scientific reports

Check for updates

OPEN

SML-IS: a multi-scale annotation approach to improve instance segmentation accuracy

Adem Kosar[1], Onur Erdem Korkmaz[1,2], Tolga Turay[1,2], Ebru Ergün[3]✉ & Cengiz Oztireli[4]

Instance segmentation requires precisely outlining object boundaries in images using polygonal masks. Conventional single-scale annotations often fail to represent multi-scale appearance variations, which limits segmentation accuracy and reduces the semantic richness available to learning algorithms. We introduce *Small–Medium–Large Instance Segmentation (SML-IS)* labelling, a multi-scale polygon annotation strategy that provides three different annotation scales for each instance within the same image. Instead of assigning a single mask, SML-IS replicates each image and applies small, medium, and large masks for every instance, thereby tripling the number of segmentation masks while keeping the original imagery unchanged. Unlike data-augmentation methods that modify raw images, SML-IS enriches only the supervision, improving label quality and scale diversity without altering the dataset size. On a custom dataset, SML-IS consistently improves mean Average Precision (mAP) by approximately 5–6% across representative models such as Mask R-CNN and YOLOv7-seg, regardless of architectural paradigm. For domain-specific verification, the proposed method was further applied to the Kvasir-SEG colorectal polyp segmentation dataset, where SML-IS demonstrated 7% mAP gain, confirming its robustness and effectiveness in medical imaging applications. Although validated in the medical domain, the method remains broadly applicable to other visual domains such as agriculture, autonomous driving, and industrial inspection, where precise boundary delineation and high-quality annotations are essential but large-scale data collection is limited.

Keywords Computer vision, Instance segmentation, Dataset labelling, Medical image analysis, Polyp segmentation

Instance segmentation is a key computer vision task that involves detecting objects and precisely outlining their boundaries at the pixel level. It plays a crucial role in applications such as autonomous driving, medical diagnostics, and precision agriculture, where detailed segmentation enhances decision-making. The performance of instance segmentation models largely depends on two factors: the quality of annotated datasets[1] and the capabilities of the underlying deep learning model[2]. Among these, dataset quality is particularly critical, as annotations (commonly referred to as "labels") provide the essential ground truth for training models in object recognition and segmentation.

A dataset's composition significantly affects a model's ability to generalize. Factors such as class diversity, object scale, occlusions, density, lighting, and background complexity influence performance. Datasets that account for these elements improve the robustness and accuracy of instance segmentation models, enabling adaptation to diverse conditions[3]. Well-annotated, diverse datasets are thus essential for effective training and real-world applicability.

Conventional instance segmentation annotation methods[1] typically rely on single-scale polygon masks, where annotators outline instances using a single set of boundary points and assign a corresponding class label. An example of this single-scale annotation approach is shown in Fig. 1, derived from the MS COCO[1] dataset, displaying an image with its corresponding segmentation masks. While effective for many tasks, this approach often fails to capture the full semantic complexity of instances, particularly in visually intricate scenes where instance appearance varies across scales or perspectives. Consequently, conventional annotation practices

[1]Department of Electrical and Electronics Engineering, Ataturk University, 25240 Erzurum, Turkey. [2]High Performance Computing Application and Research Center, Ataturk University, 25240 Erzurum, Turkey. [3]Department of Electrical and Electronics Engineering, Recep Tayyip Erdogan University, 53100 Rize, Turkey. [4]Department of Computer Science and Technology, University of Cambridge, Cambridge CB3 0FD, UK. ✉email: ebru.yavuz@erdogan.edu.tr



Fig. 1. Visual representation of detection and segmentation annotations in MS COCO (bicycle class)[1].

may limit the performance of instance segmentation models, especially in applications where precise instance boundaries and multi-scale variations are crucial[4].

The limitations of single-scale annotation methods are particularly evident in data-scarce domains such as medical imaging and agriculture, where precise segmentation is essential[5]. In medical imaging, multi-scale annotations are often required to detect small yet clinically relevant features, while agricultural applications rely on them to account for the considerable variation in object sizes, such as pests or crop diseases. Although several annotation strategies[1,6,7] have been proposed to address these challenges, developing a standardized approach that improves annotation depth without significantly increasing manual effort remains an ongoing challenge.

To overcome these limitations, this study introduces a novel annotation approach called Small-Medium-Large instance segmentation labelling (SML-IS), which enriches segmentation depth through multi-scale polygon masks. Unlike conventional methods, SML-IS duplicates each image three times and applies three distinct polygon scales per instance, effectively tripling the number of segmentation masks without increasing the raw image set.

The motivation for this approach originates from two key insights: (i) addressing dataset limitations, particularly in fields such as medical imaging, where deep learning models often struggle to improve due to insufficient training samples[8]. Enhancing annotation quality, irrespective of model architecture, is crucial for advancing segmentation performance; (ii) incorporating the human visual system's ability to perceive scenes at multiple scales. Just as humans dynamically adjust focus based on visual cues, machine learning models can leverage multi-scale information to achieve more precise segmentation across varying contexts. Similarly, recent segmentation studies have shown that incorporating multi-scale or progressively refined supervision can enhance boundary representation and segmentation accuracy[9]. As illustrated in Figure 2, this principle served as the basis for developing the SML-IS labelling approach, which applies multi-scale polygon annotations to enhance segmentation depth. Mirroring the collaboration between the human eye and brain, computer vision systems combine sensory components (e.g., cameras) with computational units (e.g., processors) to interpret and analyze visual data with greater efficiency.

Experimental analysis of the SML-IS labelling approach, performed on a custom dataset specifically designed and annotated for this study, demonstrates consistent improvements in segmentation accuracy, leading to a 5–6% increase in mean Average Precision (mAP) across diverse model architectures, including Mask R-CNN[2] and YOLO[10]. For further domain-specific validation, the proposed method was also applied to the Kvasir-SEG[11]

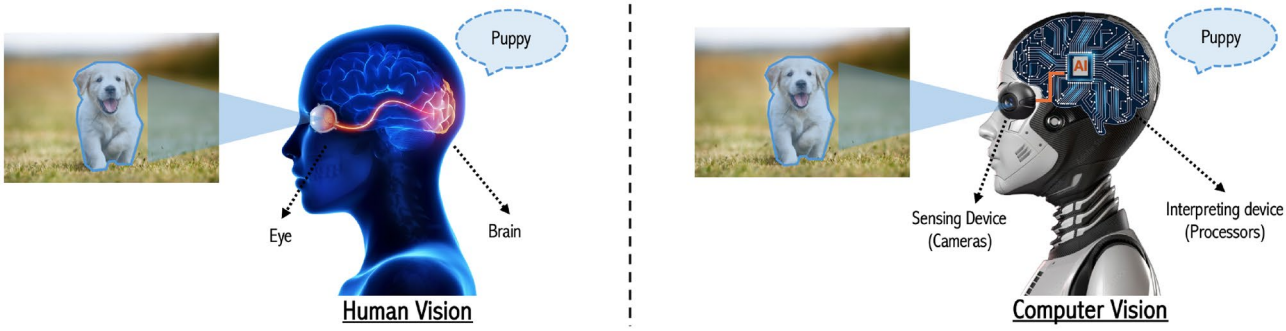


Fig. 2. Conceptual analogy between human and computer vision systems.

dataset, an open-source benchmark for colorectal polyp segmentation. When trained on the same segmentation models, SML-IS achieved an additional 7% improvement in mAP, confirming its robustness and generalizability across domains. These results highlight the flexibility of the SML-IS labelling approach and its effectiveness in data-limited environments, where high-quality annotations are critical for reliable segmentation and clinical decision-making.

This paper is structured as follows: "Related work" examines various dataset annotation techniques used in instance segmentation. "SML-IS: principles and practical integration" introduces the SML-IS Labelling approach, outlining its theoretical foundations and implementation. "Performance evaluation and discussion" presents the experimental results and their significance. "Critical reflections and future research" discusses possible future research directions, followed by a conclusion summarizing key findings and contributions.

Related work

Annotation strategies in instance segmentation

First, to contextualize the proposed SML-IS approach, annotation techniques used in instance segmentation can be broadly categorized into: (i) Manual labelling, where human annotators draw detailed segmentation masks (commonly using polygonal tools) to define object boundaries with high precision[1]; (ii) Semi-automatic labelling[12], which combines algorithmic suggestions with human verification to accelerate mask creation; (iii) Fully automatic labelling, which uses algorithms to generate masks without human input[13], offering scalability at the cost of potential inaccuracies; (iv) Weakly-supervised[14] and unsupervised[15] labelling techniques, which attempt to infer segmentation cues from image-level labels or unlabelled data, reducing annotation costs but generally producing less precise boundaries; and (v) Crowdsourcing approaches[16], enabling rapid dataset creation through distributed human labour, though often requiring post-processing to ensure consistency. Thus, the selection of an annotation method is generally influenced by factors such as the complexity of the task, the availability of data, and the balance between annotation quality and associated costs.

Manual labelling variants

As the SML-IS approach builds on manual labelling, the most prominent sub-variants of manual labelling are reviewed below.

Single-scale polygon masks: this is the most frequently used (conventional) approach in most datasets like COCO[1], where each instance is annotated once using a single polygon that outlines its visible boundary. While efficient, this method lacks scale diversity and may fail to capture smaller or more ambiguous regions of the instance in cluttered or high-variance scenes.

