KNOWING WHEN NOT TO SEGMENT: LABEL-FREE ADAPTATION IN FREE-HAND ULTRASOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-training and other label-free adaptation methods for segmentation usually assume that every image contains the target structure, but in videos and sweeps the structure is often out of view. Wireless handheld probes bring ultrasound out of the hospital, but many of their users are not medical professionals and cannot judge whether a model's prediction is correct. Segmentation models trained on public data from cart-based scanners segment structures in handheld images well when they are in view, but they do not know when not to segment. We study how to teach such a model, without target labels, to predict nothing when a structure is not in the frame. We built a dataset of free-hand thyroid sweeps with a wireless handheld probe, which we will release. It contains 20,014 frames from 29 adults, 232 of which are annotated for thyroid, carotid artery and trachea. Of these, 47% show no identifiable target structure, and two clinicians annotated the 58 held-out frames of its hidden test set. Our experiments reveal two problems. First, current evaluation measures only how accurately a structure is drawn and ignores frames without it, so it cannot show whether a model chooses not to draw when the target structure is absent. Second, because existing models are trained on datasets in which nearly every image contains the target structure, they also draw it on frames where it is absent. We therefore propose an evaluation that also scores predicting nothing. Frames with a structure are scored by how accurately it is drawn and frames without it by whether nothing is drawn, with equal weight, so predicting nothing scores 0.50. We also add anatomical constraints to self-training. Predicted structures are checked against rules on their position, shape, echogenicity and relations to the other structures, and implausible predictions are not used for training. The model thereby learns to predict nothing. On the evaluation and held-out frames, false thyroid detections fall from 99% and 90% to 40% and 29%, and the score reaches 0.59 and 0.61, significantly above predicting nothing and above the seven other adaptation methods that start from the same source models, while mean segmentation accuracy on frames with the structure does not drop.

1 INTRODUCTION

Self-training is a common way to adapt a pretrained segmentation model to unlabelled data: the model labels the new images, discards the unreliable parts of these pseudo-labels and retrains on the rest. Like other label-free adaptation methods, it is developed and evaluated almost exclusively on images that contain the target structure. In videos and sweeps, however, the structure is often out of view, and a model must also know when not to segment.

Handheld ultrasound is a clear and high-stakes instance of this problem. Wireless probes that stream images to a tablet now take ultrasound out of the hospital, into community screening, primary care and even the home. Because the imaging physics is the same, segmentation models trained on public data from cart-based hospital scanners also outline the thyroid and its neighbouring structures in handheld images reasonably well when they are in view.

Yet these models draw the structures just as readily when they are not in view. Nearly every image in public cart-based datasets contains the target, and a thyroid model trained on such data draws a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

thyroid on almost every handheld frame without one. Training on public data that include empty images does not solve this: a model that leaves 90% of such images empty on its own data still draws a thyroid on 80% of the handheld frames without one. In the hospital this matters little, because clinicians can judge whether a drawn region is really the thyroid. Many users of handheld devices, however, are not medical professionals and cannot tell right predictions from wrong ones, so a structure drawn where there is none misleads them.

Self-training does not correct this, because these false detections are exactly what its filters keep. They are confident, with a median maximum thyroid probability of 0.99 on frames without a thyroid, and they recur in consecutive frames, so confidence thresholds, augmentation agreement and agreement between adjacent frames all retain them as pseudo-labels. After self-training, the model still draws a thyroid on 68–74% of the frames without one.

Standard evaluation does not reveal this failure, because it scores only frames that contain the structure. Across 16 label-free methods on our sweeps, the ranking by Dice on such frames is almost unrelated to the ranking once frames without the structure are also scored (Kendall's τ of -0.02 and 0.20), and two common test-time adaptation methods that look like improvements under the usual evaluation score below a model that predicts nothing.

We therefore ask: *when a segmentation model trained on cart-based ultrasound data meets frames in which the target structure is absent, how can it learn, without target labels, to predict nothing?*

We argue that the decision whether to draw a structure must rest on evidence that does not come from the model's own confidence and that applies to whole structures rather than pixels. In neck ultrasound, anatomy provides such evidence: the position, shape and echogenicity of the thyroid, carotid artery and trachea, and their positions relative to each other, are largely fixed. We therefore check each structure the teacher draws against anatomical rules and exclude implausible ones from the pseudo-labels, so that the student learns to predict nothing when a structure is absent. To measure this, we score frames with a structure by Dice and frames without it by whether nothing is drawn, with equal weight, so that a model that predicts nothing scores 0.50 and any useful method must exceed it.

On free-hand thyroid sweeps that we recorded with a wireless handheld probe, the constraints reduce false thyroid detections from 99% and 90% of the frames without a thyroid to 40% and 29% on the evaluation and held-out frames, and raise balanced Dice from 0.44 and 0.45 to 0.59 and 0.61, significantly above predicting nothing and above every method without constraints that starts from the same source models, while mean Dice on frames with the structure does not drop. The findings hold with stronger source models trained on public data that include empty images. Applied at test time to other self-trained models, the same constraints give similar scores. Placing them in training yields a single per-frame network that needs no rules, plane classifier or neighbouring frames at deployment.

Contributions.

- **We will release a dataset**, to our knowledge the first public dataset of wireless handheld-probe thyroid sweeps, annotated for the thyroid, the carotid artery and the trachea, with nearly half of the annotated frames showing none of them (§4, Table 7).
- **We design a scheme** of two parts, an evaluation that also scores predicting nothing and anatomical constraints that let a pretrained model learn to predict nothing while it self-trains on new data (§3).
- **We present a systematic experimental analysis** of common label-free adaptation methods and promptable foundation models, from two sets of source models and on a second target domain, which shows that the usual evaluation cannot tell whether a model predicts nothing and that anatomical constraints teach it to (§5).

2 Related work

Neck ultrasound datasets. Most public thyroid ultrasound datasets were acquired with hospital cart-based scanners and contain still images or short cine clips with gland or nodule masks, such as TG3K and TN3K (Gong et al., 2021; 2023), DDTI (Pedraza et al., 2015) and the Stanford thyroid cine dataset (Yamashita et al., 2022). SegThy (Krönke et al., 2022) provides 3D volumes, which are likewise reconstructed from sweeps with a tracked probe on a cart-based scanner, and later work

adds carotid and jugular vein labels (Chakrabarty & Soni, 2025). These datasets come from hospital cart-based scanners operated by sonographers, and most of their images are selected so that nearly every one contains the target structure. To our knowledge, no public dataset of thyroid sweeps with a wireless handheld probe exists. Our dataset consists of free-hand sweeps with a wireless handheld probe, labels the thyroid, the carotid artery and the trachea, and nearly half of its annotated frames show no identifiable target structure.

Segmentation models for neck ultrasound. Supervised models trained on these data segment the thyroid with its nodules, the carotid artery, the jugular vein and other cervical structures (Gong et al., 2021; Webb et al., 2021; Bolomiti et al., 2026; Munsterman et al., 2024; Luo et al., 2022). When such models are used on a new ultrasound device without target labels, test-time adaptation (TTA) methods update the model while it predicts, by re-estimating normalisation statistics (Schneider et al., 2020), by making its outputs more confident (Wang et al., 2021; Niu et al., 2022; 2023) or by distilling pseudo-labels from a mean teacher (Wang et al., 2022; Döbler et al., 2023). Source-free domain adaptation (SFDA) methods need no access to the original training data: the model first labels the unlabelled images of the new device automatically, the unreliable parts of these labels are discarded and the model is retrained on the rest (Liu et al., 2021; Kundu et al., 2021; Huang et al., 2021). Medical variants filter or constrain the automatic labels by prediction uncertainty, class prototypes, image context or priors on organ size and shape (Chen et al., 2021; Wu et al., 2023; Huai et al., 2023; Bateson et al., 2022a;b). All of these models output a segmentation for every input image, and the adaptation methods aim to make each pixel more accurate or more confident, but none learns to decide that a structure is not in the image at all. Their evaluation, too, mostly uses images that contain the target structure (Yu et al., 2025).

Models that can abstain. Some models and methods can decline to answer, but none applies directly to handheld thyroid scanning. In classification, selective prediction (Geifman & El-Yaniv, 2017), partial and universal domain adaptation (Cao et al., 2018; You et al., 2019) and label-shift correction (Lipton et al., 2018; Alexandari et al., 2020) decide whether to answer from the model's own confidence and judge whole images only, without producing a segmentation. SAM 3 (Carion et al., 2026) and UltraSAM3 (Xu et al., 2026), a SAM 3 model fine-tuned on ultrasound data, first decide whether a concept is present and only then segment it, but the prompt and the threshold of this decision have to be tuned on labelled data of the target device, and without labels the score of UltraSAM3 in our evaluation falls from 0.65 and 0.64 to 0.54 and 0.56 (§5). Training a segmenter on data that include empty images is not enough either: source models that we trained on public data including empty slices still draw a thyroid on 80% and 71% of the handheld frames without one (§5). Anatomical priors have been used to make segmented shapes plausible, through learned shape models (Oktay et al., 2018; Painchaud et al., 2020), adjacency constraints (Ganaye et al., 2019; Gupta et al., 2022), constrained optimisation (Kervadec et al., 2019; El Jurdi et al., 2021) and mask correction (Larrazabal et al., 2020), but all of them assume that the structure is present. The closest work, SHAPE, scores automatically generated labels by the relative position of classes, but it needs the model's original training images during adaptation, which are often unavailable on a new device (Zhou et al., 2026). Unlike these works, we use anatomical constraints to decide whether a structure is drawn at all: each whole predicted structure is checked against rules on its position, shape, echogenicity and relations to the other structures, without relying on the model's confidence and without labels of the target device (§3).

3 Learning to predict nothing with anatomical constraints

Our scheme has two parts: an evaluation that also scores predicting nothing, and anatomical constraints in self-training that teach the model to predict nothing when the target structure is not in the frame.

An evaluation that also scores predicting nothing. Current evaluation measures how accurately a structure is segmented only on images that contain it, so it cannot reveal whether a model draws structures on frames without them. We score each structure separately: on frames that contain it, by the overlap between prediction and clinician annotation (Dice on annotated frames), and on frames that do not, by whether nothing is drawn. The two halves weigh equally in a balanced Dice, so a model that predicts nothing scores 0.50, and every method must exceed this level to be of practical

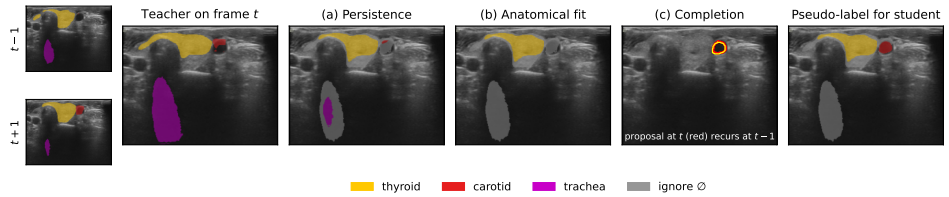


Figure 1: **Persist & fit self-training.** The composed source teacher predicts four-class probabilities on every frame of an unlabelled sweep. (a) Predictions of adjacent frames are aligned by optical flow and fused. Pixels that are not confident or not supported by a neighbour are withheld. (b) Connected components are tested against plane-conditional anatomical rules, and failures become *ignore* (∅), not background. (c) Completion: background inside vessel-like blobs is withheld (visible in (b)), and a blob next to an accepted lobe is proposed as carotid and accepted only if it recurs in an adjacent frame. A per-frame student is trained on the result and becomes the next teacher. In this transverse frame from an older participant, the teacher labels a deep acoustic shadow as trachea and only a crescent of the carotid lumen. Persistence removes most of the false trachea, the anatomical test withholds the rest and the crescent, and completion recovers the full lumen, whose proposal recurs in the previous frame (yellow outline).

use. We also report the false detection rate, the share of frames without a structure on which it is drawn, and the miss rate, the share of frames with a structure on which it is not drawn. Dice scores are averaged within participants and then over participants and structures, rates are pooled over frames, and paired differences are tested with a participant bootstrap and Holm correction (App. C). Unless stated otherwise, pairs of numbers refer to the 174 evaluation and the 58 held-out frames of §4, in this order.

Setting. The source consists of a gland model g trained on the public dataset TG3K (Gong et al., 2021), in which every training image contains a thyroid, a landmark model h for background, carotid and trachea, and a plane classifier c that tells transverse from sagittal frames, both trained on hand-held sweeps of two volunteers recorded with a different probe (App. D). g and h are composed into a teacher that gives each pixel u of frame x_t probabilities for background, thyroid, carotid and trachea,

$$p_t(u) = \big[\, h_0(1-g),\ h_0\, g,\ h_1,\ h_2 \,\big](x_t, u), \tag{1}$$

so thyroid probability is gland probability restricted to pixels the landmark model calls background. Using only unlabelled handheld sweeps, we want a per-frame segmenter f_θ that segments a structure accurately when it is in the frame and outputs nothing when it is not.

Why plain self-training does not learn to predict nothing. Self-training lets the teacher label the unlabelled frames, keeps the confident part of these labels and trains a student on them. Three obstacles stop it from learning to predict nothing. First, the teacher's false detections are confident: on frames without a thyroid, its median maximum thyroid probability is 0.99, so confidence filtering keeps these false detections as reliable labels. Second, false detections recur in consecutive frames, so requiring agreement between adjacent frames does not remove most of them: with this check alone, a thyroid is still drawn on 59% and 68% of the frames without one. Third, the filtering acts on pixels, so a stricter confidence threshold keeps fewer pixels without emptying a frame: with a median maximum thyroid probability of 0.99, even a threshold of 0.9 keeps thyroid pixels on at least half of the frames without one. The student therefore still learns to draw a thyroid on most frames without one.

Anatomical constraints decide whether to draw. Whether a structure should be drawn has to be decided for the whole structure. Anatomy gives evidence for this decision that does not come from the model's own confidence, because the position, shape and echogenicity of neck structures and their relations to each other are largely fixed and do not depend on how confident the model is. Persist & fit (Fig. 1) therefore filters the teacher's labels in three steps during self-training.

Persistence: agreement between adjacent frames. Let $\mathcal{N}(t) = \{t' \in \{t-1, t+1\} : |\tau_{t'} - \tau_t| \le \delta\}$ be the neighbours of frame t, with timestamps τ and $\delta = 0.6\,\mathrm{s}$. Optical flow $F_{t\to t'}$ (Farnebäck,

Table 1: Plane-conditional acceptance rules (a is area as a fraction of the image, W and H are image width and height, heights are measured from the top, and a component is darker than its rim if $r < 0.85$). A component failing a rule is set to ignore.

Structure	Transverse view	Sagittal view
Carotid	$0.2\% \le a \le 15\%$, circularity ≥ 0.45, darker than its rim	$a \ge 0.2\%$, elongation ≥ 2, darker than its rim
Trachea	the largest component that has its top edge in $[0.06H, 0.5H]$, $0.5\% \le a \le 35\%$ and a hyperechoic top edge (others ignored)	not expected, all ignored
Thyroid	centroid above $0.7H$. With an accepted carotid or trachea: $a \ge 0.4\%$, centroid within $0.45W$ of the nearest one, and horizontally between trachea and each carotid (margin $0.15W$) if both exist. Otherwise the two largest components with $a \ge 1.5\%$	largest component only, if $a \ge 2\%$

2003) warps each neighbour's probabilities into frame t, $\tilde{p}_{t'} = p_{t'} \circ (\mathrm{id} + F_{t \to t'})$, and the fused prediction and its label are

$$\bar{p}_t = \frac{1}{1 + |\mathcal{N}(t)|}\Big(p_t + \sum_{t' \in \mathcal{N}(t)} \tilde{p}_{t'}\Big), \qquad \hat{y}_t(u) = \arg\max_k \bar{p}_{t,k}(u). \tag{2}$$

A pixel u is reliable if $\max_k \bar{p}_{t,k}(u) \ge \gamma$ with $\gamma = 0.7$ and at least two frames of $\{t\} \cup \mathcal{N}(t)$ individually predict $\hat{y}_t(u)$. This step removes errors that appear in single frames but not errors that recur.

Anatomical fit: the anatomical test. For each 8-connected component C of class k among the reliable pixels, we compute its area fraction $a(C) = |C|/(HW)$ for an image of height H and width W, its circularity $\rho(C) = 4\pi|C|/P(C)^2$ with perimeter $P(C)$, its elongation $e(C)$, the aspect ratio of its minimum-area rectangle, its rim contrast $r(C) = \bar{I}(C)/\bar{I}(R_4(C))$, the ratio of its mean intensity to that of a 4-pixel ring around it, and the position of its centroid, with heights measured from the top. The rules of Table 1 for the plane π that c predicts give an acceptance $A_k^\pi(C) \in \{0, 1\}$. They use position, shape, echogenicity and the relations between structures but not the model's confidence, so they reject false detections that are confident and recur.

Pseudo-labels and training. The pseudo-label of pixel u is

$$y_t(u) = \begin{cases} \hat{y}_t(u) & \text{if } u \text{ is reliable and its component } C \text{ satisfies } A_k^\pi(C) = 1, \\ \varnothing & \text{otherwise,} \end{cases} \tag{3}$$

where $\varnothing$ means ignore. Failing components are ignored rather than set to background, because a failed test shows only that the teacher's class is implausible, not that the structure is absent. The student f_θ, a U-Net (Ronneberger et al., 2015) with an EfficientNet-B2 encoder (Tan & Le, 2019) initialised from g, is trained with cross-entropy plus Dice loss over non-ignored pixels and becomes the next teacher, for two rounds. At deployment persist & fit is one per-frame network that needs no rules, plane classifier or neighbouring frames, and it has learned to predict nothing when a structure is not in the frame. Carotid completion, the flow parameters and the training settings are detailed in App. E.

4 Dataset: free-hand sweeps in which structures are often absent

Table 2 summarises the dataset, App. C describes it in full, and Table 7 compares it with existing neck ultrasound datasets.

Acquisition. Research-team members recorded one continuous free-hand bilateral thyroid sweep, in transverse and sagittal planes, of each of 29 adults with one wireless handheld linear probe (SonoStar 5W) that streams to a tablet, where the sweeps were saved as screen recordings. The participants came from a laboratory session (15 adults aged 18–44) and a community screening event (13 adults aged 69–81 and one younger volunteer), and all gave written informed consent. The recordings were sampled at 4 frames per second, and 20,014 frames remained after blank and frozen frames were removed.

Table 2: The dataset at a glance.

Sweeps	Free-hand thyroid sweeps of 29 adults aged 18 to 81, 86 minutes, one wireless handheld probe (SonoStar 5W)
Unlabelled frames	20,014, used for self-training without labels
Annotated frames	232, labelled for thyroid, carotid lumen and trachea: 174 evaluation frames annotated by one clinician and 58 held-out frames of a hidden test set annotated by two clinicians
Frames without target	110 (47%) show no identifiable target structure
Splits	Unlabelled pool, evaluation frames, hidden test set, participant-disjoint protocol

Annotation. A head and neck specialist with at least ten years of thyroid ultrasound experience annotated the thyroid, the carotid lumen and the trachea on 232 frames from scratch, in a tool that shows no model prediction. 174 of them are evaluation frames. The other 58, the held-out frames, were never displayed during method development and form a hidden test set, which a second specialist with the same experience annotated again from scratch. For every structure left unannotated, the clinician also marked whether it is absent or uncertain.

Structures are often absent. Structures leave the image during probe placement, lift-off and changes of plane. 110 of the 232 frames (47%) show no identifiable target structure: on 54 the clinician marked all three structures as absent, and on 56 at least one as uncertain. By contrast, nearly every image of the public cart-based datasets contains the target structure.

Splits and release. The dataset has three splits: the unlabelled pool of 20,014 frames, which includes the images but not the labels of the annotated frames, the 174 evaluation frames for development and evaluation, and the 58 frames of the hidden test set. A participant-disjoint protocol self-trains on the sweeps of two of three fixed participant groups and tests on the third (App. C). The anatomical rules and their thresholds were set before any annotation existed. The data will be released, de-identified, under CC BY-NC 4.0, and the official track ranks label-free methods by balanced Dice on the hidden test frames of the released data (last column of Table 3, App. K and B).

