LEARNING POLYLINE GEOMETRY FOR LENGTH MEASUREMENT IN MEDICAL & BIOLOGICAL IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Object length measurement is a common objective in many biological and medical imaging applications, such as in identifying a tumor's longest axis, tracking larval growth, or measuring duodenal villi height. Conventional deep learning pipelines typically compute length as a byproduct of dense segmentation mask outputs. In this work, we question whether segmentation is even required if length measurement is the primary objective and propose polyline annotations for length measurement as an alternative. A polyline is defined as a sequence of K points along an object's principal axis; we posit that it is inexpensive to annotate yet sufficient for measurement. We introduce MEASURENET, a framework that directly predicts polylines using specialized, measurement aware loss functions. We evaluate MEASURENET across five datasets spanning brightfield and fluorescence microscopy, optical coherence microscopy, computed tomography, and H&E histopathology, including on CeDeM, a newly curated duodenal biopsy dataset with application to Celiac disease gradation. Across these measurement datasets, MEASURENET trained with polyline annotations outperforms SoTA segmentation models trained on dense segmentation masks, when controlled for the annotation budget, reducing length measurement Mean Absolute Error (MAE) by up to 30 points. Furthermore, MEASURENET outperforms the closest length measurement baselines across all datasets in measurement as well as localization, yielding up to 12.87 point MAE and 11.56 point mAP gains.

1 INTRODUCTION

Length measurement is a common task across biological and clinical domains (Rhyou et al., 2025). In clinical oncology, for instance, tumor size is a primary metric for determining cancer staging and evaluating patient response to treatment (Moon et al., 2023; Taniyama et al., 2021). In histopathology, quantifying the length of duodenal villi and crypts is essential for grading the severity of Celiac disease (Das et al., 2019). Beyond clinical applications, length measurement is also used in microbiology and developmental biology, where bacterial cell length is measured to determine growth rates and assess antibiotic response (Spahn et al., 2022), while organism length is used to monitor larval developmental stages (Fishman et al., 2023) and quantify worm size (Wählby et al., 2012).

To estimate object length, conventional deep learning pipelines (Rhyou et al., 2025; Koukoutegos et al., 2024) rely on an indirect, multistage paradigm. They first predict a dense segmentation mask or a pixel-level skeleton using segmentation model such as SegFormer (Xie et al., 2021), and then postprocess (Rhyou et al., 2025) this geometry to extract the final length (Fig. 1 (a)). This approach introduces downstream errors because even a small discontinuity in the mask propagates directly into the length estimate. Moreover, dense segmentation masks are expensive to annotate. Annotators have to mark multiple points along the boundary of an instance to approximate the actual shape, making them labor intensive. To circumvent these bottlenecks, we propose a polyline representation defined as a sequence of K points along an instance's principal axis as an efficient and accurate alternative. Here K can be varied based on the curvature of the object, with two points sufficient for straight objects and more for curved ones. A polyline is therefore cheap to annotate yet closely approximates the geometry that length measurement requires (see Fig. 1 (b)).

A possible way to leverage polyline annotations is to formulate length measurement as an object detection problem, where the goal is to predict one polyline per instance. General object detectors,

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.