Multi-polygon per instance: some datasets[1,17] incorporate multiple disconnected regions per instance in an image, which is particularly useful when the object is fragmented or partially occluded. This annotation style allows for a more accurate representation of visible object parts but does not inherently address scale variation. In parallel, Run-Length Encoding (RLE)[18] is often employed as a compact representation format for dense binary masks, especially in cases involving large or complex instances. While RLE itself is not an annotation method, it can represent multiple masks per instance at varying resolutions, thus facilitating multi-scale annotation strategies in practice. For example, crowd annotations or fine-grained segmentations in datasets like COCO[1] utilize RLE to efficiently encode multiple instance-level masks, including those of differing scales or granularities.

Resolution-aware annotation: in domains like medical imaging or satellite analysis, masks are drawn considering the resolution of the image, with finer granularity applied to high-resolution data[19,20]. This implicitly introduces scale-awareness but is not always consistent across datasets.

Multi-scale annotation across images: traditional approaches[1,6,21] to scale variation typically rely on different instances of the same object appearing at various distances in separate images. For example, a car may appear large in a close-up and small in a distant street shot, and each is annotated accordingly. However, this does not ensure consistent representation across scales within the same image. In contrast, the SML-IS method proposed in this study introduces a multi-scale annotation strategy applied directly within the same image, thereby providing consistent scale-aware supervision for each object instance. This strategy enriches annotation quality and semantic diversity without requiring additional data, ultimately yielding improved segmentation accuracy and representing a distinctive contribution to the field.

Positioning SML-IS in relation to transformer architectures and data augmentation

In the field of computer vision, instance segmentation approaches are developed either on CNN- or Transformer-based architectures, with different priorities depending on the application domain. Transformer-based models have demonstrated notable superiority in accuracy compared to traditional CNN-based models (e.g., Mask R-CNN[2], YOLO[22]). For instance, DETR[23], trained on bounding-box and polygon-mask annotations from the MS COCO dataset, has achieved strong results in both detection and segmentation tasks. Similarly, SAM[24] and its medical adaptation MedSAM[25], trained on pixel-level mask annotations, have had a significant impact on the segmentation literature. However, these Transformer-based models contain approximately 636 million parameters[24] and reach gigabyte-level sizes, restricting their usability to high-capacity GPUs and limiting their accessibility in real-time applications. In comparison, CNN-based models such as Mask R-CNN (∼41M parameters)[2] and the YOLO family (∼2.6M parameters)[26] are relatively lightweight, capable of operating with lower hardware requirements and higher resource efficiency, and are therefore widely preferred in real-time systems due to their speed[27–29]. In particular, the single-stage architecture of YOLO has made it especially attractive in both industry and academia.

In light of these developments, applying the proposed SML-IS method initially to polygon masks rather than pixel-level masks can be regarded as an appropriate starting strategy. Success at this stage would demonstrate the potential to extend SML to pixel-level annotations with alternative methodologies in subsequent studies. Since all models, whether Transformer- or CNN-based, rely on annotated ground truth for learning, SML provides annotation richness that can enhance accuracy and thus offers an integrable solution adaptable to different label types across tasks.

Furthermore, data augmentation techniques (e.g., rotation, cropping, brightness/noise modifications) increase dataset diversity but do not improve the semantic depth of annotations[30,31]. In contrast, SML enriches annotation quality by introducing scale diversity into labels without altering the raw data. Therefore, SML and data augmentation are distinct yet complementary; when used together, they can enhance both visual and annotation diversity, thereby further improving model performance.

SML-IS: principles and practical integration

The goal of SML-IS is to enrich semantic diversity during training, allowing models to better generalize across object scales and appearances. This is particularly effective for models like Mask RCNN[2] and YOLOv7-seg[22], which generate segmentation masks through region-based or dense prediction mechanisms. The following sections describe the conceptual basis and technical implementation of the SML-IS method.

Human-centric rationale for multi-scale masking

As discussed in "Introduction", the design of the SML-IS method is inspired by the human visual system, which interprets scenes through dynamic attention at multiple scales[32]. As illustrated in Fig. 2, human perception involves both localized detail and holistic understanding, enabling recognition even when objects are partially visible or viewed at different distances. In contrast, most instance segmentation models are trained using single-scale polygon masks, which restrict the model's semantic scope and adaptability.

Figure 3a shows a *Boston Dynamics* robot image, Spot®[33], annotated with a conventional tight polygon mask, reflecting the standard single-scale annotation approach. In contrast, Fig. 3b–d illustrate the SML-IS annotations at small, medium, and large scales, respectively. Each mask captures distinct semantic cues (ranging from fine-grained details to broader object context) while remaining semantically valid and independently recognizable. This demonstrates how multi-scale representations more closely mirror human visual interpretation. Even when only a coarse or partial view is available, the human visual system can often recognize the object based on scale-invariant features. By simulating this capacity, SML-IS supplies segmentation models with richer and more varied training signals, improving their ability to generalize across scale variations, occlusions, and viewpoint changes commonly encountered in real-world tasks.

The role of anchor boxes in instance segmentation

Anchor boxes are fundamental components in many instance segmentation architectures, including Mask RCNN[2] and YOLOv7-seg[22], where they serve as spatial priors for object localization. In these models, anchors help guide the detection process before mask prediction, enabling effective handling of objects at various scales and aspect ratios. However, conventional annotation practices typically provide a single mask per instance, limiting the diversity of anchor-level supervision. By introducing multiple masks at different scales for each instance, the proposed SML-IS labelling method enhances anchor coverage during training, leading to more robust mask predictions under scale variation.

Anchor boxes and proposal generation in mask RCNN

Mask RCNN[2], built upon the Faster RCNN[34] architecture, utilizes anchor boxes as the foundation for generating object proposals via its Region Proposal Network (RPN). These anchor boxes are predefined at multiple scales and aspect ratios [typically combining scales per pixel with three aspect ratios (e.g., 1:1, 1:2, 2:1)] enabling the model to hypothesize object locations across a wide range of shapes and sizes. This built-in diversity allows the RPN to cover varied object geometries effectively and contributes to the overall robustness of the detection and segmentation process.

Once proposals are generated, they are refined through bounding box regression and subsequently passed to the RoIAlign layer, which preserves spatial alignment. This enables the extraction of semantically rich features for classification and mask generation. A dedicated mask head then predicts a binary segmentation mask for each object class within these regions.

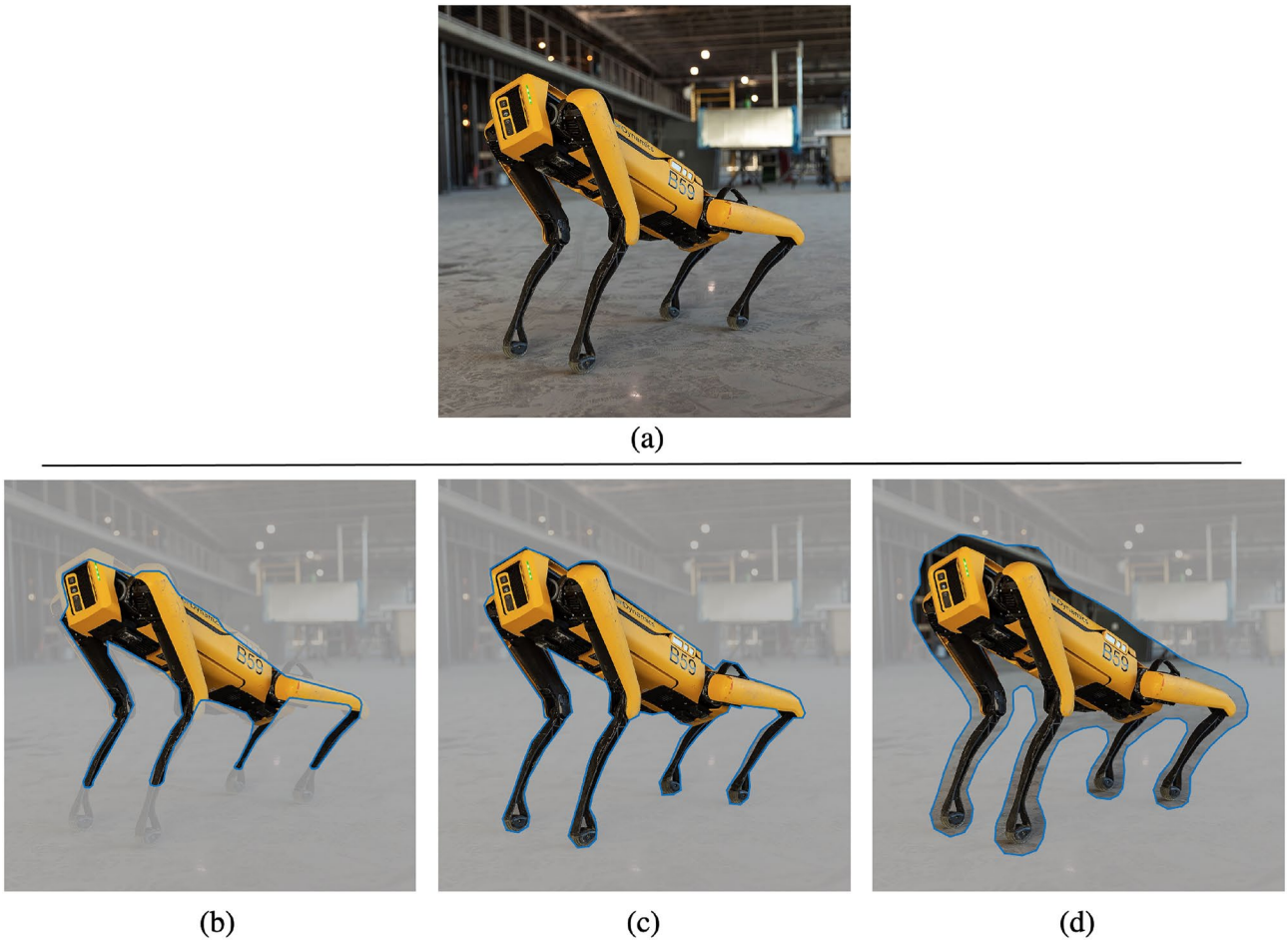


Fig. 3. Demonstration of SML-IS mask generation at multiple scales.