5 Experiments

Experimental setup. Every adaptation method starts from one of two sets of source models, neither of which has seen a label of the target device. The first set is the source of §3 (App. D describes its training, including one retraining of the landmark model). The second set replaces the gland model and the carotid part of the landmark model by models that we trained on public data only, without using any target frame (App. I). Its gland model is trained on TG3K and on slices of the SegThy ultrasound volumes (Krönke et al., 2022), including slices without a thyroid, and its carotid model on SegThy and a public carotid dataset (Momot, 2022). On held-out subjects of the public data, this gland model draws a thyroid on only 10% of the slices without one, against 78% for the TG3K model, so it has learned to predict nothing on its own data.

We compare eight label-free adaptation methods: the TTA methods BN adapt (Schneider et al., 2020), TENT (Wang et al., 2021) and CoTTA (Wang et al., 2022), the medical SFDA methods AdaMI (Bateson et al., 2022a) and DPL (Chen et al., 2021), two self-training baselines and persist & fit (γ=0.7, two rounds of six epochs). The self-training baselines keep pixels with confidence at least γ (*confidence only*, Lee, 2013) or in addition require three of four test-time augmentations to agree (*augmentation agreement*). They share the student and training schedule of persist & fit and differ from it only in how pseudo-labels are formed.

As references, we compare temporal averaging of flow-aligned source predictions without training (cf. Painchaud et al., 2022), promptable foundation models, and, as an upper reference, a supervised model trained on the labelled evaluation frames (participant-grouped five-fold cross-validation, all 174 frames for the held-out frames). MedSAM (Ma et al., 2024) is prompted by boxes from the source models' masks, SAM (Kirillov et al., 2023) by boxes that Grounding DINO (Liu et al., 2024) finds for text prompts, and UltraSAM3 (Carion et al., 2026; Xu et al., 2026) by text directly. For analysis, *rules at test time* apply the anatomical test of §3 to the final mask of any model (App. E).

Table 3: Main results. Left: Dice on annotated frames per structure (174 evaluation frames). Middle: all-frame Dice (Dice over all frames, a frame without the structure scoring 1 if nothing is drawn and 0 otherwise), balanced Dice, balanced Dice with uncertain structures excluded from the negatives (Bal. cert.), false thyroid detections and carotid misses on the 174 frames. Right: the 58 held-out frames, never displayed during development. †/‡Persist & fit with our source models is significantly better/worse (participant bootstrap, Holm-adjusted $p < 0.05$, over 40 comparisons for Bal. cert.). Rows ^{s} use source models trained on public data (App. I). Box-prompted MedSAM takes its boxes, and thus its detections, from our source models. ^{e}Prompt and threshold chosen on the labelled evaluation frames. ^{t}Trained on the labelled evaluation frames. Official score: balanced Dice on the 58 held-out frames of the released data, on which every method was re-run, marked against persist & fit on the same data (App. K).

	Dice on annotated frames			174 evaluation frames					58 held-out frames			Official
Method	Thyroid	Carotid	Trachea	All-frame	Balanced	Bal. cert.	FP thyroid	Miss carotid	All-frame	Balanced	Bal. cert.	score
Predict nothing	0.00	0.00	0.00	0.63	0.50†	0.50†	0%	100%	0.67	0.50†	0.50†	0.50†
Source models	0.62	0.15	0.17	0.50†	0.44†	0.41†	99%	71%	0.54†	0.45†	0.43†	0.45†
BN adapt (Schneider et al., 2020)	0.60	0.18	0.25	0.37†	0.35†	0.36†	93%	55%	0.36†	0.32†	0.31†	0.32†
TENT (Wang et al., 2021)	0.63	0.11	0.23	0.49†	0.44†	0.44†	85%	76%	0.47†	0.39†	0.38†	0.40†
CoTTA (Wang et al., 2022)	0.62	0.18	0.23	0.39†	0.37†	0.37†	96%	55%	0.40†	0.34†	0.33†	0.35†
AdaMI (Bateson et al., 2022a)	0.65	0.06	0.22	0.53†	0.47†	0.48†	86%	89%	0.59†	0.48†	0.49†	0.47†
UltraSAM3, text, tuned^{e} (Xu et al., 2026)	0.42	0.66	0.00	0.74‡	0.65‡	0.68	11%	23%	0.76	0.64	0.68	0.65
MedSAM, source boxes (Ma et al., 2024)	0.58	0.14	0.20	0.53†	0.47†	0.46†	85%	82%	0.60†	0.49†	0.48†	0.49†
Temporal averaging at test time	0.62	0.16	0.17	0.53†	0.46†	0.45†	96%	73%	0.56†	0.45†	0.44†	0.46†
Self-training, confidence only	0.66	0.25	0.20	0.59†	0.53†	0.54†	74%	69%	0.60†	0.49†	0.49†	0.51†
Self-training, augmentation agreement	0.65	0.22	0.20	0.60†	0.54†	0.55†	68%	69%	0.63†	0.52†	0.52†	0.51†
DPL (Chen et al., 2021)	0.64	0.24	0.20	0.60†	0.54†	0.56	69%	69%	0.63†	0.52†	0.50†	0.52†
Self-training (conf.) + rules at test time	0.66	0.21	0.20	0.65	0.58	0.60	51%	77%	0.73	0.59	0.62	0.60
Self-training (aug.) + rules at test time	0.65	0.17	0.20	0.67	0.58	0.61	46%	77%	0.73	0.60	0.62	0.59
Persist & fit	0.64	0.23	0.18	0.68	0.59	0.63	40%	66%	0.74	0.61	0.64	0.58
Source models^{s}	0.63	0.78	0.17	0.57†	0.54	0.56	80%	6%	0.56†	0.53	0.56	0.55
Source models^{s} + rules at test time	0.63	0.76	0.17	0.72	0.67‡	0.71‡	33%	10%	0.69	0.63	0.66	0.63
Self-training (conf.)^{s} + rules at test time	0.65	0.78	0.21	0.74‡	0.70‡	0.74‡	28%	8%	0.74	0.67	0.69	0.67‡
Self-training (aug.)^{s} + rules at test time	0.64	0.77	0.17	0.76‡	0.70‡	0.73‡	25%	8%	0.76	0.68	0.70	0.67‡
Persist & fit^{s}	0.64	0.76	0.17	0.75‡	0.69‡	0.74‡	25%	8%	0.73	0.65	0.69	0.68‡
Supervised reference^{t}	0.67	0.72	0.70	0.80‡	0.78‡	0.83‡	26%	11%	0.81	0.74‡	0.77‡	0.76‡

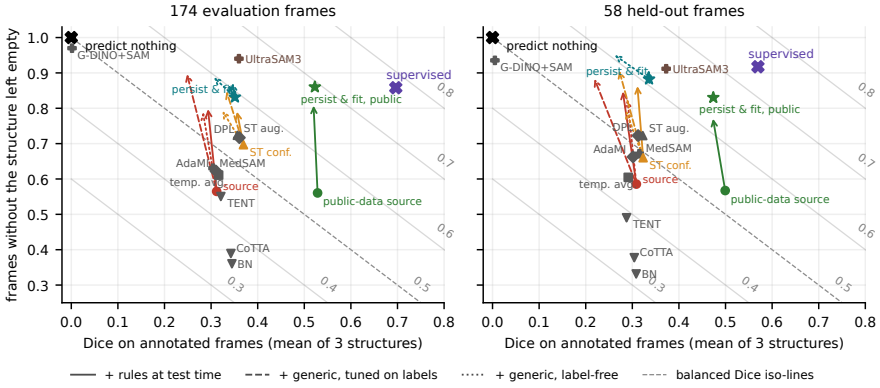


Figure 2: Every method as Dice on annotated frames (x) against the share of frames without the structure that it leaves empty (y), both averaged over the three structures, so balanced Dice is $(x + y)/2$ and constant along the grey anti-diagonals. Predict nothing sits at (0, 1). Arrows show how rules at test time (solid), generic abstention (filters on confidence and component size, App. G) tuned on the labelled evaluation frames (dashed) and with label-free thresholds (dotted) move both sets of source models (green: public-data, rules only), confidence-only self-training and persist & fit. Numbers are in Tables 3, 4, 5, 10 and 11.

Balanced Dice shows whether a model can predict nothing, and the usual evaluation does not. Measuring segmentation accuracy only on frames that contain the target misjudges what the models actually do. Ranked by Dice on annotated frames alone, the 16 label-free methods built on our source models (the 13 of Table 3 and the source models, temporal averaging and DPL with rules at test time) order almost independently of balanced Dice, with Kendall's τ of -0.02 and 0.20, and -0.16 over 23 methods on the second dataset SegThy (App. J). On the evaluation frames, BN adapt

and CoTTA rank above our source models by Dice on annotated frames and thus look like effective adaptation, but once frames without the structure are scored, they rank last, below predict nothing. The reason is that the gains and losses of adaptation come mostly from frames without the structure. Across both frame sets and both sets of source models, the eight methods change, relative to their source models, Dice on annotated frames, averaged over the three structures, by at most 0.06, and the share of frames without a structure left empty by -0.25 to $+0.30$ (Table 6 in App. A). Balanced Dice scores these gains and losses directly. Predict nothing scores exactly 0.50, so a method below 0.50 loses more by drawing structures on frames without them than it gains by segmenting them on frames with them. Fig. 2 plots the two halves separately, so it shows directly whether a method only draws, only predicts nothing, or does both. This judgement does not depend on the weight of the two halves. For any weight on Dice on annotated frames between 0.45 and 0.6, every method with anatomical constraints stays above predict nothing and BN adapt and CoTTA stay below it, on both frame sets (App. A). All-frame Dice is no remedy either: excluding the structures the clinician marked as uncertain lowers predict nothing from 0.63 to 0.45 on the evaluation frames and reverses most of the comparisons it wins. In balanced Dice, predict nothing stays at 0.50 and the ranking holds (App. L).

Models draw structures on many frames without them, and common adaptation methods do not fix this. Our source models draw a thyroid on 99% and 90% of the frames without one and stay below predict nothing, with a balanced Dice of 0.44 and 0.45. Common adaptation methods do not fix this. After BN adapt, TENT and CoTTA, the model still draws a thyroid on 85–96% of the evaluation frames without one, false trachea detections rise from 28% to 38–65% of the frames without a trachea, and balanced Dice stays at 0.44 or falls to 0.35–0.37 on the evaluation frames. AdaMI also stays below predict nothing (0.47 and 0.48). The two self-training baselines and DPL still draw a thyroid on 68–74% of the evaluation frames without one and stay within 0.04 of predict nothing on both frame sets. A source that predicts nothing on public data does not solve the problem either. On the handheld sweeps, the public-data source models still draw a thyroid on 80% and 71% of the frames without one: predicting nothing, learned on other devices, does not hold on our handheld device.

Anatomical constraints teach the model to predict nothing. With anatomical constraints in self-training, persist & fit draws a thyroid on 40% and 29% of the frames without one, down from 99% and 90% for the source models, and reaches a balanced Dice of 0.59 and 0.61, significantly above predict nothing and above every method in Table 3 that starts from our source models without anatomical constraints (Holm-adjusted $p < 0.05$). Its mean accuracy on frames with the structure does not drop: Dice on annotated frames, averaged over the three structures, rises from 0.31 for the source models on both frame sets to 0.35 and 0.34, and thyroid Dice on the evaluation frames is 0.64 against 0.62.

The anatomical test contributes most. In persist & fit, changing the pixel-confidence threshold from 0.7 to 0.5 or 0.9 changes balanced Dice by at most 0.01 (Table 22 in App. N). Added to confidence-only self-training, agreement between adjacent frames or augmentation agreement raises it by at most 0.02 and 0.03, whereas the anatomical test raises it by 0.03 and 0.08, and by 0.05 and 0.09 together with carotid completion (Table 9 in App. F).

For self-trained models, whether the constraints act in training or at test time makes no significant difference. Applied to the output of the two self-training baselines, the same constraints give a balanced Dice of 0.58 on the evaluation frames and 0.59–0.60 on the held-out frames, and the paired differences to persist & fit have 95% confidence intervals between -0.03 and $+0.05$ (-0.05 and $+0.02$ on the released data). The advantage of persist & fit is that it deploys a single network without rules or a plane classifier.

On our sweeps, for the two self-training baselines and DPL, the model's own confidence is a weaker basis for this decision than anatomy. As a test, we drop a structure from a frame if its highest confidence falls below a threshold, and drop its components smaller than an area threshold, with both thresholds set automatically from all unlabelled frames (*label-free thresholds*, App. G). On the held-out frames, these thresholds reach 0.54–0.55 for the three methods, 0.03–0.05 below the anatomical constraints, and even the best single confidence threshold for all structures, chosen with the labels of the frames being scored, stays below the constraints for these three methods on both

frame sets. For the source models and persist & fit, label-free thresholds and anatomical constraints differ by at most 0.01.

The constraints also work with the public-data source models. Rules at test time raise their balanced Dice from 0.54 to 0.67, and from 0.53 to 0.63 on the held-out frames, and persist & fit reaches 0.69 and 0.65, against 0.65 and 0.63 for confidence-only self-training from the same source models, a significant gain on the evaluation frames only.

UltraSAM3 needs labels to reach its score. It reaches 0.65 and 0.64 with prompts and threshold tuned on the labelled evaluation frames, 0.58 and 0.59 with prompts fixed in advance and a threshold tuned on these labels, and 0.54 and 0.56 with prompts fixed in advance and label-free thresholds (App. H).

Robustness. In the participant-disjoint protocol, where no test participant is seen during self-training, persist & fit keeps its balanced Dice (0.59 and 0.62, against 0.59 and 0.61 with all participants, Table 16 in App. K). Re-running every method on the released data, with the overlay text masked, changes balanced Dice by at most 0.04 and leaves the conclusions unchanged (official score in the last column of Table 3, App. K). Excluding the structures the clinician marked as uncertain raises the balanced Dice of persist & fit to 0.63 and 0.64 and keeps the ranking (App. L). Scoring against the second clinician's labels of the held-out frames (presence Cohen's κ 0.91–0.97) changes balanced Dice by at most 0.02 (App. M). The plane classifier agrees with the clinician's plane label on 92% of the frames with a recognisable plane, and using the clinician's label in rules at test time changes balanced Dice by at most 0.007 (App. D). Finally, we run the same methods unchanged from our source models on SegThy, 3D volumes of 16 participants from a cart-based scanner (App. J). The source models still draw a thyroid on 81% of the slices without one, and every method with anatomical constraints or label-free thresholds (0.65–0.74) beats every method without them (0.56–0.63). There, label-free thresholds match or beat the constraints.

6 Discussion and future work

Open points. Our data come from one probe and 29 participants. The main conclusions rest on large differences, for example persist & fit exceeds the source models by 0.15 and 0.16 in balanced Dice, so they are clear at the current scale. Whether they hold for other probes and other populations remains to be tested with more participants and more probes.

Towards AI-guided ultrasound scanning. Wireless handheld probes bring thyroid ultrasound to community screening, primary care and potentially the home, but the person holding the probe is often not an experienced sonographer. Novices do not know which structure is in view, confuse the left and right lobes, do not know where to move the probe and cannot tell when a sweep is complete. To guide them, a system must know in every frame whether the thyroid, carotid and trachea are present and where, use the relative position of trachea and carotid to infer side and plane, and state clearly when a structure is absent, since false detections would mislead the guidance. Our results bear directly on this perception layer. A segmenter for guidance must be evaluated on frames without structures, and with anatomical constraints it can learn, without target labels, to predict nothing when a structure is not in the frame.

Next, we will connect an abstaining segmenter to real-time guidance on the tablet, which infers the current plane, the side and the covered extent of the gland and prompts the next probe movement. We will evaluate scan completeness, time and image quality with novice operators against experienced clinicians, and extend the dataset and evaluation to further probes and sites.

7 Conclusion

In free-hand sweeps where target structures are often absent, segmentation models trained on cart-based data do not know when not to segment, and from our source models common label-free adaptation methods do not fix this. Evaluation must therefore score frames without the structure. Anatomical constraints in self-training teach a model, without target labels, to predict nothing, without a significant drop in its mean segmentation accuracy. We will release the dataset as a first step towards AI-guided ultrasound scanning.

AI USE STATEMENT

We used a generative AI assistant to help with coding, experiments, writing and figures. The authors verified all results and take full responsibility for the content.

ETHICS STATEMENT

All 29 target participants gave written informed consent, on a bilingual form, to the anonymised use and sharing of their data for research, and research-team members scanned them with non-invasive ultrasound. The scans were not diagnostic, and participants were advised to have a clinical examination if a possible finding was noticed. The two volunteers whose scans trained the landmark model and plane classifier are research-team members who signed the same consent form, and their recordings will not be released. The release will remove audio and metadata, mask overlay text, replace participant codes by random codes and reduce demographics to age group and sex (App. B). All older participants self-reported Asian ethnicity and were recorded at one event, so results for this group cannot be attributed to age alone. The data and models are intended for research, are not a medical device and were not evaluated for diagnosis or operator guidance.

REPRODUCIBILITY STATEMENT

The method is specified in §3, Table 1 and App. E, the source models in Apps. D and I, the data in §4, App. C and the datasheet (App. B), the protocol in §4 and App. C, the baselines and hyperparameters in §5 and App. E, and the selection of held-out frames and the few-shot track in App. P. Appendix A gives the main methods and controls on both frame sets, with 95% participant-bootstrap confidence intervals for all-frame and balanced Dice, and Apps. F–N give the ablations, variants and remaining controls. Three seeds of persist & fit give all-frame Dice 0.676 ± 0.006 (sd) on the evaluation frames and 0.724 ± 0.014 on the held-out frames, and balanced Dice 0.587 ± 0.007 and 0.596 ± 0.018 (Table 22). An anonymised copy of the code is provided as supplementary material. It covers preprocessing, the selection of evaluation and held-out frames, landmark-model retraining, the training of the public-data source models, adaptation, all baselines, MedSAM with ground-truth boxes, ablations, rules at test time and generic abstention with tuned and label-free thresholds, the participant-disjoint protocol, the supervised reference, the few-shot track, the release-data re-run, the speed benchmark, the de-identification pipeline and the evaluation script. This script computes Dice on annotated frames, all-frame Dice, balanced Dice, miss rates, false detection rates and the predict-nothing reference, with participant-level bootstrap tests and Holm correction, and optionally excludes uncertain structures or scores against the second annotator. Scripts generate the body of every results table from the evaluation outputs. The official splits are the adaptation pool of 20,014 unlabelled frames, the 174 evaluation frames for development and evaluation and the 58 held-out frames as the hidden test set. The participant-disjoint folds and the participant draws of the few-shot track are fixed by seed. The official track ranks label-free entries by balanced Dice on the released held-out frames (§4). The de-identified sweeps and frame crops, the annotations and uncertainty flags of the evaluation frames, the split files, the code, including the evaluation script, and the source model weights are available for review at `https://huggingface.co/datasets/freehand-us-anon/freehand-thyroid-sweeps` (App. B), and we will score predictions on the hidden held-out frames on request. Overlay text is masked in the released frames, and App. K re-runs the main methods on them from scratch.

REFERENCES

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 222–232. PMLR, 2020. URL `https://proceedings.mlr.press/v119/alexandari20a.html`.

Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-free domain adaptation for image segmentation. *Medical Image Analysis*, 82:102617, 2022a. URL `https://doi.org/10.1016/j.media.2022.102617`.

Mathilde Bateson, Hervé Lombaert, and Ismail Ben Ayed. Test-time adaptation with shape moments for image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 2022b. URL https://arxiv.org/abs/2205.07983.

Maria Bolomiti, Joel Burman, Keyur Radiya, Olav Inge Håskjold, Karl Øyvind Mikalsen, and Vegard Heimly Brun. Automated segmentation of thyroid tissue and carotid artery in ultrasound videos using expert-in-the-loop deep learning: a foundational step toward AI-assisted thyroid diagnostics. *Updates in Surgery*, 2026. URL https://doi.org/10.1007/s13304-026-02782-9. Online first, 31 July 2026.

Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Satrajit Chakrabarty and Ravi Soni. Comparing SAM 2 and SAM 3 for zero-shot segmentation of 3D medical data. *arXiv preprint arXiv:2511.21926*, 2025. URL https://arxiv.org/abs/2511.21926.

Cheng Chen, Quande Liu, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In *Medical Image Computing and Computer Assisted Intervention*, 2021. URL https://arxiv.org/abs/2109.09735.

Mario Döbler, Robert A. Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2211.13081.

Rosana El Jurdi, Caroline Petitjean, Paul Honeine, Veronika Cheplygina, and Fahed Abdallah. High-level prior-based loss functions for medical image segmentation: A survey. *Computer Vision and Image Understanding*, 210:103248, 2021. URL https://doi.org/10.1016/j.cviu.2021.103248.

Gunnar Farnebäck. Two-Frame Motion Estimation Based on Polynomial Expansion. In *Scandinavian Conference on Image Analysis*, pp. 363–370, 2003.

Pierre-Antoine Ganaye, Michael Sdika, Bill Triggs, and Hugues Benoit-Cattin. Removing segmentation inconsistencies with semi-supervised non-adjacency constraint. *Medical Image Analysis*, 58:101551, 2019. URL https://doi.org/10.1016/j.media.2019.101551.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *IEEE International Symposium on Biomedical Imaging*, 2021. URL https://ieeexplore.ieee.org/document/9434087/.

Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Guanbin Li, and Fei Chen. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in Biology and Medicine*, 155:106389, 2023. URL https://pubmed.ncbi.nlm.nih.gov/36812810/.

Saumya Gupta, Xiaoling Hu, James Kaan, Michael Jin, Mutshipay Mpoy, Katherine Chung, Gagandeep Singh, Mary Saltz, Tahsin Kurc, Joel Saltz, Apostolos Tassiopoulos, Prateek Prasanna, and Chao Chen. Learning topological interactions for multi-class medical image segmentation. In *European Conference on Computer Vision*, pp. 701–718, 2022. URL https://arxiv.org/abs/2207.09654.

Zheang Huai, Xinpeng Ding, Yi Li, and Xiaomeng Li. Context-aware pseudo-label refinement for source-free domain adaptive fundus image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 2023. URL https://arxiv.org/abs/2308.07731.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2110.03374.

Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Constrained-CNN losses for weakly supervised segmentation. *Medical Image Analysis*, 54:88–99, 2019. URL https://doi.org/10.1016/j.media.2019.02.009.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, Lemonia Konstantinidou, Marcus Makowski, James Nagarajah, Nassir Navab, Wolfgang Weber, and Thomas Wendler. Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *PLOS ONE*, 17(7):e0268550, 2022. URL https://doi.org/10.1371/journal.pone.0268550.

Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R. Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2108.11249.

Agostina J. Larrazabal, Cesar Martinez, Ben Glocker, and Enzo Ferrante. Post-DAE: Anatomically plausible segmentation via post-processing with denoising autoencoders. *IEEE Transactions on Medical Imaging*, 39:3813–3820, 2020. URL https://arxiv.org/abs/2006.13791.

Dong-Hyun Lee. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2103.16372.

Shao-Yuan Lo, Poojan Oza, Sumanth Chennupati, Alejandro Galindo, and Vishal M. Patel. Spatio-temporal pixel-level contrastive learning-based source-free domain adaptation for video semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2303.14361.

Hongxia Luo, Laifa Ma, Xiangqiong Wu, Guanghua Tan, Hui Zhu, Senmin Wu, Kenli Li, Yan Yang, and Shengli Li. Deep learning-based ultrasonic dynamic video detection and segmentation of thyroid gland and its surrounding cervical soft tissues. *Medical Physics*, 49(1):382–392, 2022. URL https://doi.org/10.1002/mp.15332.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024. URL https://pubmed.ncbi.nlm.nih.gov/38253604/.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. MedSAM2: Segment anything in 3D medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025. URL https://arxiv.org/abs/2504.03600.

Agata Momot. Common carotid artery ultrasound images, 2022. Version 1, CC BY 4.0.

Roxane Munsterman, Tim Boers, Sicco J. Braak, Jelmer M. Wolterink, Michel Versluis, and Srirang Manohar. Deep learning-based segmentation of 3D ultrasound images of the thyroid. *WFUMB Ultrasound Open*, 2(2):100055, 2024. URL https://www.sciencedirect.com/science/article/pii/S2949668324000235.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2204.02610.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.12400.

Joo Young Oh, Euihwan Jeong, Moo Jin Jeong, Joo Young Lee, and Hoon-Hee Park. NNUS: Large dataset for normal neck organ detection in ultrasound. GitHub repository joo053/Neck-Ultrasound-Dataset, 2023. Accessed September 2026, repository without image or label files.

Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A. Cook, Antonio de Marvao, Timothy Dawes, Declan P. O'Regan, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation. *IEEE Transactions on Medical Imaging*, 37(2): 384–395, 2018. URL https://doi.org/10.1109/TMI.2017.2743464.

Nathan Painchaud, Youssef Skandarani, Thierry Judge, Olivier Bernard, Alain Lalande, and Pierre-Marc Jodoin. Cardiac segmentation with strong anatomical guarantees. *IEEE Transactions on Medical Imaging*, 39:3703–3713, 2020. URL https://arxiv.org/abs/2006.08825.

Nathan Painchaud, Nicolas Duchateau, Olivier Bernard, and Pierre-Marc Jodoin. Echocardiography segmentation with enforced temporal consistency. *IEEE Transactions on Medical Imaging*, 41 (10):2867–2878, 2022. URL https://arxiv.org/abs/2112.02102.

Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, volume 9287 of *Proc. SPIE*, pp. 92870W. SPIE, 2015. URL https://doi.org/10.1117/12.2073532.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. URL https://doi.org/10.1007/978-3-319-24574-4_28.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.16971.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.10726.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2203.13591.

Wenhan Wang, Jiale Zhou, Cheng Zhang, Weiwei Xing, Shangchun Fan, and Xiaolei Qu. Prototype bank-driven test-time adaptation for medical ultrasound image segmentation. *Medical Physics*, 53:e70280, 2026. URL https://pubmed.ncbi.nlm.nih.gov/41549639/.

Jeremy M. Webb, Duane D. Meixner, Shaheeda A. Adusei, Eric C. Polley, Mostafa Fatemi, and Azra Alizad. Automatic deep learning semantic segmentation of ultrasound thyroid cineclips using recurrent fully convolutional networks. *IEEE Access*, 9:5119–5127, 2021. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7978237/.

Jianghao Wu, Guotai Wang, Ran Gu, Tao Lu, Yinan Chen, Wentao Zhu, Tom Vercauteren, Sébastien Ourselin, and Shaoting Zhang. UPL-SFDA: Uncertainty-aware pseudo label guided source-free domain adaptation for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42 (12):3932–3943, 2023. URL https://arxiv.org/abs/2309.10244.

Bo Xu, Quanhao Zhu, Rui Lin, et al. Ultrasam3: A concept-driven foundation model for universal ultrasound image segmentation. *arXiv preprint arXiv:2607.29200*, 2026. URL https://arxiv.org/abs/2607.29200.

Rikiya Yamashita, Tara Kapoor, Minhaj Nur Alam, Alfiia Galimzianova, Saad Ali Syed, Mete Ugur Akdogan, Emel Alkim, Andrew Louis Wentland, Nikhil Madhuripan, Daniel Goff, Victoria Barbee, Natasha Diba Sheybani, Hersh Sagreiya, Daniel L. Rubin, and Terry S. Desser. Toward reduction in false-positive thyroid nodule biopsies with a deep learning-based risk stratification system using US cine-clip images. *Radiology: Artificial Intelligence*, 4(3):e210174, 2022. URL https://doi.org/10.1148/ryai.210174.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Wenjing Yu, Shuo Jiang, Yifei Chen, et al. A large scale benchmark for test time adaptation methods in medical image segmentation. *arXiv preprint arXiv:2512.02497*, 2025. URL https://arxiv.org/abs/2512.02497.

Linkuan Zhou, Yinghao Xia, Yufei Shen, Xiangyu Li, Wenjie Du, Cong Cong, Leyi Wei, Ran Su, and Qiangguo Jin. SHAPE: Structure-aware hierarchical unsupervised domain adaptation with plausibility evaluation for medical image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30000–30010, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhou_SHAPE_Structure-aware_Hierarchical_Unsupervised_Domain_Adaptation_with_Plausibility_Evaluation_for_CVPR_2026_paper.html.

A Full results

Weight of balanced Dice. With weight w on Dice on annotated frames and $1 - w$ on the share of frames without the structure left empty, predict nothing scores $1 - w$ and every score is linear in w (Fig. 2). For every $w \geq 0.45$, all methods with rules at test time and persist & fit with both sets of source models stay above predict nothing, and for every $w \leq 0.6$, BN adapt and CoTTA with our source models and Grounding DINO + SAM stay below it, on both frame sets. For w from 0.3 to 0.75, persist & fit has the highest score among the methods of Table 3 built on our source models.

Table 4: Main methods and controls on the 174 evaluation frames: Dice on annotated frames, miss rates (annotated structure without a prediction), false detection rates, all-frame Dice and balanced Dice with 95% participant-bootstrap CIs.

Method	Dice (annotated)			Miss		FP			
	Thy.	Car.	Tra.	Car.	Tra.	Thy.	Tra.	All-frame	Balanced
Predict nothing	0.00	0.00	0.00	100%	100%	0%	0%	0.63 [0.57,0.69]	0.50 [0.50,0.50]
Source models	0.62	0.15	0.17	71%	61%	99%	28%	0.50 [0.47,0.54]	0.44 [0.41,0.47]
BN adapt (Schneider et al., 2020)	0.60	0.18	0.25	55%	28%	93%	65%	0.37 [0.33,0.41]	0.35 [0.32,0.38]
TENT (Wang et al., 2021)	0.63	0.11	0.23	76%	57%	85%	38%	0.49 [0.45,0.52]	0.44 [0.40,0.47]
CoTTA (Wang et al., 2022)	0.62	0.18	0.23	55%	33%	96%	57%	0.39 [0.35,0.44]	0.37 [0.33,0.40]
AdaMI (Bateson et al., 2022a)	0.65	0.06	0.22	89%	69%	86%	22%	0.53 [0.50,0.57]	0.47 [0.44,0.50]
Grounding DINO + SAM, text (Liu et al., 2024)	0.00	0.00	0.00	100%	96%	5%	4%	0.61 [0.55,0.67]	0.49 [0.48,0.49]
UltraSAM3, text, tuned (Xu et al., 2026)	0.42	0.66	0.00	23%	93%	11%	0%	0.74 [0.71,0.78]	0.65 [0.63,0.67]
MedSAM, source boxes (Ma et al., 2024)	0.58	0.14	0.20	82%	65%	85%	26%	0.53 [0.49,0.57]	0.47 [0.43,0.50]
Temporal averaging at test time	0.62	0.16	0.17	73%	61%	96%	22%	0.53 [0.51,0.56]	0.46 [0.44,0.49]
Self-training, confidence only	0.66	0.25	0.20	69%	67%	74%	21%	0.59 [0.55,0.63]	0.53 [0.50,0.56]
Self-training, augmentation agreement	0.65	0.22	0.20	69%	65%	68%	17%	0.60 [0.57,0.64]	0.54 [0.51,0.57]
DPL (Chen et al., 2021)	0.64	0.24	0.20	69%	69%	69%	18%	0.60 [0.56,0.64]	0.54 [0.51,0.57]
Persist & fit	0.64	0.23	0.18	66%	70%	40%	8%	0.68 [0.63,0.73]	0.59 [0.56,0.63]
Source + rules	0.62	0.09	0.17	89%	70%	52%	8%	0.63 [0.59,0.67]	0.55 [0.52,0.58]
Temporal averaging + rules	0.61	0.14	0.17	81%	69%	47%	6%	0.65 [0.61,0.69]	0.56 [0.54,0.59]
Self-training (conf.) + rules	0.66	0.21	0.20	77%	70%	51%	10%	0.65 [0.62,0.69]	0.58 [0.55,0.60]
Self-training (aug.) + rules	0.65	0.17	0.20	77%	70%	46%	7%	0.67 [0.63,0.71]	0.58 [0.56,0.61]
DPL + rules	0.64	0.22	0.20	74%	70%	45%	7%	0.67 [0.63,0.71]	0.59 [0.56,0.62]
AdaMI + rules	0.63	0.06	0.22	95%	69%	60%	11%	0.61 [0.57,0.65]	0.53 [0.50,0.57]
Persist & fit + rules	0.63	0.19	0.18	73%	72%	32%	5%	0.70 [0.65,0.74]	0.60 [0.58,0.63]
Self-training (conf.) + generic	0.55	0.25	0.20	69%	70%	29%	12%	0.68 [0.63,0.72]	0.59 [0.56,0.62]
Self-training (aug.) + generic	0.58	0.22	0.19	69%	72%	28%	12%	0.68 [0.64,0.72]	0.60 [0.57,0.63]
Persist & fit + generic	0.63	0.23	0.17	66%	74%	32%	3%	0.70 [0.66,0.75]	0.61 [0.58,0.65]
MedSAM, persist & fit boxes	0.59	0.23	0.19	69%	70%	38%	7%	0.68 [0.63,0.73]	0.59 [0.56,0.63]
SAM, source boxes	0.53	0.15	0.19	82%	65%	86%	26%	0.52 [0.49,0.56]	0.45 [0.42,0.49]
Grounding DINO + SAM, default threshold	0.07	0.01	0.07	71%	22%	66%	78%	0.23 [0.19,0.26]	0.19 [0.17,0.22]
UltraSAM3, a-priori prompts, default threshold	0.60	0.37	0.00	61%	83%	54%	6%	0.63 [0.59,0.67]	0.55 [0.52,0.59]
UltraSAM3, a-priori prompts, tuned threshold	0.52	0.29	0.00	68%	87%	21%	2%	0.68 [0.64,0.73]	0.58 [0.55,0.62]
Source models^s	0.63	0.78	0.17	6%	61%	80%	28%	0.57 [0.54,0.60]	0.54 [0.51,0.58]
Source models^s + rules	0.63	0.76	0.17	10%	70%	33%	8%	0.72 [0.69,0.75]	0.67 [0.64,0.70]
Self-training (conf.)^s	0.66	0.79	0.21	6%	69%	38%	22%	0.68 [0.65,0.71]	0.65 [0.62,0.68]
Self-training (conf.)^s + rules	0.65	0.78	0.21	8%	69%	28%	7%	0.74 [0.71,0.77]	0.70 [0.67,0.73]
Self-training (aug.)^s	0.64	0.79	0.17	5%	70%	35%	15%	0.70 [0.67,0.74]	0.66 [0.63,0.69]
Self-training (aug.)^s + rules	0.64	0.77	0.17	8%	74%	25%	6%	0.76 [0.72,0.79]	0.70 [0.67,0.73]
Persist & fit^s	0.64	0.76	0.17	8%	70%	25%	7%	0.75 [0.71,0.79]	0.69 [0.66,0.72]
Persist & fit^s + rules	0.64	0.71	0.17	15%	72%	20%	3%	0.76 [0.73,0.80]	0.70 [0.66,0.73]
Supervised reference	0.67	0.72	0.70	11%	7%	26%	9%	0.80 [0.77,0.83]	0.78 [0.74,0.81]

Table 5: Main methods and controls on the 58 held-out frames, same columns as Table 4.

Method	Dice (annotated) Thy.	Car.	Tra.	Miss Car.	Tra.	FP Thy.	Tra.	All-frame	Balanced
Predict nothing	0.00	0.00	0.00	100%	100%	0%	0%	0.67 [0.57,0.77]	0.50 [0.50,0.50]
Source models	0.59	0.11	0.23	80%	40%	90%	33%	0.54 [0.50,0.59]	0.45 [0.41,0.48]
BN adapt (Schneider et al., 2020)	0.61	0.12	0.20	67%	20%	100%	70%	0.36 [0.31,0.41]	0.32 [0.28,0.36]
TENT (Wang et al., 2021)	0.63	0.07	0.17	73%	47%	87%	40%	0.47 [0.42,0.53]	0.39 [0.35,0.43]
CoTTA (Wang et al., 2022)	0.62	0.11	0.18	53%	27%	100%	53%	0.40 [0.34,0.46]	0.34 [0.29,0.39]
AdaMI (Bateson et al., 2022a)	0.65	0.06	0.19	93%	60%	71%	23%	0.59 [0.53,0.65]	0.48 [0.44,0.52]
Grounding DINO + SAM, text (Liu et al., 2024)	0.00	0.00	0.02	100%	87%	13%	5%	0.64 [0.54,0.74]	0.47 [0.44,0.50]
UltraSAM3, text, tuned (Xu et al., 2026)	0.60	0.52	0.00	40%	100%	10%	2%	0.76 [0.69,0.84]	0.64 [0.58,0.71]
MedSAM, source boxes (Ma et al., 2024)	0.59	0.12	0.24	80%	40%	68%	28%	0.60 [0.55,0.65]	0.49 [0.46,0.53]
Temporal averaging at test time	0.59	0.08	0.20	87%	60%	87%	28%	0.56 [0.50,0.61]	0.45 [0.40,0.49]
Self-training, confidence only	0.64	0.15	0.17	80%	67%	77%	26%	0.60 [0.54,0.65]	0.49 [0.45,0.53]
Self-training, augmentation agreement	0.63	0.16	0.18	80%	67%	65%	21%	0.63 [0.58,0.68]	0.52 [0.48,0.57]
DPL (Chen et al., 2021)	0.64	0.12	0.19	80%	67%	65%	21%	0.63 [0.58,0.69]	0.52 [0.48,0.56]
Persist & fit	0.60	0.22	0.19	67%	67%	29%	5%	0.74 [0.66,0.81]	0.61 [0.55,0.67]
Source + rules	0.56	0.08	0.19	87%	67%	32%	9%	0.70 [0.62,0.77]	0.57 [0.52,0.62]
Temporal averaging + rules	0.57	0.05	0.19	93%	67%	23%	7%	0.72 [0.64,0.79]	0.58 [0.54,0.63]
Self-training (conf.) + rules	0.62	0.15	0.16	80%	73%	35%	2%	0.73 [0.66,0.79]	0.59 [0.54,0.64]
Self-training (aug.) + rules	0.62	0.16	0.18	80%	67%	32%	2%	0.73 [0.66,0.80]	0.60 [0.56,0.65]
DPL + rules	0.62	0.10	0.19	87%	67%	32%	5%	0.72 [0.65,0.79]	0.59 [0.53,0.64]
AdaMI + rules	0.62	0.06	0.13	93%	73%	42%	2%	0.70 [0.63,0.77]	0.56 [0.52,0.61]
Persist & fit + rules	0.59	0.19	0.19	73%	67%	23%	5%	0.74 [0.66,0.82]	0.61 [0.55,0.67]
Self-training (conf.) + generic	0.48	0.15	0.17	80%	67%	13%	16%	0.71 [0.64,0.78]	0.59 [0.56,0.63]
Self-training (aug.) + generic	0.50	0.16	0.18	80%	67%	19%	16%	0.70 [0.63,0.77]	0.58 [0.54,0.62]
Persist & fit + generic	0.60	0.22	0.19	67%	67%	19%	2%	0.76 [0.68,0.83]	0.63 [0.57,0.68]
MedSAM, persist & fit boxes	0.57	0.16	0.18	73%	67%	23%	5%	0.74 [0.66,0.81]	0.60 [0.55,0.66]
SAM, source boxes	0.55	0.14	0.22	80%	40%	68%	30%	0.59 [0.54,0.64]	0.48 [0.45,0.52]
Grounding DINO + SAM, default threshold	0.05	0.01	0.02	40%	53%	77%	77%	0.23 [0.17,0.29]	0.18 [0.13,0.23]
UltraSAM3, a-priori prompts, default threshold	0.68	0.21	0.00	60%	87%	45%	2%	0.69 [0.62,0.77]	0.55 [0.50,0.61]
UltraSAM3, a-priori prompts, tuned threshold	0.61	0.14	0.00	80%	100%	13%	2%	0.73 [0.65,0.81]	0.59 [0.54,0.63]
Source models^s	0.55	0.72	0.23	13%	40%	71%	33%	0.56 [0.49,0.63]	0.53 [0.47,0.60]
Source models^s + rules	0.54	0.69	0.19	20%	67%	39%	9%	0.69 [0.63,0.75]	0.63 [0.56,0.69]
Self-training (conf.)^s	0.59	0.79	0.17	7%	67%	39%	26%	0.68 [0.61,0.74]	0.63 [0.57,0.70]
Self-training (conf.)^s + rules	0.59	0.76	0.17	13%	67%	39%	5%	0.74 [0.68,0.80]	0.67 [0.62,0.73]
Self-training (aug.)^s	0.58	0.72	0.17	13%	67%	39%	23%	0.68 [0.62,0.74]	0.62 [0.56,0.69]
Self-training (aug.)^s + rules	0.58	0.68	0.17	20%	67%	26%	0%	0.76 [0.70,0.82]	0.68 [0.62,0.74]
Persist & fit^s	0.56	0.68	0.18	13%	67%	35%	5%	0.73 [0.66,0.80]	0.65 [0.58,0.72]
Persist & fit^s + rules	0.56	0.63	0.18	27%	67%	23%	2%	0.77 [0.70,0.83]	0.68 [0.61,0.75]
Supervised reference	0.62	0.61	0.48	27%	33%	16%	7%	0.81 [0.73,0.88]	0.74 [0.66,0.82]

Table 6: Change in balanced Dice of each adaptation method relative to its source models (rows ^{s}: public-data source models), split into the change in Dice on annotated frames (mean of the three structures) and the change in the share of frames without the structure left empty. The change in balanced Dice is half the sum of the two.

