mathematics

MDPI

Review

From Weakly Supervised Segmentation to Diagnostic Decisions: An Annotation-Economics Review for Medical Imaging

Xiu Shu [1], Xinming Zhang [2,*] and Di Yuan [3,*]

[1] School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China; shuxiu@gzhu.edu.cn
[2] School of Science, Harbin Institute of Technology, Shenzhen 518055, China
[3] Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China
* Correspondence: xinmingxueshu@hit.edu.cn (X.Z.); yuandi@xidian.edu.cn (D.Y.)

Abstract

Accurate medical image segmentation and the diagnostic decisions it supports are increasingly constrained by the cost of supervision: dense voxel-level annotation by expert clinicians remains slow, expensive, and difficult to scale, leaving many anatomies, pathologies, and modalities sparsely labeled. A large literature therefore relaxes the demand for full masks, yet existing surveys often catalog broad forms of limited supervision without directly comparing the practical trade-off between label granularity, annotation burden, and segmentation quality. This review focuses on weakly and coarsely labeled settings (image-level tags, points/clicks, bounding boxes, scribbles, and coarse masks, together with gaze or partial annotations as separate, protocol-dependent forms of supervision) and organizes them through an annotation-economics perspective that considers per-case annotation cost, required expertise, scalability, and the reported performance gap relative to full supervision. The evidence suggests that box and scribble supervision can provide favorable cost–performance trade-offs in several medical segmentation studies, although the reported metrics and datasets are heterogeneous and should not be interpreted as a unified quantitative frontier. We next examine the foundation-model turn, namely the Segment Anything Model and its medical derivatives, and argue that promptable models re-encode rather than eliminate human supervision, moving part of the effort from offline labeling to inference-time interaction. Finally, we extend the analysis beyond Dice to ask whether weakly produced masks support downstream clinical tasks: aggregate measurements may be relatively tolerant of moderate boundary error, whereas texture and high-order radiomic features can be substantially more sensitive. We close by identifying benchmark fragmentation, annotation variability, and the still-limited evidence connecting weak segmentation to clinical outcomes as open problems and propose a unified "weak label → task" benchmark for future evaluation.

Keywords: weakly supervised segmentation; annotation economics; medical imaging; foundation models; promptable segmentation; clinical decision support; uncertainty estimation

MSC: 68T07; 92C55

Check for updates

Academic Editor: Wei-Shih Du

Received: 29 September 2026
Revised: 18 October 2026
Accepted: 20 October 2026
Published: 22 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Medical image segmentation, which partitions scans into anatomically and pathologically meaningful regions at the pixel or voxel level, underpins computer-aided diagnosis, surgical planning, and longitudinal monitoring, and the algorithmic capacity to

produce accurate masks has matured rapidly over the past decade. From the convolutional U-Net [1] and its 3D [2] and attention-augmented [3] descendants, through nested designs such as UNet++ [4], to self-configuring pipelines like nnU-Net [5] and Transformer-based encoders such as TransUNet [6], Swin UNETR [7], and nnFormer [8], task-specific models now routinely surpass once-formidable benchmarks [5,7]. What has not kept pace is annotation: building a fully supervised training set for volumetric segmentation still demands that expert clinicians delineate every structure of interest slice by slice, a process that can take tens of minutes to hours per case [9], and that is expensive, hard to scale, and vulnerable to inter-rater variability. As a result, large densely annotated datasets exist for only a handful of well-resourced tasks, while the long tail of anatomical structures, rare pathologies, and acquisition protocols remains effectively unlabeled. A major practical bottleneck for clinical deployment is therefore the cost of supervision, alongside model architecture, data quality, and validation.

This asymmetry between capable models and scarce labels has motivated a substantial literature that relaxes the need for full pixel-level supervision. Among these directions, weakly supervised segmentation, which learns from annotations far cheaper than dense masks such as image-level tags, points/clicks, bounding boxes, scribbles, or coarse masks, is particularly relevant because it reduces direct annotation effort rather than relying only on unlabeled data [10–13]. We define weak supervision narrowly for this review: the training signal is less expensive to provide than a full mask yet still explicitly authored by a human, as opposed to being inferred solely from data. Weak labels form a spectrum of decreasing granularity. Image-level labels indicate only class presence and provide the weakest spatial signal, typically recovered via class activation [14] or seed-expand-constrain pipelines [15,16]. Bounding boxes [12,13,17] enclose the target and are often a by-product of clinical reading. Scribbles [18–21] are intuitive freehand strokes that can require substantially less effort than dense masks and have become a practical weak-label option in medicine, so much so that the scribble family now spans dozens of specialized designs, which we systematize in Section 3.3. Points and clicks [22–25] support interactive refinement, while gaze trajectories and coarse masks [26,27] provide distinct forms of supervision with different spatial characteristics. Gaze provides indirect, temporally structured cues, whereas coarse masks offer explicit but incomplete spatial annotations. Both can approach dense supervision in particular settings, but neither is equivalent to a full mask. Two neighboring paradigms are not treated as weak-label types in this review: semi-supervised learning, which leverages unlabeled data through consistency regularization such as mean-teacher [28] and its uncertainty-aware variants [29], and self-supervised pre-training on pretext tasks [30–34]. These paradigms do not, by themselves, lower the cost of human annotation, and they are therefore not part of the weak-label spectrum of Section 2. They are, however, included as structural enhancement mechanisms when they are combined with weak labels, because many weakly supervised methods borrow consistency regularization, self-ensembling, or self-supervised pre-training to stabilize pseudo-label learning. This inclusion criterion is made explicit so that the scope of this review remains consistent across Sections 3 and 5.

Several surveys already chart the broad territory of limited supervision in medical imaging [35–37] and the wider gap between weak supervision and dense prediction [38]. Valuable as they are, these works trend toward comprehensive catalogs that span image-level, box, scribble, semi-, and self-supervised methods alike [35–38], which makes it hard for a reader to answer the practically urgent question: given a fixed annotation budget, which weak label type and which method should I choose? This review focuses on weakly and coarsely labeled settings and organizes them through an annotation-economics perspective, comparing methods along annotation burden, required expertise, scalability, variability,

and reported performance relative to full supervision. This framing turns the field from a list of methods into a set of cost–quality trade-offs. We also extend the analysis beyond segmentation accuracy to ask whether weakly produced masks are sufficient to support clinical decision making, a question largely absent from method-oriented reviews and one that we treat as the ultimate test of weak supervision's value.

Against this backdrop, this review makes four contributions. First, it provides a granularity-centered taxonomy that unifies weak labels, including image-level, bounding box, scribble, point/click, and gaze/coarse categories, and maps the dominant technical mechanisms, namely pseudo-labeling, constraint losses, uncertainty weighting, and foundation-model prompting, onto each label type. This clarifies which mechanisms are label-type specific and which are cross-cutting. Second, it introduces an annotation-economics lens that combines annotation time, required expertise, scalability, variability, and reported performance loss relative to full supervision. Rather than treating heterogeneous studies as a single numerical frontier, the review uses the reported values to identify recurring cost–performance patterns and their limitations [13,17,20,21,39,40]. Third, it builds a bridge from weak segmentation to downstream clinical tasks, asking when uncertainty and segmentation error remain acceptable for quantitative measurements or diagnostic modeling [41–44], and it distinguishes aggregate measurements that may tolerate moderate boundary error from texture- and high-order-feature analyses that can be substantially more sensitive. Fourth, it examines the recent rise of promptable foundation models, including Segment Anything (SAM) [45] and its medical derivatives [46–49], within the same annotation-economics framework, treating prompts as an inference-time form of human guidance rather than as a replacement for human supervision.

Beyond the annotation-economics framing, this review differs from earlier limited-supervision surveys [35–38] in three respects: it analyzes the pair (label type; task) rather than the label type alone; it separates medical evidence from natural-image methodological context through an explicit evidence hierarchy (Section 3); and it evaluates weak labels by downstream diagnostic utility rather than by Dice (Section 5).

The scope of this review is deliberately bounded. This article is a narrative review rather than a systematic review or meta-analysis. The synthesis is based on the primary studies and review papers cited throughout Sections 2–6. Because annotation-time measurements, datasets, supervision protocols, and evaluation metrics differ substantially across studies, reported numerical values are retained as study-specific evidence rather than pooled into a common effect size. We also distinguish evidence about segmentation quality from evidence about downstream clinical utility; a high Dice score alone is not treated as evidence of diagnostic equivalence.

The remainder of this review proceeds as follows. Section 2 establishes the weak-supervision spectrum and the annotation-economics coordinate system used throughout, including a quantitative cost–accuracy table. Section 3 is the methodological core: we review techniques grouped by weak-label type (image-level, bounding box, scribble, and point/click) and then distill the cross-cutting mechanisms that transcend label type. Section 4 examines the foundation-model turn, with ScribblePrompt [49] as the canonical "foundation model as weak supervisor", and asks whether foundation models dissolve the weak-supervision problem or merely relocate it. Section 5 moves beyond segmentation to assess whether weak masks support diagnostic decisions, treating uncertainty as a usability signal. Section 6 discusses failure modes, benchmark gaps, and future directions and concludes this review.

2. Weak Supervision Spectrum and Annotation Economics

Weak supervision for medical image segmentation is best understood as a spectrum of human signals with different levels of spatial information. At the coarse end, image-level labels indicate only class presence [14–16,50,51]. Points or clicks provide sparse spatial cues, bounding boxes delimit the target region, and scribbles provide partial information about its shape and extent [12,20,25,52,53]. Coarse masks provide more explicit but incomplete spatial annotations, whereas gaze trajectories offer indirect and temporally structured cues [26,27]. Neither should be treated as equivalent to full pixel- or voxel-level supervision.

Figure 1 orders explicit spatial labels by increasing spatial information density: image-level → point/click → bounding box → scribble → coarse mask → full mask. Gaze is shown separately because it does not represent an explicit spatial region and therefore cannot be placed unambiguously on the same density axis. This ordering reflects spatial information density rather than annotation time or semantic informativeness.

The practical importance of this distinction is that annotation cost does not decrease monotonically with spatial information density. For example, a single point can be cheaper to acquire than a scribble because clicking is faster than tracing a stroke, although the point conveys less spatial information. Conversely, a coarse mask may require substantial manual effort despite providing only partial spatial coverage. The effort required to recover a full segmentation mask does not necessarily decrease in proportion to annotation effort.

The relative cost and usefulness of each label type depend on anatomy, imaging modality, annotator expertise, and the annotation protocol, as discussed in Section 2.3. Figure 1 should therefore be interpreted as a conceptual spatial-information spectrum rather than a universal ranking of annotation costs. Figure 1 summarizes this spectrum and illustrates representative medical examples.

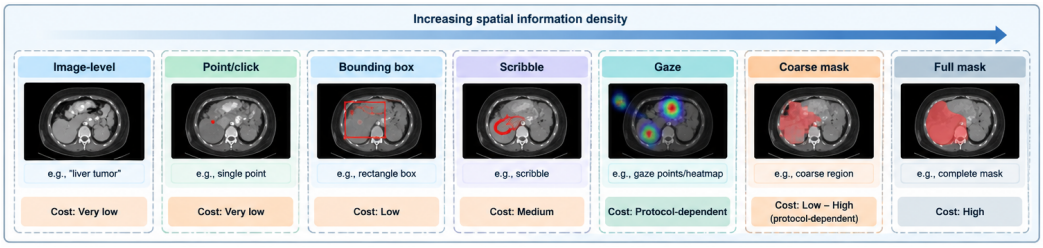


Figure 1. Weak-label spectrum ordered by spatial information density. Gaze is shown separately as an indirect, protocol-dependent form of supervision. Annotation costs are indicative ordinal levels, with point/click, bounding box, scribble, and full mask corresponding to very low, low, low to medium, and high costs, respectively. Coarse-mask and gaze costs depend on the acquisition protocol. Gaze is drawn below the annotation-density arrow to indicate that it is not placed on the same density axis; it is a separate, protocol-dependent category rather than a point on the spectrum. This is an original schematic created by the authors; no third-party material is reproduced.

Two neighboring paradigms sit deliberately outside this spectrum. Semi-supervised learning [28,29] and self-supervised pre-training [31–34,54] reduce the need for labeled data but not the cost of the labels that are used; they are therefore treated as enabling mechanisms (Section 3) rather than as weak-label types. This boundary distinguishes the present review from broader surveys that intermix all forms of limited supervision [35–38].

2.1. Annotation Cost Model

To compare weak labels on a common conceptual basis, we use four cost dimensions: annotation time, required expertise, scalability, and annotation variability. Annotation time captures how long a human takes to provide the signal; expertise captures whether the task requires specialist knowledge; scalability captures whether the signal can be produced across

many cases and modalities; and annotation variability captures how strongly differences in box tightness, scribble geometry, click placement, gaze trajectory, or partial-slice selection can affect the resulting supervision. These dimensions are complementary rather than interchangeable. A useful conceptual total-cost expression is as follows:

$$C_{\text{total}} = C_{\text{annotation}} + C_{\text{training}} + C_{\text{inference}} + C_{\text{correction}}, \quad (1)$$

where the present review quantifies human annotation burden more directly than computational cost. Among the four terms, the annotation term is the one that most directly distinguishes weak labels from full supervision. The remaining three terms are retained in the conceptual model but are not used to define the quantitative efficiency index below, because training, inference, and correction costs are reported far less consistently across the cited studies. They are discussed qualitatively in Sections 3.5 and 6. We therefore make it concrete as follows:

$$C_{\text{annotation}} = N_{\text{case}} \times t_{\text{per case}} \times c_{\text{expert}}, \quad (2)$$

where N_{case} is the number of annotated cases, $t_{\text{per case}}$ is the average annotation time per case, and c_{expert} is the unit cost of expert time. This expression makes explicit that the annotation burden depends jointly on the number of cases, the per-case time, and the required expertise, and it allows the weak-label types of Section 2 to be compared on a common basis. The unit cost c_{expert} is setting-dependent rather than universal; it varies with clinician salary, geography, expertise level, and annotation platform, as discussed below.

Inter-rater variability as an economic cost. Inter-rater variability is not only an accuracy concern but also an economic cost. When annotators disagree on a boundary, a box tightness, a scribble path, or a click placement, additional expert time is required for adjudication, consensus formation, re-annotation, or quality control, and the resulting labels may still require correction during deployment. In the total-cost expression of Equation (1), this effort appears in the correction term $C_{\text{correction}}$ and in the hidden cost of quality assurance. It also interacts with the other cost dimensions: variability increases the number of cases that must be reviewed, reduces the effective scalability of a labeling protocol, and raises the required expertise when adjudication itself needs a senior clinician. For weak labels, this effect can be substantial, because a loose box, a scribble that leaks into a neighboring structure, or a click placed near an ambiguous boundary may be internally consistent for one annotator but not for another. We therefore treat annotation variability as a first-class cost dimension alongside annotation time, required expertise, and scalability, rather than as a purely statistical property of the labels.

How per-case annotation cost is estimated. Direct measurements of per-case annotation time are available only in a minority of the cited studies. Where a study reports annotation time directly, we use the reported value. For example, Bearman et al. [10] report point annotation on the order of seconds per image versus minutes for a full mask; ScribblePrompt [49] reports a 28% reduction in annotation time relative to SAM in a 16-expert user study; and MedSAM [46] reports a greater than 82% reduction in expert 3D-tumor annotation time. For studies that do not report direct timing, we use two kinds of proxy. The first is a label-type proxy, such as the commonly cited ratio that bounding-box annotation requires roughly 1/15 of mask annotation time [17]. The second is an interaction-protocol proxy, such as seconds per click for point supervision or seconds to approximately one minute per structure for scribble supervision. These proxies are indicative rather than measured cross-study values. They are used only to compare label types on a common conceptual basis, not to assign a universal price to each label. Where a study reports a cost ratio but not an absolute time, we retain the ratio and explicitly mark the corresponding entry as estimated. This estimation strategy is also reflected in the ordinal cost scale and the evidence-type column of Table 1, which distinguish natural-image entries from medical entries and avoid treating heterogeneous cost values as directly comparable.

Table 1. Study-specific annotation economics of weak labels. Cost is expressed on an ordinal scale to avoid mixing qualitative descriptions, natural-image numbers, and medical estimates. Entries retain the metric and evidence type reported or estimated in the cited sources; values are not directly comparable across rows.

Weak Label	Ordinal Cost	Cost Detail	Evidence Type	Reported Accuracy Gap vs. Full Supervision	Example Dataset(s)	Example Method(s)
Image-level	Very low	routine reports [50]	Medical	mIoU gap ≈ 2% on histopathology [55]; CXR localization reported as AUC/IoBB, not Dice [11]	ChestX-ray8 [11]; LUAD-HistoSeg [55]	CAM [14] (natural origin), SEC [15] (natural origin), Cluster-Re-Supervision [51]
Point/extreme point	Very low	seconds per click; fewer clicks than scribbles [9,23]	Medical	Dice gap 1.0% (MSD-Spleen), 0.9% (MO-Spleen), 1.7% (MO-Liver), 4.4% (MO-Pancreas), 2.9% (MO-L-Kidney), 1.3% (MO-Gallbladder) [9]	Spleen, liver, pancreas, kidney, gallbladder [9]; placenta, brain tumor [23]	Roth et al. "extremes" [9]; interactive DeepIGeoS [23]
Bounding box	Low	≈1/15 of mask time [17]	Mixed (natural + medical)	mIoU gap ≈ 1.8 percentage points on PASCAL VOC [12] (≈62.0% under box supervision vs. ≈63.8% under full-mask supervision); ≈95% of full quality on PASCAL VOC [56]; Dice gap 7.5% (fetal brain) and 12.9% (fetal lung) [17]	Fetal MRI [17]; PASCAL VOC [12,13]	BoxSup [12], DeepCut [17], bounding-box tightness prior [13]
Scribble	Low to medium	seconds to approximately one minute per structure, estimated; not directly measured in the cited studies	Medical	Dice gap 1.4% on ACDC and −4.5% on MSCMRseg [20]; Dice gap 3.2% on ACDC and 6.3% on MSCMRseg [21]; Dice gap −0.6% on ACDC with 20% labeled data [40]	ACDC [57]; MSCMRseg [20,21]; multi-organ [40]	CycleMix [20], ScribFormer [21], SOUSA [40]
Gaze/coarse mask	Low to high (protocol-dependent)	gaze: 2.2 h vs. 18.7 h for 900 images [58]; partial: 20% annotated slices [27]	Medical	Dice gap 4.3% (polyp) and 2.9% (prostate) for gaze [26]; Dice gap −0.1% (ID) and −6.3% (OOD) for partial slices [27]	Polyp, prostate [26]; fetal MRI [27]	Multi-level gaze networks [26]; selective Dice loss [27]

Annotation costs are indicative and protocol-dependent, and the reported values are not directly comparable across studies. Gaze and coarse masks represent distinct forms of supervision: gaze provides indirect, temporally structured cues, whereas coarse masks provide explicit but incomplete spatial annotations. Negative performance gaps indicate better results than the corresponding full-supervision baseline, subject to verification of the metric definition used in each cited study. The gaze annotation-time figure is taken from [58].