Formally, the bounding box regression targets are computed as follows, where (x_a, y_a, w_a, h_a) are the anchor box parameters and (x, y, w, h) denote the corresponding ground truth box:

$$t_x = \frac{(x - x_a)}{w_a}, \quad t_y = \frac{(y - y_a)}{h_a}, \quad t_w = \log\left(\frac{w}{w_a}\right), \quad t_h = \log\left(\frac{h}{h_a}\right). \tag{1}$$

These normalized offsets facilitate efficient learning by minimizing spatial variance between anchor and ground truth boxes. The overall multi-task loss used in Mask RCNN is defined as:

$$\begin{aligned} L(\{p_i\}, \{t_i\}, \{m_i\}) = & \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(p_i, p_i^*) + \lambda_1 \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{\text{reg}}(t_i, t_i^*) \\ & + \lambda_2 \frac{1}{N_{\text{mask}}} \sum_i p_i^* L_{\text{mask}}(m_i, m_i^*) \end{aligned} \tag{2}$$

where p_i, t_i, and m_i represent predicted class probabilities, bounding box offsets, and binary masks, respectively, and p_i^*, t_i^*, and m_i^* are their ground truth counterparts.

While the architecture incorporates a built-in multi-scale mechanism via anchor box design, it remains constrained by the scale limitations of conventional annotations. By providing additional masks per instance at varying scales, the SML-IS method introduces more variation into anchor-to-mask associations during training, allowing the model to benefit from richer supervision without altering its architectural components.

Anchor boxes in YOLO-based instance segmentation

YOLO-based instance segmentation models, such as YOLOv7-seg[22], extend the original YOLO detection framework by integrating a segmentation head alongside the object detection pipeline. These models adopt a single-stage architecture in which bounding box regression, classification, and mask prediction are performed simultaneously. The input image is divided into an $S \times S$ grid, where each grid cell is responsible for predicting multiple anchor-aligned bounding boxes, class probabilities, and instance masks.

Unlike Faster RCNN's use of fixed anchor configurations, YOLOv7-seg employs a data-driven anchor generation strategy using k-means clustering[35] to adaptively match the distribution of object sizes and aspect ratios in the training dataset. Each predicted bounding box $\mathbf{b}_i$ is compared to its corresponding ground truth $\mathbf{b}_i^*$ using the Intersection over Union (IoU) criterion. If the IoU exceeds a certain threshold (typically 0.5), the anchor is assigned as positive and supervised accordingly.

The model then jointly optimizes localization, classification, objectness, and segmentation losses. For a matched anchor $_i$, the total loss can be expressed as:

$$L(\mathbf{b}_i, \mathbf{c}_i, m_i) = \lambda_{\text{bbox}} L_{\text{bbox}}(\mathbf{b}_i, \mathbf{b}_i^*) + \lambda_{\text{cls}} L_{\text{cls}}(\mathbf{c}_i, \mathbf{c}_i^*) + \lambda_{\text{obj}} L_{\text{obj}}(o_i, o_i^*) + \lambda_{\text{mask}} L_{\text{mask}}(m_i, m_i^*), \quad (3)$$

where $\mathbf{c}_i$ and m_i are the predicted class probabilities and segmentation mask, while o_i is the objectness score; each $*$-marked term denotes the corresponding ground truth.

While this structure enables efficient joint learning, conventional datasets typically provide only one mask per object, limiting the variation that anchors encounter during training. By introducing additional masks per instance at multiple scales, SML-IS modestly increases the diversity of mask-level supervision, helping the model capture object variations more effectively without altering its architecture or learning process.

Models without anchor boxes

Anchor-free instance segmentation models such as CenterMask[36] and YOLOv8-seg[37] eliminate predefined anchor boxes by directly predicting object locations and segmentation masks from dense feature maps. These methods typically employ heatmap- or keypoint-based localization together with lightweight mask heads for instance-level prediction. While this design simplifies the detection pipeline, it also increases sensitivity to the quality and diversity of ground-truth annotations.

Although anchor-based models (e.g., Mask R-CNN[2], YOLOv7-seg[22]) and anchor-free approaches differ in their detection mechanisms, both rely on high-quality annotations for effective learning. Conventional labelling provides only a single polygon mask per instance, limiting multi-scale representation. SML-IS addresses this limitation by providing multiple masks at different scales without modifying the network architecture or requiring additional image data.

Implementation procedure of SML-IS labelling

Figures 1 and 3a illustrate conventional instance segmentation annotations generated using single-scale polygon masks. In contrast, Fig. 3b–d depict the SML-IS labelling process, in which the same image is annotated three times (triple sampling), with each version containing polygon masks drawn at a distinct scale. The scale definitions were established in a standardized quantitative manner, with geometric scaling factors systematically applied relative to the Medium (M) mask. Specifically, the Small (S) masks were scaled down by approximately 10–20%, and the Large (L) masks were scaled up by approximately 10–20% relative to the original Medium (M) mask. To ensure consistency, polygon areas were calculated from the mask vertices using the Python-based Shapely geometry library, and the resulting S and L masks were manually adjusted to satisfy the target area constraints. These scaling operations were strictly confined within this range, ensuring proportional consistency while preserving consistent object representations across all annotation scales. All polygon masks were manually refined by the annotators to maintain smooth contours and uniform margin consistency throughout the dataset. Figure 3b–d correspond to these three scale levels: Small (S) annotation (Fig. 3b), Medium (M) annotation (Fig. 3c), which represents the conventional single-scale method, and Large (L) annotation (Fig. 3d). It should be further emphasized that the additional S and L annotations are introduced solely as n annotations for auxiliary supervision during training. They are not intended to replace the original M annotation or represent alternative ground-truth annotations. Instead, they provide multiple annotation scales of the same object to enrich supervision during learning, while the original M annotation remains the sole reference ground truth for model evaluation.

For images containing multiple object instances, which is common in our custom dataset, SML-IS is applied consistently across all replicated images. Each duplicated version receives new polygon annotations reflecting a specific scale for all present objects. Figure 4 demonstrates this with an image featuring three different robot classes (Spot®[33], Atlas®[38], and Handle®[39]). Figure 4a shows the standard single-scale annotations, while Figure 4b–d illustrate the multi-scale masks. A controlled blur effect is applied to these latter figures to emphasize that the annotated mask regions alone contain sufficient semantic information to recognize the objects, even without full scene context.

In quantitative terms, a dataset containing m annotated object instances will result in $3m$ polygon masks after applying SML-IS. This augmentation is also applied to the test set, maintaining consistency throughout training and evaluation pipelines for fair performance comparison across scales.

Performance evaluation and discussion

This section presents a detailed evaluation of the proposed SML-IS labelling framework conducted on two distinct datasets: a custom instance segmentation dataset developed and annotated for this study, and the Kvasir-SEG[11] dataset, a publicly available benchmark for colorectal polyp segmentation. The first part of the analysis focuses on the custom dataset, providing a systematic comparison between the proposed SML-IS approach and traditional single-scale polygon annotations[1] to assess its fundamental performance advantages. In the latter part, the method is further validated through its first domain-oriented application on the Kvasir-SEG dataset, demonstrating its contribution to practical medical imaging scenarios by evaluating its effect on colorectal polyp segmentation accuracy.