Method	174 evaluation frames Balanced	Annotated Dice	Left empty	58 held-out frames Balanced	Annotated Dice	Left empty
BN adapt	−0.09	+0.03	−0.20	−0.13	0.00	−0.25
TENT	0.00	+0.01	−0.01	−0.06	−0.02	−0.10
CoTTA	−0.07	+0.03	−0.18	−0.11	0.00	−0.21
AdaMI	+0.03	0.00	+0.06	+0.03	−0.01	+0.08
DPL	+0.10	+0.05	+0.15	+0.07	0.00	+0.14
Self-training, confidence only	+0.09	+0.06	+0.13	+0.04	+0.01	+0.07
Self-training, augmentation agreement	+0.10	+0.04	+0.16	+0.08	+0.01	+0.14
Persist & fit	+0.15	+0.04	+0.27	+0.16	+0.03	+0.30
BN adapt^s	−0.13	−0.01	−0.24	−0.10	0.00	−0.19
TENT^s	−0.02	+0.01	−0.06	0.00	+0.02	−0.03
CoTTA^s	−0.08	+0.01	−0.17	−0.08	+0.01	−0.16
AdaMI^s	+0.05	+0.01	+0.10	+0.06	−0.01	+0.13
DPL^s	+0.11	0.00	+0.23	+0.09	0.00	+0.18
Self-training, confidence only^s	+0.11	+0.02	+0.19	+0.10	+0.02	+0.18
Self-training, augmentation agreement^s	+0.12	+0.01	+0.22	+0.09	−0.01	+0.19
Persist & fit^s	+0.15	−0.01	+0.30	+0.12	−0.03	+0.26

Table 7: Public or published thyroid and neck ultrasound datasets compared with ours. A dash marks an entry we could not confirm from the cited paper.

Dataset	Device / setting	Acquisition	Participants or images	Structures annotated	Public
TG3K (Gong et al., 2021)	–	Frames from 16 videos	3,585 frames	Thyroid	Yes
TN3K (Gong et al., 2023)	Various devices	Still	3,493 images	Nodule	Yes
DDTI (Pedraza et al., 2015)	Single device	Still	299 patients (637 images)	Nodule	Yes
Stanford AIMI cine (Yamashita et al., 2022)	GE Logiq E9, Siemens S2000	Cine clips	167 patients (17,412 frames)	Nodule	Yes
SegThy (Krönke et al., 2022)	Siemens Acuson NX3 (VF12-4), Piur tracking	Tracked 3D	28 volunteers (aged 24–50) + 186 patients	Thyroid, later carotid, jugular (Chakrabarty & Soni, 2025)	Yes
NNUS (Oh et al., 2023)	GE LOGIQ P6	Still, three scan directions	25 volunteers (790 images)	Boxes: trachea, thyroid, carotid, jugular vein, oesophagus, others	Listed, no files
Luo et al. (Luo et al., 2022)	–	Video, 5 standard views	80 patients	Thyroid, carotid, trachea, jugular vein, oesophagus, others	–
Bolomiti et al. (Bolomiti et al., 2026)	–	Video	280 patients (300 videos)	Thyroid, carotid	–
Ours	SonoStar 5W wireless probe, tablet	Free-hand video	29 participants (16 aged 18–44, 13 aged 69–81)	Thyroid, carotid lumen, trachea	Planned

B Datasheet

We follow the questions of Gebru et al. (2021) in compressed form.

Motivation

The dataset was created to measure how segmentation models for the thyroid, the common carotid artery and the trachea transfer, without target labels, to free-hand wireless-probe sweeps of younger and older adults in which these structures are often absent, scoring frames with and without each structure. Public thyroid datasets are mostly still images or cine clips from cart-based scanners with gland or nodule labels, SegThy adds tracked 3D volumes with thyroid and later carotid and jugular labels, and we found no public video dataset that labels the trachea (Table 7). Author and funding details are withheld for anonymous review.

Composition

Each instance is the screen recording of one participant's free-hand bilateral thyroid examination in transverse and sagittal planes. Of the 29 participants, 16 are aged 18–44 years (15 from a laboratory session and one from a community screening event) and 13 are aged 69–81 years (all from the event). All older participants self-reported Asian ethnicity, and the younger group is predominantly Asian. The recordings total 86 minutes at 1800×1080 pixels and about 20 fps, with an imaging depth of 40 or 60 mm (100 mm in 16 sampled frames of two younger participants). From them, 20,014 frames are sampled at 4 fps for adaptation. Six frames per participant form the 174 evaluation frames and two further frames per participant form the 58 held-out frames. The evaluation frames carry 201 polygons on 94 frames and the held-out frames 57 polygons on 28 frames, and the remaining 80 and 30 frames are flagged as showing no identifiable structure. There are no nodule, diagnostic or

pathology labels. The data are lossy screen recordings of the tablet display, not DICOM images or radio-frequency data.

COLLECTION

All recordings were made with one SonoStar 5W wireless linear probe (85 mm aperture, 256 elements, 7.5 MHz thyroid preset) streaming to a tablet whose screen was recorded. Members of the research team performed all scans free-hand. Every participant gave written informed consent to the anonymised use and sharing of their data for research. The two volunteers whose scans trained the source landmark model and plane classifier are members of the research team, are not among the 29 target participants and signed the same consent form.

PREPROCESSING AND LABELLING

Frames are sampled at 4 fps and cropped to the image area, and frozen-display and blank frames are removed. The main experiments use frames that still contain the scanner app's overlay text, and the main methods were re-run on the released frames, in which it is masked (App. K). One clinician, a head and neck specialist with at least ten years of thyroid ultrasound experience, annotated all 232 labelled frames from scratch, following a written protocol, in a web tool that displays no model prediction. Polygons mark the thyroid (lobes and isthmus, including nodules), the lumen of the common carotid artery (excluding the jugular vein) and the trachea as its hyperechoic arc together with the dome it encloses. Frames without any identifiable structure are flagged. Uncertain structures were left unlabelled. After the second internal review, the same clinician marked every unannotated structure of every frame as absent or uncertain, without seeing model predictions, and the flags of the evaluation frames are released with their annotations so that uncertain cases can be excluded from the negative-side scores, while those of the held-out frames stay hidden with their labels (App. L). A second clinician with the same specialty and experience annotated the 58 held-out frames from scratch (presence κ 0.91–0.97, Dice 0.87–0.90 on frames both annotated). The first clinician also labelled the imaging plane of every labelled frame by eye as transverse, sagittal or cannot tell, without seeing model predictions, to measure the plane classifier (App. D). The 58 held-out frames were never displayed during method development and were annotated after the method was frozen.

USES

The dataset is intended for source-free and test-time adaptation, few-shot segmentation and temporal modelling of free-hand neck ultrasound. Its official adaptation pool is all 20,014 frames without labels, including the images of the labelled frames. The 174 evaluation frames serve for development and evaluation, and the 58 held-out frames are the hidden test set. The official track ranks label-free entries by balanced Dice on the held-out frames of the released data, reports entries tuned or trained on the evaluation labels outside the ranking, and each entry states which set of source models it builds on. A participant-disjoint protocol adapts on two of three fixed participant folds and predicts the third. The few-shot track trains on the labelled frames of k participants from the 174 evaluation frames and tests on the held-out frames of the other participants. Dice scores are averaged within each participant and then over participants, and false detection and miss rates are pooled over frames. The data come from one device, one site and 29 participants, all annotated frames of older participants are at 60 mm, and the labels carry no diagnostic information. Hence the dataset does not support diagnostic use or claims about other devices and populations.

DISTRIBUTION

The de-identified cropped videos, the frame crops, the annotations (labelme JSON), uncertainty flags and plane labels of the 174 evaluation frames, the split files, the source model weights and the code are available on a public research repository, anonymised for review (`https://huggingface.co/datasets/freehand-us-anon/freehand-thyroid-sweeps`). The released sweeps are re-encoded without audio or container metadata, non-imaging system screens are blacked out (563 of 107,614 frames), and automatic checks found no audio stream, creation-time tag or residual static text in any of the 29 videos. The labels of the 58 held-out frames are withheld to keep a hidden test set: their images are included, and the maintainers will score submitted predictions with

the released evaluation script. Overlay text is masked, participant codes are replaced by random codes, demographic information is reduced to age group and sex, and the released data contain no names, faces or dates. The licences are CC BY-NC 4.0 for the data and MIT for the code.

MAINTENANCE

The authors maintain the dataset and its repository. They keep the participant key that links the random codes to participants private and do not distribute it. The evaluation code will be released with the data, so results on the official splits are comparable across methods, and the maintainers will score predictions on the hidden held-out frames on request.

C DATASET, PROTOCOL AND STATISTICS

This appendix gives the full description summarised in §4.

Acquisition. All target sweeps were recorded with one wireless linear probe (SonoStar 5W, 85 mm aperture, 256 elements, 7.5 MHz thyroid preset) streaming to a tablet and captured as lossy screen recordings (1800×1080, variable frame rate of about 20 fps) without DICOM or radio-frequency data. Research-team members performed one continuous free-hand bilateral thyroid examination per participant in transverse and sagittal planes. A **laboratory session** included 15 adults aged 18–44 and a **community screening event** 13 adults aged 69–81 and one younger volunteer, giving a younger group of 16 (median age 29, 8 female) and an older group of 13 (median 75, 8 female, Table 8). All participants gave written informed consent to the anonymised use and sharing of their data for research. Sweeps last 70–334 s (median 118 s in the laboratory and 223 s at the event, 86 min in total). Imaging depth was 40, 60 or, in 16 frames, 100 mm (60 mm for all event sweeps, mid-sweep changes for three laboratory participants) and is read per frame from the image width.

Source data. The landmark model (90 frames with 26 carotid and 35 trachea boxes) and the plane classifier (clip-level plane labels) were trained on sweeps of two research-team volunteers, recorded with a different handheld probe and app. Neither volunteer is a target participant or part of the release. After the target recordings, and after viewing its predictions on unlabelled target frames, we retrained the landmark model on the same source data with stronger augmentation because it missed carotids imaged at a different depth (carotid IoU on the source validation clips rose from 0.41 to 0.61, App. D). This decision used target images, which a strict source-free protocol forbids, so Table 22 also reports the source models and persist & fit with the original landmark model. The gland model was trained on the 3,226 training images of TG3K (Gong et al., 2021), with 359 further images held out for checkpoint selection.

Frames. Frames are sampled at 4 fps, cropped to the image area and removed if the display is frozen or blank (mean intensity at most 22), leaving an adaptation pool of 20,014 frames, about 2,000 of which show no tissue because the probe is off the skin. Without using any segmentation output, we selected six *evaluation frames* per participant (174 in total) among frames with enough tissue in the image centre, four transverse and two sagittal according to the plane classifier, spread over the sweep and at least 2 s apart in nominal time (frame index divided by the average frame rate, so the true spacing can be slightly shorter). Two further *held-out frames* per participant (58 in total) were drawn with the same content criterion near the start and end of each sweep, at least 2 s in nominal time from every evaluation frame and without plane balancing. Age is confounded with depth (all 78 evaluation frames of older participants are at 60 mm), session, operator settings and self-reported ethnicity (all older participants are Asian), so we also compare the age groups at matched 60 mm depth (Table 20). The main experiments use frames that contain the scanner's overlay text, and App. K repeats the main methods on the released data, in which it is masked.

Annotation. One clinician, a head and neck specialist with at least ten years of thyroid ultrasound experience, labelled all 232 evaluation and held-out frames from scratch in a web tool that shows no model prediction and offers polygon, freehand and click-to-grow drawing. A written protocol defines the thyroid as lobes and isthmus including nodules, the common carotid artery as its lumen (excluding the jugular vein) and the trachea as the hyperechoic arc and the dome it encloses (excluding the reverberation tail). Uncertain structures were not labelled, so a missing annotation means

the structure is absent or uncertain, and correct predictions of uncertain structures count as errors. After the second internal review, the same clinician marked each unannotated structure as absent or uncertain (App. L), and a second clinician annotated the 58 held-out frames from scratch (App. M). The evaluation frames carry 201 polygons on 94 frames, with the thyroid on 79 (27 participants), the carotid on 62 (27) and the trachea on 54 (26), and the other 80 are flagged as showing no identifiable structure. Held-out frames over-weight probe placement and lift-off, with 30 of the 58 flagged and the thyroid, carotid and trachea on 27 (20 participants), 15 (13) and 15 (15), so predicting nothing on every frame (the *predict nothing* reference, §3) scores an all-frame Dice of 0.67 there (0.63 on evaluation frames). Of the 110 frames without identifiable structure, the first clinician later marked all three structures as absent on 54 and at least one as uncertain on 56.

Method development and splits. Before any annotation existed, we designed the anatomical rules and the lumen detector, set their thresholds and decided to retrain the landmark model by visually inspecting model outputs and pseudo-labels on unlabelled target frames, including the 174 evaluation frames. The configuration of persist & fit and of the TTA, temporal-averaging and two self-training baselines was fixed before annotation, and App. P lists the references, controls and analyses added afterwards. Held-out frames were never displayed during development and were annotated only after the configuration was fixed. The dataset therefore defines three splits: the adaptation pool, which includes the images of the labelled frames but not their labels (transductive adaptation), the 174 evaluation frames for development and evaluation, and the 58 held-out frames as the hidden test set. A participant-disjoint protocol adapts on two of three fixed participant folds and predicts the third, so that no test participant is seen during adaptation. The few-shot track fine-tunes on the labelled evaluation frames of k participants (six frames each) and tests on the held-out frames of the other participants (App. P). The *official track* uses the released frames and ranks label-free entries by balanced Dice on the hidden held-out frames (last column of Table 3), with all-frame Dice, miss and false detection rates alongside. Entries whose prompts or thresholds were tuned on the evaluation labels, or that were trained on them, are marked and reported outside the ranking. Both sets of source models will be released, and each entry states the set it builds on.

De-identification and release. For release, participants are mapped to random codes, demographics are reduced to age group and sex, and the full-rate sweeps and the 232 labelled frames are cropped to the ultrasound image with overlay text masked by fixed boxes. Data (CC BY-NC 4.0) and code (MIT) will be released on a public research repository, as documented in the datasheet (App. B), and Table 7 compares the dataset with public or published thyroid and neck ultrasound datasets.

Protocol. Label-free methods adapt on the 20,014 unlabelled frames and are scored on the 174 evaluation frames and the 58 held-out frames, the clean test set. Per structure we report Dice on annotated frames, the miss rate (annotated frames without any predicted pixel of the structure) and the false detection rate (frames without that annotation, including uncertain cases, with any predicted pixel). All-frame Dice scores every frame, with 1 for an empty prediction on a frame without the structure and 0 for any predicted pixel there. With 80 of 174 evaluation and 30 of 58 held-out frames showing no identifiable structure, it rewards abstention, and *predict nothing*, a mandatory reference, scores 0.63 and 0.67. Balanced Dice averages, per structure, Dice on annotated frames and the share of frames without the structure left empty, so predict nothing scores 0.50, and we report it also with uncertain structures excluded from the negatives (Bal. cert.). With a weight on Dice on annotated frames anywhere from 0.45 to 0.6 instead of 0.5, every method with rules stays above predict nothing, and BN adapt and CoTTA stay below it on both frame sets (App. A). Dice scores are averaged within participants, then over participants and structures, and miss and false detection rates are pooled over frames. Paired tests against persist & fit use a two-sided participant bootstrap (5,000 resamples), Holm-corrected over the 82 comparisons per frame set and metric, and the tables mark significant differences (App. P lists uncorrected and Wilcoxon p-values). Only the method configuration, six baselines, Dice on annotated frames and a false detection rate were fixed before annotation, and App. P dates every later addition, including balanced Dice.

All older participants self-reported Asian ethnicity. The younger group is predominantly Asian, with White, Black and multiracial participants. Five participants (four older) reported a history of thyroid nodules.

Table 8: Target dataset by age group. The one participant aged 18–44 recorded at the community event is counted with the younger group. Imaging depth is read per frame from the crop width, and three younger participants changed depth during the sweep. Two frames at the start of one older participant's sweep show a tablet system screen and are counted at neither depth, as are 16 frames of two younger participants imaged at 100 mm. Planes are those predicted by the plane classifier.

	Younger	Older
Participants	16	13
Age, range (median)	18–44 (29)	69–81 (75)
Female	8	8
Sweep duration, total	38.0 min	48.0 min
Frames used for adaptation (4 fps)	8,788	11,226
at 40 / 60 mm	5,628 / 3,144	0 / 11,224
Evaluation frames (transverse / sagittal)	96 (64 / 32)	78 (52 / 26)
at 40 / 60 mm (participants)	70 (13) / 26 (6)	0 / 78 (13)
with thyroid / carotid / trachea	45 / 31 / 27	34 / 31 / 27
flagged as showing no structure	45	35
Held-out frames (transverse / sagittal)	32 (19 / 13)	26 (16 / 10)

D SOURCE MODELS

The gland model, the plane classifier and the original landmark model were trained before the target recordings. Our experiments use a landmark model retrained afterwards on source data alone.

Gland model g. U-Net with an ImageNet-initialised EfficientNet-B2 encoder, trained on the 3,226 training images of TG3K (Gong et al., 2021) (256×256, Dice + binary cross-entropy, AdamW, learning rate 3×10^{-4}, cosine schedule, 80 epochs) with generic acquisition augmentation (scale-down with padding, aspect-ratio jitter, noise and blur). We keep the checkpoint with the best IoU on the 359 validation images (0.70).

Landmark model h. U-Net with an EfficientNet-B0 encoder (input 352×256) trained on 90 frames from sweeps of two volunteers recorded with a different handheld probe and app, with 26 carotid and 35 trachea boxes filled as masks and a clip-level train/validation split (75 / 15 frames). The original model used flip, rotation, brightness, gamma and noise augmentation and kept the best of 55 epochs on validation. Before any annotation existed, its predictions on unlabelled target frames, including frames later annotated for evaluation, showed that it often missed carotids on the new device, which we attributed to the different imaging depth. We retrained it on the same frames and split with added zoom (0.8–$1.9\times$), aspect-ratio jitter (0.7–1.5), translation and blur, a wider brightness range, four repeats per epoch and 60 epochs, keeping the final epoch. No target frame was used for training. We adopted the retrained model because carotid IoU on the source validation clips rose from 0.41 to 0.61 (6 carotid boxes from one volunteer, no trachea boxes). The carotid IoU logged on the same clips every ten epochs, averaged over batches, ranged from 0.54 to 0.72 (0.63 at the final epoch). At that time we also viewed, for both models, how often a predicted carotid and a predicted trachea passed the rules of Table 1 on 1,500 unlabelled transverse target frames. From the original to the retrained model, the carotid rate went from 0.09 to 0.05 at 40 mm and from 0.08 to 0.16 at 60 mm, and the trachea rate from 0.44 to 0.33 and from 0.16 to 0.09. With the original landmark model as teacher, the source models reach an all-frame Dice of 0.41 and our method 0.65 (Table 22).