Socio-technical factors in cost estimation. The unit cost c_{expert} in Equation (2) is not a universal constant. It depends on at least four factors. The first is clinician salary, which can differ by an order of magnitude between a senior radiologist and a trained research assistant, and which varies across health systems and institutions. The second is geography: the same nominal annotation time can translate into very different monetary costs across countries, regions, and funding models, so a cost estimate calibrated in one setting may not transfer to another. The third is expertise level: a task that a subspecialist can complete in minutes may take a general annotator substantially longer, and adjudication of ambiguous cases may require a senior clinician whose time is more expensive. The fourth is the annotation platform: a web-based interface with keyboard shortcuts, auto-completion, real-time quality checks, and structured prompts can reduce per-case time relative to a generic viewer, while a poorly designed interface can increase it. Because these factors are rarely reported in a comparable way, this review deliberately avoids assigning a universal monetary price to each label type. Instead, it uses relative cost ratios, study-specific evidence, and the ordinal cost scale of Table 1, so that the economic comparison remains meaningful across settings without implying a single global price.

For a practitioner, the quantitative and qualitative dimensions can be combined in a simple three-step procedure. First, fix the downstream error tolerance, since this constrains which label types are admissible at all. Second, estimate the annotation burden with Equation (2) and the study-specific efficiency with Equation (3), using measured times where available and the proxies of Table 1 otherwise. Third, apply expertise and variability as qualitative modifiers: a label that requires a subspecialist, or whose geometry varies strongly between annotators, should be discounted relative to a label of similar nominal cost that a general annotator can produce consistently. When the qualitative factors cannot be monetized reliably, the recommended practice is to report the comparison as a range, obtained by sensitivity analysis over the unit cost c_{expert} and the per-case time $t_{\text{per case}}$, rather than as a single number.

To express the central trade-off without implying a universal monetary price, we define a study-specific annotation efficiency as

$$\eta_{\text{ann}} = \frac{Q_{\text{weak}}/Q_{\text{full}}}{C_{\text{annotation}}/C_{\text{full}}}, \tag{3}$$

where Q denotes the task-relevant performance measure, and C denotes annotation burden. Values of η_{ann} are meaningful only within a study or matched protocol because both quantities depend on the task, metric, annotator, and interaction procedure.

Why the efficiency index focuses on annotation cost. The index in Equation (3) is an annotation-specific efficiency, not a total-cost efficiency. It focuses on annotation cost for two reasons. First, annotation is the term that most directly distinguishes weak labels from full supervision: training, inference, and correction costs can be similar across supervision types, whereas the human annotation burden is precisely what weak labels are designed to reduce. Second, the cited literature reports annotation time and annotation ratios far more consistently than it reports training time, inference latency, or correction effort, so a total-cost index would be under-determined for most studies. This focus should not be read as a claim that the other three terms are negligible. When the additional terms are available, the index can be extended to

$$\eta_{\text{total}} = \frac{Q_{\text{weak}}/Q_{\text{full}}}{\left(C_{\text{annotation}} + C_{\text{training}} + C_{\text{inference}} + C_{\text{correction}}\right)/C_{\text{full}}}. \tag{4}$$

In practice, most studies do not report enough information to estimate η_{total}, so this review uses η_{ann} as a transparent but partial proxy. Computational cost, which enters the denominator through C_{training} and $C_{\text{inference}}$, is discussed qualitatively in Sections 3.5 and 6,

where efficient architectures are treated as a secondary cost dimension rather than as direct evidence of better annotation economics.

When computational cost becomes decisive. Computational cost becomes the decisive term when two methods have similar annotation requirements. If two scribble-supervised methods require the same scribble effort per case, or if two box-supervised methods use the same bounding-box protocol, then the annotation term $C_{\text{annotation}}$ no longer separates them, and the comparison shifts to C_{training} and $C_{\text{inference}}$. In that setting, a method with a heavier backbone, a longer training schedule, or a higher inference latency may be less economical overall even if its annotation burden is identical. This is particularly relevant for weakly supervised and promptable pipelines, where the annotation cost has already been reduced and the remaining cost is increasingly dominated by computation. It is also why efficient architectures, including the Mamba-based and lightweight models discussed in Section 6, are treated as a secondary cost dimension in this review: they do not lower the annotation burden, but they lower the total cost at which a given weak-label strategy can be deployed. Conversely, when annotation requirements differ substantially, computational cost is usually a second-order consideration, because the human annotation term dominates the total cost.

To illustrate how η_{ann} can be estimated in practice, we compute proxy values for representative weak-label methods using the Dice ratios and annotation-time ratios reported in the cited studies. These values are illustrative rather than a unified benchmark, because the underlying studies differ in dataset, protocol, and annotator.

The quality ratios in Table 2 are taken from the most favorable reported subset for each method. DeepCut, for example, reports a Dice gap of 7.5% on fetal brain but 12.9% on fetal lung under the same box-supervision protocol [17]; the table uses the fetal brain value. Roth et al. reports per-organ gaps ranging from 0.9% to 4.4%, with the largest gap on the pancreas [9]; the table uses the best-organ value. The efficiency values should therefore be read as upper bounds rather than as representative averages.

Table 2. Illustrative best-case proxy estimates of annotation efficiency η_{ann} for representative weak-label methods. Values are derived from reported Dice ratios and annotation-time ratios and should not be interpreted as a unified benchmark. The annotation-time ratios follow the estimation strategy described in Section 2.1: direct measurements are used when available, and label-type or interaction-protocol proxies are used otherwise. The quality ratios use the most favorable reported subset for each method, so the resulting efficiency values should be read as upper bounds rather than averages. η_{ann} is an annotation-specific index; it does not include training, inference, or correction costs. See Section 2.1 for the rationale and for the extended total-cost form η_{total}.

Method (Label Type)	$Q_{\text{weak}}/Q_{\text{full}}$	$C_{\text{annotation}}/C_{\text{full}}$	η_{ann}	Evidence Type
DeepCut (box) [17]	≈0.925	≈0.067	≈13.8	Medical
CycleMix (scribble) [20]	≈0.986	≈0.10	≈9.9	Medical
Roth et al. (point) [9]	≈0.99	≈0.05	≈19.8	Medical

All three entries are derived from medical datasets. The CycleMix annotation-cost ratio is a label-type proxy rather than a measured value; the Roth et al. ratio is an interaction-protocol proxy derived from seconds-per-click assumptions. Natural-image studies used elsewhere in the review are marked as methodological context and are not pooled with medical entries.

Natural-image studies provide useful anchors: Bearman et al. [10] report point annotation on the order of seconds per image versus minutes for a full mask, while box-supervision studies report substantially shorter annotation times than dense labeling [13,17,56]. Medical studies show the same general ordering, but the absolute time depends on modality, anatomy, annotator, and interaction protocol [17,49]. The economic lens therefore does not assume a universal price for each label; instead, it asks which supervision strategy provides an acceptable task-specific error at a realistic total cost. Classical variational segmentation offers a complementary annotation-free perspective; the survey of Shu et al. [59]

reviews region-based, edge-based, and hybrid variational models for readers interested in segmentation without learned supervision.

2.2. Quantitative Economics: The Accuracy Gap

The key quantity for any weak label is the task-specific performance loss relative to an appropriate full-supervision reference. Because the primary literature reports different metrics—mAP for bounding boxes on PASCAL VOC [12], mIoU and FwIoU for image-level labels on histopathology [55], and Dice for scribbles, points, gaze, and partial annotations on medical datasets [9,20,21,26,27,40]—Table 1 retains the metric used by each source rather than forcing heterogeneous results into a single Dice scale. Within individual studies, box and scribble supervision often show favorable cost–performance trade-offs, but the magnitude of the trade-off remains dependent on the dataset, anatomy, modality, annotation protocol, and baseline.

Qualifying the favorable trade-off. The claim that box and scribble supervision provide favorable cost–performance trade-offs should be read as a study-specific pattern rather than a universal rule, because the reported evidence varies substantially along three axes. First, anatomy matters. DeepCut reports a Dice gap of 7.5% on fetal brain but 12.9% on fetal lung under the same box-supervision protocol [17], showing that the same weak-label type can behave very differently across structures of different size, contrast, and boundary complexity. Second, modality matters. Scribble results on ACDC cardiac cine MRI [20,21] do not automatically transfer to CT, ultrasound, or histopathology, where contrast, resolution, noise characteristics, and boundary definition differ; the reported scribble gap on MSCMRseg, for example, differs from that on ACDC within the same study [20,21]. Third, benchmark split matters. ACDC scribble numbers are split-sensitive, and values obtained under the standard 70/15/15 partition are not directly comparable with values obtained under a different partition or under the full 70-subject training set [21,60]. A further complication is that several of the strongest box-supervision results are reported on natural-image benchmarks such as PASCAL VOC [12,56], where the annotation protocol and target statistics differ from medical imaging. For these reasons, Table 1 is presented as an evidence map rather than a statistically pooled frontier, and the numerical gaps should be interpreted within each cited study rather than as a universal medical benchmark. The practical implication is that a favorable trade-off observed for one anatomy, modality, and split should be re-validated before it is assumed to hold for another.

The accuracy column reports the metric used by each source; it is not a unified Dice column, and values across rows are therefore not directly comparable. Negative gaps indicate that the weak label matched or exceeded full supervision in that study, including scribble on MSCMRseg [20], SOUSA on ACDC [40], and partial slices on FIESTA ID/OOD [27]. The scribble annotation-time entry is an approximate range and should not be treated as a measured cross-study value. The bounding-box row combines natural-image and medical evidence [12,13,17]; its numerical gaps should not be interpreted as a universal medical benchmark. Gaze and partial-slice entries are protocol-specific and should be interpreted only within the cited studies. Annotation-time entries are either directly measured in the cited study or estimated using the label-type and interaction-protocol proxies described in Section 2.1. They are indicative and should not be read as universal monetary or temporal costs. More generally, the reported gaps vary substantially with anatomy, modality, and benchmark split, so the table is intended as an evidence map rather than a ranking of weak-label types. The gaze and coarse-mask entries are grouped in the same row only because both sit at the fine end of the spectrum; they differ in acquisition mechanism and information density and should not be treated as interchangeable. Entries marked as natural-image evidence are included only to explain the origin of a mechanism and should not be cross-read as medical cost or performance

estimates. The ordinal cost scale is used precisely because monetary cost estimates depend on clinician salary, geography, expertise level, and annotation platform, and are therefore not comparable across settings.

2.3. Why an Economics Lens

The value of the economics lens is that it turns a fragmented literature into a task-oriented comparison. Prior reviews of limited supervision [35–38] are indispensable as catalogs, whereas the present review emphasizes the supervision burden associated with each label type. The framework should not be interpreted as a universal price list: if a reader has a fixed annotation budget, the appropriate comparison depends on the target anatomy, annotator expertise, interaction protocol, and acceptable downstream error. The lens also exposes a less obvious point: the cheapest label is not necessarily the most economical, because expertise, variability, correction effort, training cost, and deployment constraints can offset time savings. In particular, when two methods have comparable annotation requirements, computational cost can become the deciding factor, as discussed in Section 2.2. This framing also motivates the later bridge to diagnosis, where the relevant question is not simply how much Dice is lost, but whether the residual error changes the downstream measurement or decision.

The spectrum is task-conditional. The ordering in Figure 1 should be read as a label-type ordering under a generic task, not as a fixed position for each label in every clinical setting. The same weak-label type can occupy different positions on the economic spectrum depending on anatomy and modality. A bounding box around a large, high-contrast organ such as the liver may support a much smaller accuracy gap than a box around a small, low-contrast lesion. Similarly, scribble supervision on cardiac cine MRI [20,21] may behave differently from scribble supervision on CT, ultrasound, or histopathology, where contrast, resolution, and boundary definition differ. Point supervision may be highly economical for a well-circumscribed structure but less so for a diffuse or multi-focal pathology, where a single point provides little spatial guidance. The relevant unit of analysis is therefore the pair (label type; task), not the label type alone. This is why Section 5 argues that the appropriate weak-label budget should be determined by the required downstream tolerance rather than by a generic ranking of label types. It also follows that the spatial-density ordering of Figure 1 should not be read as an acquisition-time ordering or a semantic-informativeness ordering: the three dimensions are correlated but not identical, and a label that is cheap in spatial-density terms may still be expensive in acquisition time or limited in semantic informativeness for a specific task.

Rare-disease and common-organ segmentation differ systematically along these cost dimensions. For common organs, large public benchmarks, abundant cases, and cheaper annotator pools make full supervision feasible, so the advantage of weak labels is mainly one of cost. For rare diseases, the number of cases is small, expert annotators are scarcer and more expensive, inter-rater variability is larger, and a full-supervision reference set is often unavailable; the relative benefit of weak labels is therefore larger, but validation against a reliable reference is harder. In practice, the evidence burden for rare-disease weak supervision shifts toward external validation and uncertainty reporting rather than toward a lower Dice gap on a single dataset.

3. Methodological Landscape: Learning from Sparse Human Signals

This section organizes methods by the type of weak human signal they consume, namely image-level (Section 3.1), bounding box (Section 3.2), scribble (Section 3.3, the largest and most mature family), and point/click/interactive (Section 3.4), and then extracts the mechanisms that cut across these families in Section 3.5. Figure 2 provides a naviga-

tional map of this taxonomy. Each method is presented in a uniform format: mechanism, representative method, position on the annotation-economics axes (cost; accuracy gap), and limitation. This format keeps the section from becoming a method dictionary and supports this review's central argument that the supervision signal is a major practical bottleneck alongside architecture, data, and validation.

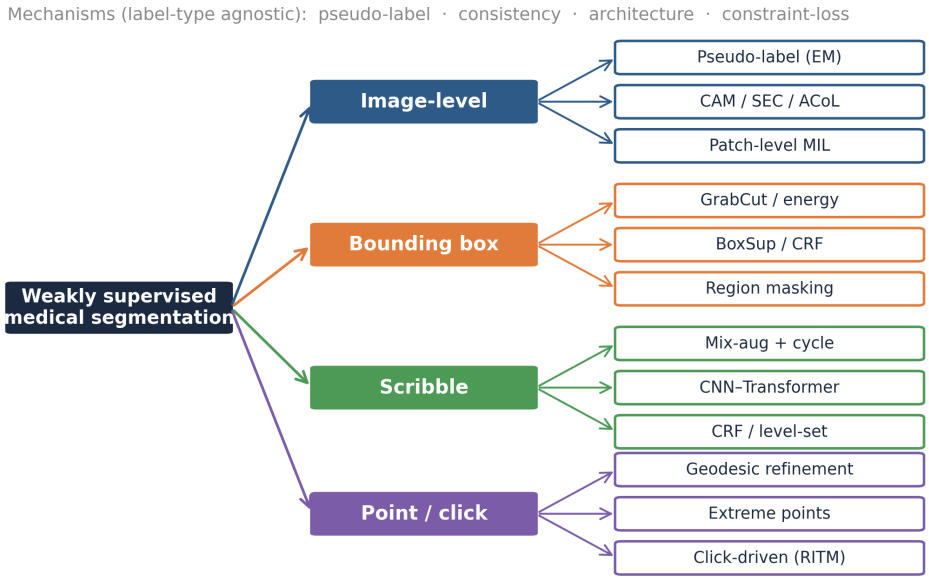


Figure 2. Method family tree for weakly supervised medical segmentation. The root splits by weak-label type (image-level/box/scribble/point), with each branch further splitting by mechanism (pseudo-label, consistency, architecture, constraint-loss).

Evidence hierarchy. Throughout this review, we distinguish between primary medical evidence and natural-image evidence used only as methodological context. Primary medical evidence refers to studies evaluated on medical or biomedical datasets, including ACDC, MSCMRseg, BraTS, ChestX-ray8, LUAD-HistoSeg, and the Medical Segmentation Decathlon. Natural-image evidence refers to methods such as CAM [14], SEC [15], BoxSup [12], and Simple Does It [56], which are cited to explain the origin of a mechanism rather than to support a medical performance claim. When a table or paragraph mixes the two, the natural-image origin is stated explicitly. Readers should not interpret natural-image mAP or mIoU values as medical segmentation benchmarks. This distinction matters for the annotation-economics argument: a favorable cost–performance trade-off observed on a natural-image benchmark does not by itself establish that the same trade-off holds for a medical anatomy, modality, or clinical task. Within primary medical evidence, we further distinguish two levels: E1 denotes medical evidence with external validation, i.e., evaluation on data independent of training, across centers, modalities, or public challenge test sets; E2 denotes medical evidence from a single dataset or a single split, where the reported value is not externally validated. Natural-image evidence is labeled E3 and is included only to explain the origin of a mechanism.

3.1. Image-Level Supervision: From a Class Tag to a Mask

Image-level labels are often among the cheapest signals on the spectrum (Section 2): a clinician records only which structures or pathologies are present, without requiring additional pixel-level annotation when suitable labels can be derived from routine reports [50]. The main algorithmic challenge is therefore to expand a global label into a pixel map. Nearly all methods inherit the class-activation machinery from natural images: class activation

maps (CAM) [14], the seed-expand-constrain pipeline (SEC) [15], adversarial complementary learning (ACoL) [61], deep seeded region growing (DSRG) [62], and AffinityNet's pixel-level affinity refinement [63]; expectation-maximization over pseudo-masks [16], regularized losses [64], cross-image semantic mining [65], and constrained deep weak supervision for histopathology [66] further populate the image-level toolbox. These methods originate in natural-image segmentation and are cited here as mechanism sources; their reported mIoU or localization values are not medical segmentation benchmarks. The medical adaptations discussed below are the primary evidence for image-level supervision in this review. Formally, for a feature map $F \in \mathbb{R}^{H\times W\times K}$ and a class weight vector $w_c \in \mathbb{R}^K$, the class activation map for class c is as follows:

$$M_c(x,y) = \sum_{k=1}^{K} w_{c,k} f_k(x,y), \tag{5}$$

where $f_k(x,y)$ is the activation of unit k at spatial location (x,y), and $w_{c,k}$ is the weight of unit k for class c [14]. A pseudo-mask is then obtained by thresholding or normalizing M_c:

$$\hat{Y}_c(x,y) = \mathbb{1}[M_c(x,y) \geq \tau_c], \tag{6}$$

where τ_c is a class-specific threshold.