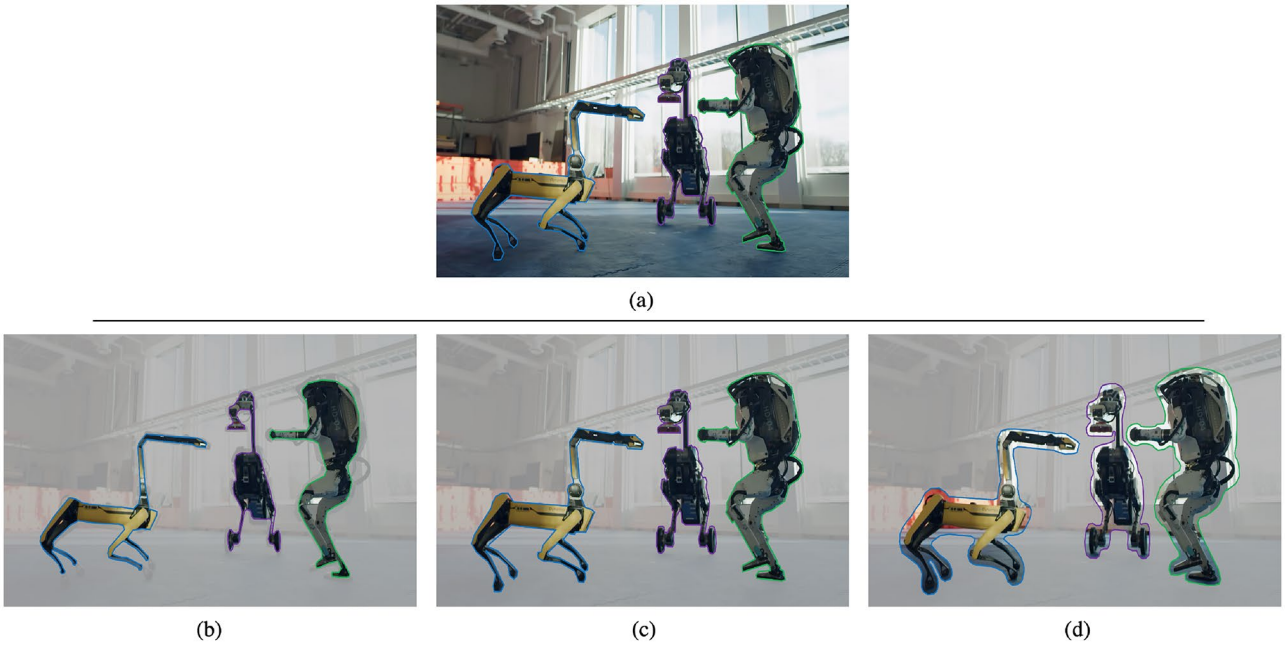


Fig. 4. Demonstration of SML-IS multi-scale annotation application across diverse robot instances.

Class occurrences	Training	Test
	Image numbers | class distributions	Image numbers | class distributions
1 Class (1 instance)	240 | atlas:80, handle:80, spot:80	60 | atlas:20, handle:20, spot:20
2 Classes (2 instances)	81 | atlas:46, handle:54, spot:62	18 | atlas:10, handle:12, spot:14
3 Classes (3 instances)	64 | atlas:60, handle:63, spot:69	16 | atlas:15, handle:15, spot:18
3 Classes (4 instances)	10 | atlas:12, handle:16, spot:12	3 | atlas:4, handle:4, spot:4
Instances with RLE[18]	5 | atlas:2, handle:1, spot:2	3 | atlas:1, handle:1, spot:1
Total number	400 | atlas:200, handle:214, spot:225	100 | atlas:50, handle:52, spot:57

Table 1. Distribution of object classes across training and test splits.

For comprehensive assessment, two advanced instance segmentation architectures (Mask R-CNN[2] and YOLO[22]) were utilized in the experiments. Their performances were measured using mAP across multiple Intersection over Union (IoU) thresholds. The findings from both datasets indicate that the SML-IS labelling approach significantly improves segmentation performance across various object scales and boundary structures, consistently surpassing conventional single-scale annotation methods.

Evaluations on custom dataset

A custom dataset (https://www.kaggle.com/datasets/ademkosar/custom-instance-segmentation-dataset) was carefully constructed to support the evaluation of the proposed SML-IS labelling strategy, with particular emphasis on improving model generalization and maintaining compatibility with leading instance segmentation frameworks such as Mask RCNN[2] and YOLOv7-seg[22]. All images were uniformly resized to a resolution of 640×640 pixels to ensure computational consistency across training platforms, including Detectron2[40] and Ultralytics[41].

The dataset adheres to the MS COCO format[1] and contains both bounding box and polygon mask annotations. While bounding boxes were included to maintain compatibility with detection frameworks, all experiments in this study rely exclusively on polygon annotations for segmentation. The dataset features three object categories developed by Boston Dynamics: Spot®[33], Atlas®[38], and Handle®[39]. Figure 3 presents a single-class sample (Spot®), and Fig. 4 depicts an example containing all three classes.

To ensure a balanced representation of object categories and instance compositions, the dataset was intentionally designed to include images with varying numbers of object classes and instances, as summarized in Table 1. Specifically, the dataset contains 300 images with a single object class (one instance), 99 images containing two object classes (two instances), and 93 images containing all three object classes, including both three-instance and four-instance configurations. Additionally, 8 images containing occluded instances were annotated using Run-Length Encoding (RLE)[18], which is widely adopted in segmentation benchmarks such as MS COCO, further enhancing the dataset's ability to evaluate robustness under challenging visual conditions.

Although limited in scale with only 500 base images, the dataset was designed to reflect a broad range of object interactions and conditions. By applying SML, the number of annotations is effectively tripled without the need to increase the image count. This results in richer training data and more informative ground truth supervision, enabling reliable segmentation performance evaluations while avoiding the annotation overhead associated with large-scale datasets.

Multi-scale dataset variants

The dataset used in this study was expanded into multiple variants, each reflecting a different polygon-based labelling strategy. The baseline variant, denoted as M, represents conventional single-scale annotation. Additional variants (S, L, SM, SL, ML, and SML-IS) were generated to capture different combinations of polygon mask scales. For instance, the SML-IS-labelled dataset includes three distinct polygon masks per object, corresponding to small (S), medium (M), and large (L) scales.

To conduct a comprehensive assessment of the proposed SML-IS method, all dataset variants were derived from the same base set of 500 images (Table 1). This ensured a consistent comparison across different labelling strategies by isolating the effect of annotation scale from other variables. While the image content remained unchanged, the SML-IS approach increased the total number of instance masks threefold by introducing scale-specific polygon annotations for each object. This demonstrates the adaptability of SML-IS in enhancing segmentation quality without necessitating any additional image data. Table 2 provides a breakdown of the image distribution for training and testing across all annotation variants.

Details of training process

The effectiveness of the proposed SML-IS labelling method was assessed using the dataset variants listed in Table 2. For this evaluation, both two-stage and one-stage instance segmentation architectures were selected. The two-stage models included Mask RCNN R101-FPN and X101-FPN backbones, implemented using the Detectron2 framework[40]. For one-stage segmentation, we employed YOLOv7-seg[22], a high-performance anchor-based model, and YOLOv8l-seg[37], which adopts an anchor-free design. This diverse model selection allows us to examine the applicability of SML-IS across distinct learning paradigms, particularly between anchor-based and anchor-free approaches.

All images were resized to 640×640 pixels to maintain uniform input dimensions. Each model was trained for 50 epochs with a batch size of 10. The initial learning rate was set to 0.001 and decayed by a factor of 10 after 30 epochs via a step scheduler. Weight decay was fixed at 10^{-4}, and optimization was performed using stochastic gradient descent (SGD) with a momentum of 0.9. For YOLO-based models implemented via the Ultralytics framework[41], training hyperparameters were modified from their defaults to align with the Mask RCNN training regime, ensuring a balanced comparison.

Models were trained using S, M, L, SM, SL, ML, and SML-IS-labelled datasets and evaluated on both the M-labelled and SML-IS-labelled test sets. This setup enabled a controlled comparison of conventional versus multi-scale annotation strategies. Model performance was quantified using mAP across multiple Intersection over Union (IoU) thresholds, following the COCO evaluation standard[1]. The evaluation included mAP@[.5:.95], mAP@.5, and mAP@.75 metrics, offering insights into both general and fine-grained segmentation performance.

Training and convergence behavior

Figure 5 presents the convergence behavior of the Mask R-CNN R101-FPN and YOLOv7-seg models trained with M-labelled and SML-IS-labelled datasets. The comparison clearly demonstrates the effect of the proposed SML-IS annotation approach during training. Both models show smooth convergence trends, and in each case, the SML-IS-labelled datasets achieve lower final loss values than the conventional M-labelled ones. The learning-rate decay introduced at the 30th epoch triggers a stabilization phase, confirming convergence. After this point, the SML-IS-labelled configuration exhibits improved responsiveness to fine-grained features, emphasizing the advantages of multi-scale representation. Mask R-CNN attains a slightly lower final loss (≈ 0.97) compared to YOLOv7-seg (≈ 1.03) when trained on the SML-IS-labelled dataset.

In the early epochs (especially the first 10), YOLOv7-seg shows stronger oscillations in total loss than Mask R-CNN. This behavior stems from its one-stage architecture, which simultaneously optimizes classification and localization, making it more sensitive to learning-rate changes and early gradient instability. Additionally, the anchor-based bounding-box assignment in YOLOv7-seg varies more before stabilizing, while the region-proposal mechanism in Mask R-CNN provides a more structured and steady gradient flow from the beginning.

Generated dataset names	Training (image number)	Test (image number)
S	400	100
M	**400**	**100**
L	400	100
SM	800	200
SL	800	200
ML	800	200
SML-IS	**1200**	**300**

Table 2. Dataset variants derived from annotation scale combinations. Significant values are in bold.