Plane classifier c. EfficientNet-B0 (timm), 224×224, transverse vs. sagittal, trained on the same two volunteers. Per-frame accuracy is 96.5% on a clip-level validation split and 85.0% in leave-one-subject-out evaluation across the two volunteers (88.9% and 81.1% for the two folds, 94.6% with a clip-level majority vote). To measure it on the target device, the first clinician labelled the plane of the 232 annotated frames by eye as transverse, sagittal or cannot tell, in random order and in a tool that showed no prediction. The 100 frames marked cannot tell (43%) are mostly frames without an identifiable structure, as 92 of them have none of the three structures annotated. On the other 132 frames the classifier matches the manual plane on 121 (91.7%, Wilson 95% CI 85.7–95.3%, Cohen's κ 0.81), with 96.6% of transverse and 82.2% of sagittal frames correct (91.3% on

the evaluation and 92.9% on the held-out frames). Running rules at test time with the manual plane on the 11 frames where the two disagree changes the balanced Dice of each of 16 methods by at most 0.007 on the evaluation frames and leaves the held-out results unchanged. Our method uses the per-frame prediction without smoothing to select the rule set of Table 1, and the violation rate uses it for its plane-dependent checks (App. O). It also balanced planes when evaluation frames were chosen, so the transverse/sagittal split of those frames is the predicted one.

E IMPLEMENTATION DETAILS

Ignore or background. A component that fails a rule is set to ignore, because a failed test shows only that the teacher's class is implausible, not that the structure is absent, and relabelling it as background could teach the student to suppress the structure wherever the teacher was wrong. In the ablation that relabels failing components, together with candidate lumens and unconfirmed proposals, as background, thyroid Dice on annotated frames falls from 0.64 to 0.59 and all-frame and balanced Dice change by at most 0.01 (App. F).

Completion, verified by persistence. Filtering alone can only remove, so a structure the teacher usually misses, such as the carotid, keeps a background label and gets few positive examples. Two steps in transverse frames address this. Background inside *candidate lumens*, dark round blobs enclosed by brighter tissue, is withheld so that missed carotids are not taught as background. Completion adds positives by proposing, when a thyroid lobe but no carotid is accepted, the candidate lumen that best fits the layout beside the lobe (see Vessel candidates and completion below). A proposal in frame t is accepted if a proposal in a frame of $\mathcal{N}(t)$ covers at least 30% of it within a 4-pixel tolerance (without flow alignment) and is set to $\varnothing$ otherwise. With persistence, neither withholding and completion (column C of Table 9) nor flow alignment changes carotid or all-frame Dice significantly on either frame set.

Self-training. The student is a U-Net (Ronneberger et al., 2015) with an EfficientNet-B2 encoder (Tan & Le, 2019), initialised from g and trained with cross-entropy plus Dice loss over non-ignored pixels and with photometric and geometric augmentation. After each round the student becomes the teacher, pseudo-labels are regenerated with the same tests, and a new student is trained from the same initialisation. Because the tests only shape the pseudo-labels, persist & fit distils them into one per-frame network (10.0M parameters) that abstains without rules, plane classifier or neighbouring frames at inference. The remaining paragraphs give the parameters of each step.

Frames. Videos are sampled at 4 fps. The image region of each frame is located by the column-intensity profile of the display area. Its width gives the imaging depth of each frame (about 1,575 pixels at 40 mm and 1,080 at 60 mm in the 1800×1080 recording), and the region is about 1.9 times as wide as it is high at 40 mm and 1.3 times at 60 mm. Near-duplicate samples (mean absolute difference below 0.6 grey levels from the previous sample on a 160×120 thumbnail), which include frozen display, are removed. Frames with mean intensity at most 22 are blank and are not used for adaptation. About 2,000 of the remaining frames show no tissue because the probe is off the skin, and they are kept. Evaluation and held-out frames were drawn only from frames whose central $60\%\times60\%$ region has mean intensity at least 28 and at least 12% of pixels brighter than 45. The students, the gland model and the rules work on the region resized to 256×256 without preserving its aspect ratio (App. N describes the consequence for the carotid at 40 mm).

Persistence. Farnebäck optical flow is computed from frame t to each neighbour on 256×256 grey-scale frames (pyramid scale 0.5, 3 levels, window 21, 3 iterations, polynomial neighbourhood 5, $\sigma = 1.1$). Neighbours further than $\delta = 0.6$ s apart are not used, and $\gamma = 0.7$. A frame with no neighbour within δ (sweep ends and gaps left by removed frames) is tested for confidence only. The no-flow ablation uses the neighbours' probabilities without warping and changes nothing else.

Anatomical tests. Rules are applied to 8-connected components of the reliable labels. Unreliable pixels count as background during the tests and are ignored afterwards. Circularity is $4\pi A/P^2$ of the largest contour, and elongation is the aspect ratio of the minimum-area rectangle. "Darker than its rim" means the mean intensity inside is below $0.85\times$ that of a 4-pixel ring around it. A trachea

top edge is hyperechoic if the mean intensity of a band spanning its bounding box, from two rows above its top down to 20% of its height, is at least $1.25\times$ that of its pixels in the lower 60% of its height (components under 6 rows fail). Trachea and thyroid components are tested in decreasing order of area, and only the first trachea component that passes is kept. When a carotid or trachea is accepted, a transverse thyroid component also needs $a \geq 0.4\%$, and with both present it must lie between the trachea and each accepted carotid.

Vessel candidates and completion. In transverse frames, candidate vessel lumens are components of pixels darker than $0.65\times$ their 41×41 local mean and below the 60th intensity percentile of the frame, after median and Gaussian smoothing and morphological opening. Each candidate needs area 0.3–8%, a bounding box at least $0.03W$ from the left edge, at least $0.05W$ from the right edge and ending above $0.85H$, a centroid at least $0.12H$ below the top, circularity ≥ 0.6, elongation ≤ 1.8 (on the 256×256 grid) and, in all four quadrants of a 5-pixel ring, a mean intensity of at least $1.4\times$ the lumen mean. Background-labelled pixels of every candidate are set to ignore unless the candidate becomes a confirmed completion. Completion runs in transverse frames with an accepted thyroid and no accepted carotid. It discards candidates that overlap accepted thyroid or trachea by more than 20% of their area and pairs each remaining candidate with every thyroid component of $a \geq 0.4\%$. A pair qualifies if the horizontal gap between their bounding boxes is below $0.08W$, their vertical extents overlap once the lobe's extent is widened by $0.1H$ on each side, and the candidate lies on the side of the lobe away from the accepted trachea, if there is one. The qualifying candidate with the highest circularity minus gap$/W$ is proposed. A proposal is confirmed if a neighbour within δ has proposal pixels inside the 9×9 dilation of the proposal amounting to at least 30% of its area (no flow alignment), and all its pixels then become carotid. Without the persistence test (ablation rows without T), every proposal is confirmed.

Student training. Each round trains a new student from the source initialisation for six epochs over every second frame (AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, cosine schedule, batch 16), with horizontal flip, brightness/contrast, gamma, Gaussian noise, and scale (0.85–1.15) and translation ($\pm8\%$) augmentation. Ignored pixels contribute to neither loss term. One adaptation run takes 25–50 min on one L40S GPU, and the 10.0M-parameter student takes 35 ms per frame on four server CPU threads (not timed on the tablet).

Ablation variants. Column C of Table 9 switches candidate withholding and completion on together. In the reject-to-background variant, components that fail a rule become background, and candidates and unconfirmed proposals keep the teacher's background label. The no-flow variant is described above, and the γ variants set the confidence threshold to 0.5 or 0.9. Other variants drop the hyperechoic-top test, stop after one round or run three, or weight the cross-entropy of structure class k by $(f_0/f_k)^{1/2}$ clipped to $[1, 5]$, where f_k is its pseudo-label pixel frequency and f_0 that of background. The one-round variant is the first-round student of the default run.

Rules at test time. Table 10 applies the anatomical test to a method's final argmax mask as post-processing. The mask (by nearest neighbour) and the frame are resized to 256×256, and the rules of Table 1 run with the same thresholds and the same per-frame plane prediction as in pseudo-label generation, with three differences: failing components become background, candidate withholding and completion are switched off, and no confidence or persistence test is applied. Each 8-connected component of each class in the native-resolution mask is then kept if more than half of its pixels survive at 256×256 and removed otherwise, so the post-processing deletes whole components and never moves a boundary. It removes true detections along with false ones and, for example, raises the thyroid miss rate of persist & fit on the evaluation frames from 1% to 5%.

Baselines. BN adapt, TENT and CoTTA adapt the two source networks and predict with the composed teacher $p^{(0)}$ (above). For BN adapt, batch-norm statistics are re-estimated on every third target frame. TENT updates batch-norm affine parameters by entropy minimisation (Adam, 10^{-4}, one pass over every second frame, batch 16) and then re-estimates statistics. CoTTA trains the network for one pass over every second frame (Adam, 10^{-5}, batch 16) on the soft predictions of an EMA teacher ($\alpha = 0.999$), which are averaged over a flip and $0.9\times$ and $1.1\times$ rescalings when the source anchor's mean confidence is below 0.9, and applies stochastic restoration ($p = 0.01$). The final teacher, with batch-norm statistics re-estimated on target frames, is used for prediction. Tempo-

ral averaging fuses flow-warped source probabilities of adjacent frames exactly as in the persistence test, without training. Augmentation agreement averages teacher probabilities over the identity, a horizontal flip and rescalings to 224^2 and 288^2, and keeps a pixel if the averaged confidence is at least γ and at least three of the four augmented predictions agree. Confidence-only self-training uses only the confidence threshold γ. Both self-training baselines share the student, schedule and rounds of our method.

Generic abstention. Each control acts per structure on a method's native-resolution argmax mask, computed from its 256×256 probabilities as for all other methods, and only removes predictions. Minimum area removes 8-connected components smaller than a fraction $\tau_a \in \{0, 0.1, 0.2, 0.5, 1, 2, 3, 5, 8, 12, 16, 20, 25\}\%$ of the image, largest component keeps only the largest component, and the presence threshold drops the structure from the frame if its maximum probability is below $\tau_p \in \{0, 0.5, \ldots, 0.99999\}$. The best generic control searches all combinations of the three. For every method, control and structure, the setting with the highest balanced Dice of that structure on the 174 evaluation frames is kept (ties go to the milder setting) and applied unchanged to the held-out frames.

Shape tests in true proportions. This post-hoc variant multiplies horizontal coordinates by the width-to-height ratio of the frame's image region (about 1.9 at 40 mm and 1.3 at 60 mm) before computing circularity and elongation in the rules and the lumen detector, and changes nothing else.

Participant-disjoint protocol. The 29 participants are stratified by age group, permuted with a fixed seed and dealt into three folds of 10, 10 and 9 participants (split file released with the code). For each fold, adaptation uses the unlabelled sweeps of the other two folds only, and the resulting student predicts the annotated frames of the fold, so every annotated frame is predicted by a model that never saw its participant. The rules and thresholds were designed before this protocol existed, while viewing frames of all participants.

Supervised reference. The supervised reference trains the student architecture from the gland model's weights, with a new four-class head, using cross-entropy plus multi-class Dice loss for 60 epochs (AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, cosine schedule, batch 8, the student's augmentation), and keeps the final epoch. Pixels outside the annotated polygons, including uncertain structures, are background. On the evaluation frames it is cross-validated over five participant-grouped folds stratified by age group. For the held-out frames, one model is trained on all 174 evaluation frames.

F Ablation

Table 9: Ablation of the acceptance tests. T: persistence. A: anatomical test with failures withheld as ignore. C: withholding of background in candidate lumens and carotid completion. Mean: Dice on annotated frames averaged over thyroid, carotid and trachea. *Without T, completions are not confirmed by an adjacent frame. ‡Rejected regions become background. §Neighbours used without optical-flow alignment. [a]Run after annotation.

			Dice on annotated frames (174)			174 frames		58 held-out	
T	A	C	Thyroid	Carotid	Mean	All-frame	Balanced	All-frame	Balanced
			0.66	0.25	0.37	0.59	0.53	0.60	0.49
✓			0.66	0.19	0.34	0.63	0.55	0.63	0.52
	✓		0.64[a]	0.23	0.36	0.64	0.56	0.70	0.57
	✓	✓*	0.65	0.31	0.40	0.66	0.59	0.71	0.58
✓	✓		0.65	0.23	0.35	0.68	0.59	0.72	0.59
✓	✓	✓	0.64	0.23	0.35	0.68	0.59	0.74	0.61
✓	✓‡	✓‡	0.59	0.23	0.33	0.69	0.60	0.73	0.61
✓§	✓	✓	0.64[a]	0.24	0.35	0.68	0.60	0.73	0.60

Persistence (T) and the anatomical test (A) each raise all-frame Dice over confidence-only self-training (0.63 and 0.64 against 0.59, Table 9), and persist & fit exceeds both (all $p<0.002$). Optical-flow alignment makes no measurable difference at this frame rate (all-frame Dice without it 0.68 and 0.73 against 0.68 and 0.74, $p\geq 0.42$). Completion (C) raises carotid Dice only without persistence (A 0.23, A+C 0.31, T+A+C 0.23, $p=0.031/0.039$ against A+C) and adds nothing measurable to T+A (all-frame $p=0.54/0.70$). Withholding rejected regions, candidate lumens and unconfirmed proposals as ignore instead of background keeps thyroid Dice on annotated frames higher (0.64 against 0.59, $p=0.007/0.04$, held-out 0.60 against 0.56, $p=0.004/0.06$) with no significant change in all-frame Dice ($p=0.10/0.39$). Setting γ to 0.5 or 0.9 changes all-frame and balanced Dice by at most 0.02 (Table 22).

G ABSTENTION CONTROLS

Table 10: Abstention controls: balanced Dice of each method as is, with the anatomical rules applied at test time (failing components removed), with the best generic abstention (minimum area, largest component and presence threshold per structure) tuned on the labelled 174 evaluation frames, whose scores on those frames are therefore optimistic, and with generic presence and area thresholds chosen without labels by Otsu's method on the unlabelled adaptation frames. †/‡Persist & fit (as is) is significantly better/worse (Holm-adjusted $p < 0.05$, label-free column not tested in this family). Table 11 gives each control separately and all-frame Dice.

	As is		+ Anatomical rules		+ Generic, tuned on labels		+ Generic, label-free	
Method	174	58	174	58	174	58	174	58
Source models	0.44†	0.45†	0.55	0.57†	0.58	0.55	0.54	0.56
Temporal averaging	0.46†	0.45†	0.56	0.58	0.58	0.57	–	–
Self-training, confidence only	0.53†	0.49†	0.58	0.59	0.59	0.59	0.56	0.54
Self-training, augmentation agreement	0.54†	0.52†	0.58	0.60	0.60	0.58	0.58	0.55
DPL	0.54†	0.52†	0.59	0.59	0.60	0.57	0.58	0.55
Persist & fit	0.59	0.61	0.60	0.61	0.61‡	0.63	0.60	0.61

The tuned thresholds are extreme. For the thyroid, the best generic control keeps a prediction only if its maximum probability reaches 0.995–0.9999 for every method, often together with a minimum area of 2–3%. On the trachea it removes components below 3–12% of the image, and on the carotid it mostly applies no threshold, because carotid false detections are already rare. The largest-component control alone never helps, because it keeps one component on every frame with a prediction. With the same setting for all structures and no tuning (minimum area 0.2% or 1%, or presence threshold 0.9 or 0.99), held-out balanced Dice is 0.52–0.54 for the two self-training baselines, 0.48–0.53 for the source models and 0.60–0.62 for persist & fit (function `apriori_table` in `abstention.py`). Label-free operating points use the same controls with thresholds chosen on the 20,014 unlabelled adaptation frames: per method and structure, Otsu's method on the histogram of the logit of the per-frame maximum probability (over the frames on which the structure is predicted) sets the presence threshold, and Otsu's method on the logarithm of the predicted area sets the minimum area. This procedure was fixed before any of its labelled results was examined. The presence thresholds lie between 0.93 and 0.99997 and are highest for the self-trained models, the area thresholds between 0.2% and 4% of the image. On the held-out frames the label-free control stays 0.03–0.05 below the rules for the two self-training baselines and DPL ($p=0.010$, <0.001 and 0.12) and within 0.01 of them for the source models and persist & fit, and the presence test alone does almost all of the work. Figure 3 sweeps one presence threshold for all structures: even the best threshold, chosen on the frames being scored, stays below the rules for the two self-training baselines and DPL on both frame sets. Table 12 gives the threshold-free view, the frame-level AUROC with which the maximum probability separates frames with and without each structure.

Table 11: Every abstention control on both frame sets. Generic controls are tuned on the 174 evaluation frames (App. E), except the label-free rows, whose thresholds come from the unlabelled adaptation frames. FP: false thyroid detections, Miss: carotid misses, both on the 174 frames.

		174 evaluation frames				58 held-out frames	
Method	Control	All-frame	Balanced	FP thyroid	Miss carotid	All-frame	Balanced
Source models	as is	0.50	0.44	99%	71%	0.54	0.45
	anatomical rules	0.63	0.55	52%	89%	0.70	0.57
	minimum area	0.64	0.55	38%	79%	0.66	0.53
	largest component	0.50	0.43	99%	71%	0.54	0.45
	presence threshold	0.68	0.57	19%	71%	0.70	0.56
	best generic	0.68	0.58	13%	71%	0.68	0.55
	presence, label-free	0.63	0.54	44%	89%	0.68	0.56
	area, label-free	0.56	0.48	79%	82%	0.61	0.50
	both, label-free	0.63	0.54	44%	89%	0.68	0.56
Temporal averaging	as is	0.53	0.46	96%	73%	0.56	0.45
	anatomical rules	0.65	0.56	47%	81%	0.72	0.58
	minimum area	0.65	0.56	38%	73%	0.66	0.53
	largest component	0.53	0.46	96%	73%	0.56	0.45
	presence threshold	0.68	0.57	28%	73%	0.72	0.57
	best generic	0.68	0.58	20%	73%	0.69	0.57
Self-training, confidence only	as is	0.59	0.53	74%	69%	0.60	0.49
	anatomical rules	0.65	0.58	51%	77%	0.73	0.59
	minimum area	0.65	0.58	42%	69%	0.68	0.56
	largest component	0.59	0.53	74%	69%	0.60	0.49
	presence threshold	0.67	0.58	32%	69%	0.71	0.59
	best generic	0.68	0.59	29%	69%	0.71	0.59
	presence, label-free	0.64	0.56	46%	84%	0.66	0.54
	area, label-free	0.63	0.56	59%	71%	0.65	0.53
	both, label-free	0.64	0.56	45%	84%	0.67	0.54
Self-training, augmentation agreement	as is	0.60	0.54	68%	69%	0.63	0.52
	anatomical rules	0.67	0.58	46%	77%	0.73	0.60
	minimum area	0.65	0.57	49%	69%	0.67	0.55
	largest component	0.60	0.54	68%	69%	0.63	0.52
	presence threshold	0.68	0.59	28%	69%	0.70	0.58
	best generic	0.68	0.60	28%	69%	0.70	0.58
	presence, label-free	0.66	0.57	38%	79%	0.67	0.55
	area, label-free	0.64	0.57	53%	73%	0.67	0.55
	both, label-free	0.66	0.58	38%	79%	0.68	0.55
DPL	as is	0.60	0.54	69%	69%	0.63	0.52
	anatomical rules	0.67	0.59	45%	74%	0.72	0.59
	minimum area	0.66	0.58	41%	69%	0.68	0.56
	largest component	0.60	0.54	69%	69%	0.63	0.52
	presence threshold	0.67	0.59	37%	71%	0.69	0.56
	best generic	0.68	0.60	36%	71%	0.69	0.57
	presence, label-free	0.66	0.58	37%	84%	0.68	0.55
	area, label-free	0.63	0.56	57%	73%	0.65	0.54
	both, label-free	0.67	0.58	37%	84%	0.68	0.55
Persist & fit	as is	0.68	0.59	40%	66%	0.74	0.61
	anatomical rules	0.70	0.60	32%	73%	0.74	0.61
	minimum area	0.70	0.61	29%	66%	0.76	0.63
	largest component	0.68	0.59	40%	66%	0.74	0.61
	presence threshold	0.70	0.61	32%	66%	0.76	0.63
	best generic	0.70	0.61	32%	66%	0.76	0.63
	presence, label-free	0.69	0.60	26%	77%	0.75	0.61
	area, label-free	0.69	0.60	35%	73%	0.75	0.62
	both, label-free	0.69	0.60	26%	77%	0.75	0.61

Table 12: Frame-level presence AUROC of the maximum probability of each structure over all frame-structure pairs. Excluding uncertain pairs changes the means by at most 0.04 and raises the thyroid AUROC by up to 0.12 (source models 0.96 and 0.89).