CAM threshold selection and sensitivity. The threshold τ_c in Equation (6) is a critical but often under-reported design choice. Several strategies are used in practice. A fixed threshold, such as 0.5 after min–max normalization, is the simplest but rarely optimal across classes. A percentile-based threshold selects the top q% of activations and is easier to transfer across images, though it assumes a known target size. A validation-set search optimizes τ_c against a segmentation metric such as Dice or IoU, which is more reliable but requires dense validation masks that are often unavailable in weak-supervision settings. An adaptive method such as Otsu thresholding chooses τ_c from the activation histogram, and joint optimization with a post-processing step such as a conditional random field can further refine the boundary. Segmentation quality is sensitive to this choice, especially for small targets. A threshold that is appropriate for a large organ may suppress a small lesion almost entirely, because the lesion occupies a small fraction of the activation histogram; conversely, a threshold low enough to preserve a small lesion may over-segment adjacent tissue. For this reason, image-level supervision should be treated primarily as a localization cue rather than as a substitute for dense supervision, and any quantitative use of CAM-derived masks should report the threshold rule and a sensitivity analysis over τ_c.

Recent work has begun to explicitly model medical-specific distributional discrepancies for image-level weak supervision. DISCERN [67] identifies three medical characteristics that hinder image-level WSSS—distributional discrepancy, ambiguous boundaries, and structural interference—and proposes a contrastive convolution module plus a commonality-specificity attention module to localize medical regions of interest from image-level labels alone, consistently outperforming state-of-the-art WSSS methods on pneumonia, skin cancer, and nuclear boundary datasets.

In medical settings this lineage has been adapted to two dominant regimes. One is 2D classification to segmentation, where a global label such as "pneumothorax present" is back-projected into a localization map. Wang et al. [11] established ChestX-ray8 as the canonical benchmark for this regime, and subsequent work improved activation completeness through equivariant and uncertainty-minimizing objectives [37]. Another is patch or slide-level MIL to pixel, dominant in histopathology and volumetric scans where dense labeling is infeasible. Han et al. [55] introduced multi-layer pseudo-supervision that lifts patch-level classification labels to tissue masks; Kuang et al. [51] ("Cluster-Re-Supervision") explicitly bridges the image-level-to-pixel gap by clustering features and re-deriving dense pseudo-labels; and INSIGHT

[68] embeds the heatmap as an inductive bias inside a MIL aggregator, so that interpretability is produced by the architecture rather than reconstructed post hoc, improving both classification and weakly labeled segmentation on CT and whole-slide images.

A complementary line of work replaces the single CAM-based seeding step with a composite loss [15]. SEC, for instance, minimizes as follows:

$$\min_{\theta} \sum_{(X,T)\in\mathcal{D}} \left[L_{\text{seed}} + L_{\text{expand}} + L_{\text{constrain}} \right], \tag{7}$$

where the seeding loss is as follows:

$$L_{\text{seed}}(f(X), T, S_c) = -\frac{1}{\sum_{c\in T} |S_c|} \sum_{c\in T} \sum_{u\in S_c} \log f_{u,c}(X), \tag{8}$$

with S_c the set of locations weakly localized as class c, and the expansion loss uses global weighted rank pooling,

$$G_c(f(X), d_c) = \frac{1}{Z(d_c)} \sum_{j=1}^{n} (d_c)^{j-1} f_{i_j,c}(X), \quad Z(d_c) = \sum_{j=1}^{n} (d_c)^{j-1}. \tag{9}$$

The constrain-to-boundary loss aligns the network output with a fully connected CRF,

$$L_{\text{constrain}}(X, f(X)) = \frac{1}{n} \sum_{u=1}^{n} \sum_{c\in\mathbb{C}} Q_{u,c}(X, f(X)) \log \frac{Q_{u,c}(X, f(X))}{f_{u,c}(X)}. \tag{10}$$

Figure 3 illustrates this CAM-based mechanism on histopathology data, where progressive dropout attention enlarges the activated regions across training epochs [55]; the representative results are summarized in Table 3. Two cautions frame this family. The discriminative-region bias is structural: CAM-family methods activate the most class-evidence region and under-cover the rest, which is tolerable for localization but harmful for volumetry [36]. The under-coverage has three structural causes. First, the classification objective rewards the most discriminative region rather than the full extent of the object, so the activation concentrates on the part of the lesion that best separates the class from its background. Second, global average pooling further compresses the spatial response into class weights, which favors the most informative local pattern and suppresses weaker but still lesion-bearing regions. Third, backbone downsampling reduces the spatial resolution of the activation map, so fine boundaries and thin structures are lost before thresholding is applied. This problem is generally more severe for small medical targets. A small lesion occupies few pixels in the feature map, so its activation signal is more easily diluted by surrounding tissue; low contrast and ambiguous boundaries, which are common in medical imaging, shrink the discriminative region relative to the full lesion; and downsampling can reduce a small lesion to only a few feature-map cells, leaving little spatial resolution for thresholding. The result is that CAM-derived masks tend to under-segment small, low-contrast, or heterogeneous lesions more than large, high-contrast organs. This is consistent with the observation that image-level supervision is most useful for localization and classification and least reliable when precise volumetry or boundary-sensitive radiomics are required. Backbone resolution modulates the severity of this under-coverage. Stronger downsampling reduces the spatial resolution of the feature map, so a small target occupies fewer cells, and the activation that survives thresholding covers a coarser and smaller region; low-contrast boundaries compound this because the discriminative response is weaker relative to surrounding tissue. Higher-resolution or multi-scale features can alleviate the effect, but they cannot remove it, because the bias originates in the classification objective and global pooling rather than in resolution alone. A radiologist audit of Grad-CAM on CheXpert found that thresholded saliency only loosely overlaps expert reference segmentations, indicating that image-level signals are localization cues rather than segmentation

substitutes [69]. The reported accuracy gap for image-level supervision also varies considerably across targets and metrics (Table 1, Section 2), although the available evidence is too heterogeneous to rank it as the most volatile weak-label type. The histopathology study reports an mIoU gap of approximately 2%, while CXR localization is reported as AUC/IoBB rather than Dice, so image-level supervision does not provide a uniform accuracy gap across targets and metrics. We therefore position image-level supervision primarily as a localization or classification-oriented signal. When exact boundaries or quantitative measurements are required, additional spatial supervision or validation is generally needed.

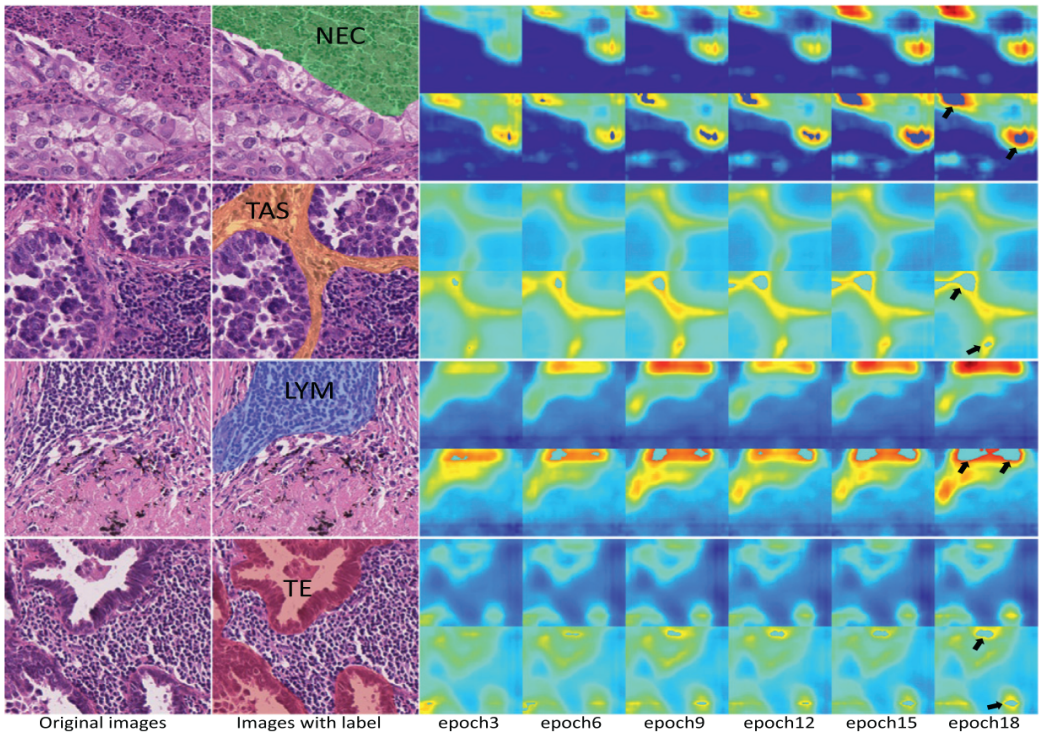


Figure 3. Class activation maps (CAMs) of each tissue category in histopathology, with and without Progressive Dropout Attention (PDA), across training epochs. The deactivated areas (black arrows) enlarge as training proceeds, pushing the CAMs to cover more complete tissue regions. Reproduced from [55] under CC BY-NC-ND 4.0.

Table 3. Representative image-level medical segmentation results. Values are reported under the metric used by each source; they are not directly comparable across rows.

Approach	Mechanism	Reported Accuracy/Gap	Dataset
CAM [14]	Classification activation	top-5 localization error 37.1% on ILSVRC; natural-image origin	Generic (natural)
SEC [15]	Seed, expand, constrain	mIoU improvement 13.5% (test) and 14.1% (val) on PASCAL VOC; natural-image origin	Generic (natural)
Han et al. [55]	Multi-layer pseudo-supervision	mIoU 0.6892, FwIoU 0.73745; ≈2% gap to fully-sup	Histopathology (LUAD-HistoSeg)
Kuang et al. [51]	Cluster re-supervision	DC 68.51% (BraTS2019), 73.13% (prostate); +2.39–10.03% over CMER	Brain tumor; prostate

Table 3. *Cont.*

Approach	Mechanism	Reported Accuracy/Gap	Dataset
SeeNet, SEAM, AffinityNet, CIAN (surveyed in [38])	Seed refinement by cross-pixel similarity, cross-view consistency, and cross-image relation	Qualitative improvements in seed completeness; survey-level summary, no unified Dice gap	PASCAL VOC
Grad-CAM localization audit [69]	Post hoc saliency	mIoU and hit rate both below human benchmark; worst for small, complex lesions	CheXpert (CXR)
DISCERN [67]	Discrepancy learning (image-level)	Outperforms SOTA WSSS on COVID-19 pneumonia, skin cancer, nuclear boundaries	COVID-19, ISIC, nuclei

Rows above the first midrule are natural-image evidence; rows below it are primary medical evidence. The two groups are separated to prevent natural-image localization or mIoU values from being read as medical segmentation benchmarks. SeeNet, SEAM, AffinityNet, and CIAN are image-level weakly supervised methods included as seed-refinement baselines to illustrate mechanisms also used under weak labels; they are not semi-supervised baselines. Their citation details are not reproduced here; the reader is referred to the survey [38] for the original references. CAM, SEC, SeeNet, SEAM, AffinityNet, and CIAN are natural-image methods included as methodological context only; their mIoU and localization values are not medical segmentation benchmarks. The remaining rows report primary medical evidence. CAM-based entries are sensitive to the activation threshold τ_c, and their reported values should be interpreted with the threshold rule in mind, as discussed in Section 3.1.

3.2. Bounding-Box Supervision: Cheap Geometry, Expensive Boundaries

A bounding box requires roughly an order of magnitude less annotation time than a dense mask [56] yet supplies a strong geometric prior: the object is somewhere inside the box and nowhere outside it. This inside/outside constraint is the workhorse of box supervision. BoxSup [12] and the "simple does it" baseline [56] treat the box as a region from which a CRF proposes a mask; in medical imaging, DeepCut [17] is the foundational method, as it generalizes GrabCut with a learned CNN classifier and a dense conditional random field, optimizing an energy over pixel labels and iteratively updating the training targets, and was demonstrated on fetal brain and lung MRI.

A box provides only an inside/outside constraint, so distinguishing multiple adjacent anatomical structures relies on additional cues: appearance-based pairwise terms such as the dense CRF in DeepCut, class-specific boxes that assign different targets to different boxes, box-tightness priors, and multi-instance constraints. The main failure mode is the case in which two structures share a box or are tightly adjacent with similar intensity; there, the inside/outside constraint alone cannot separate them, and additional supervision or a boundary-aware prior is required.

BoxSup formulates box-supervised segmentation as an alternating optimization problem [12]. Let S denote a candidate segment and B a ground-truth bounding box. The overlapping objective is as follows:

$$\mathcal{E}_o = \frac{1}{N} \sum_S (1 - \mathrm{IoU}(B, S)) \delta(l_B, l_S), \tag{11}$$

where $\mathrm{IoU}(B, S)$ is the intersection-over-union between B and the tight bounding box of S, and $\delta(\cdot, \cdot)$ is one when the semantic labels agree and zero otherwise. The regression objective is as follows:

$$\mathcal{E}_r = \sum_p e(X_\theta(p), l_S(p)), \tag{12}$$

and the combined objective is $\min_{\theta, \{l_S\}} \sum_i (\mathcal{E}_o + \lambda \mathcal{E}_r)$, with λ representing a fixed weighting parameter.

Earlier, fully supervised organ segmentation work such as DeepOrgan [70] is included here only as historical context; it uses full pixel-level labels and is not a weak-label method.

In contrast, RAUNet's limited-annotation COVID-19 lung-infection segmentation [71] sits at the weakly or coarsely labeled end of the spectrum, while box-only instance methods such as BoxSnake [72] show that a box can drive polygon-level masks, albeit so far outside the medical domain. Later medical variants tighten the box into a mask via region masking [73] or bounding-box-controlled training [74]. More recent weakly supervised 3D pipelines, such as the probabilistic-aware framework of Jiang et al. [75], convert sparse 3D point labels into dense pseudo-labels and incorporate annotator confidence into both the network and the loss, achieving up to 18.1% Dice improvement on CT and MRI datasets and matching fully supervised performance on some organs.

DeepCut formulates box-supervised segmentation as energy minimization over a densely connected CRF [17]. The energy is as follows:

$$E(f) = \sum_i \psi_u(f_i) + \sum_{i<j} \psi_p(f_i, f_j), \tag{13}$$

where the pairwise term is as follows:

$$\psi_p(f_i, f_j) \propto \exp\left(-\frac{(I_i - I_j)^2}{2\theta_\beta}\right), \tag{14}$$

and the unary term is the negative log-likelihood of a CNN,

$$\psi_u(f_i) = -\log P(y_i \mid x, \Theta). \tag{15}$$

The key empirical question is how much accuracy the box's missing interior boundary costs. The answer is strongly organ- and dataset-dependent: BoxSup reports an mIoU gap of approximately 1.8 percentage points on PASCAL VOC [12], with approximately 62.0% under box supervision and 63.8% under full-mask supervision. Simple Does It reaches ≈ 95% of full quality [56], and DeepCut reports Dice gaps of 7.5% on fetal brain and 12.9% on fetal lung [17]. The gap is therefore not uniform across targets, and generic box-supervision surveys obscure this by aggregating over easy organs.

Beyond box supervision, a related line of work replaces box annotations with sparse 3D point annotations and transforms them into dense confidence maps; Figure 4 summarizes the probability-aware framework of Jiang et al. [75]. While this framework uses sparse 3D points rather than boxes, it is included here as a representative of the broader sparse-supervision family.

3.3. Scribble Supervision: A Practical Middle Ground

Scribbles are free-hand strokes that trace object interiors [18]. They provide a useful compromise between annotation burden and spatial information: they are substantially sparser than a full mask while directly specifying interior pixels. Several studies report scribble methods approaching, matching, or in some cases exceeding the reported fully supervised reference under their own evaluation protocols [20,21,39,40]; these results should not be generalized across datasets without checking the split and baseline.

The scribble family can be analyzed along several mechanistic dimensions: pseudo-label generation, consistency regularization, architecture design, volumetric extension, classical constraint losses, and application-specific variants. These dimensions are not mutually exclusive. Many methods combine several of them, and the boundaries between them are often blurred by implementation choices. We therefore review them as connected strands rather than as separate categories.

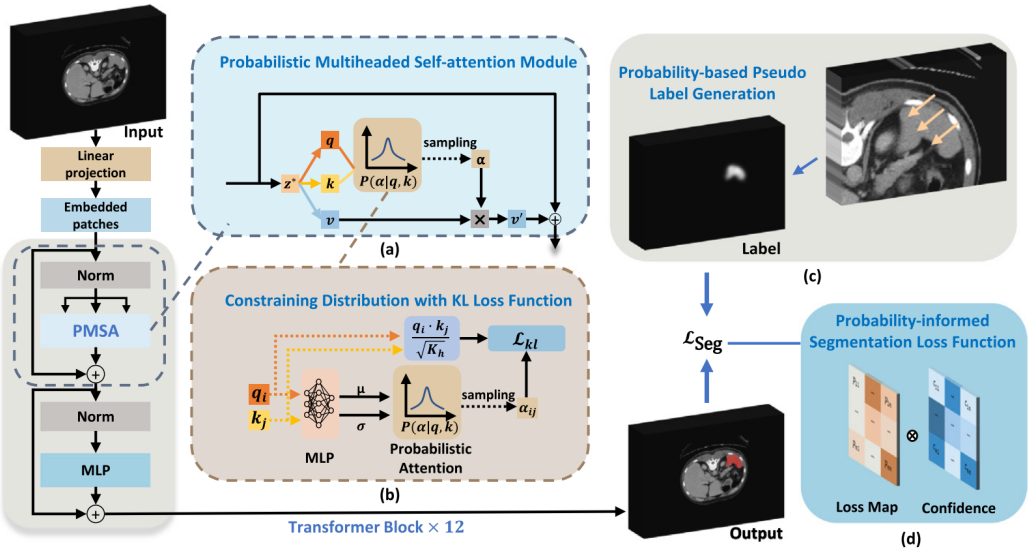


Figure 4. Overview of the probability-aware learning framework for weakly supervised 3D medical image segmentation. (**a**) Probability-based pseudo-label generation transforms sparse 3D point annotations into dense Gaussian confidence maps. (**b**) The probabilistic multi-headed self-attention (PMSA) module models the dependency scores as a Gaussian distribution and samples them via the reparameterization trick. (**c**) The KL loss constrains the learned attention distribution by anchoring it to a data-informed prior. (**d**) The probability-informed segmentation loss combines Dice loss with a confidence-weighted cross-entropy term. Reproduced from [75] under CC BY-NC 4.0.