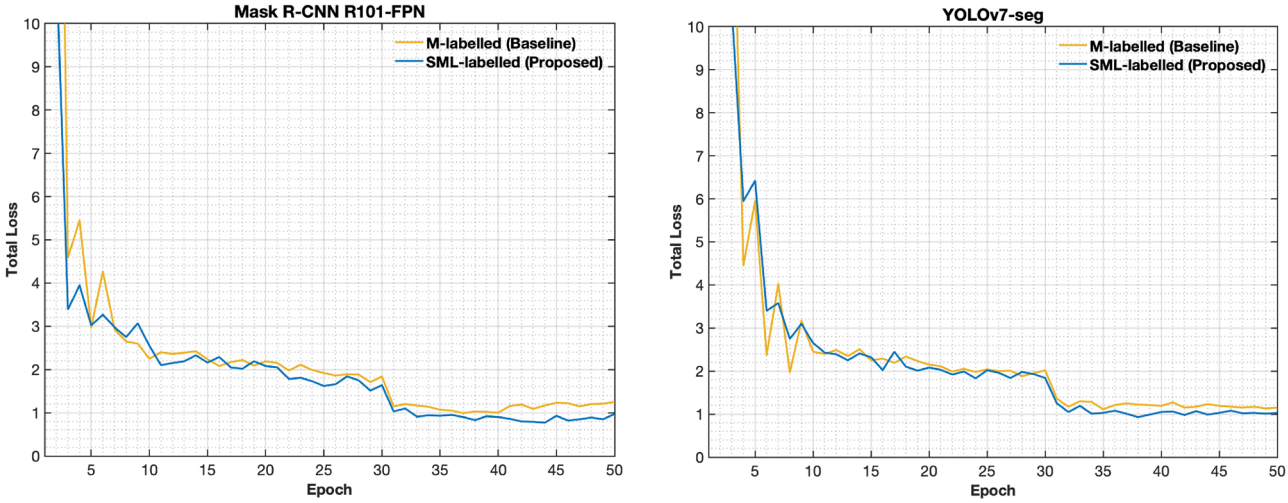


Fig. 5. Comparison of total loss convergence across different labelling strategies for mask R-CNN R101-FPN and YOLOv7-seg models.

Dataset	Mask R-CNN R101-FPN			Mask R-CNN X101-FPN		
	mAP@[.5, .95]	**mAP@.5**	**mAP@.75**	**mAP@[.5, .95]**	**mAP@.5**	**mAP@.75**
S	26.6	33.9	27.7	28.4	35.0	30.1
M	**38.9**	**53.9**	**41.6**	**42.2**	**57.5**	**44.3**
L	34.2	47.2	36.0	37.0	50.2	39.8
SM	41.2	54.6	47.7	45.0	59.9	51.5
SL	38.5	54.4	42.4	42.4	57.4	44.3
ML	44.4	54.9	50.3	47.4	61.0	52.8
SML-IS	**45.7**	**55.1**	**52.2**	**48.7**	**59.0**	**55.6**
	YOLOv7-seg			YOLOv8l-seg		
	mAP@[.5, .95]	mAP@.5	mAP@.75	mAP@[.5, .95]	mAP@.5	mAP@.75
S	26.3	32.6	27.4	27.8	34.2	30.3
M	**40.6**	**55.0**	**42.9**	**43.3**	**58.9**	**45.6**
L	34.5	47.7	35.8	38.3	49.8	39.9
SM	42.2	55.1	48.5	44.5	58.8	50.1
SL	40.1	54.1	42.5	43.3	58.5	45.9
ML	43.6	55.4	49.1	46.1	59.2	51.9
SML-IS	**45.2**	**56.1**	**51.0**	**48.5**	**59.4**	**54.4**

Table 3. Performance comparison on M-labelled test set for models trained with varying annotation strategies. Significant values are in bold.

Performance on M-labelled test dataset

The evaluation findings on the M-labelled test set, as detailed in Table 3, highlight a notable outcome: the marked enhancement in segmentation accuracy achieved through SML-IS labelling when compared to conventional (M) labelling. For example, the Mask RCNN R101-FPN model trained on the SML-IS-annotated dataset demonstrated a 6% increase in mAP@[.5, .95], a 1% gain in mAP@.5, and an 11% rise in mAP@.75 relative to the same model trained with M-labelled data. These results demonstrate the effectiveness of SML-IS as a data-centric strategy for improving segmentation performance without requiring additional image data. The performance improvements were consistently observed across all models, with Mask RCNN architectures showing slightly greater benefits compared to YOLO-based models (approximately 6% versus 4–5% in mAP@[.5, .95], respectively). This stronger performance in Mask RCNN models may be attributed to their anchor box-driven architecture, which appears to benefit more substantially from the enriched contextual information offered by multi-scale annotations. On the other hand, YOLO's efficient anchor implementation appears comparatively limited in fully utilizing the added semantic diversity introduced through SML-IS labelling (refer to "SML-IS: principles and practical integration").

The impact of SML-IS labelling was most evident in the mAP@.75 metric, where Mask RCNN models exhibited improvements of 10–11%, and YOLO-based models achieved gains of 8–9%. In contrast, the increase observed in mAP@.5 remained relatively modest at 1–2% across all models. This divergence is attributed to

the lower IoU threshold of mAP@.5, which imposes less stringent mask alignment requirements and therefore benefits less from the richer supervision provided by SML-IS. These results highlight the capacity of SML-IS to improve segmentation performance particularly under more demanding evaluation conditions, emphasizing its utility in scenarios that require precise mask delineation.

A closer inspection of the results shows that training on the S-labelled dataset leads to significantly lower performance than the L-labelled version, as small-scale polygon annotations lack the semantic richness necessary for accurate instance segmentation. As seen in Table 3, the Mask R-CNN R101-FPN model drops by approximately 12% in mAP for S, while the decline for L is around 5%. This trend is consistent across mAP@.5 and mAP@.75 for all models. Among the combined-scale variants, ML yields higher accuracy than SM, further underscoring the contribution of large-scale masks (e.g., Mask R-CNN X101-FPN improves by 5% with ML, compared to 3% with SM). Interestingly, the SL-labelled dataset performs almost identically to the M-labelled one: Mask R-CNN X101-FPN shows a 0.02% increase, Mask R-CNN R101-FPN and YOLOv7-seg show a slight 0.05% decrease, while YOLOv8l-seg records identical results. These observations confirm the robustness of M-labelling as a baseline while showing that small-scale masks alone lack sufficient semantic information to outperform conventional medium-scale annotations.

Another key insight from Table 3 is that SML-IS labelling yields only modest gains in the mAP@.5 metric (limited to 1–2% across all models), whereas more substantial improvements are observed in mAP@[.5, .95] and mAP@.75. This modest increase is attributed to the relatively lenient IoU threshold of 0.5, which reduces the impact of refined annotation strategies such as SML-IS. These findings indicate that the advantages of SML-IS become more evident under stricter IoU thresholds.

Extended accuracy metrics: object size and class-level evaluation

To gain further insights into the effectiveness of the dataset, we examined the extended accuracy metrics summarized in Table 4, based on evaluations conducted with the M-labelled test set. These metrics include mAP-S, mAP-M, and mAP-L, which correspond to performance across object sizes: small $(under32 \times 32pixels)$, medium $(32 \times 32to96 \times 96pixels)$, and large (over 96×96 pixels). Additionally, class-specific scores (mAP-Atlas, mAP-Handle, and mAP-Spot) offer a detailed perspective on per-class performance. The relatively uniform mAP@[.5, .95] values among classes suggest a balanced dataset in terms of class representation. Results indicate that segmentation models consistently perform better on medium-sized objects, while performance on small-scale instances is substantially lower than on large ones. This pattern reaffirms the limitations of the S-labelled dataset, previously noted, and reinforces the need for multi-scale and semantically rich annotation strategies to improve segmentation across diverse object sizes.

In addition to these findings, the most striking observation in Table 4 is the performance behavior of YOLOv8l-seg, particularly its unexpectedly high mAP-S score and comparatively lower mAP-L score relative to other models. As detailed in "Models without anchor boxes", this can be attributed to the anchor-free nature of YOLOv8l-seg, which detects object centers directly rather than relying on anchor box overlap. This architectural characteristic improves recall for small or densely packed objects, such as those that are partially occluded or clustered. However, it may lead to diminished precision when segmenting larger objects, contributing to the observed drop in mAP-L performance. It is also important to clarify that YOLOv8l-seg achieving the highest overall mAP does not imply that anchor-free models are inherently superior. Recent studies[42,43] demonstrate that Mask R-CNN (when equipped with more advanced backbones beyond R101-FPN or X101-FPN) continues to outperform in producing top-tier mAP scores, maintaining its position at the forefront of instance segmentation benchmarks.

Dataset	**Mask R-CNN R101-FPN**
	mAP/mAP@.5/mAP@.75/mAP-S/mAP-M/mAP-L/mAP-Atlas/mAP-Handle/mAP-Spot
M	38.9 / 53.9 / 41.6 / 13.5 / 77.8 / 42.8 / 38.0 / 40.3 / 38.9
SML-IS	45.7 / 55.1 / 52.2 / 17.6 / 81.6 / 46.6 / 44.5 / 47.1 / 43.9
	Mask R-CNN X101-FPN
	mAP/mAP@.5/mAP@.75/mAP-S/mAP-M/mAP-L/mAP-Atlas/mAP-Handle/mAP-Spot
M	42.2 / 57.5 / 44.3 / 15.4 / 80.0 / 44.3 / 43.1 / 41.9 / 41.1
SML-IS	48.7 / 59.0 / 55.6 / 18.0 / 83.2 / 47.6 / 49.7 / 47.5 / 48.5
	YOLOv7-seg
	mAP/mAP@.5/mAP@.75/mAP-S/mAP-M/mAP-L/mAP-Atlas/mAP-Handle/mAP-Spot
M	40.6 / 55.0 / 42.9 / 14.5 / 79.2 / 44.1 / 41.1 / 40.2 / 41.1
SML-IS	45.2 / 56.1 / 51.0 / 17.1 / 82.6 / 46.7 / 46.1 / 45.6 / 44.3
	YOLOv8l-seg
	mAP/mAP@.5/mAP@.75/mAP-S/mAP-M/mAP-L/mAP-Atlas/mAP-Handle/mAP-Spot
M	43.3 / 58.9 / 45.6 / 26.9 / 81.1 / 39.8 / 43.3 / 43.7 / 44.0
SML-IS	48.5 / 59.4 / 54.4 / 29.7 / 83.9 / 40.9 / 48.9 / 48.2 / 50.1

Table 4. Comparative extended performance metrics for conventional (M) and multi-scale (SML-IS) annotations.

