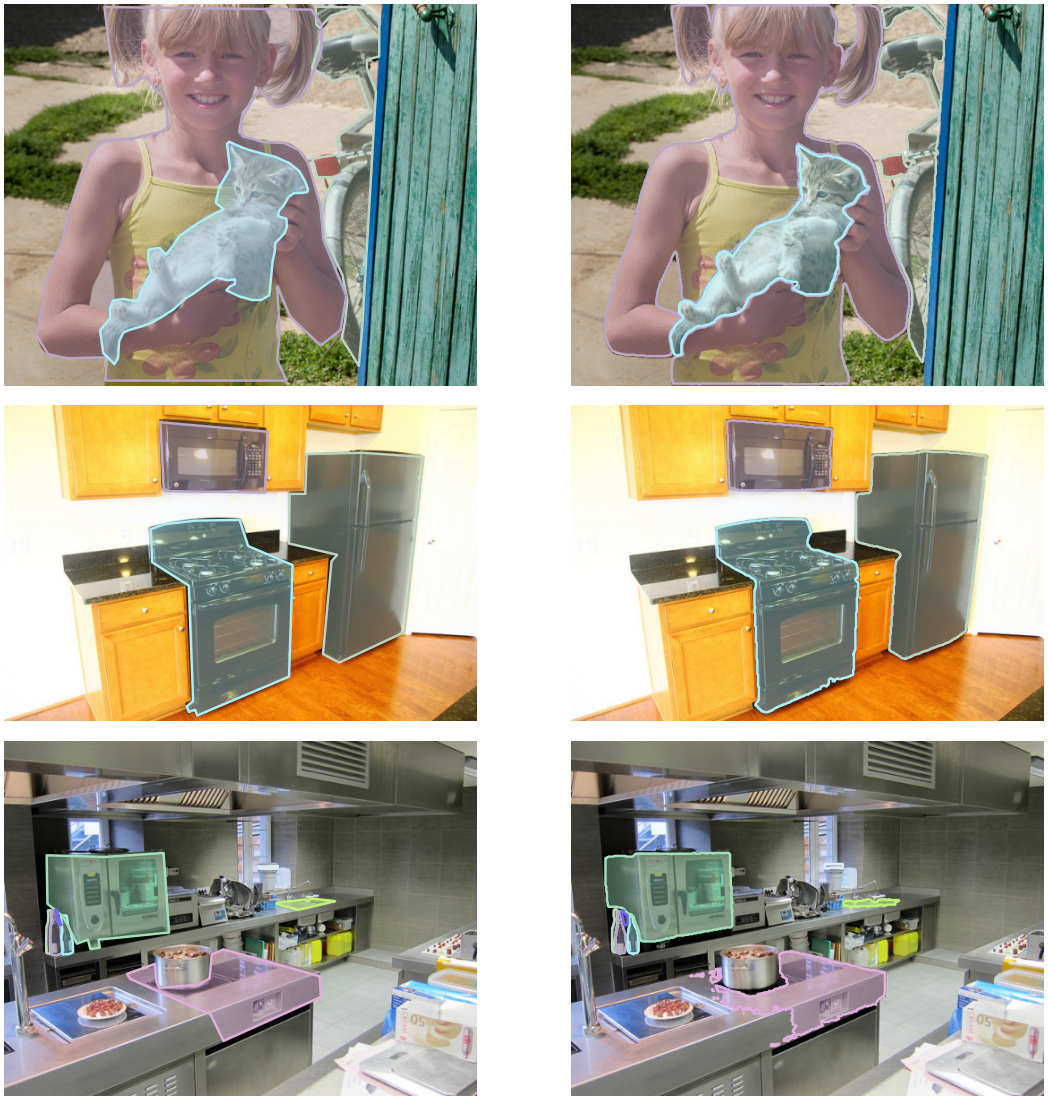


Figure 14: Pairs of COCO annotations (left) and **COCO-WAN** easy annotations (right). Top pair shows high fidelity annotations for **COCO-WAN**, compared to the original noisy counterparts. The bottom example examine that when color changes by little, even with bounding box prompts, SAM confuses due to color biases in segments and can not capture the desired segments such as stovetop or sink.

While SAM is familiar with the concept of a stove-top, it lacks the contextual knowledge of what it should be within the specific context of the COCO dataset, leading to poor masking.