The dominant strategy is to grow the scribble into a full pseudo-mask and train on it. The lineage runs from ScribbleSup [18] in natural images to medical adaptations such as Scribble2Label [19], which generates self-training pseudo-labels with consistency, MAAG [39], which uses multi-scale adversarial attention gates, and DMPLS [76], which introduces a dual-branch network with dynamically mixed pseudo-labels.

ScribbleSup formulates scribble-supervised learning as a graphical model over superpixels [18],

$$\sum_i \psi_i(y_i \mid X, S) + \sum_{i,j} \psi_{ij}(y_i, y_j \mid X), \tag{16}$$

where ψ_i combines a scribble-based unary term and a network-based unary term, and ψ_{ij} is a pairwise appearance term between adjacent superpixels. The recurring failure mode is pseudo-label noise: erroneously grown regions reinforce themselves. Consistency and uncertainty mechanisms are therefore often added to suppress that noise.

Let $S \subset \Omega$ denote the scribble region and y_p the scribble label at pixel $p \in S$. The partial cross-entropy loss is [18,19]

$$\mathcal{L}_{\text{pCE}} = -\sum_{p \in S} \log p\big(y_p \mid x_p\big), \tag{17}$$

which is the loss used by ScribbleSup, Scribble2Label, and most subsequent scribble-supervised methods.

Because the scribble covers only part of the object, models are trained to make consistent predictions under perturbations. CycleMix [20] is the clearest instantiation of this idea: it performs mix-based augmentation with random occlusion, incrementing and decrementing scribbles, plus global and local cycle-consistency losses, and on ACDC it reaches parity with fully supervised methods.

CycleMix adds two consistency terms [20]. Its global consistency loss is

$$\mathcal{L}_{\text{con-g}} = \frac{1}{2}\left[\mathcal{L}_{\text{ncs}}(p_{12}, q_{12}) + \mathcal{L}_{\text{ncs}}(p_{21}, q_{21})\right], \tag{18}$$

where $\mathcal{L}_{\text{ncs}}$ is the negative cosine similarity, p_{12} is the mixed segmentation, and q_{12} is the segmentation of the mixed image. Its local consistency loss is

$$\mathcal{L}_{\text{con-l}} = \frac{1}{2}[\mathcal{L}_{\text{ncs}}(\hat{y}_1, C(\hat{y}_1)) + \mathcal{L}_{\text{ncs}}(\hat{y}_2, C(\hat{y}_2))], \tag{19}$$

where $C(\cdot)$ is a morphological operation that outputs the largest connected area of each non-background class. Later consistency refinements include ScribbleVC's vision-class embedding [77], superpixel-guided scribble walking [78], SRPA's scribble-match pixel alignment [79], PCLMix's pixel-level contrastive learning [80], and SDT-Net's dynamic teacher switching with hierarchical consistency [60], which reaches 90.8% on ACDC and 90.0% on MSCMRseg.

Architecture design provides a complementary route. Recovering global shape from sparse scribbles remains challenging, and ScribFormer [21] addresses this with a triple-branch CNN, Transformer, and ACAM design that fuses local CNN features with global Transformer representations and an attention-guided class-activation map, outperforming prior scribble methods and most fully supervised ones at lower annotation cost. HELPNet [52] reaches comparable ACDC accuracy via hierarchical perturbation consistency and entropy-guided ensembling. These works show that architecture and supervision interact: stronger backbones can improve the use of sparse labels, but the reported gains cannot be attributed to label type alone.

Volumetric extension is harder because annotating every slice is impractical. Most scribble research remains 2D, typically on cardiac cine MRI. Han et al. [81] propose a triple-branch multi-dilated network with pixel- and class-wise consistency for 3D multi-organ scribble segmentation; Chen et al. [82] ("Scribble2D5") learn volumetric segmentation from sparse 2D scribbles; and reference-set shape priors [83] regularize 3D scribble learning. This remains an under-served but high-value sub-area: most clinical targets are 3D, yet the scribble literature remains 2D-dominated.

Classical constraint losses remain competitive in this setting. Before deep pseudo-labeling took over, scribbles were regularized with graph, CRF, and level-set constraints. Gated CRF [84] remains a strong, lightweight supervision loss, and normalized-cut and geodesic formulations [77] are still used where datasets are small. These remain useful as regularizers that modern deep methods combine with pseudo-labels, and they matter most in the low-data regime where learned pseudo-labels are unreliable.

Beyond these strands, a long tail of specialized designs targets particular settings rather than proposing a new mechanism. Hierarchical and PU-learning formulations address brain-tumor MRI under scribble supervision [85,86]; interaction- and efficiency-oriented networks, including online-likelihood designs and regularized triangular flows, target the low-compute regime [87]; application-specific studies cover COVID-19 infection on chest CT and breast cancer in DCE-MRI [88,89]; and non-iterative pacing schedules, masked context modeling, and query-based MaxViT backbones revisit how pseudo-masks are generated and refined [90–92]. Recent surveys on label-noise learning further warn that pseudo-label noise, which is pervasive in weak supervision, can silently degrade segmentation and downstream diagnosis [93]. Weakly supervised multi-task designs such as WML-Net [94] show that segmentation and classification can reinforce each other even without dense mask annotations: a dual-branch pseudo-mask generation module plus a feature enhancement module improves carotid plaque classification by roughly 3% while eliminating the need for manual segmentation masks. These works populate rather than reorder the strands above, but they are frequently the closest match to a practitioner's actual modality and organ. The corresponding ACDC results are summarized in Table 4, while Figure 5 illustrates two representative scribble-learning designs.

Table 4. Scribble-supervised methods on ACDC (average Dice over LV/RV/MYO). All values are reported under the standard 70/15/15 ACDC partition. The CycleMix value (84.8%) is the number reported in the ScribFormer and SDT-Net benchmark tables under the standard 70/15/15 split, not the value in the original CycleMix paper, whose main ACDC results use a 35-image training setting (84.8%) and the full 70-image training set (85.5%). The Scribble2Label value (82.0%) is likewise taken from the ScribFormer benchmark under the 70/15/15 partition. The HELPNet value (88.8%) is the SDT-Net benchmark result under the same split. These three values are therefore re-reported benchmark values under a common split rather than the original authors' primary evaluations, and are comparable within this table but not with results obtained under other partitions.

Method	Avg Dice (%)	Mechanism Axis
$UNet_{pce}$ [21]	56.2	Partial-CE baseline (no pseudo-label)
$UNet_{rw}$ [21]	79.1	Classical Random-Walker pseudo-label
Scribble2Label [19]	82.0	Label-filtered pseudo-label
CycleMix [20]	84.8	Mix-aug + cycle consistency
DMPLS [76]	87.2	Dual-branch dynamic mix
ScribbleVC [77]	88.4	Vision-class embedding
ScribFormer [21]	88.8	CNN, Transformer and ACAM
HELPNet [52]	88.8	Perturbation + entropy ensemble
SDT-Net [60]	90.8	Dynamic teacher switching + hierarchical consistency
nnU-Net [5]	92.0	Upper-bound reference

All values in Table 4 are reported under the standard 70/15/15 ACDC partition. Results obtained under different partitions, including the full 70-subject training set, are excluded from the main table to ensure comparability. The 70/15/15 convention is independently corroborated by the recent SDT-Net benchmark [60], which reports 88.8% for ScribFormer, 88.4% for ScribbleVC, and 87.2% for DMPLS under exactly this split, matching the values tabulated above. The lower-bound rows ($UNet_{pce/rw}$, S2L) and the nnU-Net reference row come from ScribFormer's ablation and benchmark tables [21] on the same partition. The CycleMix value of 84.8% is likewise the re-reported benchmark value under the standard 70/15/15 split rather than the original CycleMix paper's primary result, whose ACDC evaluation used a different training protocol. The DMPLS value of 87.2% is taken from the SDT-Net benchmark under the standard 70/15/15 partition, so that all values in the table are reported under the same split. For CycleMix, Scribble2Label, and HELPNet, the tabulated values are therefore re-reported benchmark numbers from ScribFormer and SDT-Net rather than the original authors' primary evaluations, and they are comparable within this table but not with results obtained under other partitions. Because ACDC scribble numbers are split-sensitive, the numerical gap should be interpreted within each cited benchmark rather than as a universal 2–4% rule.

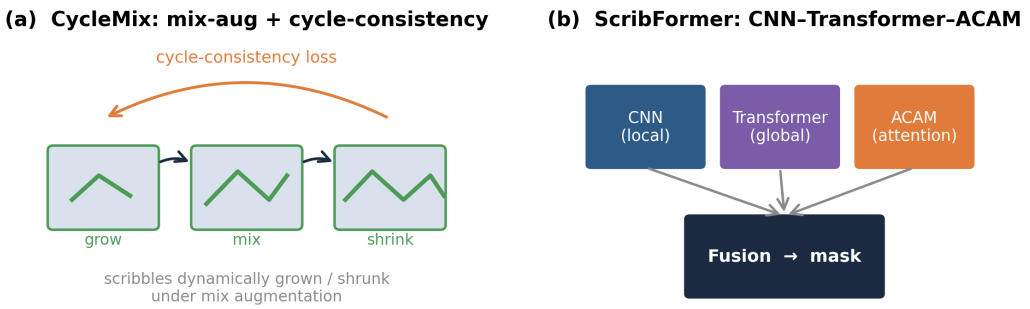


Figure 5. Scribble family, two panels. (**a**) A schematic of CycleMix's mix-augmentation with cycle-consistency [20]: scribbles are dynamically grown and shrunk. (**b**) A schematic of ScribFormer's triple-branch CNN–Transformer–ACAM architecture [21].

3.4. Point, Click, and Interactive Supervision

Points and clicks are among the most explicitly human-in-the-loop forms of sparse supervision: a user provides localized corrections or prompts, and the model updates its segmentation accordingly. This family spans classic interactive segmentation, including f-CN [53], BIFSeg [22], DeepIGeoS [23] (geodesic refinement), and Roth et al.'s "extreme points" regime [9], to modern click-driven models such as RITM [24] and SimpleClick [25]. Beyond clicks, partial-points annotation, as demonstrated for nuclei segmentation [95], shows that even a sparse set of labeled points can supervise a full mask. The defining economic property is asymptotic accuracy: with enough clicks, interactive methods can approach full supervision, but the cost is the click budget, which must be counted on the same economic axes as other labels [9]. Representative point- and click-based results are collected in Table 5. As in Table 1, cost is reported on an ordinal scale, and natural-image entries are marked as methodological context rather than medical evidence.

Table 5. Representative point-, click-, and interactive-supervision results. Cost is expressed on the same ordinal scale used in Table 1 (Very low, Low, Medium, High, and Very High), and evidence type distinguishes primary medical evidence from natural-image methodological context. Values are reported under the metric used by each source and are not directly comparable across rows.

Method	Prompt Type	Ordinal Cost	Reported Accuracy	Evidence Type
DeepIGeoS [23]	Scribble + clicks	Low to medium	2D placenta: 89.31% vs. 85.86% before refinement; 3D brain tumor: 89.93% vs. 87.55%	Medical
Roth et al. "extremes" [9]	Extreme points	Very low	Dice 0.948 (MSD-Spleen) down to 0.703 (MO-Pancreas); gap 1.0–4.4% vs. full	Medical
BIFSeg [22]	Box + clicks	Low	Interactive refinement of box-initialized masks; image-specific fine-tuning	Medical
f-CN [53]	Clicks	Low	Classic interactive object selection	Natural (context)
RITM [24]/ SimpleClick [25]	Clicks	Low	SOTA interactive segmentation on natural-image benchmarks	Natural (context)
ScribblePrompt-UNet [49]	Scribble, box, click	Low to medium	0.84 Dice on ACDC scribbles; HD95 1.80; ≈28% less time than SAM	Medical
MedSAM [46]	Box/point prompt	Low (box)	Avg. Dice 0.85 over 86 + 60 tasks; 3D tumor annotation time −82%	Medical
Gaze + VLM [58]	Gaze + text	Medium	80.78/80.53/84.22 Dice on Kvasir-SEG/NCI-ISBI/ISIC	Medical

Rows between the first and second midrules are natural-image methodological context; all other rows are primary medical evidence. Natural-image entries (f-CN, RITM, and SimpleClick) are included as methodological context only and should not be read as medical performance evidence. Cost entries are study-specific and indicative rather than universal.

Two shifts matter for this review. Interactive prompting and weak supervision are closely related but not identical: a scribble can serve as a sparse training annotation or as an inference-time prompt, depending on the pipeline. Foundation models have also redefined the baseline (detailed in Section 4). SAM [45] and MedSAM [46] treat clicks, scribbles, and boxes as prompts, delivering strong zero-shot masks; ScribblePrompt [49] shows that an interactive U-Net trained on heterogeneous weak prompts matches ScribFormer's Dice (0.84 vs. 0.85, $p = 0.70$) on ACDC scribbles while needing less annotation time and yielding lower HD95 (1.80 vs. 4.05) [49], an economics-framed comparison that previews Section 4. The 0.85 value is the ScribFormer Dice reported inside the ScribblePrompt study under its own evaluation setting. Under the standard 70/15/15 ACDC partition, ScribFormer reports 88.8% in its original paper (Table 4). The two values are therefore

not directly comparable, and the reader should use the standard-partition value of 88.8% when comparing ScribFormer with other scribble methods on ACDC. Recent gaze-based and VLM-assisted interactive methods push this further. The teacher–student framework of Chen et al. [58] fuses clinician gaze with vision–language model explanations, achieving Dice values of 80.78%, 80.53% and 84.22% on Kvasir-SEG, NCI-ISBI, and ISIC, improving 3–5% over gaze-only baselines without increasing annotation burden.

3.5. Cross-Cutting Mechanisms: Semi- and Self-Supervised Techniques as Structural Enhancements to Weak Labels

Having organized methods by label type, we now extract the mechanisms that are label-type-agnostic, that is, the techniques a practitioner reaches for regardless of whether the weak signal is a tag, a box, or a scribble. Semi-supervised and self-supervised techniques are discussed here as structural enhancements to weak-label learning, not as weak-label types themselves. Framing them separately allows Section 5 to ask which mechanism best serves a diagnostic goal, rather than which method achieves the highest score on a given dataset.

Pseudo-labeling and self-training grow sparse labels into dense ones and are common across weak-supervision settings (Sections 3.1 and 3.3). Their main failure mode is error accumulation, which motivates the use of additional regularization. Constraint losses such as DenseCRF [96] (the dense-CRF of DeepLab [97]), entropy minimization [16], and Gated CRF [84] encode the prior that nearby pixels with similar appearance share a label; they therefore act as regularizers on pseudo-labels and can be important in the low-data regime. Consistency and self-ensembling provide another transferable mechanism. Mean-Teacher [28] and uncertainty-aware variants (UA-MT) [29] encourage agreement under perturbation. Although this principle originated in semi-supervision, which we hold at the boundary (Section 1), it is routinely borrowed by weakly supervised methods, with CycleMix's cycle consistency [20] and SDT-Net's hierarchical consistency providing direct examples. Medical instantiations include Interpolation Consistency Training [98], Cross Pseudo Supervision [99], Dual-Task Consistency [100], Mutual Consistency (MC-Net+) [101], Uncertainty Rectified Pyramid Consistency [102], the self-supervised SS-Net [103], pseudo-label guided contrastive learning [104], causality-inspired learning [105], and the Uncertainty-guided Collaborative Mean-Teacher (UCMT) [106]. Uncertainty weighting further uses model uncertainty to suppress unreliable pseudo-label regions, as in UA-MT [29] and HELPNet's entropy-guided ensemble [52]; the probabilistic-aware 3D framework of Jiang et al. [75] additionally integrates annotator confidence into the loss. Representation learning acts as a force multiplier rather than a new label type: self-supervised pre-training such as Models Genesis [31], TransVW [32], and SwinMM [33] can improve the starting representation and thereby reduce the amount of weak supervision needed to reach a given performance level. Synthetic and generative data provide another emerging mechanism, because diffusion and generative models can manufacture training pairs that partially compensate for scarce annotations; this direction is revisited in Section 6.

These mechanisms use unlabeled or self-generated signal to compensate for labels that were not collected. They are the reason the cost–accuracy frontier of Section 2 (Table 1) keeps moving toward lower annotation time and the reason the next section asks whether foundation models have made the whole weak-supervision problem moot.

4. Foundation Models and the Re-Encoding of Weak Supervision

This section examines how the rise of segmentation foundation models (FMs), above all the Segment Anything Model (SAM) and its medical derivatives [45–49], reshapes the weak-supervision problem rather than abolishing it. We keep the lens introduced in

Section 2 (annotation economics) and the taxonomy built in Section 3 (labels ordered by granularity) and examine which parts of the supervision burden are reduced, retained, or shifted by FMs.

4.1. The Foundation-Model Shift: From Task-Specific Training to Promptable Segmentation

Medical image segmentation has traditionally relied on models trained end-to-end on a specific dataset for a specific organ. The weak-supervision literature of Sections 2 and 3 was a response to the resulting labeling cost: if dense masks are unaffordable, one can train on boxes, scribbles, or image-level tags instead. The Segment Anything Model (SAM) changed this premise [45]. SAM is trained on SA-1B, comprising 11 million natural images and roughly 1 billion masks [45], to perform promptable segmentation: given visual prompts such as points, boxes, or masks, it returns a segmentation without task-specific retraining. This reframes supervision as an interaction rather than a dataset. Formally, promptable segmentation learns a mapping:

$$\hat{M} = f_\theta(I, P), \qquad P \in \mathcal{P}, \tag{20}$$

where I is the input image, P is a prompt, and $\mathcal{P}$ is the prompt space that includes points, boxes, scribbles, and masks. SAM predicts multiple masks for a single prompt to remain ambiguity-aware, and the final mask is selected by an estimated IoU score [45].

A SAM prompt can be interpreted as weak human guidance, but it is not necessarily a training weak label. A bounding-box prompt is analogous to the box supervision reviewed in Section 3.2, while clicks and scribbles can be used either as training annotations or as inference-time prompts. This distinction matters economically: classical weak supervision incurs human annotation costs primarily during dataset construction, whereas promptable foundation models may incur additional interaction costs at inference time. Although pre-trained models can reuse similar interaction types across tasks, medical-domain adaptation may still require substantial supervision. As illustrated in Figure 6, the correspondence between traditional weak labels and foundation-model prompts is conceptual rather than strictly one-to-one. Similar spatial signals may supervise model training or guide segmentation at inference time, but their roles and associated costs differ. Figure 6 summarizes the relationship between the weak-label spectrum and prompt-based inference.