Performance on SML-IS-labelled test dataset

Tables 3 and 4 present the primary evaluation results of this study, obtained on the conventional M-labelled test set, which follows the standard evaluation protocol and reflects practical deployment scenarios where only a single ground-truth annotation is available. To complement these primary results and provide additional insight into the behavior of the proposed annotation strategy under a more challenging evaluation setting, further analyses were also conducted using the SML-IS-labelled test set (in Table 5).

As shown in Table 5, testing on the SML-IS-labelled test set consistently results in lower mAP scores for both M-trained and SML-IS-trained models compared to their performance on the M-labelled test set. For example, YOLOv7-seg trained with M-labelled data achieves an mAP of 40.6% on the M test set but drops to 32.6% when evaluated on the SML-IS-labelled test set. Similarly, when trained on SML-IS-labelled data, its mAP declines from 45.2% to 41.3%. This performance drop highlights the increased complexity of the SML-IS-labelled test set, which includes multi-scale polygon annotations (S, M, and L) for each instance, thereby posing a more challenging evaluation scenario. Notably, the reduction in accuracy is more pronounced for M-trained models (an 8.0% drop for YOLOv7-seg) than for SML-IS-trained models (a 3.9% drop), emphasizing the enhanced generalization capacity of models trained with multi-scale annotations.

Although achieving higher performance on the SML-IS-labelled test set would further substantiate the effectiveness of SML-IS, its demonstrated superiority on the M-labelled test set (prepared using standard labelling practices) already serves as a robust benchmark. This confirms that SML-IS delivers substantial performance gains over conventional annotations, reinforcing its practical value and potential for real-world instance segmentation tasks. These improvements are attributable to the complementary nature of multi-scale annotations: small masks emphasize fine-grained object details, large masks capture contextual surroundings, and medium masks provide a balanced representation. Their integration allows the model to jointly learn object-specific details and contextual cues, showing that the observed accuracy gains are driven by genuine semantic diversity rather than marginal statistical effects.

Medical imaging application: Kvasir-SEG dataset

To further assess the effectiveness of the proposed method within a medical context, the Kvasir-SEG dataset[11] was utilized as a benchmark for colorectal polyp segmentation and is publicly available at https://datasets.simula.no/kvasir-seg/. The Kvasir-SEG dataset was chosen for three key reasons. First, it represents a challenging medical imaging domain where accurate segmentation is vital for early diagnosis and clinical decision-making. Second, medical datasets are typically limited in scale, making the SML-IS-labelling method particularly advantageous because it improves supervisory quality without expanding the dataset size. Third, the small and morphologically diverse structure of colorectal polyps provides an excellent benchmark for validating the method's ability to capture subtle spatial variations that single-scale annotations often overlook.

The dataset contains 1000 endoscopic images along with their corresponding masks, illustrating colorectal polyps captured under realistic clinical conditions. The original binary mask annotations were first converted into contour-based polygon annotations and stored in a COCO-compatible JSON format to enable their use with instance segmentation models while ensuring alignment with the experimental framework described earlier. Manual annotations (S and L) were then generated according to the SML-IS labelling protocol. A representative example of a conventional M-labelled image is shown in Fig. 6a, while Fig. 6b–d illustrate the corresponding Small (S), Medium (M), and Large (L) scale masks produced using the proposed approach. A deliberate blur effect was applied to Fig. 6b–d to emphasize that the information contained within the polygon mask annotations alone is sufficient for class identification, as similarly demonstrated in Fig. 4. This visualization underscores how multi-scale annotations enhance spatial precision and semantic richness in medical imagery.

For experimental evaluation, the dataset was divided into 800 training and 200 testing samples, following the same 4:1 ratio used for the custom dataset. The conventional M-labelled dataset was directly derived from the original masks, whereas the SML-IS-labelled dataset was generated using the same scaling policy described earlier, resulting in three times more annotations per image across S, M, and L scales.

The training configuration for the Kvasir-SEG experiments followed the same setup described in "Details of training process". The training convergence behavior is illustrated in Fig. 7. The results from the M-labelled test set are reported in Table 6, while those corresponding to the SML-IS-labelled test set are summarized in Table 7 (presented as a complementary analysis, whereas the M-labelled test set in Table 6 serves as the primary evaluation following the standard evaluation protocol).

Dataset	Mask R-CNN R101-FPN			Mask R-CNN X101-FPN		
	mAP@[.5, .95]	**mAP@.5**	**mAP@.75**	**mAP@[.5, .95]**	**mAP@.5**	**mAP@.75**
M	29.9	52.8	23.4	31.7	56.6	25.0
SML-IS	40.4	54.3	37.1	43.6	58.7	40.3
	YOLOv7-seg			YOLOv8l-seg		
	mAP@[.5, .95]	mAP@.5	mAP@.75	mAP@[.5, .95]	mAP@.5	mAP@.75
M	32.6	54.4	26.5	35.3	57.0	29.5
SML-IS	41.3	55.5	38.1	44.7	58.7	41.6

Table 5. Performance comparison on SML-IS-labelled test set for models trained with varying annotation strategies.

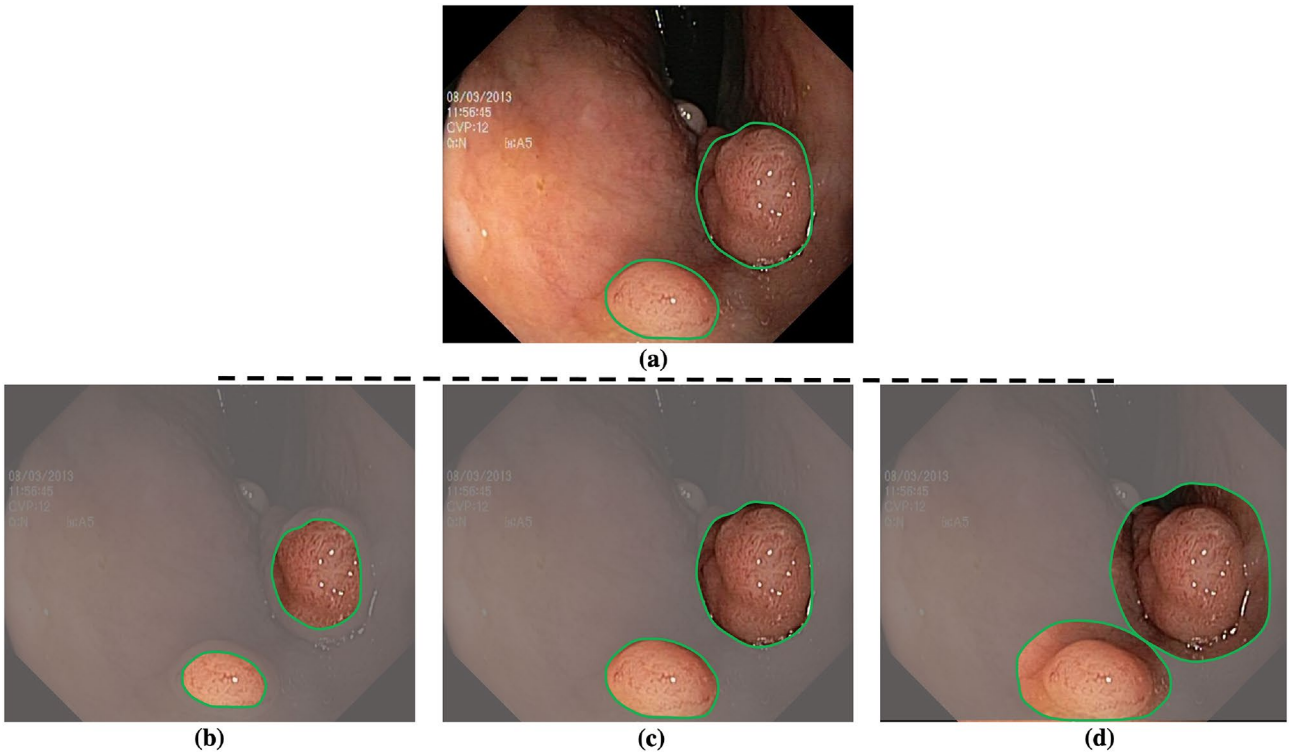


Fig. 6. Application of SML-IS labelling method on a sample from the Kvasir-SEG dataset.