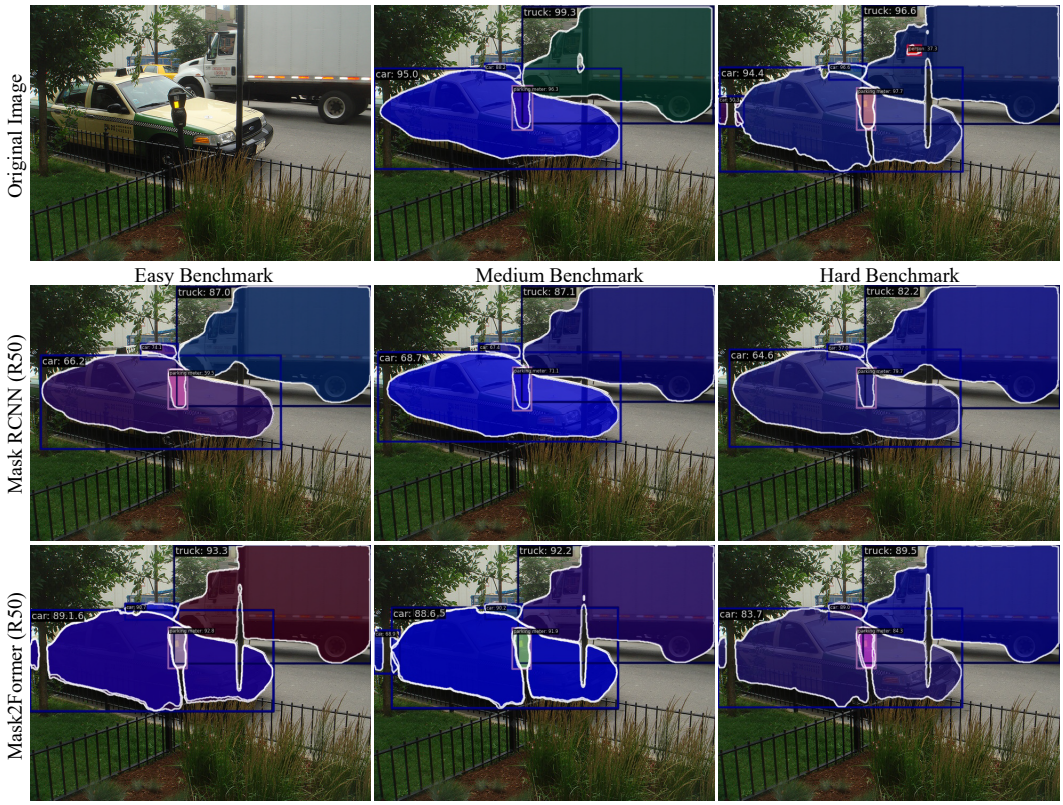


Figure 15: Comparison Mask RCNN and Mask2Foramer models predictions. Top row (left to right): original image M-RCNN and M2F predictions on clean COCO. Middle: M-RCNN predictions on Easy, Medium and Hard **COCO-N**. Bottom: M2F predictions on Easy, Medium and Hard **COCO-N**.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have made sure the abstract and introduction capture the main contributions of this paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: refer to Section 5.

   Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: No theoretical results or proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, not only is the method reproducible (with code intended for release upon acceptance), but also the models evaluated are all publicly available from online repositories cited in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code to reproduce all the benchmarks creation in https://anonymous.4open.science/r/noisy_labels-0C70/README.md while the models and training recipies from https://github.com/open-mmlab/mmdetection and https://github.com/facebookresearch/segment-anything.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to last question and Appendix D for all implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The error bars of multiple runs are not a common practice in segmentation. However, many of the ablations show significant of effects. note that we run all experiments with same random seed and data splits to validate any hypothesis.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Refer to Hardware details in Appendix D.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have read the NeurIPS Code of Ethics, and our research conforms to it in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we discussed it in the introduction and final discussion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: All masks retained from publically avalible datasets. We refer to it in the limitations Section 5.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All code and resources not created by the authors was properly credited in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: We made publicly available the code that create the benchmark suit, yet a dataset, including the synthetic data will be release upon acceptance (with the proper documentation).

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: NA: No human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.