	174 evaluation frames				58 held-out frames			
Method	Thy.	Car.	Tra.	Mean	Thy.	Car.	Tra.	Mean
Source models	0.83	0.59	0.52	0.65	0.85	0.61	0.69	0.72
Self-training, confidence only	0.84	0.71	0.63	0.73	0.86	0.71	0.53	0.70
Self-training, augmentation agreement	0.84	0.59	0.61	0.68	0.84	0.53	0.44	0.60
DPL	0.83	0.74	0.53	0.70	0.89	0.68	0.57	0.71
Persist & fit	0.86	0.76	0.78	0.80	0.93	0.73	0.73	0.80

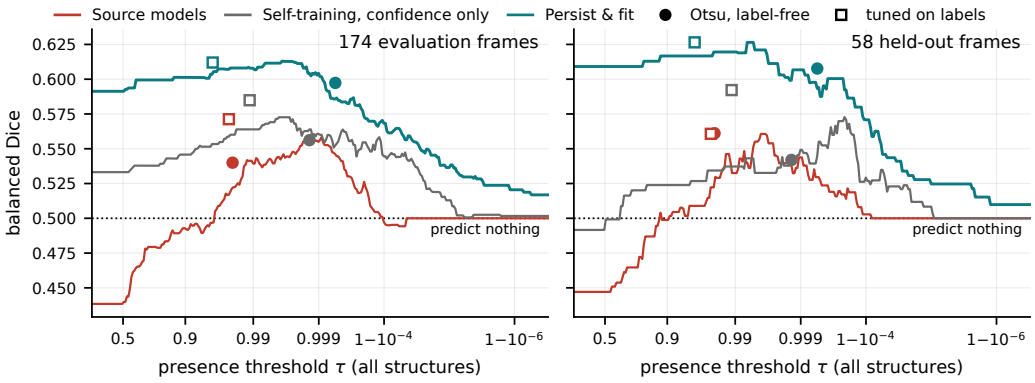


Figure 3: Balanced Dice against one presence threshold applied to all structures (logit scale), on the evaluation (left) and held-out frames (right). Circles: label-free operating points (Otsu on unlabelled frames). Squares: presence thresholds tuned on the labelled evaluation frames. Both are per structure and plotted at the mean logit of their three thresholds.

H MEDICAL SFDA AND FOUNDATION-MODEL BASELINES

DPL. DPL (Chen et al., 2021) replaces only the pseudo-label filter of our self-training pipeline, so it shares the student, schedule and two rounds of the other self-training baselines. The teacher's prediction is averaged over ten Monte-Carlo dropout passes, with dropout of 0.5 on the deepest encoder features and 0.1 before the segmentation head, placed as in the official code because our U-Nets have no dropout layers. A pixel is labelled structure k if $p_k \geq 0.75$ and background otherwise, and it is kept only if the standard deviation of its class probability over the passes is below 0.05 (temperature 2, as in the code) and its decoder feature is closer to the per-image prototype of its own pseudo-class than to the other prototypes. The other pixels are ignored. All hyperparameters follow the paper and its code.

AdaMI. AdaMI (Bateson et al., 2022a) minimises a class-weighted entropy plus the KL divergence between a class-ratio prior and each image's predicted class ratio, with weight 1. The priors come from source labels only: a thyroid fraction of 0.091 from the TG3K training masks and carotid and trachea fractions of 0.054 and 0.41 from the landmark model's source boxes. Because image-level tags would be target labels, we use the paper's unsupervised variant, in which the source prediction decides whether a structure is present in an image and is re-estimated once at two thirds of the adaptation. All parameters of each source network are updated (Adam, learning rate 10^{-6}, the paper's schedule) for as many images as our self-training students see (six passes over the adaptation frames), and the composed networks then predict. AdaMI ends slightly above the source models in balanced Dice (0.47 and 0.48 against 0.44 and 0.45) and below predict nothing, and it nearly removes the carotid (Dice 0.06, misses on 89% and 93% of annotated frames). With rules at test time its balanced Dice is 0.53 and 0.56.

Box-prompted foundation models. Every 8-connected component covering at least 0.2% of the image in the source models' mask gives one tight box. The model returns one mask per box, masks of the same class are united, and a pixel claimed by boxes of different classes goes to the box with the higher predicted IoU. MedSAM uses the released checkpoint with fine-tuned encoder and decoder (Ma et al., 2024) and its official preprocessing (direct resize to 1024×1024, min-max scaling, probability above 0.5), and SAM uses ViT-B with its standard preprocessing (Kirillov et al., 2023). Nothing is tuned. A box-prompted model cannot decide whether a structure is present, so it predicts a structure only where the box source has a component covering at least 0.2% of the image, and only the shape inside each box changes. With boxes from persist & fit instead of the source models, MedSAM matches persist & fit in balanced Dice (0.59 and 0.60 against 0.59 and 0.61, not significant) and does not raise its Dice on annotated frames.

Ground-truth boxes (ceiling, not label-free). Prompted with one tight box per annotated component, MedSAM reaches a Dice on annotated frames of 0.81, 0.88 and 0.80 for thyroid, carotid and trachea on the evaluation frames and 0.79, 0.89 and 0.67 on the held-out frames, against 0.67, 0.72 and 0.70 and 0.62, 0.61 and 0.48 for the supervised reference. The lower held-out trachea value comes from the overlap rule: the thyroid box spans the isthmus and its mask takes over the trachea on four frames, and scoring each structure's own masks gives a trachea Dice of 0.90 on both frame sets. Box-prompted pipelines are therefore limited by the detections they receive.

Text-prompted pipeline. Grounding DINO (base) proposes boxes for the prompts "thyroid gland.", "carotid artery." and "trachea." (text threshold 0.25, non-maximum suppression at IoU 0.5 per prompt), and SAM ViT-B segments each box, with overlaps resolved as above. On these frames the highest-scoring box for each prompt usually covers most of the image (80–99% of its area), so with the default box threshold of 0.35 the pipeline predicts structures on most frames without them (held-out balanced Dice 0.18). Tuned on the evaluation frames, the box threshold rises to 0.60, the top of the grid, where almost nothing is predicted (held-out balanced Dice 0.47, below predict nothing).

UltraSAM3. UltraSAM3 (Xu et al., 2026), a SAM 3 model fine-tuned on 37 public ultrasound datasets (public code and checkpoint, which holds the complete fine-tuned SAM 3 image model), segments each structure from a text prompt, and a structure is predicted if an instance scores at least a threshold. With the prompts "thyroid", "carotid artery" and "trachea" fixed a priori, the

default threshold of 0.5 gives balanced Dice of 0.55 on both frame sets and a threshold tuned on the evaluation frames (0.70) gives 0.58 and 0.59. Choosing the prompt per structure and the threshold jointly on the evaluation frames ("common_carotid_artery", 0.75) gives 0.65 and 0.64, the variant in Table 3, whose evaluation-frame score is optimistic. No variant segments the trachea, a concept absent from UltraSAM3's training data. Table 13 combines UltraSAM3 with rules at test time, label-free thresholds and self-training. Rules at test time leave the tuned variant almost unchanged (0.65 and 0.64), because it already leaves most frames without a structure empty (false thyroid detections 11%). With the a-priori prompts and thresholds chosen without labels (the Otsu protocol of App. G applied to its instance scores over the adaptation frames), UltraSAM3 reaches 0.54 and 0.56, so its lead over the methods built on our source models in Table 3 rests on the prompt and threshold tuned on labelled frames. As a label-free teacher (rows ^{u}), its soft masks without any threshold (per pixel, the maximum over instances of instance score times mask probability) replace the gland model and the carotid channel, composed with the trachea channel of the landmark model as in App. I. Without any target label, self-training and persist & fit on this teacher with rules at test time reach 0.63 on the evaluation and 0.67 on the held-out frames, level with the tuned UltraSAM3, with a trachea Dice of 0.18–0.21 against 0.00, but find fewer carotids (Dice 0.34–0.35 against 0.66).

Table 13: UltraSAM3 with rules at test time, with label-free thresholds and as a label-free teacher (rows ^{u}). Dice on annotated frames, false thyroid detections, all-frame and balanced Dice on the 174 evaluation frames, and all-frame and balanced Dice on the 58 held-out frames, with Bal. cert. for both sets. †The tuned UltraSAM3 is significantly better (participant bootstrap, Holm over 12 comparisons).

	Dice (annotated)			FP	174 evaluation frames			58 held-out frames		
Method	Thy.	Car.	Tra.	thyroid	All-frame	Balanced	Bal. cert.	All-frame	Balanced	Bal. cert.
UltraSAM3, prompts and threshold tuned on labels	0.42	0.66	0.00	11%	0.74	0.65	0.68	0.76	0.64	0.68
+ rules at test time	0.41	0.65	0.00	11%	0.74	0.65†	0.67	0.77	0.64	0.68
UltraSAM3, a-priori prompts, label-free thresholds	0.31	0.38	0.00	34%	0.63	0.54†	0.56	0.70	0.56†	0.58
+ rules at test time	0.31	0.38	0.00	32%	0.64	0.55†	0.57	0.70	0.56†	0.58
UltraSAM3 teacher^{u}	0.55	0.37	0.17	52%	0.59	0.54†	0.56	0.64	0.55†	0.56
+ rules at test time	0.54	0.37	0.17	39%	0.66	0.59†	0.64	0.72	0.60	0.62
Self-training (conf.)^{u}	0.55	0.35	0.21	33%	0.65	0.59†	0.62	0.72	0.63	0.65
+ rules at test time	0.54	0.34	0.21	18%	0.71	0.63	0.67	0.78	0.67	0.69
Persist & fit^{u}	0.49	0.38	0.18	19%	0.70	0.62	0.67	0.76	0.65	0.68
+ rules at test time	0.46	0.35	0.18	13%	0.71	0.63	0.67	0.78	0.67	0.68

EATA and SAR (Niu et al., 2022; 2023), UPL-SFDA (Wu et al., 2023), shape-moment TTA (Bateson et al., 2022b), prototype-bank TTA (Wang et al., 2026), video SFDA (Lo et al., 2023), SAM 3 without ultrasound fine-tuning and MedSAM2 video propagation (Ma et al., 2025) were not run.

I STRONGER SOURCE MODELS FROM PUBLIC DATA

To test whether the conclusions of §5 depend on our source models, we trained two further source models on public data only, without viewing any target frame or prediction. The gland model sees the 3,226 TG3K training images (Gong et al., 2021) and transverse slices of the SegThy ultrasound volumes (Krönke et al., 2022), including slices without thyroid, so it can learn to predict nothing. A separate carotid model sees SegThy slices with and without the carotid and 900 of the 1,100 transverse images of a public carotid dataset with expert lumen masks (Momot, 2022). Both are EfficientNet-B2 U-Nets initialised from the TG3K gland model and trained with strong geometric augmentation (zoom 0.6–1.6, aspect ratio 0.6–1.8, rotation) and photometric augmentation, and the checkpoints were chosen on held-out volunteers and subjects of the public data (three SegThy volunteers, two carotid-dataset subjects). On those, the new gland model predicts thyroid on 10% of SegThy slices without thyroid and never on more than 1% of the image, against 78% and 41% for the TG3K model, at a Dice of 0.85 on slices with thyroid (TG3K model 0.77). The carotid model reaches a Dice of 0.90 on SegThy and 0.96 on the carotid dataset, against 0.57 and 0.66 for the carotid channel of our landmark model. We found no public ultrasound dataset with pixel labels of the trachea, so the new teacher keeps the trachea probability of the landmark model and composes the four classes as $p_{\text{tra}} = h_2$, $p_{\text{car}} = (1 - h_2)\, c$, $p_{\text{thy}} = (1 - h_2)(1 - c)\, g'$ and background as the rest, where c and g' are the new models. Every other part of the pipeline (rules, thresholds, student, schedule) is unchanged, and rows marked ^{s} in Tables 3, 4, 5 and 19 use this teacher. On the target

frames, the public-data source models reach a carotid Dice of 0.78 and 0.72 on annotated frames with 6% and 13% misses, against 0.15 and 0.11 with 71% and 80% misses for our source models, while thyroid Dice changes little (0.63 and 0.55 against 0.62 and 0.59) and false thyroid detections fall only from 99% and 90% to 80% and 71%. The conclusions about abstention hold with this teacher: rules at test time raise the balanced Dice of the source models from 0.54 to 0.67 (held-out 0.53 to 0.63), self-training with rules reaches 0.70 and 0.67–0.68, persist & fit 0.69 and 0.65, and applying the rules in training or at test time again makes no significant difference (bootstrap $p \geq 0.21$). Trachea Dice stays at 0.17–0.23, because the trachea channel is unchanged. Table 14 repeats the TTA, SFDA and self-training baselines with this teacher, in a Holm family of its own with persist & fit^s as the reference. BN adapt, TENT and CoTTA again lower balanced Dice below that of the source models (0.42–0.52 against 0.54 on the evaluation frames), BN adapt and CoTTA to below predict nothing, whereas AdaMI, DPL and both self-training baselines raise it to 0.60–0.66, and rules at test time bring all of them to 0.59–0.70. With this teacher AdaMI no longer loses the carotid (Dice 0.77).

Table 14: TTA, SFDA and self-training baselines with the public-data source models (rows ^{s}): carotid Dice on annotated frames, all-frame and balanced Dice on the 174 evaluation frames, balanced Dice with uncertain structures excluded (Bal. cert.), balanced Dice on the 58 held-out frames, and balanced Dice with rules at test time. †Persist & fit^s is significantly better (Holm over the 18 comparisons of this family, $p < 0.05$).

	174 evaluation frames				58 frames	+ rules, balanced	
Method	Carotid	All-frame	Balanced	Bal. cert.	Balanced	174	58
Predict nothing	0.00	0.63†	0.50†	0.50†	0.50†	–	–
Source models^s	0.78	0.57†	0.54†	0.56†	0.53†	0.67	0.63
BN adapt^s	0.71	0.40†	0.42†	0.46†	0.43†	0.59†	0.56
TENT^s	0.77	0.53†	0.52†	0.58†	0.53†	0.62†	0.59
CoTTA^s	0.76	0.46†	0.46†	0.51†	0.46†	0.60†	0.59
AdaMI^s	0.77	0.62†	0.60†	0.65†	0.59	0.68	0.64
Self-training, confidence only^s	0.79	0.68†	0.65†	0.70†	0.63	0.70	0.67
Self-training, augmentation agreement^s	0.79	0.70†	0.66†	0.71†	0.62	0.70	0.68
DPL^s	0.76	0.70†	0.66†	0.71†	0.62	0.70	0.66
Persist & fit^s	0.76	0.75	0.69	0.74	0.65	0.70	0.68

J A SECOND DEVICE AND TARGET DOMAIN

To test whether the hand-set rules transfer beyond our probe, we ran the transverse rules of Table 1 unchanged on the ground-truth labels of 2,591 transverse SegThy slices (every fifth slice of the 32 volumes, excluding slices less than a quarter scanned, cart-based scanner), with each slice and its labels resized to 256×256 as in our pipeline. A ground-truth component counts as rejected if the rules remove more than half of its pixels. The rules reject 6.6% of the 2,289 thyroid components and 3.3% of the 2,482 carotid components, which make up 0.4% and 0.6% of the ground-truth pixels. Most rejected components are small pole or isthmus fragments below the area floors and partial or low-contrast carotid lumens at the edge of the scanned region. On the 116 evaluation frames that the plane classifier calls transverse, the same rules reject only two fragments smaller than 20 pixels (one thyroid and one carotid component, no trachea), and shape measures in true proportions change no decision on either dataset. The main sensitivity is anchoring: on the 68 SegThy thyroid components of slices without any accepted carotid, the fallback area rule rejects about half. SegThy has no trachea annotation, so the trachea test and the rule that the thyroid lies between trachea and carotid could not be checked on the second device.

A second target domain. To test whether abstention also decides label-free performance elsewhere, we ran the methods unchanged from our source models (the TG3K gland model and the retrained landmark model, neither trained on SegThy) on the transverse slices of the 32 SegThy volumes of 16 participants, treating each volume as a sweep. The adaptation pool holds the 12,969 slices at least a quarter scanned, and the 2,591 slices above are scored for thyroid and carotid (SegThy has no trachea labels), averaged within and then over participants and bootstrapped over

participants as in §3. Slices are cropped to the scanned region and resized to 256×256 like our frames, all are treated as transverse, and consecutive slices are spaced by the median frame gap of our sweeps (0.244 s), so that they are neighbours for persistence. No SegThy label influenced any choice. Only 14% of the scored slices lack a thyroid and 5% a carotid, so predict nothing scores an all-frame Dice of 0.09. The source models predict a thyroid on 81% of the slices without one, and their thyroid balanced Dice (0.47) stays at the level of predict nothing, as do those of temporal averaging, BN adapt and CoTTA (0.46–0.48), while the mean over the two structures of all four (0.56, Table 15) exceeds it because they predict a carotid on only 8–10% of the slices without one. Every method with rules, in training or at test time, or with label-free thresholds reaches a balanced Dice of 0.65–0.74, above every method without them (0.56–0.63). Rules at test time significantly raise each of the seven methods run with them, by 0.02–0.10 and mostly by removing false thyroids, and persist & fit has a higher all-frame Dice than every other method in the upper block of the table (Holm-adjusted p=0.003). Unlike on our sweeps, the label-free thresholds of App. G beat the rules for the source models and both self-training baselines, by 0.03–0.07 (Holm-adjusted $p<0.01$), and do not differ significantly from them for persist & fit (p=0.11). Ranked by Dice on annotated slices alone, the 23 label-free methods run on SegThy (the 21 of Table 15 and the presence-only and area-only label-free thresholds of the source models) order almost independently of their balanced Dice (Kendall's $\tau = -0.16$). The maximum probability of the source models separates slices with and without the thyroid and the carotid with AUROCs of 0.96 and 0.81 on SegThy, against 0.83–0.85 and 0.59–0.61 on our annotated frames, consistent with recalibrated thresholds sufficing there. Slices of a reconstructed volume are far denser than our 4 fps frames, so the persistence test removes little on SegThy.

Table 15: Second target domain: the methods run unchanged from our source models on SegThy volumes (2,591 scored slices, 16 participants, thyroid and carotid). Dice on annotated slices, false thyroid detections, carotid misses, all-frame Dice, thyroid balanced Dice and balanced Dice (mean over the two structures) with 95% participant-bootstrap CIs. Label-free thresholds use Otsu's method on the unlabelled slices (App. G). †/‡Persist & fit is significantly better/worse (Holm over the 13 comparisons of the upper block and the 11 of the lower block, which also include predict nothing and the presence-only and area-only label-free thresholds of the source models).