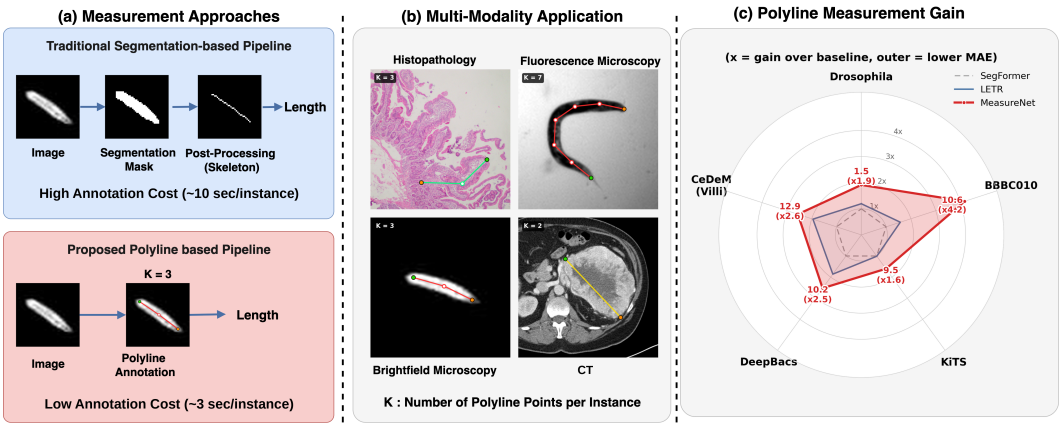

Figure 1: **(a)** In conventional pipeline a dense segmentation mask (with $\sim$ 10 sec/instance for annotation) is skeletonized, and then length is measured. MEASURENET directly predicts a K-point polyline ($\sim$ 3 sec/instance to annotate). **(b)** Examples across modalities: villus (H&E, K=3), C. elegans (fluorescence, K=7), Drosophila larva (brightfield, K=3), and kidney lesion (CT, K=2). **(c)** Relative length error of SegFormer, LETR, and MEASURENET across five datasets. Greater distance from the center indicates lower error. MEASURENET achieves the lowest error on all datasets.

such as DETR (Carion et al., 2020) or DINO-DETR (Zhang et al., 2022), cannot be directly applied because they localize instances using bounding boxes rather than polylines. To bridge this gap, we introduce MEASURENET, a measurement aware polyline detector that combines standard localization losses (Zhang et al., 2022) with three measurement specific loss functions, which align the network optimization objective with the final measurement task. First, the Earth Mover's Distance loss ($\mathcal{L}_{\text{EMD}}$) (Rubner et al., 2000) matches predicted and ground-truth points via a bijective optimal transport plan, treating the polyline as a coherent structure to preserve its geometry and curvature. Second, the length loss ($\mathcal{L}_{\text{L}}$) directly penalizes error in total polyline length. Third, a part length loss ($\mathcal{L}_{\text{PL}}$) penalizes errors in individual segment lengths, thereby preserving curvature as K grows.

We evaluate MEASURENET across five diverse datasets spanning bright-field and fluorescence microscopy, optical coherence microscopy, computed tomography (CT), and H&E histopathology. In addition to experimenting with four publicly available datasets – DeepBacs (Spahn et al., 2022), Drosophila (Fishman et al., 2023), KiTS19 (Heller et al., 2021), and BBBC010 (Wählby et al., 2012) – we introduce CeDeM, a new duodenal biopsy dataset. CeDeM comprises 752 H&E images containing 2,617 villi, 4,211 crypts, 872 villus shoulder and crypt border polyline annotations, all annotated by expert pathologists (Fig. 5). Across all datasets, under the same annotation time as our polyline annotation, a segmentation model trained on dense masks underperforms by up to 30 Mean Absolute Error (MAE) points. Even when segmentation models are trained using an unrestricted annotation budget (the same number of annotated instances), MEASURENET surpasses them on most datasets. This shows that dense segmentation provides little additional benefit despite requiring several times more labeling effort (Fig. 1 (c)). Furthermore, MEASURENET outperforms the closest polyline detection baseline by up to 12.87 points in MAE and 11.56 points in mAP.

In summary, our contributions are fourfold: first, we demonstrate that for object length measurement, polyline representations are more time efficient and accurate than dense segmentation masks. Second, we propose MEASURENET, a dedicated polyline detection architecture driven by measurement aware losses like Earth Mover's Distance, length, and part length. Third, we introduce and release CeDeM, a clinical benchmark of duodenal biopsy for villi and crypt measurement with polyline annotation. Fourth, we evaluate MEASURENET across five datasets with different modalities, showing that MEASURENET consistently achieves strong measurement performance. Our code is available at `https://anonymous.4open.science/r/MeasureNet_Polyline-2163`.

2 RELATED WORK

Segmentation based measurement. Most measurement pipelines segment structures and derive length from the predicted masks. General segmenters such as SegFormer (Xie et al., 2021) and

BEiT (Bao et al., 2021), along with task specific methods (Ning et al., 2024; Galdran et al., 2022; Koukoutegos et al., 2024; Rhyou et al., 2025), predict dense masks or skeletons that are converted to medial axes. This approach has two limitations. Weak boundaries can fragment masks and propagate errors to length estimates, while pixel level annotation is labor intensive. For touching or overlapping objects, Omnipose (Cutler et al., 2022) provides accurate instance masks for bacteria and worms and is used as our dense mask baseline in the annotation cost study (Sec. 5.1).