How to classify SAM-based systems. SAM-based systems do not fit into a single category. Their classification depends on where the supervision comes from and when it is applied. If a human provides a point, box, or scribble at inference time, the system is best described as interactive segmentation and, under this review's definition, as an inference-time form of weak supervision, because the prompt is a cheap signal explicitly authored by a human. If the prompt is generated automatically by another model with no human interaction at inference time, the system is better described as transfer learning or foundation-model adaptation, as discussed in Section 4.4. If SAM is used offline to generate pseudo-labels for training a downstream segmentation model, then SAM acts as a pseudo-labeler, and the overall pipeline remains a weak-supervision pipeline, even though the prompt itself may be automatically generated. Finally, if SAM is fine-tuned on a large medical corpus and then used with prompts, it is a separate foundation-model paradigm that inherits both transfer learning and interactive segmentation. We therefore treat SAM-based systems as a cross-cutting family rather than as a single weak-label type. The relevant question for the annotation-economics framework is not which label the system belongs to but where the human supervision enters the pipeline, how much it costs at that point, and whether the resulting mask is accurate enough for the downstream task.

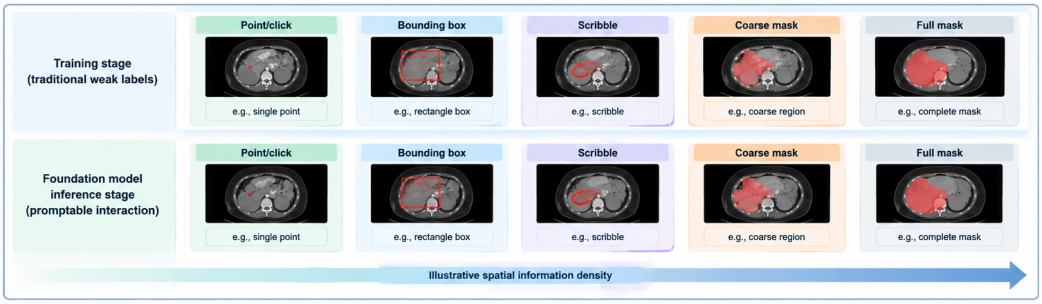


Figure 6. Prompts as re-encoded weak labels. The upper row shows traditional weak labels used during training, while the lower row shows corresponding prompts used for foundation-model inference. The prompt types are ordered by increasing spatial information density, consistent with Figure 1. Annotation costs are indicative ordinal levels rather than measured times. Human-provided prompts are treated as inference-time weak supervision, whereas automatically generated prompts are considered transfer learning or foundation-model adaptation. This is an original schematic created by the authors; no third-party material is reproduced.

4.2. How FMs Re-Encode Weak Labels: Prompts as a New Supervision Currency

Repeated inference-time prompts occupy a distinct position on the weak-label spectrum. A single prompt is an inference-time analogue of a sparse weak label, but when a clinician issues several corrective prompts in sequence, the supervision is cumulative: each round adds spatial information and expert time, and the total cost grows with the number of rounds. In the spectrum of Figure 1, repeated prompting therefore sits between a one-shot weak label and full supervision, and its position depends on the number of interaction rounds rather than on the label type alone. Economically, this cost enters the inference and correction terms of Equation (1), not the training-time annotation term, which is why prompt efficiency (the number of rounds needed to reach an acceptable mask) is itself an annotation-economics quantity. The economic logic of Section 2 extends to the foundation-model setting, but the unit of account shifts from annotation time per training example to interaction cost per inference instance. Table 6 summarizes representative prompt types and their study-dependent costs, while Figure 6 illustrates their conceptual correspondence with traditional weak labels. The figure emphasizes the spatial information conveyed by each prompt type rather than providing a quantitative ranking of interaction time. The reported times and segmentation results in Table 6 should therefore be interpreted in the context of their original studies and annotation protocols. Sparser prompts generally provide less spatial information, and zero-shot performance varies substantially with target size, contrast, modality, and prompt quality (Table 7 and Figure 7). At the same time, foundation models do not eliminate human involvement; they shift some of the effort from offline dataset annotation to online inference. The resulting annotation economics therefore depend not only on the initial cost of training data but also on the frequency and effort of user interactions during inference.

Table 6. Prompts as a re-encoded weak label (annotation economics; cf. Section 2 and Table 1). Time values are study-dependent and should be interpreted as indicative rather than universal; the bounding-box entry is derived from reported ratios in [13,17]. Prompt types are classified according to Section 4.1: human-provided prompts count as inference-time weak supervision, whereas automatically generated prompts are treated as transfer learning or foundation-model adaptation.

Prompt Type	Weak-Label Analogue	Approx. Time/Instance	Information Content	Representative Medical Dice
Point (1–few clicks)	point/click supervision [9,23–25]	seconds	very sparse (seed)	SAM ≈ 0.50 zero-shot [107]; SAM-Med2D 1 pt = 70.0% [47]
Bounding box	box weak label [12,13,17]	study-dependent	region-level	representative zero-shot/adapted results in [46,107]
Scribble	scribble weak label [10,20,21] (Section 3.3)	seconds to approximately one minute, estimated	curve-level	ScribblePrompt: Dice +15%, time −28% vs. SAM [49]
Full mask	full supervision	minutes–hours (3D)	dense	nnU-Net ceiling [5]

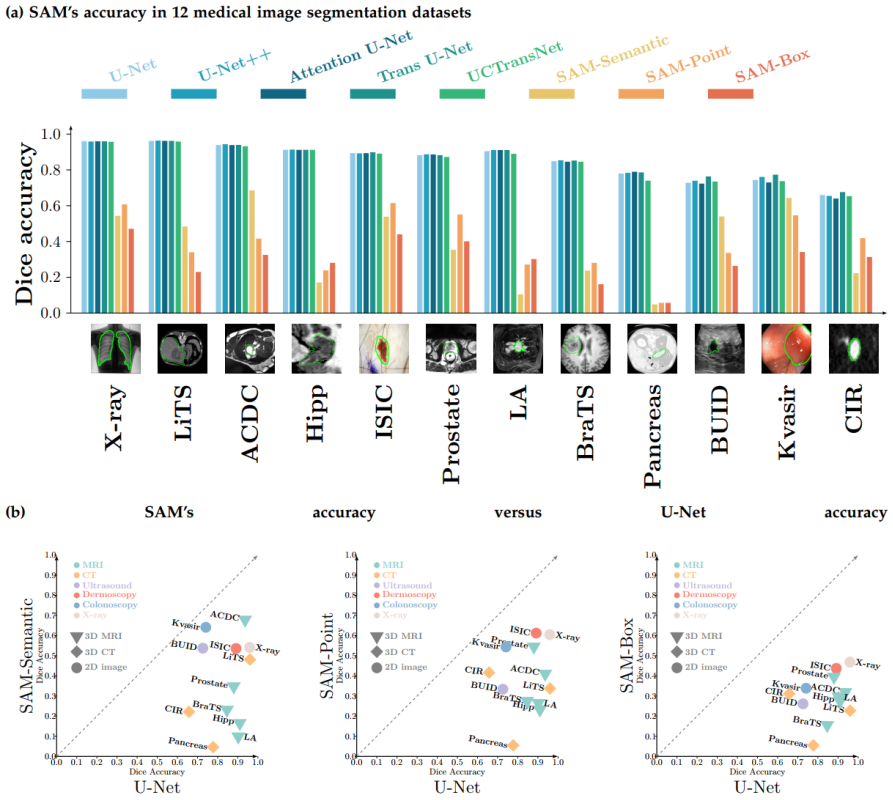


Figure 7. The zero-shot gap of vanilla SAM on medical images. (**a**) Dice overlaps of five medical-image-specific algorithms and three SAM variants. (**b**) Scatter plots of SAM's Dice versus U-Net's Dice across datasets, for SAM-Semantic (**left**), SAM-Point (**middle**), and SAM-Box (**right**). SAM's zero-shot performance is consistently below medical-image-specific models, and the gap is largest for point prompts and small targets. Reproduced from [107] under CC BY 4.0.

Table 7. Foundation-model family on medical segmentation. Numbers are not from a single benchmark; frame each cell as representative of the cited study, not a head-to-head leaderboard. The classification of each system follows Section 4.1: interactive use with human prompts is treated as inference-time weak supervision, while automatic prompting and large-corpus fine-tuning are treated as transfer learning or foundation-model adaptation.

Model	Training Scale	Prompt(s) Tested	Representative/Mean Dice
Vanilla SAM [45]	SA-1B: 11 M natural imgs, ~1 B masks	point/box	zero-shot point ≈ 0.5, box ≈ 0.6 (12 datasets) [107]; BTCV 1/3/10 pt = 0.548/0.631/0.704 [108]
MedSAM [46]	1.57 M medical pairs, 10 modalities	box	avg 0.85 over 86 + 60 tasks; 3D tumor annotation time −82%
MedSAM + TTA [109]	MedSAM + random circular shifts	box	best case up to 93.6% DSC, 97% NSD on MSD/KiTS/COVID-19-20
SAM-Med2D [47]	4.6 M imgs, 19.7 M masks, 31 organs	box/point	test box 79.3% vs. SAM 61.6%; 1 pt 70.0% vs. SAM 18.9%; 9-challenge gen. 81.9%
SAM-Med3D [48]	143 K-mask corpus; 131 K training masks	point (volumetric)	10 pt 60.9% vs. SAM 17.0%, SAM-Med2D 42.8%
ScribblePrompt [49]	65 datasets, ~54 k imgs	scribble/click/box	user study: Dice +15%, time −28% vs. SAM; 0.27 s/pred on CPU
ABP-SAM [110]	SAM + GV-Adapter + Auto Bbox Prompter	auto box	no manual prompt; qualitative improvements on complex lesions (no Dice reported)
MedSAM-2 [111]	SAM 2-tiny fine-tune	box + sparse (one-prompt)	BTCV 88.57% (box+sparse); +0.70% over MedSegDiff; 5% vs. full prompt −2%

4.3. The Medical Adaptation Ladder

SAM's zero-shot transfer to medical images is highly variable and is generally weaker on low-contrast, small, or complex targets, as reported by several benchmarks [107,108,112] and visualized in Figure 7. The field's response has been a ladder of medical adaptation, trading additional medical supervision for improved prompt-driven accuracy.

A further distinction concerns prompt provenance. A model can be evaluated with ground-truth-derived prompts, with human-drawn prompts, or with simulated prompts generated from existing masks; these are not equivalent. Ground-truth-derived boxes or points are an evaluation shortcut that removes prompt-placement error, whereas a clinician must draw the prompt in practice, and simulated prompts occupy an intermediate position. Studies typically state the prompt source in their evaluation protocol; where they do not, the prompt should be assumed to be ground-truth-derived if the benchmark reports box-prompt Dice against a reference mask. Dice obtained with ground-truth-derived prompts should therefore be read as an optimistic upper bound relative to realistic clinical prompting.

Vanilla SAM [45] performs poorly on many medical datasets: point-prompt mean Dice is approximately 0.5 and box-prompt Dice is approximately 0.6 across 12 datasets [107]; on BTCV, even 10 points reach only 0.704 [108]. The failure modes are systematic for small targets, low contrast, and 3D consistency because SAM is natively 2D [107,112]. MedSAM [46] fine-tunes the SAM image encoder and mask decoder on 1,570,263 medical image–mask pairs spanning 10 modalities and more than 30 structures while retaining the box prompt. It reports an average Dice of 0.85 over 86 internal and 60 external tasks and, in a user study, reduced expert 3D-tumor annotation time by more than 82%. MedSAM provides a strong example that a box prompt, when paired with medical-domain adaptation, can substantially reduce manual dense-labeling effort for many structures. Subsequent work

has improved MedSAM's robustness at test time: Nazzal et al. [109] combine MedSAM with test-time augmentation using random circular shifts, reaching up to 93.6% DSC and 97% NSD on MSD, KiTS, and COVID-19-20 CT datasets, with notable gains on pancreas and colon boundaries. SAM-Med2D [47] pushes scale further, with 4.6 M images and 19.7 M masks spanning 10 modalities and 31 organs, together with adapter layers and interactive correction-point training. On its test set, box Dice is 79.3% versus SAM's 61.6%, and a single point reaches 70.0% versus SAM's 18.9%; on nine MICCAI-2023 challenges, the box-prompt generalization average is 81.9%. SAM-Med3D [48] restores volumetric consistency: its SA-Med3D-140K corpus contains about 143 K masks, with 131 K used for training, covering 247 categories; it was evaluated on 15 volumetric datasets. With 10 prompt points it reaches 60.9% Dice, versus 42.8% for SAM-Med2D applied slice-wise and 17.0% for vanilla SAM, directly quantifying the 2D-to-3D gap that Section 3.3's scribble methods also struggle with. ScribblePrompt [49] is particularly relevant to this survey's thesis. Trained on 65 open-access datasets ($\sim$ 54 k images) with simulated scribbles, clicks, and boxes, it accepts the same cheap labels studied in Sections 2 and 3 through one lightweight network that runs on a single CPU (0.27 s/prediction). In a 16-expert user study, it reduced annotation time by 28% and improved Dice by 15% versus SAM, at approximately 1.5 min per segmentation versus SAM's more than 2 min. ScribblePrompt is therefore an example of weak supervision combined with a foundation model as an interactive tool. ABP-SAM [110] removes manual prompts by integrating a Gated Vision Adapter into SAM's frozen encoder and an Auto Bbox Prompter that generates bounding boxes from encoded features, reducing clinical interaction while preserving SAM's generalization. This is a concrete instance of the FM-as-weak-supervisor idea in which the prompt is produced by a lightweight module rather than directly by a clinician. MedSAM-2 [111] and SAM 2 [113] extend the paradigm to video and 3D by treating a volume as a sequence and propagating a single prompt across slices. MedSAM-2 reaches 88.57% Dice on BTCV with box and sparse prompts, reported as +0.70% over the fully supervised diffusion model MedSegDiff and +36.97% over zero-shot SAM 2; with 5% prompt coverage, the reported Dice drop is only 2% on BTCV, compared with 7.5% for SAM 2 [111]. Table 7 summarizes the representative foundation-model results, while Figure 8 visualizes the corresponding medical adaptation ladder.

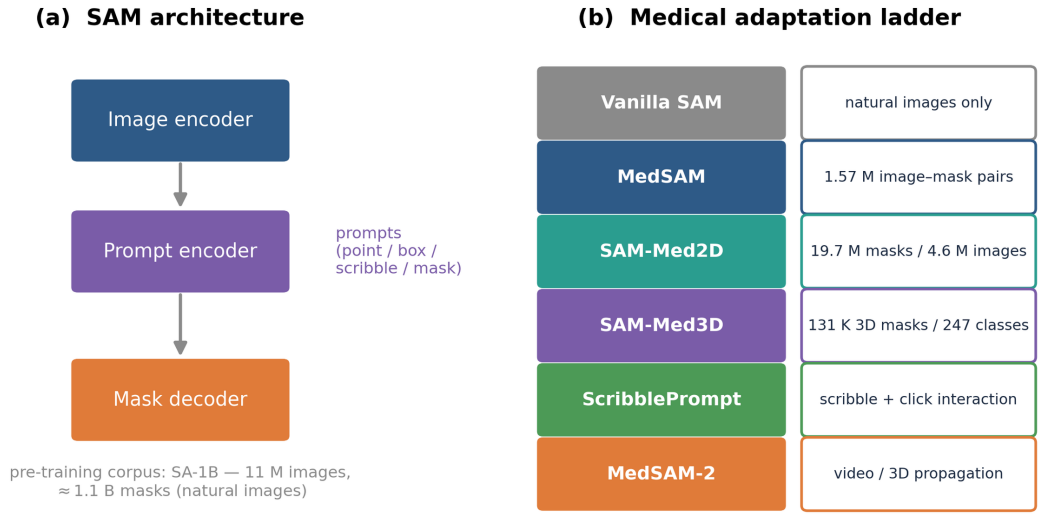


Figure 8. The adaptation ladder. (**a**) MedSAM fine-tunes SAM's image encoder and mask decoder on a large medical corpus. Reproduced from [46] under CC BY 4.0. (**b**) The medical adaptation ladder with training-corpus scale, showing where the 1.57 M-pair MedSAM corpus [46] sits relative to SA-1B and its successors.

4.4. Do Foundation Models Dissolve the Weak-Supervision Problem?

The central observation of this section is that foundation models change the location and form of human supervision rather than simply eliminating it. Following the classification in Section 4.1, we treat human-prompted SAM use as an inference-time form of weak supervision, automatic prompting as transfer learning or foundation-model adaptation, and SAM-based pseudo-labeling as part of a weak-supervision pipeline. The evidence cuts in both directions, and a survey that only celebrated FMs would mislead the reader interested in the annotation-economics argument of Section 2.

FMs can substantially reduce annotation time: MedSAM's $>$ 82% reduction in 3D tumor labeling [46] and ScribblePrompt's 28% saving [49] are large in the cited settings, and for high-contrast, large structures such as liver against fat or kidney against retroperitoneum, a box or a few clicks can already yield useful masks [46,49]. Generalization can also reduce task-specific training, since a single FM can serve multiple modalities and structures and shift part of the labeling effort toward prompt-based interaction, although upstream adaptation and validation costs remain [46,47]. The prompt itself can be a low-burden supervision unit as well: Table 6 shows that box or scribble prompts may require only seconds to minutes of interaction in the cited studies and can approach supervised references for selected targets.

Automatic prompts and the boundary of weak supervision. If a prompt is generated entirely by another model and no human interaction occurs at inference time, it does not satisfy this review's definition of weak supervision, because the inference-time signal is not explicitly authored by a human. Such a system is better described as transfer learning, foundation-model adaptation, or auto-prompting. This distinction matters for annotation economics. An auto-prompting module can reduce clinician interaction, but it does not remove supervision; it moves the supervision upstream, because the module itself must be trained, validated, and maintained. ABP-SAM [110] is a representative example: its Gated Vision Adapter and Auto Bbox Prompter reduce manual prompting, but the auto-prompter is itself trained on annotated data, so the human supervision is shifted upstream rather than removed. We therefore treat automatic prompting as a boundary case that can be part of a weakly supervised pipeline without being a weak-label type in its own right. The relevant question for the annotation-economics framework is not whether the prompt is generated automatically but where the human supervision enters the pipeline and how much it costs at that point.