	Dice (annotated)		FP	Miss		Balanced	
Method	Thyroid	Carotid	thyroid	carotid	All-frame	Thyroid	Mean
Predict nothing	0.00	0.00	0%	100%	0.09†	0.50	0.50 [0.50,0.50]†
Source models	0.73	0.34	81%	39%	0.51†	0.47	0.56 [0.53,0.59]†
+ rules at test time	0.72	0.32	40%	51%	0.53†	0.67	0.65 [0.63,0.68]†
+ label-free thresholds	0.69	0.28	4%	59%	0.52†	0.82	0.72 [0.69,0.75]
Temporal averaging	0.74	0.34	79%	39%	0.52†	0.48	0.56 [0.53,0.59]†
+ rules at test time	0.72	0.32	38%	52%	0.53†	0.67	0.66 [0.63,0.68]†
BN adapt	0.71	0.36	83%	35%	0.51†	0.46	0.56 [0.52,0.59]†
TENT	0.74	0.27	52%	48%	0.51†	0.63	0.62 [0.59,0.65]†
CoTTA	0.72	0.36	82%	36%	0.51†	0.46	0.56 [0.53,0.58]†
Self-training, confidence only	0.74	0.33	54%	51%	0.53†	0.59	0.62 [0.58,0.66]†
+ rules at test time	0.72	0.32	33%	55%	0.53†	0.69	0.67 [0.64,0.70]†
Self-training, augmentation agreement	0.74	0.33	52%	52%	0.53†	0.60	0.62 [0.59,0.66]†
+ rules at test time	0.72	0.32	30%	54%	0.54†	0.71	0.68 [0.65,0.71]
Persist & fit	0.73	0.45	38%	38%	0.60	0.68	0.70 [0.67,0.73]
AdaMI	0.70	0.25	69%	53%	0.47†	0.51	0.56 [0.54,0.59]†
+ rules at test time	0.68	0.23	28%	64%	0.48†	0.71	0.66 [0.63,0.68]†
DPL	0.77	0.30	50%	60%	0.53†	0.63	0.63 [0.60,0.67]†
+ rules at test time	0.75	0.30	28%	62%	0.54†	0.73	0.68 [0.66,0.71]
Self-training (conf.) + label-free thresholds	0.71	0.32	19%	55%	0.54†	0.75	0.70 [0.66,0.73]
Self-training (aug.) + label-free thresholds	0.71	0.32	16%	55%	0.54†	0.76	0.70 [0.67,0.73]
Persist & fit + rules at test time	0.72	0.45	28%	41%	0.60	0.72	0.72 [0.69,0.75]‡
Persist & fit + label-free thresholds	0.65	0.44	9%	44%	0.57†	0.76	0.74 [0.70,0.77]‡

Table 16: Transductive adaptation (all participants, as in Table 3) against the participant-disjoint protocol, in which each annotated frame is predicted by a student adapted without its participant (three folds). †Transductive persist & fit is significantly better (Holm-adjusted $p < 0.05$).

		Transductive				Participant-disjoint			
		174 frames		58 frames		174 frames		58 frames	
Method	Rules	All-frame	Balanced	All-frame	Balanced	All-frame	Balanced	All-frame	Balanced
Self-training, confidence only	as is	0.59†	0.53†	0.60†	0.49†	0.58†	0.52†	0.61†	0.50†
	+ rules	0.65	0.58	0.73	0.59	0.66	0.58	0.71	0.58
Self-training, augmentation agreement	as is	0.60†	0.54†	0.63†	0.52†	0.59†	0.52†	0.62†	0.51†
	+ rules	0.67	0.58	0.73	0.60	0.66	0.58	0.73	0.60
Persist & fit	as is	0.68	0.59	0.74	0.61	0.67	0.59	0.73	0.62

Table 17: Released data: every method of Table 3 re-run from scratch (adaptation, threshold tuning where a method has one, and prediction) on frames with the overlay text masked exactly as in the release. The balanced Dice on the 58 frames is the official score. †/‡Persist & fit on the released data is significantly better/worse (Holm over the 26 comparisons of this table, $p < 0.05$). Rows ^{s} use the public-data source models, and ^{e}/^{t} mark entries tuned or trained on the labelled evaluation frames (Table 3). The last two columns repeat the balanced Dice on the original frames (Tables 4 and 5).

	174 frames, released data			58 frames, released data		Original data, balanced	
Method	All-frame	Balanced	Miss carotid	All-frame	Balanced	174	58
Predict nothing	0.63	0.50†	100%	0.67	0.50†	0.50	0.50
Source models	0.52†	0.46†	68%	0.55†	0.45†	0.44	0.45
BN adapt	0.37†	0.35†	56%	0.36†	0.32†	0.35	0.32
TENT	0.48†	0.43†	77%	0.49†	0.40†	0.44	0.39
CoTTA	0.41†	0.38†	58%	0.40†	0.35†	0.37	0.34
Temporal averaging	0.54†	0.48†	71%	0.56†	0.46†	0.46	0.45
Self-training, confidence only	0.60†	0.54†	66%	0.61†	0.51†	0.53	0.49
Self-training, augmentation agreement	0.60†	0.53†	69%	0.62†	0.51†	0.54	0.52
Persist & fit	0.67	0.59	71%	0.71	0.58	0.59	0.61
Source + rules	0.64	0.55†	87%	0.71	0.58	0.55	0.57
Temporal averaging + rules	0.64	0.56†	81%	0.72	0.58	0.56	0.58
Self-training (conf.) + rules	0.67	0.59	71%	0.73	0.60	0.58	0.59
Self-training (aug.) + rules	0.67	0.59	76%	0.72	0.59	0.58	0.60
Persist & fit + rules	0.68	0.59	79%	0.72	0.57	0.60	0.61
AdaMI	0.54†	0.47†	87%	0.58†	0.47†	0.47	0.48
DPL	0.59†	0.52†	68%	0.63†	0.52†	0.54	0.52
Grounding DINO + SAM, text^{e}	0.61†	0.49†	100%	0.63	0.46†	0.49	0.47
UltraSAM3, text, tuned^{e}	0.72‡	0.63‡	31%	0.77	0.65	0.65	0.64
MedSAM, source boxes	0.54†	0.47†	81%	0.59†	0.49†	0.47	0.49
Source models^{s}	0.59†	0.57	5%	0.56†	0.55	0.54	0.53
Source models^{s} + rules	0.71	0.67‡	10%	0.70	0.63	0.67	0.63
Self-training (conf.)^{s}	0.66	0.63‡	6%	0.67	0.62	0.65	0.63
Self-training (conf.)^{s} + rules	0.74‡	0.69‡	10%	0.74	0.67‡	0.70	0.67
Self-training (aug.)^{s}	0.70	0.67‡	6%	0.69	0.63	0.66	0.62
Self-training (aug.)^{s} + rules	0.76‡	0.70‡	8%	0.75	0.67‡	0.70	0.68
Persist & fit^{s}	0.74‡	0.69‡	6%	0.75	0.68‡	0.69	0.65
Supervised reference^{t}	0.80‡	0.78‡	6%	0.81	0.76‡	0.78	0.74

K PARTICIPANT-DISJOINT ADAPTATION AND RELEASED DATA

For the release-data re-run, the 20,014 adaptation frames were masked with the boxes of the release (App. B), so the annotated crops equal the released ones up to JPEG coding and the other frames match the released videos up to video coding, and every method in Table 17 was adapted, tuned and evaluated again from scratch. Balanced Dice changes by at most 0.04 against the original frames. The largest drops are those of persist & fit on the held-out frames (0.61 to 0.58, 0.57 with rules), which there falls level with self-training with rules (0.60 and 0.59, not significant), while every method without abstention built on our source models stays significantly below it on both frame sets. The 95% confidence intervals of the differences from self-training with rules are -0.05 to $+0.01$ and -0.04 to $+0.01$ on the held-out frames. UltraSAM3, with its threshold re-tuned on the released evaluation frames (0.70 instead of 0.75), reaches 0.65 there (bootstrap p=0.006 against

persist & fit, Holm-adjusted p=0.06), and the public-data rows with rules or persist & fit reach 0.63–0.68. Table 17 gives the reference scores for the released data.

L UNCERTAIN STRUCTURES

After the second internal review, the annotating clinician went through the 232 frames in a tool that showed the existing polygons and no model prediction, in which every unannotated structure starts as absent, and changed it to uncertain where the structure was possibly visible but not drawn. Of 444 unannotated structures, 188 were marked uncertain: 158 of 327 on the evaluation frames (thyroid 60 of 95, carotid 40 of 112, trachea 58 of 120) and 30 of 117 on the held-out frames. Among the frames flagged as showing no identifiable structure, 46 of 80 evaluation frames and 10 of 30 held-out frames have at least one uncertain structure. Table 18 removes the uncertain frame-structure pairs from every negative-side score (false detections, the abstention half of balanced Dice and all-frame Dice), so that a correct prediction of an uncertain structure is no longer an error, and leaves Dice on annotated frames unchanged. With uncertain pairs removed, balanced Dice rises for persist & fit (0.63 and 0.64) and for the supervised reference (0.83 and 0.77), stays at 0.50 for predict nothing and falls slightly for the source models (0.41 and 0.43). The ranking of Table 10 holds. On the held-out frames, self-training and DPL with rules at test time reach 0.62 and do not differ significantly from persist & fit, whereas the same three methods with tuned generic abstention reach 0.56–0.59, so for these methods the rules transfer from the evaluation to the held-out frames better than thresholds tuned on the former.

Table 18: Scores with uncertain frame-structure pairs removed from the negative side. †Persist & fit is significantly better on the scores with uncertain pairs removed (participant bootstrap, Holm over the 40 comparisons of this analysis, $p < 0.05$). The all-pairs columns are not marked. FP: false thyroid detections on the 174 frames without the uncertain cases.

Method	174 frames, all pairs		174 frames, uncertain removed		58 frames, balanced		
	All-frame	Balanced	All-frame	Balanced	All pairs	Uncertain removed	FP thyroid
Predict nothing	0.63	0.50	0.45†	0.50†	0.50	0.50†	0%
Source models	0.50	0.44	0.47†	0.41†	0.45	0.43†	100%
BN adapt (Schneider et al., 2020)	0.37	0.35	0.41†	0.36†	0.32	0.31†	91%
TENT (Wang et al., 2021)	0.49	0.44	0.49†	0.44†	0.39	0.38†	80%
CoTTA (Wang et al., 2022)	0.39	0.37	0.42†	0.37†	0.34	0.33†	91%
AdaMI (Bateson et al., 2022a)	0.53	0.47	0.52†	0.48†	0.48	0.49†	74%
Grounding DINO + SAM, text (Liu et al., 2024)	0.61	0.49	0.43†	0.47†	0.47	0.47†	11%
UltraSAM3, text, tuned (Xu et al., 2026)	0.74	0.65	0.66	0.68	0.64	0.68	3%
MedSAM, source boxes (Ma et al., 2024)	0.53	0.47	0.49†	0.46†	0.49	0.48†	80%
Temporal averaging at test time	0.53	0.46	0.51†	0.45†	0.45	0.44†	94%
Self-training, confidence only	0.59	0.53	0.56†	0.54†	0.49	0.49†	63%
Self-training, augmentation agreement	0.60	0.54	0.57†	0.55†	0.52	0.52†	54%
DPL (Chen et al., 2021)	0.60	0.54	0.58	0.56	0.52	0.50†	54%
Persist & fit	0.68	0.59	0.64	0.63	0.61	0.64	17%
Self-training (conf.) + rules	0.65	0.58	0.62	0.60	0.59	0.62	37%
Self-training (aug.) + rules	0.67	0.58	0.63	0.61	0.60	0.62	29%
Persist & fit + rules	0.70	0.60	0.64	0.63	0.61	0.65	14%
Self-training (conf.) + generic	0.68	0.59	0.62	0.62	0.59	0.59	11%
Self-training (aug.) + generic	0.68	0.60	0.63	0.62	0.58	0.57	11%
Persist & fit + generic	0.70	0.61	0.65	0.65	0.63	0.65	11%

M SECOND ANNOTATOR

A second clinician, also a head and neck specialist with at least ten years of thyroid ultrasound experience, annotated the 58 held-out frames with the same written protocol and tool and without model predictions. On first opening the tool, the second annotator briefly saw the first annotator's outlines, which the browser had restored from its storage. All frames were then annotated from scratch in a clean copy of the tool, and none of the 1,734 polygon vertices coincides with a vertex of the first annotator. The two annotators agree on whether each structure is present with Cohen's κ of 0.97, 0.91 and 0.91 for thyroid, carotid and trachea, disagreeing on 1, 2 and 2 of the 58 frames, each time a structure that only the second annotator drew. On frames where both drew a structure, their Dice is 0.90, 0.87 and 0.90. This agreement lies far above the supervised reference on the same frames (0.62, 0.61 and 0.48), so annotation noise does not explain its lower Dice. Scored against

the second annotator, the balanced Dice of every method changes by at most 0.02, and the same comparisons with persist & fit are significant (Table 19).

Table 19: Held-out frames scored against the first and the second annotator. †/‡Persist & fit is significantly better/worse against the second annotator (Holm over 24 comparisons, $p < 0.05$).

Method	First annotator		Second annotator		Second, Dice on annotated		
	All-frame	Balanced	All-frame	Balanced	Thy.	Car.	Tra.
Predict nothing	0.67	0.50	0.64	0.50†	0.00	0.00	0.00
Source models	0.54	0.45	0.52†	0.44†	0.57	0.10	0.21
TENT	0.47	0.39	0.47†	0.39†	0.61	0.07	0.18
AdaMI	0.59	0.48	0.58†	0.48†	0.64	0.06	0.19
Grounding DINO + SAM	0.64	0.47	0.61	0.47†	0.00	0.00	0.02
UltraSAM3	0.76	0.64	0.76	0.66	0.58	0.56	0.00
MedSAM, source boxes	0.60	0.49	0.58†	0.49†	0.57	0.11	0.21
Temporal averaging	0.56	0.45	0.54†	0.44†	0.58	0.07	0.19
Self-training, confidence only	0.60	0.49	0.58†	0.49†	0.62	0.15	0.17
Self-training, augmentation agreement	0.63	0.52	0.62†	0.52†	0.61	0.15	0.18
DPL	0.63	0.52	0.60†	0.51†	0.61	0.11	0.18
Persist & fit	0.74	0.61	0.71	0.60	0.58	0.21	0.19
Self-training (conf.) + rules	0.73	0.59	0.70	0.58	0.60	0.15	0.16
Self-training (aug.) + rules	0.73	0.60	0.70	0.60	0.59	0.15	0.18
Source models^{s} + rules	0.69	0.63	0.68	0.62	0.52	0.67	0.18
Supervised reference	0.81	0.74	0.79	0.73‡	0.59	0.55	0.49

N DEPTH, VARIANTS AND ROBUSTNESS CHECKS

Table 20 splits the evaluation frames by imaging depth, with only 26 frames from 6 younger participants in its middle column, and Table 22 lists seeds and variants of persist & fit. Separate paragraphs then cover age groups and depth, seeds and rounds, the confidence threshold, motion compensation, the carotid at 40 mm, the overlay text and qualitative examples.

Table 20: Results by imaging depth (assigned per frame) on the 174 frames: Dice on annotated frames per structure and balanced Dice. ^{b}Post hoc, shape tests in true proportions.

Method	Younger, 40 mm (70 frames)				Younger, 60 mm (26)				Older, 60 mm (78)			
	Thy.	Car.	Tra.	Bal.	Thy.	Car.	Tra.	Bal.	Thy.	Car.	Tra.	Bal.
Predict nothing	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50
Source models	0.62	0.10	0.41	0.47	0.71	0.02	0.15	0.45	0.59	0.22	0.04	0.43
Self-training, confidence only	0.64	0.12	0.42	0.57	0.73	0.00	0.18	0.47	0.65	0.42	0.07	0.53
Persist & fit	0.63	0.19	0.34	0.61	0.74	0.00	0.17	0.51	0.62	0.33	0.09	0.62
Persist & fit, true proportions^{b}	0.64	0.19	0.27	0.61	0.74	0.00	0.14	0.51	0.63	0.27	0.09	0.60
Supervised reference	0.64	0.65	0.56	0.75	0.68	0.53	0.74	0.73	0.69	0.82	0.78	0.80

Age groups and depth. Adaptation raises balanced Dice in younger and older adults (persist & fit 0.57 and 0.62, source models 0.45 and 0.43, Table 21). In older adults, trachea Dice is at most 0.11 for every label-free method in Table 21 (younger adults: source models 0.31, persist & fit 0.27), and adapting on the younger laboratory session alone never finds the carotid in either age group. The resize to 256×256 stretches 57% of the annotated carotids at 40 mm beyond the lumen detector's elongation cutoff, but measuring shape in true proportions leaves carotid Dice there at 0.19.

Seeds and rounds. The default configuration (seed 0) was fixed before annotation, and the runs with seeds 1 and 2, the class-weighted loss, three rounds, no hyperechoic-top test and reject-to-background (Table 9) were launched before annotation as variants. We ran the variants with the original landmark model after annotation and the γ and no-flow variants after an internal review.

Table 21: Results by age group on the 174 evaluation frames, with adaptation on both sessions or on one session only. Younger: 16 participants aged 18–44. Older: 13 participants aged 69–81. Dice on annotated frames per structure and balanced Dice.

		Younger (18–44)				Older (69–81)			
Method	Adapted on	Thy.	Car.	Tra.	Bal.	Thy.	Car.	Tra.	Bal.
Predict nothing	–	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50
Source models	–	0.65	0.08	0.31	0.45	0.59	0.22	0.04	0.43
Self-training, confidence only	both	0.67	0.09	0.33	0.54	0.65	0.42	0.07	0.53
Persist & fit	both	0.66	0.13	0.27	0.57	0.62	0.33	0.09	0.62
Persist & fit	lab only	0.65	0.00	0.20	0.54	0.64	0.00	0.11	0.55
Persist & fit	event only	0.63	0.09	0.11	0.51	0.52	0.33	0.10	0.61
Supervised reference	–	0.65	0.62	0.62	0.75	0.69	0.82	0.78	0.80

Table 22: Seeds and variants of persist & fit. Only the default configuration was fixed before annotation. [b]Post hoc, circularity and elongation measured in the frame's true proportions (App. E). [c]Strictly source-free: the original landmark model, trained before any target frame was viewed.

	Dice on annotated frames (174)			174 frames		58 held-out	
Variant	Thyroid	Carotid	Trachea	All-frame	Balanced	All-frame	Balanced
Default (seed 0)	0.64	0.23	0.18	0.68	0.59	0.74	0.61
Seed 1	0.62	0.20	0.21	0.68	0.59	0.73	0.60
Seed 2	0.63	0.22	0.18	0.67	0.58	0.71	0.58
One round	0.64	0.19	0.16	0.64	0.55	0.69	0.56
Three rounds	0.63	0.21	0.20	0.69	0.61	0.73	0.60
$\gamma = 0.5$	0.66	0.25	0.17	0.67	0.59	0.72	0.60
$\gamma = 0.9$	0.64	0.23	0.17	0.68	0.59	0.74	0.62
Class-weighted loss	0.68	0.25	0.17	0.66	0.57	0.71	0.58
No hyperechoic-top test	0.65	0.23	0.18	0.67	0.58	0.70	0.57
Shape tests in true proportions[b]	0.65	0.20	0.16	0.68	0.58	0.73	0.61
Original landmark model as teacher[c]	0.62	0.18	0.11	0.65	0.56	0.71	0.57
Source with original landmark model[c]	0.62	0.14	0.17	0.41	0.38	0.49	0.38

Over the three seeds, all-frame Dice is 0.676 ± 0.006 (sd) on the evaluation frames and 0.724 ± 0.014 on the held-out frames, and balanced Dice is 0.587 ± 0.007 and 0.596 ± 0.018. A second round of self-training raises all-frame Dice by 0.045 on the evaluation frames ($p<0.001/<0.001$, Holm-adjusted $p=0.016$) and by 0.043 on the held-out frames ($p=0.008/0.052$, adjusted $p=0.42$). Its gain in balanced Dice, 0.041 and 0.048, survives correction on both sets (adjusted $p=0.016$ and 0.023). Adding a third round raises all-frame Dice by 0.014 on the evaluation frames ($p=0.011/0.11$, adjusted $p=0.43$) and changes it by 0.001 on the held-out frames ($p=0.87/0.74$, adjusted $p=1.00$).