Detection based measurement. Object detectors (Zhang et al., 2022) localize instances with bounding boxes and cannot be directly applied to polyline detection. Lane detectors (Pan et al., 2018; Tabelini et al., 2021a;b) and map detectors such as VectorMapNet (Liu et al., 2023) and MapTR (Liao et al., 2025) predict curves as masks, point sequences or polynomials, but rely on driving scene priors such as border anchors or near vertical lanes. Yolino (Meyer et al., 2021) detects generic polylines, but composes them bottom up from short segments predicted per grid cell. Line segment detectors (Von Gioi et al., 2008; Xue et al., 2020; Pautrat et al., 2023) respond to image edges rather than complete structures. Closest to our work, LETR (Xu et al., 2021) casts line detection as DETR style endpoint prediction, and RANK-LETR (Tong et al., 2025) re-ranks its predictions using geometric cues. Both represent an object by two endpoints and cannot follow curved structures. Landmark detectors such as MedSapiens (Elbatel et al., 2026) predict a fixed set of landmarks per image rather than a set of polylines. None of these methods optimize the measured length. MEASURENET builds on DINO-DETR, and we compare it with LETR, RANK-LETR, Yolino, and MapTR (Sec. 5.2).

Point set distance metrics. Comparing a predicted polyline to ground truth requires a distance between point sets. An index wise ℓ_1 loss assumes a fixed point order, while the permutation invariant distance metrics like Chamfer distance (Barrow et al., 1977; Wu et al., 2021) allow multiple predictions to collapse to one target point. The Earth Mover's (Wasserstein) distance (Rubner et al., 2000) avoids both issues by coupling the two sets through a bijective transport plan. We adopt it as a loss and compare it with the other distances in Sec. 5.3.

3 METHOD

For object length measurement, we propose polyline annotations instead of dense segmentation masks and formulate the problem as a detection task. Formally, given an image $I \in \mathbb{R}^{H\times W\times 3}$ where H and W denote the image height and width, the goal is to predict a set of polylines

$$\mathcal{P} = \{(P_i, c_i)\}_{i=1}^{N_I}, \qquad P_i = \{p_{i,k}\}_{k=1}^{K}, \qquad p_{i,k} \in \mathbb{R}^2, \qquad c_i \in \mathcal{C},$$

where P_i denotes the polyline of the i-th instance, represented by K points $p_{i,k} \in \mathbb{R}^2$, $c_i \in \mathcal{C}$ is the associated class, and N_I is the number of instances. The length of a polyline is computed as $\text{len}(P_i) = \sum_{k=1}^{K-1} \|p_{i,k+1} - p_{i,k}\|_2$. Each training object is annotated with a polyline along its principal axis, with a fixed number of points K per class (see Fig. 1 (b)).

To solve this formulation, MEASURENET (Fig. 2) builds on the DINO-DETR (Zhang et al., 2022) architecture and replaces bounding box predictions with polyline predictions. Each query predicts a K point polyline, while the classification head remains unchanged. Each polyline prediction is matched to a ground truth polyline (Sec. 3.1), and losses are computed over the matched pairs.

In addition to localization losses, MEASURENET introduces three measurement specific losses that align training with the measurement objective (Sec. 3.2, Fig. 2(ii)). The Earth Mover loss aligns predicted and ground truth polylines, preserving their overall geometry. The length loss supervises the total length, while the part length loss matches individual segment lengths to preserve curvature. Together, these losses directly optimize the measurement rather than relying on localization alone.

3.1 ADAPTING DINO-DETR FOR POLYLINE DETECTION

DINO-DETR (Zhang et al., 2022) decodes a fixed set of object queries into instance predictions, each with a class and a bounding box, and assigns them to the ground truth by Hungarian matching (Kuhn, 1955) on a cost that combines classification, box ℓ_1 and GIoU (Rezatofighi et al., 2019) terms. We replace the box head with a polyline head, so each query predicts a class c_i and a K point polyline P_i. Let $\mathcal{P}^* = \{(P_i^*, c_i^*)\}_{i=1}^{N_I}$ denote the ground truth. We update the Hungarian matching cost accordingly, replacing GIoU with a Chamfer distance (Barrow et al., 1977) alongside