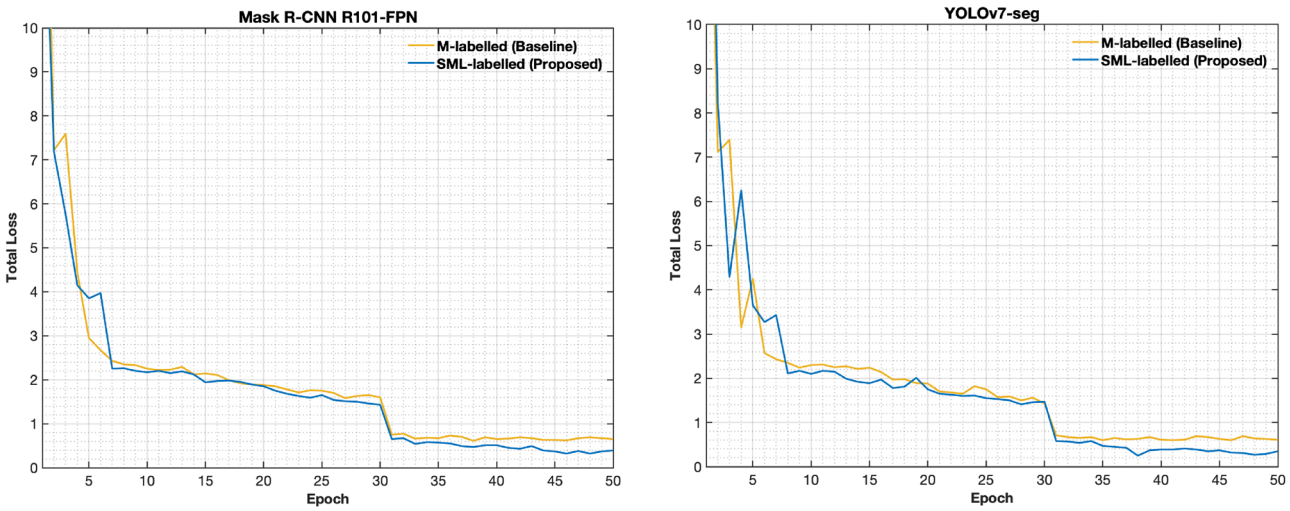


Fig. 7. Training-loss convergence patterns on the Kvasir-SEG dataset for mask R-CNN and YOLO models using M- and SML-IS-labelled training data.

As shown in Fig. 7, both Mask R-CNN R101-FPN and YOLOv7-seg models exhibited smooth and stable training-loss convergence throughout the learning process on the Kvasir-SEG dataset. Compared with the custom dataset, the overall loss curves were flatter and more stable due to the homogeneous texture and reduced semantic complexity of medical images. For the Mask R-CNN R101-FPN model, the training loss gradually decreased to approximately 0.65 with M-labelled data, whereas the SML-IS-labelled configuration achieved faster convergence, dropping to around 0.39. This faster and smoother decline highlights the improved supervision efficiency provided by the SML-IS annotation. The YOLOv7-seg model demonstrated a similar pattern, with final loss values of about 0.61 for M-labelled and 0.35 for SML-IS-labelled datasets. These results indicate that the proposed labelling strategy accelerates convergence and enhances learning stability once models reach their plateau phase, emphasizing its effectiveness in medical-image segmentation tasks.

Dataset	Mask R-CNN R101-FPN				Mask R-CNN X101-FPN			
	mAP@[.5,.95]	mAP@.5	mAP@.75	mDice	mAP@[.5,.95]	mAP@.5	mAP@.75	mDice
M	65.6	82.8	67.7	80.3	68.1	85.8	69.5	81.8
SML	72.4	83.9	76.7	84.2	75.7	86.3	79.3	86.9
	YOLOv7-seg				YOLOv8l-seg			
	mAP@[.5,.95]	mAP@.5	mAP@.75	mDice	mAP@[.5,.95]	mAP@.5	mAP@.75	mDice
M	67.5	82.3	68.9	81.1	69.1	86.5	69.5	82.0
SML	73.7	83.8	75.8	86.5	75.7	87.9	77.9	87.6

Table 6. M-labelled test set evaluations of mask R-CNN and YOLO models trained with Kvasir-SEG dataset. The mean Dice coefficient (mDice) is expressed as a percentage (%).

Dataset	Mask R-CNN R101-FPN				Mask R-CNN X101-FPN			
	mAP@[.5,.95]	mAP@.5	mAP@.75	mDice	mAP@[.5,.95]	mAP@.5	mAP@.75	mDice
M	58.8	75.9	53.4	73.8	60.8	77.1	56.8	74.7
SML	69.3	76.9	67.5	81.8	72.3	79.1	69.7	82.6
	YOLOv7-seg				YOLOv8l-seg			
	mAP@[.5,.95]	mAP@.5	mAP@.75	mDice	mAP@[.5,.95]	mAP@.5	mAP@.75	mDice
M	59.6	76.5	54.7	74.5	61.1	78.3	57.3	75.9
SML	68.1	77.2	66.4	82.2	70.1	79.9	70.6	83.6

Table 7. SML-IS-labelled test set evaluations of mask R-CNN and YOLO models trained with Kvasir-SEG dataset. The mean Dice coefficient (mDice) is expressed as a percentage (%).

Tables 6 and 7 quantitatively confirm these observations. Consistent with the convergence behavior, models trained with SML-IS-labelled data outperformed those trained with conventional M-labelled data across all evaluated COCO metrics. In Table 6, the Mask R-CNN R101-FPN achieved a 7% improvement in mAP (mAP@[.5, .95] = 72.4 vs. 65.6) on the M-labelled test set, while comparable increases were observed for the remaining Mask R-CNN and YOLO-based architectures. This improvement is particularly significant considering the limited size of the Kvasir-SEG dataset, where data scarcity typically constrains model generalization. Likewise, Table 7 indicates that evaluations on the SML-IS-labelled test set produced performance trends consistent with those observed on the custom dataset, further demonstrating the robustness and generalizability of the proposed annotation strategy.

In addition to the standard COCO evaluation metrics, the mean Dice coefficient (mDice) was also reported to facilitate direct comparison with previous medical image segmentation studies based on the Kvasir-SEG dataset. Although mDice is more commonly employed in semantic segmentation literature, it was included here as a complementary evaluation metric. Consistent with the mAP results, the proposed SML-IS annotation strategy also achieved consistently higher mDice scores than the conventional M annotation strategy across all evaluated instance segmentation models, further confirming that the proposed approach improves segmentation accuracy while preserving its effectiveness under the standard instance segmentation evaluation protocol.

Critical reflections and future research

While the proposed SML-IS labelling method demonstrated promising outcomes, certain limitations persist that warrant detailed exploration in future research.

First, assessing the generalizability of SML-IS across diverse instance segmentation datasets is essential for defining its broader applicability. As detailed in "SML-IS: principles and practical integration", the method can be applied to existing datasets with polygon-based annotations. However, a significant limitation lies in the increased manual effort required for generating multi-scale polygon masks, especially in large-scale datasets (such as MS COCO[1] and Pascal VOC[21]). This process, which followed consistent standards throughout this study, proved to be more time-intensive than conventional single-scale annotation. Based on our annotation experience, preparing SML-IS annotations typically required approximately twice the annotation time of conventional single-scale labelling. This increase is considerably lower than a threefold increase because the three polygon masks are generated sequentially on the same object, allowing each annotation to guide the next rather than requiring three independent annotation processes. Consequently, automating the multi-scale annotation workflow and integrating it into semi-supervised or assisted labelling pipelines will be a key direction for future work to reduce annotation effort while preserving the benefits of the proposed strategy. Furthermore, SML-IS introduces additional computational overhead during model training due to the enlarged dataset. This higher demand in training time and GPU resources must be addressed through architectural optimizations or more efficient data handling techniques.

Second, an interesting anomaly observed in Table 3 is the increasing gap between mAP and mAP@.75 scores when using SML-IS. While the gap for M-labelled datasets is around 2%, this difference grows to approximately 6–7% for all models trained with SML-IS. For instance, YOLOv7-seg shows a 2.3% gap for M labelling and a

5.8% gap for SML-IS. Comparable trends are evident in Tables 6 and 7 for the Kvasir-SEG dataset, indicating that the observed gap between mAP and mAP@.75 remains consistent, thereby reinforcing the reliability of this pattern across datasets. Although this may be partially explained by the higher IoU threshold of mAP@.75, which enforces stricter spatial alignment, the consistent discrepancy across models suggests a more complex interaction between label structure and model sensitivity to fine-grained accuracy. Unpacking these dynamics will be critical for understanding how multi-scale labels affect learning across varying levels of localization precision.

Further insights can be drawn from comparisons between the M-labelled and SML-IS-labelled test sets. As seen in Table 3, models generally perform better on the M-labelled test set in terms of mAP@.75 than mAP, whereas in Table 5, this trend reverses slightly for the SML-IS-labelled test set. The lower mAP@.75 in the SML-IS-labelled test set could be related to increased annotation complexity, but the exact reasons remain unclear. Likewise, minor declines in mAP@.5 across test sets are observed. A similar but less pronounced trend can also be observed in the Kvasir-SEG results presented in Tables 6 and 7, providing further validation of this behavior across datasets. While these results do not present a practical limitation (given that real-world evaluation typically relies on standard M-labelled data), they highlight an important methodological question that future work should address to improve interpretability of results under varying label structures.