Confidence threshold and motion compensation. Neither $\gamma = 0.5$ nor $\gamma = 0.9$ changes all-frame or balanced Dice significantly on either frame set (Table 22, all $p>0.05$). On the held-out frames, $\gamma = 0.9$ gives a higher Dice on annotated frames than the default $\gamma = 0.7$ (0.36 against 0.34, $p<0.001/0.03$, adjusted $p=0.016$), mainly for the carotid (0.26 against 0.22). Persistence without optical-flow alignment (the row marked § in Table 9) is indistinguishable from the default, with all-frame Dice of 0.68 in both on the evaluation frames ($p=0.92/1.00$) and 0.73 against 0.74 on the held-out frames ($p=0.87/0.42$). At 4 fps and with neighbours at most 0.6 s apart, agreement with unwarped neighbours serves as well as agreement with flow-aligned ones.

Carotid at 40 mm. The rules and the lumen detector measure shape on the 256×256 grid, on which a round lumen appears about 1.9 times taller than wide at 40 mm and 1.3 times at 60 mm, because the image region is resized without preserving its aspect ratio (App. E). Candidate lumens, which withholding and completion need, must have an elongation of at most 1.8. After evaluation we measured the elongation of the annotated carotid components on this grid. At 40 mm the median is 1.81 and 57% of 21 components exceed the cutoff, against a median of 1.24 and 2% of 47 components at 60 mm. Most carotids at 40 mm therefore cannot become candidate lumens, and neither

withholding nor completion acts on them. This matches the low carotid Dice of younger participants at 40 mm (Table 20) and the carotid Dice of 0 in both age groups after adaptation on the laboratory session alone, whose frames are mostly at 40 mm (Table 21). It does not explain the 60 mm frames of younger participants, on which carotid Dice is at most 0.02 for every label-free method in Table 20. The post-hoc variant that measures circularity and elongation in true proportions (App. E) removes this distortion from the rules and the lumen detector but leaves carotid Dice at 40 mm at 0.19 (Table 20), so the distorted shape tests do not explain the failure. With the public-data source models, the same pipeline and rules reach a carotid Dice of 0.73 at 40 mm and 0.56 in younger participants at 60 mm (source models 0.78 and 0.61), so the failure lies in the carotid channel of our landmark model. We did not change the default method.

Overlay text. The frames used for adaptation and evaluation in the main experiments contain the scanner app's overlay text, part of which, such as the depth setting, differs between depths and sessions and could serve as a shortcut. As a test-time check, we inpainted the overlay text of the 232 annotated frames (bright glyphs in the four corners of the image region, detected against a local median) and re-ran the source models, confidence-only self-training and persist & fit without retraining. All three score slightly lower on both frame sets and keep their ranking. For persist & fit, all-frame Dice goes from 0.680 to 0.668 on the evaluation frames and from 0.736 to 0.719 on the held-out frames, and balanced Dice from 0.591 to 0.582 and from 0.609 to 0.591. The students were trained with the overlay present, so this check alone cannot exclude an overlay shortcut. Appendix K re-runs the main methods from scratch on frames masked with the fixed boxes of the release, which changes balanced Dice by at most 0.04 and keeps every method without abstention built on our source models significantly below persist & fit.

Qualitative examples. Figure 4 shows ten evaluation frames selected by a fixed rule.

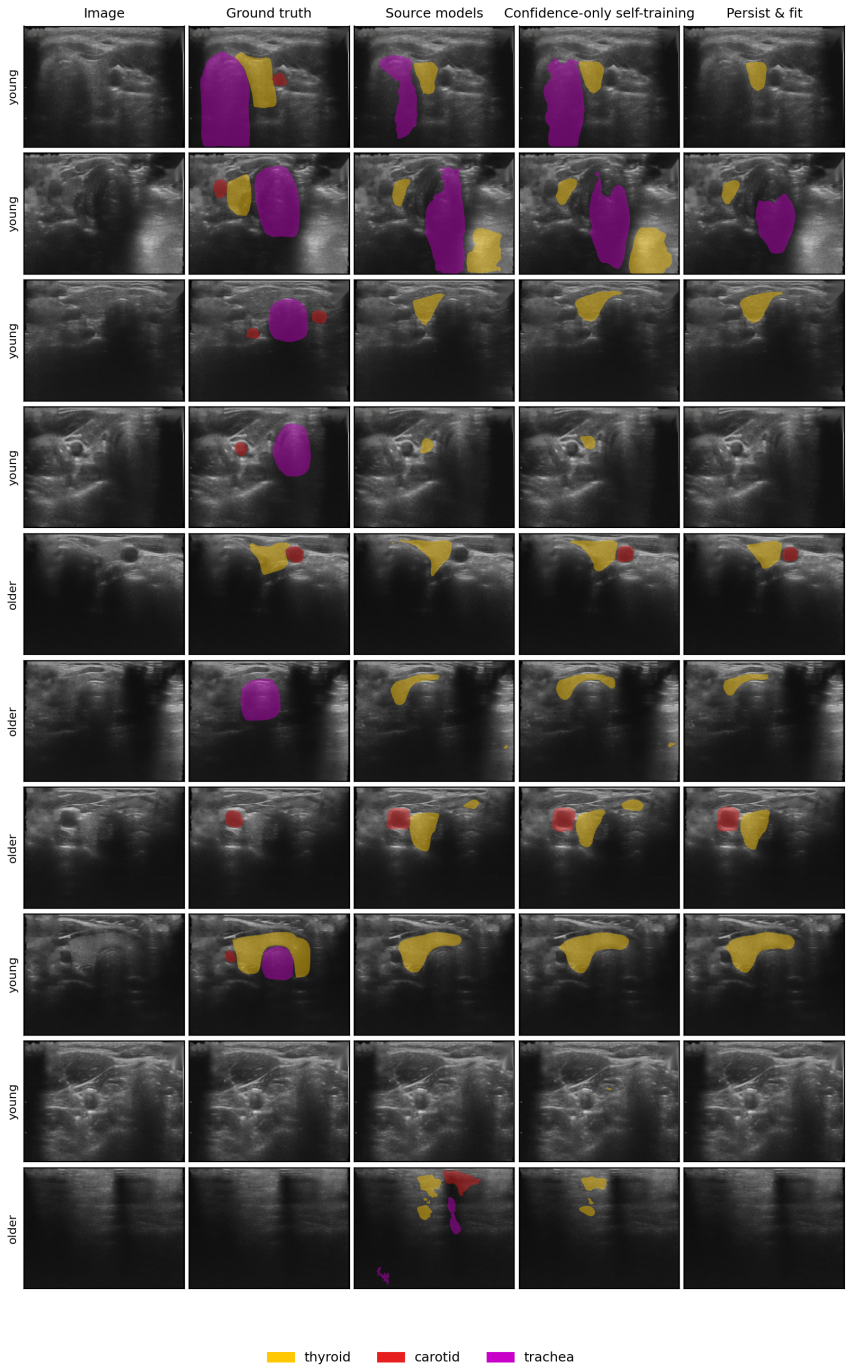


Figure 4: Predictions on ten of the 174 evaluation frames (columns: image, ground truth, source models, confidence-only self-training, persist & fit). Rows 1–8 follow a fixed rule: the earliest frame predicted as transverse that has an annotated structure, from every fourth participant, with participants ordered by session and numeric ID. The last two rows are frames flagged as showing no structure, one per age group (the 6th and 46th of the 80 flagged frames in the order of the frame manifest). Row labels give the age group (young: 18–44, older: 69–81 years). On-screen text is masked.

O LABEL-FREE DIAGNOSTICS

Table 23 reports two quantities computed on the argmax maps (256×256) of all 20,014 adaptation frames. The rows after augmentation agreement add the named tests to confidence-only self-training. Temporal consistency is the IoU of each structure between a frame's prediction and the flow-warped prediction of the next frame, over consecutive pairs at most 0.6 s apart in which either prediction contains the structure, averaged within each participant and then over participants. The violation rate is the fraction of predicted thyroid, carotid and trachea pixels, pooled over all frames, that fail one of three checks (thyroid pixels enter only the denominator): trachea in a frame the plane classifier calls sagittal, transverse trachea outside the largest trachea component, and transverse carotid components with circularity below 0.45. Neither quantity penalises missed structures or predicting nothing, because pairs in which both predictions are empty are skipped and empty frames add no pixels, so both are undefined for the empty predictor and cannot rank methods without the labelled metrics. The violation rate reuses three of our rules and therefore favours methods trained with the anatomical test. Temporal consistency favours methods that fuse adjacent frames, at test time or through the persistence test, and temporal averaging is smooth by construction because adjacent outputs share two of their three inputs. Figure 5 shows predicted areas along one minute of a sweep.

Table 23: Label-free diagnostics on all adaptation frames.

	Temporal consistency			
Method	Thyroid	Carotid	Trachea	Violation rate
Source models	0.512	0.279	0.225	8.4%
BN adapt	0.499	0.219	0.399	19.4%
TENT	0.527	0.189	0.318	11.9%
CoTTA	0.509	0.238	0.336	15.2%
Temporal averaging	0.669	0.514	0.418	5.6%
Self-training, confidence only	0.592	0.439	0.373	6.5%
Self-training, augmentation agreement	0.607	0.364	0.374	4.4%
+ persistence (T)	0.659	0.490	0.518	3.6%
+ anatomy and completion (A, C)	0.605	0.471	0.303	0.8%
T + A	0.668	0.461	0.362	0.6%
Persist & fit (T + A + C)	0.662	0.473	0.432	0.6%

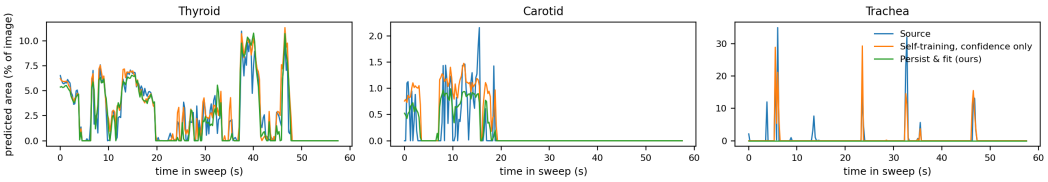


Figure 5: Predicted structure area along 60 s of the sweep of an older participant for the source models, confidence-only self-training and persist & fit. Stretches at zero area are frames in which a method predicts nothing. In this stretch the source models and confidence-only self-training produce isolated single-frame trachea detections covering up to a third of the image, and persist & fit predicts no trachea. Whether a trachea was visible was not annotated.

P CHRONOLOGY, HELD-OUT FRAMES AND FEW-SHOT TRACK

Selection. Because the rules and thresholds were chosen while inspecting unlabelled frames that include the 174 evaluation frames, the held-out frames are the clean test set. The 58 held-out frames, two per participant, were drawn by the same script and content criterion as the evaluation frames (App. E), without plane balancing. For each sweep, the script takes the frames that pass the content criterion and, at the 8th and 92nd percentiles of their times, picks the nearest frame at least 2 s from every evaluation frame, so the held-out frames lie near the start and end of the sweep. The six

evaluation frames per participant were picked in the same way at fixed percentiles of the frames the plane classifier calls transverse (four) and sagittal (two), at least 2 s apart. These times are nominal (frame index divided by the average frame rate), and because the recordings have a variable frame rate the true spacing can be slightly shorter. None of the held-out frames was displayed during method development. The same clinician annotated them with the same protocol and tool after the method was frozen, and like the evaluation frames their images belong to the unlabelled adaptation pool (App. C).

Composition. Of the 58 frames, 30 (52%) show no identifiable structure, against 80 of the 174 evaluation frames (46%). The thyroid is annotated on 27 frames of 20 participants, the carotid on 15 frames of 13 participants and the trachea on 15 frames of 15 participants. Frames near the start and end of a sweep over-weight probe placement and lift-off, and the larger share of empty frames raises the all-frame Dice of the empty predictor to 0.67 (0.63 on the evaluation frames) and of most methods. Higher all-frame Dice on the held-out frames therefore does not indicate better generalisation, whereas balanced Dice gives the empty predictor 0.50 on both sets. With 13 to 20 participants per structure, held-out comparisons resolve only large differences.

Chronology. Fixed before annotation: the method configuration (seed 0), the baselines BN adapt, TENT, CoTTA, temporal averaging and the two self-training baselines, Dice on annotated frames and the false detection rate. All-frame Dice was defined in the evaluation script's documentation but first computed after annotation, and the seeds, class-weighted, three-round, no-top and reject-to-background variants were launched before annotation. Added after annotation: the A-only ablation, the original-landmark-teacher runs and the held-out supervised reference. After a first internal review we added predict nothing, balanced Dice, miss rates, Holm correction, rules at test time, the γ and no-flow ablations, the few-shot track, the overlay check and the aspect-ratio diagnosis. Added after a second internal review: one presence rule for all metrics (any predicted pixel, where miss and false detection rates had used more than 0.2% of the image), generic abstention, AdaMI, DPL and the box-prompted and Grounding DINO baselines, the participant-disjoint protocol, few-shot scoring on participants not used for fine-tuning with matched steps and the confidence-only initialisation, the release-data re-run, the aspect-corrected variant, the uncertainty flags and the hidden test set. A third internal review added the public-data source models with the two self-training baselines, persist & fit and rules at test time, UltraSAM3, MedSAM with ground-truth boxes, the second annotator, generic abstention with label-free thresholds, the presence-threshold sweep, the presence AUROC and Fig. 2. The fourth added BN adapt, TENT, CoTTA, AdaMI and DPL with these source models, the rule-transfer test on SegThy and the manual plane labels. The fifth added the release-data re-run of every row of Table 3 and the official track, UltraSAM3 with rules, with label-free thresholds and as a teacher, confidence intervals of paired differences, the weight analysis of balanced Dice and the second target domain. The sixth added the split of each change in balanced Dice into its parts on annotated frames and on frames without the structure (Table 6), the rank correlations between Dice on annotated frames and balanced Dice, and the confidence and false-detection statistics cited in §3.

Results. Here and in App. N, p-values (bootstrap/Wilcoxon), including all per-structure and ablation ones, are uncorrected unless marked as adjusted. Adjusted p is the bootstrap p after Holm correction over the 82 comparisons with persist & fit on the same frame set, computed only for mean Dice on annotated frames, all-frame Dice and balanced Dice. Tables 3, 10, 9 and 5 report the held-out frames. Persist & fit reaches an all-frame Dice of 0.74 and a balanced Dice of 0.61, against 0.60 and 0.49 for confidence-only self-training ($p<0.001/0.001$ for all-frame Dice), 0.63 and 0.52 for augmentation agreement ($p<0.001/0.005$) and 0.54 and 0.45 for the source models ($p<0.001/<0.001$). Its all-frame margin over the empty predictor ($+0.06$, $p=0.04/0.11$) does not survive Holm correction (adjusted $p=1.00$), while its balanced margin ($+0.11$) does (adjusted $p=0.016$). All-frame Dice also fails to separate persist & fit from the supervised reference (0.74 against 0.81, $p=0.057/0.035$, adjusted $p=1.00$), and balanced Dice separates them (0.61 against 0.74, adjusted $p=0.023$). With rules at test time, confidence-only self-training (0.73 all-frame, 0.59 balanced) and augmentation agreement (0.73 and 0.60) cannot be told apart from persist & fit ($p=0.53/0.75$ and $p=0.85/0.89$ for all-frame Dice, bootstrap $p=0.21$ and 0.70 for balanced Dice). Mean Dice on annotated frames does not differ from confidence-only self-training (0.34 against 0.32, $p=0.65/0.60$), thyroid Dice is lower (0.60 against 0.64, $p=0.007/0.048$), and false detection rates fall from 77% to 29% for the thyroid and from 26% to 5% for the trachea. Persist & fit misses

the carotid and the trachea on 67% of the frames where each is annotated. The supervised reference, trained on all 174 evaluation frames, misses them on 27% and 33% and reaches thyroid, carotid and trachea Dice of 0.62, 0.61 and 0.48, against 0.60, 0.22 and 0.19 for persist & fit.

Ablation on the held-out frames. Here the anatomical test alone raises all-frame Dice over confidence-only self-training more than persistence alone (0.70 and 0.63 against 0.60, Table 9). Only differences from the full model (0.74) were tested. Persistence alone is lower ($p<0.001/0.008$, adjusted $p=0.016$), the anatomical test alone is lower by the bootstrap test only ($p=0.008/0.18$, adjusted $p=0.43$), and T+A (0.72) does not differ ($p=0.10/0.39$). Withholding rejected regions as ignore instead of background keeps thyroid Dice on annotated frames higher (0.60 against 0.56, $p=0.004/0.06$) at unchanged all-frame Dice (0.74 against 0.73, $p=0.75/0.18$).

Few-shot track. The few-shot track measures how far labelled target frames raise Dice on annotated frames and balanced Dice. For each $k \in \{1, 2, 4, 8, 16\}$, a four-class student is trained on the labelled evaluation frames of k participants (six per participant, including frames flagged as showing no structure, which are all background) and tested on the held-out frames of the other $29 - k$ participants ($58 - 2k$ frames), together with the label-free references on the same frames. Participants are drawn three times for each k, each draw taking the first k of a random permutation of the 29 participants, so the sets are nested in k within a draw. Three initialisations share the participant draws and training seeds (the draw index): supervised training from the gland model (weights loaded where shapes match, with a new four-class head, as for the supervised reference) and fine-tuning of all weights of the persist & fit student or of the confidence-only student. Training follows the supervised reference (batch 8, 256×256, final epoch kept, App. E) for 60 epochs or at least 720 optimisation steps, whichever is longer, so that no initialisation is under-trained.

Fine-tuning the persist & fit student gives a higher mean balanced Dice than training from the gland model at every k (0.59 against 0.51 at $k = 1$, 0.65 against 0.58 at $k = 2$, 0.71 against 0.69 at $k = 4$, 0.73 against 0.66 at $k = 8$ and 0.76 against 0.73 at $k = 16$), and a higher Dice on annotated frames (0.55 against 0.47 at $k = 4$). The fine-tuned confidence-only student gives a similar Dice on annotated frames (0.58 at $k = 4$) and a lower balanced Dice up to $k = 8$ (0.54, 0.60, 0.69 and 0.70), and both reach 0.76 at $k = 16$. With one labelled participant, balanced Dice (0.59) does not exceed that of the label-free persist & fit student on the same frames (0.61), so one participant does not reliably help.

In the first version of the track, every run had 60 epochs and all 58 held-out frames were scored, including those of the training participants. The gland-initialised model then had only 60 steps at $k = 1$ and was under-trained, which inflated its gap (balanced Dice 0.60 against 0.15 at $k = 1$). In both versions, the label-free students were adapted on the unlabelled sweeps of all 29 participants, including the images of the held-out frames, which favours them over the gland model.

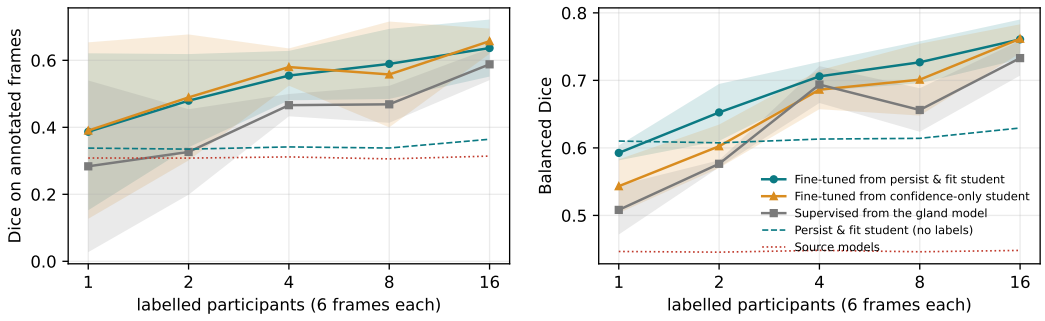


Figure 6: Few-shot track, scored on the held-out frames of the participants not used for fine-tuning ($58 - 2k$ frames): Dice on annotated frames (left) and balanced Dice (right) against the number k of labelled participants (six evaluation frames each). Every run has at least 720 optimisation steps. Dashed and dotted: label-free references on the same frames. Bands: ±1 sd over three participant draws.