The problem, however, is relocated rather than eliminated. A prompt is still human supervision, because someone must place the box, draw the scribble, or click the seed; the economy of Section 2 is therefore shifted from offline labeling to online interaction, and prompt design becomes part of supervision design [107,112]. The zero-shot gap also remains large for hard cases: vanilla SAM sits at $\approx$ 0.5 Dice under points and fails on tumors, small structures, and low-contrast boundaries [107,108,112], and even adapted FMs regress on unseen MRI sequences and small lesions (SAM-Med3D unseen-lesion Dice $\approx$ 48% vs. seen $\approx$ 50–67%) [48]. This residual gap is precisely the gap that Sections 2 and 3 tried to close with smarter weak labels. In addition, the upstream annotation cost is itself substantial: MedSAM's 1.57 M pairs [46] and SAM-Med2D's 19.7 M masks [47] are dense or densely curated annotations, so a large part of the effort is shifted to the construction and curation of medical training corpora. Evaluation can also hide the real cost, because most FM benchmarks prompt with a ground-truth-derived box [46,107], whereas a clinician must draw the box in practice, and prompt quality strongly affects the result [107]; the resulting Dice can therefore be more optimistic than performance under realistic clinician-generated prompts. Finally, 3D consistency remains open, since natively 2D FMs segment slices independently and 3D-specialized models, including the

3D scribble networks discussed in Section 3.3, still handle volumetric continuity better per FLOP [48,112].

Taken together, FMs transform weak supervision from a training-time optimization into an interaction-time one and raise the floor for easy structures, but the three questions this survey cares about persist: what does the cheap label cost, what accuracy does it buy, and is the residual gap clinically acceptable. FMs also make the Section 5 question, whether the weak segmentation is good enough for diagnosis, more urgent, because the same model now produces both the segmentation and, increasingly, the diagnostic suggestion. FMs and weak supervision are therefore complementary: FMs supply the backbone and generalization, while the cheap-label taxonomies of Sections 2 and 3 supply the interaction vocabulary and the economic yardstick. Neither retires the other.

4.5. Open Problems and Forward Look

Several open questions remain. Prompt-quality variance is under-reported: almost no FM study quantifies Dice as a function of realistic, non-ground-truth prompts [107]. The distribution of upstream supervision also deserves attention: the 1.57 M-pair MedSAM and 19.7 M-mask SAM-Med2D corpora [46,47], together with the approximately 143 K-mask SA-Med3D-140K corpus (131 K training masks) [48], concentrate substantial capability in large medical training corpora, while lightweight adaptations such as MobileSAM [114], EfficientViT [115], and MCP-MedSAM [116] provide one route toward lower deployment cost. Three-dimensional native prompting remains less mature than slice-wise 2D FM interaction [48]. Most importantly for this survey, the diagnostic closure of Section 5 remains insufficiently tested for FM outputs, since it is not yet established whether a 0.85-Dice FM mask carries the same diagnostic error profile as a fully supervised one or whether prompt-induced failures concentrate on clinically important lesions.

5. From Weak Segmentation to Diagnostic Decisions

Dice is not a sufficient endpoint because it is an aggregate, direction-agnostic overlap measure: a boundary shift that barely changes Dice can still alter the downstream measurement, as shown by Zhovannik et al. [117], where modest contour perturbations reduced the probability of a significant prognostic stratification from 90% to 28.5%. Dice is therefore used here as a segmentation-quality descriptor, not as evidence of diagnostic equivalence. The preceding sections showed that weak labels can retain substantial segmentation performance at lower annotation burden; this section asks whether the residual accuracy gap changes a downstream measurement or decision. We treat segmentation as the first stage of a quantitative-imaging pipeline and trace how annotation economy (Section 2), method choice (Section 3), and foundation models (Section 4) propagate, or fail to propagate, into diagnostic and prognostic output.

5.1. The Decision Chain: Why Segmentation Quality Is a Clinical Variable

Segmentation is rarely the clinical endpoint. In the dominant workflow it is an intermediate stage: image acquisition, pre-processing, segmentation, feature extraction or measurement, classification or risk stratification, and clinical action [117,118]. Error at any stage compounds downstream [118]. The same weak labels reviewed in Sections 2 and 3, such as a box around a fetal organ [17], a scribble on cardiac cine MRI [20,21], or a point annotation on a target structure [9], are the inputs to this chain, so the annotation economy of Section 2 is the first link in a diagnostic argument. Figure 9 illustrates this error-propagation chain from acquisition and segmentation to downstream assessment.

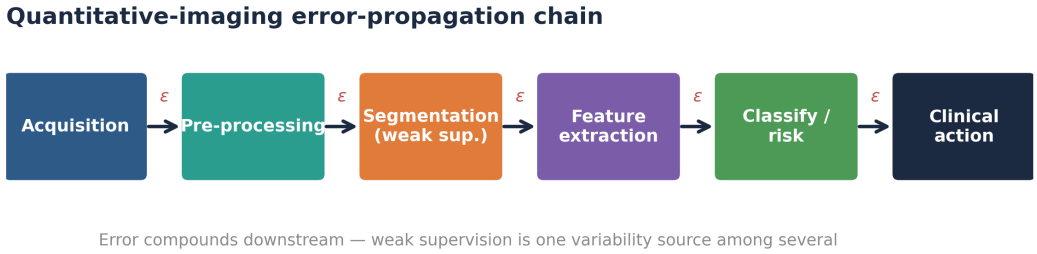


Figure 9. The quantitative-imaging error-propagation chain, inspired by the workflow discussion in a 2025 critical review of AI in cancer MRI [118]: variability at acquisition, registration, and segmentation propagates to lesion-response and outcome assessment. The figure's message for this survey is that weak supervision is one source of segmentation variability among several, and its clinical impact must be measured at the outcome end, not the Dice end.

5.2. Is Weak Segmentation Good Enough? Two Regimes

The evidence is better described as two task-dependent regimes rather than as a universal clinical threshold. Table 8 summarizes the downstream uses considered here and the corresponding sensitivity to segmentation error.

Table 8. Downstream tolerance to segmentation error (regime map). Each row pairs a clinical use with its sensitivity to the boundary noise that weak labels introduce.

Downstream Use	Representative Metric	Sensitivity to Seg. Error	Evidence
Aggregate volume/ diameter	tumor volume, LV EF, RECIST	Low to moderate (integral)	weak-supervision studies report high-80s to low-90s Dice on ACDC [20,21,39]; direct validation of volume/diameter error remains necessary
Shape descriptors	sphericity, compactness	Moderate	radiomics review: ±2 mm dilation or erosion shifts shape features measurably [119]
Texture (GLCM/GLRLM/GLSZM)	entropy, homogeneity, dissimilarity	High	Zhovannik [117]: high-sensitivity features significant in only 28.5% of perturbed realizations; GLSZM least stable in [120]
Survival/prognostic model	log-rank stratification, C-index	High if texture-based	Zhovannik [117]: 28.5% vs. 90% significant realizations across sensitivity groups

Regime A: aggregate geometric measurements may be relatively tolerant of moderate boundary error. Volume, diameter, and derived indices such as left-ventricular ejection fraction integrate information over the segmented region. The cited scribble studies report Dice values in the high-80s to low-90s on ACDC [20,21,39]; however, Dice alone does not establish that the resulting clinical measurements are equivalent to those obtained from manual masks. Aggregate measurements should therefore be validated directly against manual references and clinically relevant tolerances before weak supervision is used in practice.

Regime B: high-order and texture features can be boundary-sensitive. Here, the residual segmentation error may propagate strongly. Leijenaar et al. [120] reported substantial feature variability under test–retest and inter-observer delineation, showing that even expert contours can affect radiomic measurements. Zhovannik et al. [117] further used Monte-Carlo perturbations of gross-tumor-volume contours and found that, for a sensitivity-

defined feature group, only 1428 of 5000 realizations (28.5%) retained a significant prognostic stratification in one experimental setup, compared with 4508 of 5000 (90%) for the robust group. These results do not directly measure weak supervision, but they show that segmentation perturbations that are modest in Dice can still alter downstream radiomic or prognostic conclusions.

Downstream tolerance is therefore task-specific: aggregate measurements may be less sensitive to moderate boundary perturbations, whereas texture- and high-order-feature-based analyses can be substantially more sensitive. The appropriate weak-label budget should therefore be determined by the required downstream tolerance, not by Dice alone.

5.3. Mechanisms That Close the Loop: From Mask to Decision

The downstream evidence can be separated into direct and indirect links. A box- or scribble-segmented region can directly yield volume or diameter [17,20,21]. This is the most immediate path from weak segmentation to a quantitative endpoint, but the clinical validity of the measurement still needs to be checked against manual references and application-specific tolerances. Handcrafted or deep features can be extracted from a weak mask [55]. This route is more sensitive to contour perturbation and should be accompanied by feature-stability analysis, for example ICC under repeated or perturbed segmentations [117,120].

End-to-end joint segmentation and classification provide more direct evidence because the weakly supervised segmentation and downstream classifier are evaluated within the same pipeline. CardioRadNet [121] is particularly relevant: its point-guided segmentation achieved a Dice of 0.78, and 88% of radiomic features extracted from semiautomatic ROIs showed ICC > 0.6 relative to manual ROIs; the resulting benign-versus-malignant classifier achieved a balanced accuracy of 0.85. WML-Net [94] similarly couples weak segmentation and carotid-plaque classification without manual segmentation masks. These studies are included to illustrate downstream task evaluation rather than to compare weak-label types. These studies support the feasibility of the bridge, but they are application-specific and should not be generalized to all organs or diagnostic tasks. The representative links between spatial supervision, segmentation quality, and downstream endpoints are summarized in Table 9. The added rows in Table 9 strengthen the bridge in three ways. Leijenaar et al. [120] show that feature-level stability, not Dice alone, determines radiomic usability: 71% of features had high test–retest ICC and 91% had high inter-observer ICC, while GLSZM-based texture features were the least stable. The gaze-plus-VLM framework of Chen et al. [58] shows that a weakly supervised segmentation method can reach Dice values of 80.78, 80.53, and 84.22 on Kvasir-SEG, NCI-ISBI, and ISIC while using 2.2 h of annotation time instead of 18.7 h for full supervision. MSIM [122] shows that an interpretability module can improve downstream-oriented segmentation quality, with ablation gains of 1.6, 1.62, and 2.36 DSC on Melanoma, Lymphoma, and Lung cancer and improvements of 1.46, 1.27, and 1.93 over the best competing method.

Vision–language models such as CheXagent [123] demonstrate that image interpretation and report generation can be performed without an explicit segmentation stage. Its reader study reported a 36% time saving for residents using model-drafted reports, while attending-radiologist editing time did not differ significantly from resident-drafted reports. This is relevant as an adjacent endpoint rather than direct evidence that weak segmentation improves diagnosis. The gaze-plus-VLM framework of Chen et al. [58] is likewise a middle path linking human visual attention, VLM explanation, and weakly supervised segmentation.

Table 9. Representative evidence linking segmentation or weak spatial supervision to downstream clinical tasks. Studies differ in task, supervision, and endpoint and should not be read as a head-to-head comparison.

Study	Supervision of Segmentation	Seg. Quality	Downstream Task	Metric/Finding
Zhovannik et al. [117]	GTV manual, Monte-Carlo perturbed	surface dice tolerance ≤ 1.18 mm	NSCLC prognosis (Cox)	high-η group: 1428/5000 realizations with log-rank $p < 0.05$ (28.5%); low-η group: 4508/5000 (90%)
Leijenaar et al. [120]	Manual, test–retest and inter-observer	feature-level variability	Radiomic feature stability	71% high test–retest ICC, 91% high inter-observer ICC; GLSZM-based texture features least stable
IntelliCardiac [124]	(supervised, ACDC)	acc 92.6%	5-class cardiac disease	acc 98%
CardioRadNet [121]	point-guided (weak)	Dice 0.78	benign vs. malignant cardiac mass	balanced acc 0.85; 88% of ROI features ICC > 0.6 vs. manual
WML-Net [94]	weak (no segmentation masks)	pseudo-masks from attention	carotid plaque classification	$\approx$+2.98% accuracy over baseline
CheXagent [123]	VLM (no explicit mask)	—	CXR report drafting	resident time −36%; efficiency +81%/+61%
Gaze + VLM [58]	Gaze + text	weakly supervised segmentation	Kvasir-SEG/NCI-ISBI/ISIC	Dice 80.78/80.53/84.22; +3–5% over gaze baselines; 2.2 h vs. 18.7 h annotation time as reported in [58]
MSIM [122]	supervised PET-CT	ablation +1.6/1.62/2.36 DSC; +1.46/1.27/1.93 over best baseline	Melanoma/ Lymphoma/Lung cancer	multi-scale interpretability module

IntelliCardiac and CardioRadNet are included as downstream reference models to illustrate the diagnostic bridge, not as weak-supervision methods. MSIM is included as an interpretability-oriented downstream example and is fully supervised. Leijenaar et al. is included as evidence that feature-level stability, not Dice alone, determines downstream radiomic usability. The 2.2 h versus 18.7 h annotation-time figures in the Gaze + VLM row are taken from [58].

The downstream sensitivity can be expressed conceptually by treating the clinical quantity of interest as a function of the predicted mask. If M denotes the reference mask, $\hat{M}$ the weakly supervised mask, and $g(\cdot)$ the downstream measurement or decision score, then the induced downstream error is as follows:

$$E_{\text{down}} = d\big(g(\hat{M}), g(M)\big). \tag{21}$$

For example, if $g(\cdot)$ is tumor volume, then E_{down} is the absolute volume error; if $g(\cdot)$ is a radiomic feature, then E_{down} can be expressed as an intraclass correlation coefficient between the weak-mask and full-mask feature values. If $g(\cdot)$ is a prognostic score, then E_{down} can be expressed as the change in C-index or in log-rank stratification stability. This expression does not define a universal clinical threshold; rather, it makes explicit why the same segmentation error can be tolerable for one endpoint and unacceptable for another. In practice, E_{down} should be evaluated with an endpoint-specific metric such as volume error, ICC, AUC, or C-index stability, depending on the task.

Worked numerical example. To make Equation (21) concrete, consider the prognostic-stratification experiment of Zhovannik et al. [117]. Gross-tumor-volume contours were

perturbed with Monte-Carlo shifts within a clinically relevant surface dice tolerance of at most 1.18 mm, and a Cox proportional hazards model with median stratification was fitted for each of 5000 augmented realizations. For the high-η feature group, only 1428 of 5000 realizations (28.5%) achieved a log-rank $p < 0.05$ in the test set, whereas the low-η feature group achieved a log-rank $p < 0.05$ in 4508 of 5000 realizations (90%). If we define the downstream endpoint as the stability of prognostic stratification, then the induced downstream error from Equation (21) can be written as

$$E_{\text{down}}^{\text{prog}} = \left|\Pr(\text{log-rank } p < 0.05 \mid \hat{M}) - \Pr(\text{log-rank } p < 0.05 \mid M)\right|. \tag{22}$$

In this example, Equation (22) evaluates to $|0.285 - 0.90| = 0.615$ for the high-η group, compared with a much smaller value for the low-η group. This does not define a universal clinical threshold; it shows how a downstream endpoint can be used to quantify the clinical cost of weak or noisy segmentation. The same segmentation perturbation that is modest in surface Dice can therefore be tolerable for an aggregate volume measurement and harmful for a texture-driven prognostic model. The result also illustrates the value of the η ranking introduced in that study: features with low η are more robust to contour perturbation and are therefore more reliable for downstream prognostic modeling.

To show what the proposed protocol would change in practice, the evidence already summarized in Table 9 can be re-read under it. Ranked by Dice alone, the weakly supervised entries are close to their supervised references; ranked by a Regime-B endpoint, the same entries separate sharply. Zhovannik et al. show that a contour perturbation within a clinically relevant tolerance preserves prognostic stratification in only 28.5% of realizations for a sensitive feature group, against 90% for a robust group; Leijenaar et al. show that the same delineation variability affects GLSZM-based texture features far more than the most stable features; and CardioRadNet shows that a point-guided mask with Dice 0.78 can still yield ICC > 0.6 for 88% of ROI features and a downstream balanced accuracy of 0.85. Under the protocol, the first two studies would be scored as high downstream risk and the third as acceptable for its specific endpoint, a ranking that Dice alone does not produce. This is the sense in which the review offers a benchmark rather than a catalog.

5.4. Uncertainty as Diagnostic Usability

The weakest link in the bridge is knowing when to trust it, and here weak supervision meets the foundation models of Section 4. Bayesian Dropout [41], heteroscedastic uncertainty [42], and medical-specific reviews [43] provide established approaches for estimating uncertainty in segmentation models; these approaches can be applied to weakly supervised or FM-prompted outputs [41–44]. This has two consequences for the survey. Uncertainty can gate the human-in-the-loop: a weak mask with high boundary uncertainty on a lesion is precisely the case to refer to an expert; models that expose this, such as the interactive FM tools of Section 4.3 [46,49], turn the residual accuracy gap into a routed gap rather than a silent error. Uncertainty should also be reported at the outcome level rather than only at the Dice level. Recent methodological work [117–119] emphasizes that segmentation studies disclose feature-level agreement versus manual delineation and consistency of downstream task performance, because these, not Dice, reflect clinical impact. This is the reporting standard the present survey recommends for every weak-supervision claim.

5.5. Critical Evaluation: Where the Bridge Breaks

The bridge is not free, and several failure modes are particularly important. Texture and small-lesion failure remain important limitations. Regime B (Section 5.2) means weak segmentation can degrade prognostic and texture-driven diagnoses; the Zhovannik result [117] provides a quantitative warning. Image-level supervision provides limited spatial

information. Image-level signals are primarily localization or classification cues rather than dense segmentations [69]. They may support downstream tasks that do not require precise spatial measurements, but they are insufficient on their own when reliable volumetry, boundary-sensitive radiomics, or other spatially explicit measurements are required. Foundation-model prompt variance is another limitation (Section 4.4). A box prompt derived from ground truth is not a clinical prompt; under realistic, noisy prompts the FM mask, and therefore the downstream decision built on it, degrades in ways current benchmarks under-report. The recurring lesson is that segmentation error can propagate into downstream error, so the value of weak supervision should be assessed through measurable and auditable downstream effects rather than assumed to disappear.