Another promising future research direction is to conduct a comprehensive comparative study between SML-IS and representative mask perturbation methods[44,45] under identical experimental settings, including the same datasets, model architectures, training protocols, and evaluation metrics, to enable a fair and systematic performance comparison.

Finally, model evaluations in this study were limited to pre-annotated test sets, either M-labelled or SML-IS-labelled. While necessary for standardized benchmarking, this approach may not fully reflect real-time deployment conditions. In real-world applications where ground truth labels are unavailable, models trained with SML-IS are expected to deliver more consistent and adaptive segmentation outputs. With appropriate confidence thresholds, such models could offer increased resilience to noise and variability in unseen environments. Therefore, future research should empirically validate the practical effectiveness of SML-IS in live segmentation scenarios, potentially confirming its superiority as a data-centric enhancement strategy for real-time applications.

Conclusion

This study presented the SML-IS labelling method, a multi-scale polygon annotation strategy designed to enrich training data and improve instance segmentation performance. Unlike conventional single-scale annotations, SML-IS provides three complementary masks for each object instance, thereby increasing semantic depth without requiring additional images. Empirical evaluations demonstrated that SML-IS consistently improves mAP across different segmentation architectures, confirming its model-agnostic effectiveness. A notable distinction is that SML-IS differs from data augmentation: while augmentation creates new samples by transforming raw images, SML-IS retains the original images and instead enhances the quality and diversity of labels. This characteristic makes SML-IS particularly advantageous in data-scarce domains such as medical imaging, agriculture, and autonomous systems, where high-quality annotations are critical. Overall, SML-IS represents a practical and scalable data-centric solution that contributes both to academic research and to real-world applications requiring robust instance segmentation.

Data availability

The custom dataset used in this study is publicly available at: https://www.kaggle.com/datasets/ademkosar/custom-instance-segmentation-dataset. The Kvasir-SEG dataset used in this study, originally introduced by Jha et al. [9], is publicly available for research purposes at: https://datasets.simula.no/kvasir-seg/.

Received: 5 March 2026; Accepted: 3 August 2026
Published online: 16 August 2026

References

1. Lin, T..-Y. et al. Microsoft coco: Common objects in context. In *Computer Vision - ECCV 2014*. 740–755 (Springer, 2014).
2. He, K., Gkioxari, G., Dollár, P. & Girshick, R.: Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2980–2988. https://doi.org/10.1109/ICCV.2017.322 (2017).
3. Sharma, R., Saqib, M., Lin, C.T. & Blumenstein, M. A survey on object instance segmentation. *SN Comput. Sci.* **3**(6). https://doi.org/10.1007/s42979-022-01407-3 (2022).
4. Bai, M.& Urtasun, R. Deep watershed transform for instance segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2858–2866. https://doi.org/10.1109/CVPR.2017.305 (2017).
5. Zhao, Z.-Q., Zheng, P., Xu, S.-T. & Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(11), 3212–3232. https://doi.org/10.1109/TNNLS.2018.287686 (2019).
6. Cordts, M. et al. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
7. Li, S., Zhang, Z., Li, B. & Li, C. Multiscale rotated bounding box-based deep learning method for detecting ship targets in remote sensing images. *Sensors* **18**(8). https://doi.org/10.3390/s18082702 (2018).
8. Xu, X. et al. Cares-UNet: Content-aware residual UNet for lesion segmentation of COVID-19 from chest CT images. *Med. Phys.* **48**(11), 7127–7140. https://doi.org/10.1002/mp.15231 (2021).
9. Tang, Z., Chen, B., Zeng, A., Liu, M. & Zhao, S. Progressive deep snake for instance boundary extraction in medical images. *Exp. Syst. Appl.* **249**, 123590. https://doi.org/10.1016/j.eswa.2024.123590 (2024).
10. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. https://arxiv.org/abs/1506.02640 (2016).
11. Jha, D. et al. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling* (eds Ro, Y. M. et al.) 451–462 (Springer, 2020).

12. 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371 (2023).
13. Gutierrez-Cabello, G.S., Talavera, E., Iglesias, G., Clavijo, M. & Jimenez, F. A novel method to generate auto-labeled datasets for 3D vehicle identification using a new contrast model. *Appl. Sci.* **13**(7). https://doi.org/10.3390/app13074334 (2023).
14. Lan, S. et al. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 3386–3396 (2021).
15. Wang, X., Girdhar, R., Yu, S.X. & Misra, I. Cut and learn for unsupervised object detection and instance segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3124–3134. https://doi.org/10.1109/CVPR52729.2023.00305 (2023).
16. Sorokin, A. & Forsyth, D. Utility data annotation with amazon mechanical turk. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 1–8. https://doi.org/10.1109/CVPRW.2008.4562953 (2008).
17. Gupta, A., Dollar, P. & Girshick, R. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019).
18. Li, J., Chen, Z., Ni, C. & Lei, P. Run length encoding based weld seam detection from point clouds of ship stiffened panel. *J. Ocean Eng. Sci.* https://doi.org/10.1016/j.joes.2024.07.002 (2024).
19. Liu, Y., Piramanayagam, S., Monteiro, S.T. & Saber, E. Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order CRFs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1561–1570. https://doi.org/10.1109/CVPRW.2017.200 (2017).
20. Sherrah, J. Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery. https://arxiv.org/abs/1606.02585 (2016).
21. Everingham, M., Gool, L., Williams, C. K., Winn, J. & Zisserman, A. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4 (2010).
22. Wang, C..-Y., Bochkovskiy, A. & Liao, H..-Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).
23. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. & Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020. Lecture Notes in Computer Science* (Vedaldi, A., Bischof, H. et al. eds.). Vol. 12346. 213–229. https://doi.org/10.1007/978-3-030-58452-8_13 (Springer, 2020).
24. Kirillov, A. et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4015–4026 (2023).
25. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**(1), 654. https://doi.org/10.1038/s41467-024-44824-z (2024).
26. Jocher, G. & Qiu, J. Ultralytics Team: Ultralytics Yolo Docs. Accessed 09 Mar 2025 (2025).
27. Zeng, H. et al. Foveated instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025).
28. Yan, B., Sundermeyer, M., Tan, D.J., Lu, H. & Tombari, F. Towards real-time open-vocabulary video instance segmentation. arXiv preprint arXiv:2412.04434 (2024).
29. Jodogne-Del Litto, K. Centerdisks: Real-time instance segmentation with disk covering. arXiv preprint arXiv:2403.03296 (2024).
30. Chen, Z., Fang, H., Wang, W. & Zhang, Y. Scale-aware automatic augmentation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9563–9572. https://doi.org/10.1109/CVPR46437.2021.00944 (2021).
31. Fang, H., Wang, Z., Xu, T. & Lin, D. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1472–1481. https://doi.org/10.1109/WACV57674.2024.00153 (2024).
32. Feng, X., Jiang, Y., Yang, X., Du, M. & Li, X. Computer vision algorithms and hardware implementations: A survey. *Integration* **69**, 309–320. https://doi.org/10.1016/j.vlsi.2019.07.005 (2019).
33. Boston Dynamics: Spot Robot. https://bostondynamics.com/products/spot/. Accessed 13 Jan 2025 (2025).
34. Ren, S., He, K., Girshick, R. & Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Vol. 1. NIPS'15. 91–99. (MIT Press, 2015).
35. Jin, X. & Han, J. *K-Means Clustering* (eds. Sammut, C. & Webb, G. I.). 563–564 (Springer, 2010).
36. Lee, Y. & Park, J. Centermask: Real-time anchor-free instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13903–13912. https://doi.org/10.1109/CVPR42600.2020.01392 (2020).
37. Jocher, G., Chaurasia, A. & Qiu, J. *Ultralytics yolov8* (2023).
38. Boston Dynamics: Atlas Robot. https://bostondynamics.com/atlas/. Accessed 13 Jan 2025 (2025).
39. Boston Dynamic: Handle Robot. https://bostondynamics.com/legacy/. Accessed 13 Jan 2025 (2025).
40. Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y. & Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2. Accessed 13 Jan 2025 (2019).
41. Ultralytics: Ultralytics Official Website. Online. https://ultralytics.com. Accessed 13 Jan 2025 (2025).
42. Zong, Z., Song, G. & Liu, Y. Detrs with collaborative hybrid assignments training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 6725–6735. https://doi.org/10.1109/ICCV51070.2023.00621 (2023).
43. Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J. & Ling, H. Cbnet: A composite backbone network architecture for object detection. *IEEE Trans. Image Process.* **31**, 6893–6906. https://doi.org/10.1109/TIP.2022.3216771 (2022)
44. Karimi, D., Dou, H., Warfield, S.K. & Gholipour, A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med. Image Anal.* **65**, 101759. https://doi.org/10.1016/j.media.2020.101759 (2020) .
45. Heller, N., Dean, J. & Papanikolopoulos, N. Imperfect segmentation labels: How much do they matter? *CoRR*. https://doi.org/10.48550/arXiv.2111.13673 (2018).

Author contributions

A.K. conducted the experiments and prepared the dataset. O.E.K. and T.T. contributed to the methodology and analysis. E.E. validated the results the study. C.O. provided methodological guidance. All authors reviewed and approved the final manuscript.

Funding

This research was financially supported by the Recep Tayyip Erdogan University Development Foundation (Grant number: 02026002024140). We sincerely appreciate their support, which contributed significantly to the completion of this study.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to E.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026