5.6. Forward Look: Weak Segmentation to Diagnosis as a Benchmark

The field lacks a standard protocol that scores weak-supervision methods on downstream tasks rather than Dice. We propose it as the survey's concrete forward agenda: evaluate representative methods from Sections 2–4 on a Regime-A metric (volume or diameter error vs. manual) and a Regime-B metric (feature-level ICC and downstream AUC or C-index stability under perturbation, per [117,120]). Foundation models (Section 4) should be scored the same way, so that prompt type, budget, and downstream tolerance become a unified buying guide. Vision–language systems [123] suggest the longer arc: segmentation, weak or strong, may dissolve into the diagnostic act itself, but the accountability question of which region, which confidence, and which error survives the dissolution remains. The proposed benchmark should also include feature-stability endpoints of the kind reported by Leijenaar et al. [120], where the same delineation perturbation produces different ICC values across feature classes, and downstream-stratification endpoints of the kind reported by Zhovannik et al. [117], where the probability of a significant log-rank result changes from 90% to 28.5% depending on the sensitivity of the selected features. Concretely, the protocol requires each evaluated method to report the following: (i) the label type and the number of annotated cases; (ii) the data split and whether evaluation is external (E1) or single-dataset (E2); (iii) the prompt provenance for foundation-model methods; (iv) a Regime-A endpoint (volume or diameter error against a manual reference); (v) a Regime-B endpoint (feature-level ICC and downstream AUC or C-index stability under perturbation); and (vi) an uncertainty estimate at the outcome level. This checklist is what makes the protocol auditable rather than programmatic.

6. Challenges and Future Directions

Despite the rapid maturity of weakly supervised methods reviewed in Sections 2–5, the field is not yet at the point where cheap labels can be chosen on purely technical merit. Several gaps persist, each mapping back to an argument developed earlier.

6.1. Open Challenges

Benchmark fragmentation remains a major practical obstacle because the field lacks a shared evaluation protocol. As Table 4 (Section 3.3) shows, scribble methods on ACDC swing by several Dice points purely because of train, validation, and test partitions, and dataset-specific splits [57,125–128] make cross-paper comparison unreliable. Annotation economics (Section 2) cannot become a decision tool until cost and accuracy are reported under one protocol.

The diagnosis gap remains unresolved. Section 5 drew the distinction between Regime A (aggregate measures such as volume or ejection fraction, whose robustness to imprecise masks still requires separate validation) and Regime B (texture and high-order features, where weak masks measurably erode downstream signal [117]). Inter-observer variability

has been shown to affect radiomic features [120], but this does not establish that aggregate volume or ejection-fraction measurements are robust to weak or imprecise masks; that remains an open question and is precisely why we proposed the weak-to-task benchmark in Section 5.6. Present methods close the Dice gap far more convincingly than the diagnostic gap; image-level supervision, in particular, remains primarily a localization cue rather than a substitute for dense segmentation when spatially precise measurements are required [69].

Foundation models relocate rather than remove supervision. The analysis of Section 4, that promptable FMs shift rather than eliminate human supervision, carries a concrete cost: prompt quality and the upstream 1.57 M-pair and 19.7 M-mask supervision behind medical FMs [46,47], together with the approximately 143 K-mask SA-Med3D-140K corpus (131 K training masks) [48], are rarely accounted for in the cheap-label framing, and 3D prompt consistency is still open. The economic lens must therefore be extended to prompt cost, not just offline annotation. Trust and deployment remain central concerns. Clinical use requires calibrated, per-voxel uncertainty [41–44] reported at the outcome level, and most weak-supervision work still reports only Dice. Where data are scarce and privacy-sensitive, federated weak supervision [129–131] remains nascent and introduces additional integration challenges for promptable FMs.

Efficiency and multisensor deployment also require explicit consideration. As weak and foundation models move toward clinical deployment, computational cost becomes a first-class constraint, especially in multisensor and multimodal settings. The survey of Shu et al. [132] organizes efficient medical image segmentation into four streams—Mamba/state-space models, efficient foundation-model adaptation, lightweight architectures, and data-efficient learning—and provides a granular accuracy-versus-cost comparison across benchmarks. Representative efficient backbones, including EVSSD [133], HCMNet [134], LightSegMamba [135], MSTransBTS [136], DA-TransResUNet [137], and DPF-Net [138], are fully supervised inference-efficiency works rather than annotation-economics studies. They are cited here only because they reduce the training and inference terms of Equation (1), and therefore the total cost at which a given weak-label strategy can be deployed; they do not lower the cost of human annotation and should not be read as direct evidence of better annotation economics. For interpretability, MSIM [122] adds a multi-scale interpretability module to PET-CT tumor segmentation, improving Dice by 1.6–2.4 points while exposing the features driving each prediction. MFPNet [139] is likewise a fully supervised skin-lesion segmentation method and is cited here only as a supervised comparison, not as image-level weak-supervision evidence.

6.2. Future Directions

A unified weak-to-task benchmark would adopt the protocol proposed in Section 5.6: score representative methods on Regime-A volume error and Regime-B feature ICC and downstream stability [117,120], with FMs scored identically. This turns the annotation-economics buying guide built across Sections 2–4 into an operational standard. The checklist in Section 5.6 and the retrospective re-ranking in Section 5.3 together make this proposal concrete: a study can be scored against the six reporting items without new experiments, and the resulting ranking can be compared across label types and across foundation-model prompting strategies.

Human-in-the-loop and FM fusion provide a natural next step. ScribblePrompt-style interactive prompting [49] shows that a low-burden supervision unit can increasingly be an inference-time prompt; the productive research line is tight loops where an FM proposes and a clinician corrects, minimizing total expert time rather than offline masks. Gaze-plus-VLM frameworks [58] and auto-prompting modules such as ABP-SAM [110] are early examples of how human attention and machine semantics can be combined to reduce the

prompt burden. Federated weak supervision could combine the label-scarcity relief of weak labels with privacy-preserving training [129–131], especially for the long-tail anatomies and rare pathologies.

Active and similarity-aware annotation selection can further reduce the annotation budget. When even weak labels are costly, active learning can reduce the annotation budget further. Shu et al. [140] propose an image-similarity-based active learning framework for skin-lesion segmentation that selects the most representative images for expert annotation, maintaining segmentation accuracy even when the training set is reduced to 12.5% of the original. This is complementary to weak supervision: the former decides which images to label, the latter decides how cheaply to label them. Trustworthy reporting should make per-voxel uncertainty and outcome-level error a mandatory reporting axis, so that a residual accuracy gap is expressed as a routable diagnostic risk [41–44] rather than a silent accuracy loss.

Efficient and interpretable weak pipelines should combine the efficiency gains of Mamba/state-space and lightweight architectures [132–135] with the interpretability requirements of clinical diagnosis [122], so that weak-label pipelines are cheap to annotate, cheap to run, and easy to audit.

7. Conclusions

This review examined weakly and coarsely labeled medical image segmentation through an annotation-economics lens, emphasizing annotation time, expertise, scalability, variability, and task-specific performance rather than treating heterogeneous studies as a single numerical frontier. The cited evidence indicates that box and scribble supervision can provide useful cost–performance trade-offs in several settings, but the magnitude of the benefit depends on anatomy, modality, protocol, and evaluation metric. In particular, the reported gaps are sensitive to benchmark split, so a favorable trade-off observed in one study should be re-validated for each new anatomy, modality, and evaluation protocol rather than assumed to transfer. Foundation models further shift the problem toward prompt-based interaction: they can reduce repeated annotation effort, while still relying on human prompts and substantial upstream medical supervision. Finally, the downstream analysis shows that segmentation quality and diagnostic utility are not interchangeable. Aggregate measurements may be relatively tolerant of moderate boundary error, whereas radiomic, texture-sensitive, and small-lesion applications require stronger validation and uncertainty assessment. The resulting framework therefore treats weak supervision as a task-specific allocation problem: the appropriate label is the one whose human and computational cost is compatible with the required downstream error tolerance. Because the same weak-label type can occupy different positions on the economic spectrum depending on anatomy and modality, this allocation should be re-evaluated for each new task rather than fixed once for a label type.

Author Contributions: Conceptualization, X.S.; methodology, X.S. and X.Z.; validation, X.S. and X.Z.; formal analysis, X.S.; investigation, X.S.; resources, X.S. and D.Y.; data curation, X.S.; writing—original draft preparation, X.S. and X.Z.; writing—review and editing, X.S., D.Y. and X.Z.; visualization, X.S.; supervision, X.S. and X.Z.; project administration, X.S. and D.Y.; funding acquisition, X.S. and D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Humanities and Social Sciences Youth Foundation of Ministry of Education of China under Grant No. 25YJCZH216, by the Tertiary Education Scientific research project of Guangzhou Municipal Education Bureau under Grant No. 2024312222, by the National Natural Science Foundation of China under Grant No. 62202362, by the China Postdoctoral Science Foundation under Grant Nos. 2022TQ0247 and 2023M742742, by the Guangdong Basic and

Applied Basic Research Foundation under project grant Nos. 2024A1515011626 and 2025A1515012949, and by the Science and Technology Projects in Guangzhou under Grant No. 2023A04J0397.

Institutional Review Board Statement: Ethical review and approval were waived for this study because it is a narrative review of previously published literature and did not involve new human subjects, human material, human tissues, or identifiable human data.

Informed Consent Statement: Informed consent was waived for this study because no new human participants were recruited and no identifiable human data were collected or analyzed; all discussed studies are previously published and publicly available.

Data Availability Statement: No new datasets were created or analyzed in this review. Data sharing is not applicable to this article. The datasets and data sources discussed are reported in the cited studies and are identified in the corresponding sections of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
2. Çiçek, Ö; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Athens, Greece, 17–21 October 2016; pp. 424–432.
3. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. In Proceedings of the Medical Imaging with Deep Learning (MIDL), Amsterdam, The Netherlands, 4–6 July 2018; pp. 1–10.
4. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A nested U-Net architecture for medical image segmentation. In Proceedings of the Deep Learning in Medical Image Analysis (DLMIA/MICCAI), Granada, Spain, 20 September 2018; pp. 3–11.
5. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
6. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **2024**, *97*, 103280. [CrossRef]
7. Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.R.; Xu, D. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI. In Proceedings of the International MICCAI Brainlesion Workshop, Virtual, 27 September 2021; pp. 272–284.
8. Zhou, H.Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; Yu, Y. nnFormer: Volumetric medical image segmentation via a 3D transformer. *IEEE Trans. Image Process.* **2023**, *32*, 4036–4045. [CrossRef]
9. Roth, H.R.; Yang, D.; Xu, Z.; Wang, X.; Xu, D. Going to extremes: Weakly supervised medical image segmentation. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 507–524. [CrossRef]
10. Bearman, A.; Russakovsky, O.; Ferrari, V.; Fei-Fei, L. What's the point: Semantic segmentation with point supervision. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 549–565.
11. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2097–2106.
12. Dai, J.; He, K.; Sun, J. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1635–1643.
13. Hsu, C.C.; Hsu, K.J.; Tsai, C.C.; Lin, Y.Y.; Chuang, Y.Y. Weakly supervised instance segmentation using the bounding box tightness prior. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 6582–6593.
14. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NA, USA, 26 June–1 July 2016; pp. 2921–2929.
15. Kolesnikov, A.; Lampert, C.H. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 695–711.

16. Papandreou, G.; Chen, L.C.; Murphy, K.P.; Yuille, A.L. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1742–1750.
17. Rajchl, M.; Lee, M.C.; Oktay, O.; Kamnitsas, K.; Passerat-Palmbach, J.; Bai, W.; Damodaram, M.; Rutherford, M.A.; Hajnal, J.V.; Kainz, B.; et al. DeepCut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Trans. Med. Imaging* **2017**, *36*, 674–683. [CrossRef]
18. Lin, D.; Dai, J.; Jia, J.; He, K.; Sun, J. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3159–3167.
19. Lee, H.; Jeong, W.K. Scribble2Label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Virtual, 4–8 October 2020; pp. 14–23.
20. Zhang, K.; Zhuang, X. CycleMix: A holistic strategy for medical image segmentation from scribble supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11646–11655.
21. Li, Z.; Zheng, Y.; Shan, D.; Yang, S.; Li, Q.; Wang, B.; Zhang, Y.; Hong, Q.; Shen, D. Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation. *IEEE Trans. Med. Imaging* **2024**, *43*, 2254–2265. [CrossRef]
22. Wang, G.; Li, W.; Zuluaga, M.A.; Pratt, R.; Patel, P.A.; Aertsen, M.; Doel, T.; David, A.L.; Deprest, J.; Ourselin, S.; et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Trans. Med. Imaging* **2018**, *37*, 1562–1573. [CrossRef]
23. Wang, G.; Zuluaga, M.A.; Li, W.; Pratt, R.; Patel, P.A.; Aertsen, M.; Doel, T.; David, A.L.; Deprest, J.; Ourselin, S.; et al. DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1559–1572. [CrossRef]
24. Sofiiuk, K.; Petrov, I.A.; Konushin, A. Reviving iterative training with mask guidance for interactive segmentation. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 3141–3145.
25. Liu, Q.; Xu, Z.; Bertasius, G.; Niethammer, M. SimpleClick: Interactive image segmentation with simple vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 22233–22243.
26. Zhong, Y.; Tang, C.; Yang, Y.; Qi, R.; Zhou, K.; Gong, Y.; Heng, P.A.; Hsiao, J.H.; Dou, Q. Weakly-supervised medical image segmentation with gaze annotations. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Marrakesh, Morocco, 6–10 October 2024; pp. 530–540.
27. Fadida, B.S.; Sourani, D.L.; Sira, L.B.; Miller, E.; Bashat, D.B.; Joskowicz, L. Partial annotations for the segmentation of large structures with low annotation cost. In Proceedings of the MICCAI Workshop on Medical Image Learning with Limited and Noisy Data (MILLanD), Singapore, 22 September 2022; pp. 13–22.
28. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 1195–1204.
29. Yu, L.; Wang, S.; Li, X.; Fu, C.W.; Heng, P.A. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Shenzhen, China, 13–17 October 2019; pp. 605–613.
30. Chen, L.; Bentley, P.; Mori, K.; Misawa, K.; Fujiwara, M.; Rueckert, D. Self-supervised learning for medical image analysis using image context restoration. *Med. Image Anal.* **2019**, *58*, 101539. [CrossRef]
31. Zhou, Z.; Sodha, V.; Rahman Siddiquee, M.M.; Feng, R.; Tajbakhsh, N.; Gotway, M.B.; Liang, J. Models Genesis: Generic autodidactic models for 3D medical image analysis. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Shenzhen, China, 13–17 October 2019; pp. 384–393.
32. Haghighi, F.; Taher, M.R.H.; Zhou, Z.; Gotway, M.B.; Liang, J. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Trans. Med. Imaging* **2021**, *40*, 2857–2868. [CrossRef]
33. Tang, Y.; Yang, D.; Li, W.; Roth, H.R.; Landman, B.; Xu, D.; Nath, V.; Hatamizadeh, A. Self-supervised pre-training of Swin transformers for 3D medical image analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 20698–20708.
34. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 12–18 July 2020; pp. 1597–1607.
35. Cheplygina, V.; de Bruijne, M.; Pluim, J.P.W. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* **2019**, *54*, 280–296. [CrossRef]

36. Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J.N.; Wu, Z.; Ding, X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med. Image Anal.* **2020**, *63*, 101693. [CrossRef]
37. Kumari, S.; Singh, P. Learning with less: A survey of deep learning in medical imaging under varying supervision levels. *Artif. Intell. Med.* **2026**, *175*, 103375. [CrossRef]
38. Shen, W.; Peng, Z.; Wang, X.; Wang, H.; Cen, J.; Jiang, D.; Xie, L.; Yang, X.; Tian, Q. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9284–9305. [CrossRef]
39. Valvano, G.; Leo, A.; Tsaftaris, S.A. Learning to segment from scribbles using multi-scale adversarial attention gates. *IEEE Trans. Med. Imaging* **2021**, *40*, 1990–2001. [CrossRef]
40. Gao, F.; Hu, M.; Zhong, M.E.; Feng, S.; Tian, X.; Meng, X.; Huang, Z.; Lv, M.; Song, T.; Zhang, X.; et al. Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images. *Med. Image Anal.* **2022**, *80*, 102515. [CrossRef]
41. Gal, Y. Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
42. Kendall, A.; Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5574–5584.
43. Lambert, B.; Forbes, F.; Doyle, S.; Dehaene, H.; Dojat, M. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. *Artif. Intell. Med.* **2024**, *150*, 102830. [CrossRef]
44. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R.; et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* **2021**, *76*, 243–297. [CrossRef]
45. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 3992–4003.
46. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nat. Commun.* **2024**, *15*, 654. [CrossRef]
47. Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. SAM-Med2D[EB/OL]. *arXiv* **2023**, arXiv:2308.16184.
48. Wang, H.; Guo, S.; Ye, J.; Deng, Z.; Cheng, J.; Li, T.; Chen, J.; Su, Y.; Huang, Z.; Shen, Y.; et al. SAM-Med3D: A vision foundation model for general-purpose segmentation on volumetric medical images. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 17599–17612. [CrossRef]
49. Wong, H.E.; Rakic, M.; Guttag, J.; Dalca, A.V. ScribblePrompt: Fast and flexible interactive segmentation for any biomedical image. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 207–229.
50. Johnson, A.E.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.Y.; Mark, R.G.; Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **2019**, *6*, 317. [CrossRef]
51. Kuang, Z.; Yan, Z.; Zhou, H.; Yu, L. Cluster-Re-Supervision: Bridging the gap between image-level and pixel-wise labels for weakly supervised medical image segmentation. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 4890–4901. [CrossRef]
52. Zhang, X.; Wu, S.; Zhang, P.; Jin, Z.; Xiong, X.; Bu, Q.; Chen, J.; Feng, J. HELPNet: Hierarchical perturbations consistency and entropy-guided ensemble for scribble supervised medical image segmentation. *Med. Image Anal.* **2025**, *103*, 103719. [CrossRef]
53. Xu, N.; Price, B.; Cohen, S.; Yang, J.; Huang, T.S. Deep interactive object selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NA, USA, 26 June–1 July 2016; pp. 373–381.
54. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1422–1430.
55. Han, C.; Lin, J.; Mai, J.; Wang, Y.; Zhang, Q.; Zhao, B.; Chen, X.; Pan, X.; Shi, Z.; Xu, Z.; et al. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Med. Image Anal.* **2022**, *80*, 102487. [CrossRef]
56. Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; Schiele, B. Simple does it: Weakly supervised instance and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1665–1674.
57. Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.A.; Cetin, I.; Lekadir, K.; Camara, O.; Ballester, M.A.G.; et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans. Med. Imaging* **2018**, *37*, 2514–2525. [CrossRef]
58. Chen, J.; Duan, H.; Zhang, X.; Gao, B.; Grau, V.; Han, J. From gaze to insight: Bridging human visual attention and vision language model explanation for weakly-supervised medical image segmentation. *IEEE Trans. Med. Imaging* **2026**, *45*, 1051–1061. [CrossRef]
59. Shu, X.; Li, Z.; Chang, X.; Yuan, D. Variational methods with application to medical image segmentation: A survey. *Neurocomputing* **2025**, *639*, 130260. [CrossRef]

60. Nguyen, T.H.; Cao, H.L.; Chung, D.T.; Vu, M.A.; Nguyen, T.M.; Le, M.; Huynh, P.K.; Bagci, U. Scribble-supervised medical image segmentation with dynamic teacher switching and hierarchical consistency. In *Proceedings of the 2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*; IEEE: New York, NY, USA, 2026; pp. 1–5.
61. Zhang, X.; Wei, Y.; Feng, J.; Yang, Y.; Huang, T.S. Adversarial complementary learning for weakly supervised object localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1325–1334.
62. Huang, Z.; Wang, X.; Wang, J.; Liu, W.; Wang, J. Weakly-supervised semantic segmentation network with deep seeded region growing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7014–7023.
63. Ahn, J.; Kwak, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4981–4990.
64. Tang, M.; Perazzi, F.; Djelouah, A.; Ben Ayed, I.; Schroers, C.; Boykov, Y. On regularized losses for weakly-supervised CNN segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 507–522.
65. Sun, G.; Wang, W.; Dai, J.; Van Gool, L. Mining cross-image semantics for weakly supervised semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 347–365.
66. Jia, Z.; Huang, X.; Eric, I.; Chang, C.; Xu, Y. Constrained deep weak supervision for histopathology image segmentation. *IEEE Trans. Med. Imaging* **2017**, *36*, 2376–2388. [CrossRef]
67. Su, G.; Zheng, K.; Wang, Y.; Chen, J.; Bing, S.; Deng, S.; Jiang, H. DISCERN: Discrepancy Learning for Weakly Supervised Medical Image Segmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2026; pp. 6966–6970.
68. Zhang, W.; Chen, J.; Kanan, C. INSIGHT: Explainable weakly-supervised medical image analysis. In Proceedings of the 10th Machine Learning for Healthcare Conference, Rochester, MN, USA, 15–16 August 2025; Volume 298, pp. 1–28.
69. Saporta, A.; Gui, X.; Agrawal, A.; Pareek, A.; Truong, S.Q.; Nguyen, C.D.; Ngo, V.D.; Seekins, J.; Blankenberg, F.G.; Ng, A.Y.; et al. Benchmarking saliency methods for chest X-ray interpretation. *Nat. Mach. Intell.* **2022**, *4*, 867–878. [CrossRef]
70. Roth, H.R.; Lu, L.; Farag, A.; Shin, H.C.; Liu, J.; Turkbey, E.B.; Summers, R.M. DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 556–564.
71. Chen, X.; Yao, L.; Zhang, Y. RAU: An Interpretable Automatic Infection Diagnosis of COVID-19 Pneumonia with Residual Attention U-Net. In Proceedings of the International Conference on Web Information Systems Engineering, Melbourne, Australia, 26–29 October 2021; pp. 122–136.
72. Yang, R.; Song, L.; Ge, Y.; Li, X. BoxSnake: Polygonal instance segmentation with box supervision. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 766–776.
73. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3131–3140.
74. Redekop, E.; Chernyavskiy, A. Medical image segmentation with imperfect 3D bounding boxes. In Proceedings of the MICCAI Workshop on Data Augmentation, Labelling, and Imperfections (DALI), Virtual, 1 October 2021; pp. 193–200.
75. Jiang, R.; Fan, Z.; Wu, J.; Zhu, L.; Huang, X.; Wang, T.; Huang, H.; Xu, M. Enhancing weakly supervised 3D medical image segmentation through probabilistic-aware learning. *Pattern Recognit.* **2026**, *180*, 113973. [CrossRef]
76. Luo, X.; Hu, M.; Liao, W.; Zhai, S.; Song, T.; Wang, G.; Zhang, S. Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Singapore, 18–22 September 2022; pp. 528–538.
77. Li, Z.; Zheng, Y.; Luo, X.; Shan, D.; Hong, Q. ScribbleVC: Scribble-supervised medical image segmentation with vision-class embedding. In Proceedings of the ACM International Conference on Multimedia (ACM MM), Ottawa, ON, Canada, 29 October–3 November 2023; pp. 3384–3393.
78. Zhou, M.; Xu, Z.; Zhou, K.; Tong, R.K.Y. Weakly supervised medical image segmentation via superpixel-guided scribble walking and class-wise contrastive regularization. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Vancouver, BC, Canada, 8–12 October 2023; pp. 137–147.
79. Liu, T.; Qiao, L.; Shu, Y.; Li, W.; Gao, X. SRPA: ScribbleMatch and reliable-guided pixel alignment for scribble-supervised medical image segmentation. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 5–8 December 2023; pp. 1325–1330.

80. Lei, Y.; Luo, H.; Wang, L.; Zhang, Z.; Zhang, L. PCLMix: Weakly supervised medical image segmentation via pixel-level contrastive learning and dynamic mix augmentation. In Proceedings of the International Conference on Intelligent Computing (ICIC), Tianjin, China, 5–8 August 2024; pp. 62–73.
81. Han, M.; Luo, X.; Liao, W.; Zhang, S.; Zhang, S.; Wang, G. Scribble-based 3D multiple abdominal organ segmentation via triple-branch multi-dilated network with pixel- and class-wise consistency. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Vancouver, BC, Canada, 8–12 October 2023; pp. 33–42.
82. Chen, Q.; Hong, Y. Scribble2D5: Weakly-supervised volumetric image segmentation via scribble annotations. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Singapore, 18–22 September 2022; pp. 234–243.
83. Chen, J.; Huang, W.; Zhang, J.; Debattista, K.; Han, J. Addressing inconsistent labeling with cross image matching for scribble-based medical image segmentation. *IEEE Trans. Image Process.* **2025**, *34*, 842–853. [CrossRef]
84. Obukhov, A.; Georgoulis, S.; Dai, D.; Van Gool, L. Gated CRF loss for weakly supervised semantic image segmentation. *arXiv* **2019**, arXiv:1906.04651.
85. Ji, Z.; Shen, Y.; Ma, C.; Gao, M. Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Shenzhen, China, 13–17 October 2019; pp. 175–183.
86. Zhang, K.; Zhuang, X. ShapePU: A new PU learning framework regularized by global consistency for scribble supervised cardiac segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Singapore, 18–22 September 2022; pp. 162–172.
87. Asad, M.; Fidon, L.; Vercauteren, T. ECONet: Efficient convolutional online likelihood network for scribble-based interactive segmentation. In Proceedings of the Medical Imaging with Deep Learning (MIDL), Zurich, Switzerland, 6–8 July 2022; pp. 35–47.
88. Zhong, Y.; Wang, Y. SimPLe: Similarity-aware propagation learning for weakly-supervised breast cancer segmentation in DCE-MRI. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Vancouver, BC, Canada, 8–12 October 2023; pp. 567–577.
89. Liu, X.; Yuan, Q.; Gao, Y.; He, K.; Wang, S.; Tang, X.; Tang, J.; Shen, D. Weakly supervised segmentation of COVID-19 infection with scribble annotation on CT images. *Pattern Recognit.* **2022**, *122*, 108341. [CrossRef]
90. Yang, Z.; Lin, D.; Ni, D.; Wang, Y. Non-iterative scribble-supervised learning with pacing pseudo-masks for medical image segmentation. *Expert Syst. Appl.* **2024**, *238*, 122024. [CrossRef]
91. Wang, Z.; Ye, Y.; Chen, Z.; Shu, M.; Zhang, Y.; Xia, Y. From few to more: Scribble-based medical image segmentation via masked context modeling and continuous pseudo labels. *IEEE J. Biomed. Health Inform.* **2026**, *30*, 2419–2431. [CrossRef]
92. Nguyen-Tat, T.B.; Vo H.A.; Dang, P.S. QMaxViT-Unet+: A query-based MaxViT-Unet with edge enhancement for scribble-supervised segmentation of medical images. *Comput. Biol. Med.* **2025**, *187*, 109762. [CrossRef]
93. Shi, J.; Zhang, K.; Guo, C.; Yang, Y.; Xu, Y.; Wu, J. A survey of label-noise deep learning for medical image analysis. *Med. Image Anal.* **2024**, *95*, 103166. [CrossRef]
94. Gan, H.; Liu, L.; Wang, F.; Yang, Z.; Zhou, R.; Xu, X. WML-Net: Weakly supervised multi-task learning network for carotid plaque segmentation and classification. *Biomed. Signal Process. Control* **2026**, *127*, 111057. [CrossRef]
95. Qu, H.; Wu, P.; Huang, Q.; Yi, J.; Yan, Z.; Li, K.; Riedlinger, G.M.; De, S.; Zhang, S.; Metaxas, D.N. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE Trans. Med. Imaging* **2020**, *39*, 3655–3666. [CrossRef]
96. Krähenbühl, P.; Koltun, V. Efficient inference in fully connected CRFs with Gaussian edge potentials. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Granada, Spain, 12–14 December 2011; pp. 109–117.
97. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–13.
98. Verma, V.; Kawaguchi, K.; Lamb, A.; Kannala, J.; Solin, A.; Bengio, Y.; Lopez-Paz, D. Interpolation consistency training for semi-supervised learning. *Neural Netw.* **2022**, *145*, 90–106. [CrossRef]
99. Chen, X.; Yuan, Y.; Zeng, G.; Wang, J. Semi-supervised semantic segmentation with cross pseudo supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2613–2622.
100. Luo, X.; Chen, J.; Song, T.; Wang, G. Semi-supervised medical image segmentation through dual-task consistency. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Virtual, 2–9 February 2021; pp. 8801–8809.
101. Wu, Y.; Ge, Z.; Zhang, D.; Xu, M.; Zhang, L.; Xia, Y.; Cai, J. Mutual consistency learning for semi-supervised medical image segmentation. *Med. Image Anal.* **2022**, *81*, 102530. [CrossRef]
102. Luo, X.; Wang, G.; Liao, W.; Chen, J.; Song, T.; Chen, Y.; Zhang, S.; Metaxas, D.N.; Zhang, S. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Med. Image Anal.* **2022**, *80*, 102517. [CrossRef]

103. Wu, Y.; Wu, Z.; Wu, Q.; Ge, Z.; Cai, J. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Singapore, 18–22 September 2022; pp. 34–43.
104. Basak, H.; Yin, Z. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 19786–19797.
105. Miao, J.; Chen, C.; Liu, F.; Wei, H.; Heng, P.A. CauSSL: Causality-inspired semi-supervised learning for medical image segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 21369–21380.
106. Shen, Z.; Cao, P.; Yang, H.; Liu, X.; Yang, J.; Zaiane, O.R. Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 19–25 August 2023; pp. 4199–4207.
107. He, S.; Bao, R.; Li, J.; Stout, J.; Bjornerud, A.; Grant, P.E.; Ou, Y. Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets. *arXiv* **2023**, arXiv:2304.09324.
108. Wu, J.; Wang, Z.; Hong, M.; Ji, W.; Fu, H.; Xu, Y.; Xu, M.; Jin, Y. Medical SAM adapter: Adapting segment anything model for medical image segmentation. *Med. Image Anal.* **2025**, *102*, 103547. [CrossRef]
109. Nazzal, W.; Thurnhofer-Hemsi, K.; López-Rubio, E. Improving medical image segmentation using test-time augmentation with medsam. *Mathematics* **2024**, *12*, 4003. [CrossRef]
110. Xiong, Y.; Shu, X.; Zhu, H.; Guo, G.; Wei, P.; Yuan, D. ABP-SAM: Auto bbox prompter for medical image segmentation with segment anything model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2026; pp. 6741–6745.
111. Zhu, J.; Hamdi, A.; Qi, Y.; Jin, Y.; Wu, J. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv* **2024**, arXiv:2408.00874.
112. Mazurowski, M.A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; Zhang, Y. Segment anything model for medical image analysis: An experimental study. *Med. Image Anal.* **2023**, *89*, 102918. [CrossRef]
113. Ravi, N.; Gabeur, V.; Hu, Y.T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. Sam 2: Segment anything in images and videos. In Proceedings of the International Conference on Learning Representations, Singapore, 24–28 April 2025; pp. 28085–28128.
114. Zhang, C.; Han, D.; Qiao, Y.; Kim, J.U.; Bae, S.H.; Lee, S.; Hong, C.S. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv* **2023**, arXiv:2306.14289.
115. Cai, H.; Li, J.; Hu, M.; Gan, C.; Han, S. EfficientViT: Lightweight multi-scale attention for high-resolution dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 17256–17267.
116. Lyu, D.; Gao, R.; Staring, M. MCP-MedSAM: A powerful lightweight medical segment anything model trained with a single GPU in just one day. *arXiv* **2024**, arXiv:2412.05888.
117. Zhovannik, I.; Bontempi, D.; Romita, A.; Pfaehler, E.; Primakov, S.; Dekker, A.; Bussink, J.; Traverso, A.; Monshouwer, R. Segmentation uncertainty estimation as a sanity check for image biomarker studies. *Cancers* **2022**, *14*, 1288. [CrossRef]
118. Wu, C.; Andaloussi, M.A.; Hormuth, D.A.; Lima, E.A.; Lorenzo, G.; Stowers, C.E.; Ravula, S.; Levac, B.; Dimakis, A.G.; Tamir, J.I.; et al. A critical assessment of artificial intelligence in magnetic resonance imaging of cancer. *npj Imaging* **2025**, *3*, 15. [CrossRef]
119. Neha, F.; Shukla, D.K. Radiomics in Medical Imaging: Methods, Applications, and Challenges. *J. Imaging* **2026**, *12*, 220. [CrossRef]
120. Leijenaar, R.T.; Carvalho, S.; Velazquez, E.R.; Van Elmpt, W.J.; Parmar, C.; Hoekstra, O.S.; Hoekstra, C.J.; Boellaard, R.; Dekker, A.L.; Gillies, R.J.; et al. Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability. *Acta Oncol.* **2013**, *52*, 1391–1397. [CrossRef]
121. Ferretti, M.; Pagliaccia, M.; Baggiano, A.; Lovato, L.; Angeli, F.; Armillotta, M.; Bergamaschi, L.; Pizzi, C.; Pontone, G.; Corino, V.D. CardioRadNet: Cardiac mass diagnosis through integrated segmentation and radiomic analysis. *Comput. Methods Programs Biomed.* **2026**, *284*, 109448. [CrossRef]
122. Yang, D.; Wang, Y.; Ma, Y.; Yang, H. A multi-scale interpretability-based PET-CT tumor segmentation method. *Mathematics* **2025**, *13*, 1139. [CrossRef]
123. Chen, Z.; Varma, M.; Delbrouck, J.B.; Paschali, M.; Blankemeier, L.; Van Veen, D.; Valanarasu, J.M.J.; Youssef, A.; Cohen, J.P.; Reis, E.P.; et al. CheXagent: Towards a foundation model for chest X-ray interpretation. In Proceedings of the AAAI 2024 Spring Symposium on Clinical Foundation Models, Stanford, CA, USA, 25–27 March 2024.
124. Tsai, T.Y.; Yu, A.; Maadugundu, M.S.; Mohima, I.J.; Barsha, U.H.; Chen, M.H.F.; Prabhakaran, B.; Chana, M.C. IntelliCardiac: An intelligent platform for cardiac image segmentation and classification. In Proceedings of the IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 6–8 August 2025; pp. 109–115.
125. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2014**, *34*, 1993–2024. [CrossRef]

126. Codella, N.; Rotemberg, V.; Tsch, l.P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv* **2019**, arXiv:1902.03368.
127. Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; Klein, A. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In Proceedings of the MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, Munich, Germany, 9 October 2015; Volume 5, p. 12.
128. Bilic, P.; Christ, P.; Li, H.B.; Vorontsov, E.; Ben-Cohen, A.; Kaissis, G.; Szeskin, A.; Jacobs, C.; Mamani, G.E.H. The liver tumor segmentation benchmark (lits). *Med. Image Anal.* **2023**, *84*, 102680. [CrossRef]
129. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
130. Sheller, M.J.; Edwards, B.; Reina, G.A.; Martin, J.; Pati, S.; Kotrotsou, A.; Milchenko, M.; Xu, W.; Marcus, D.; Colen, R.R.; et al. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **2020**, *10*, 12598. [CrossRef]
131. Xu, X.; Deng, H.H.; Chen, T.; Kuang, T.; Barber, J.C.; Kim, D.; Gateno, J.; Xia, J.J.; Yan, P. Federated cross learning for medical image segmentation. In Proceedings of the Medical Imaging with Deep Learning (MIDL), Nashville, TN, USA, 10–12 July 2023; pp. 1441–1452.
132. Shu, X.; Xiong, Y.; Ma, Z.; Zhang, X.; Yuan, D. Efficient medical image segmentation in multisensor imaging: A survey in the era of Mamba and foundation models. *Sensors* **2026**, *26*, 2558. [CrossRef]
133. Yuan, D.; Xiong, Y.; Shu, X.; Xi, Y.; Chang, X. EVSSD: Efficient visual state space decoding for 2D medical image segmentation. *Signal Process.* **2026**, *249*, 110760. [CrossRef]
134. Xiong, Y.; Shu, X.; Liu, Q.; Yuan, D. HCMNet: A hybrid CNN-Mamba network for breast ultrasound segmentation for consumer assisted diagnosis. *IEEE Trans. Consum. Electron.* **2025**, *71*, 8045–8054. [CrossRef]
135. Kim, J.; Kim, J.; Dharejo, F.A.; Abbas, Z.; Lee, S.W. Lightweight mamba model for 3D tumor segmentation in automated breast ultrasounds. *Mathematics* **2025**, *13*, 2553. [CrossRef]
136. Ngu, J.Q.; Nisar, H.; Tsai, C.Y. MSTransBTS—A novel integration of Mamba with Swin Transformer for 3D brain tumour segmentation. *Mathematics* **2025**, *13*, 1117. [CrossRef]
137. Wang, K.; Lu, X.; Li, J.; Lu, Y. DA-TransResUNet: Residual U-Net Liver Segmentation Model Integrating Dual Attention of Spatial and Channel with Transformer. *Mathematics* **2026**, *14*, 575. [CrossRef]
138. Shi, L.; He, W.; Wang, G. A Dual-Path Fusion Network with Edge Feature Enhancement for Medical Image Segmentation. *Mathematics* **2026**, *14*, 55. [CrossRef]
139. Xiong, Y.; Yuan, D.; Li, L.; Shu, X. Mfpnet: Mixed feature perception network for automated skin lesion segmentation. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Shanghai, China, 18–20 October 2024; pp. 105–117.
140. Shu, X.; Li, Z.; Tian, C.; Chang, X.; Yuan, D. An active learning model based on image similarity for skin lesion segmentation. *Neurocomputing* **2025**, *630*, 129690. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.