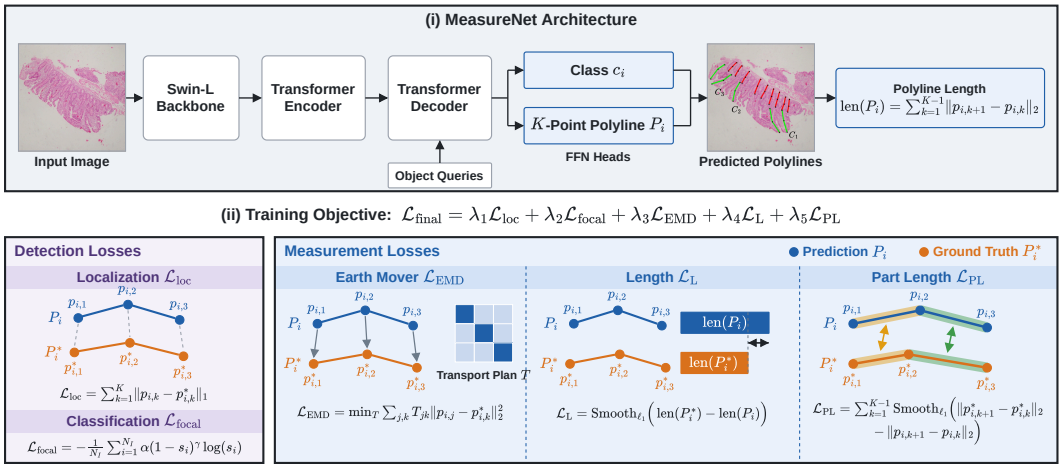

Figure 2: **MEASURENET predicts one polyline per instance and is trained to measure.** **(i)** The MeasureNet Architecture, which includes a Swin backbone and a DINO-DETR encoder-decoder. **(ii)** The training objective $\mathcal{L}_{\text{final}}$ combines five terms, where we include measurement-specific losses.

the classification and direction invariant ℓ_1 terms (App. B.2). Each matched pair (P_i, P_i^*) is then trained with the localization and classification losses.

The localization loss $\mathcal{L}_{\text{loc}}$ supervises the coordinates of the matched polyline by computing the ℓ_1 distance between corresponding predicted and ground truth points:

$$\mathcal{L}_{\text{loc}} = \sum_{k=1}^{K} \left\| p_{i,k} - p^*_{i,k} \right\|_1 . \tag{1}$$

While computing $\mathcal{L}_{\text{loc}}$, we take the minimum over P_i^* and its reversal $\overleftarrow{P}_i^*$, so reversed predictions are not penalized for opposite annotation order, making it invariant to the annotation direction.

Classification loss ($\mathcal{L}_{\text{focal}}$). To address class imbalance, we use the focal loss (Lin et al., 2020) for object classification. For the i-th matched prediction with ground truth class c_i^*, let $s_i = \Pr(c_i = c_i^*)$ be the predicted probability of that class. Summed over the N_I matched predictions, the loss is

$$\mathcal{L}_{\text{focal}} = -\frac{1}{N_I} \sum_{i=1}^{N_I} \alpha(1 - s_i)^\gamma \log(s_i), \tag{2}$$

where α and γ are the focal loss parameters (Sec. 4); unmatched queries are trained toward the no object class. The focal loss down weights well classified examples and emphasizes hard examples.

3.2 MEASUREMENT SPECIFIC LOSSES

Earth Mover loss ($\mathcal{L}_{\text{EMD}}$). The localization loss supervises each polyline point independently but does not enforce the overall polyline geometry. To preserve the global shape, MEASURENET introduces an Earth Mover's Distance (EMD) loss (Rubner et al., 2000), which aligns the predicted and ground truth polylines as point sets through an optimal transport plan. For a matched pair of polylines (P_i, P_i^*), the loss is defined as

$$\begin{aligned} \mathcal{L}_{\text{EMD}}(P_i, P_i^*) &= \min_T \sum_{j=1}^{K}\sum_{k=1}^{K} T_{jk} \, \|p_{i,j} - p^*_{i,k}\|_2^2, \\ \text{subject to} \quad & T_{jk} \geq 0, \qquad \sum_{k=1}^{K} T_{jk} = \frac{1}{K}, \qquad \sum_{j=1}^{K} T_{jk} = \frac{1}{K}, \end{aligned} \tag{3}$$

where $T \in \mathbb{R}^{K \times K}$ is the optimal transport matrix that establishes a one to one correspondence between the predicted and ground truth points (Fig. 2(ii)). Compared to the ℓ_1 loss, EMD computes

a bijective, permutation invariant correspondence between predicted and ground truth points. Consequently, the loss is independent of the order in which the K points are predicted and prevents multiple predictions from collapsing onto the same ground truth point, ensuring correct polyline structure. In practice, we use the differentiable Sinkhorn formulation of EMD (Cuturi, 2013).

Length loss ($\mathcal{L}_{\text{L}}$). While the localization and EMD losses supervise point positions and overall geometry, the length loss directly optimizes the measurement by applying a Smooth ℓ_1 loss (Girshick, 2015) between predicted and ground truth polyline lengths, with $\text{len}(\cdot)$ defined earlier:

$$\mathcal{L}_{\text{L}}(P_i, P_i^*) = \text{Smooth}_{\ell_1}(\text{len}(P_i^*) - \text{len}(P_i)) \,. \tag{4}$$

Part length loss ($\mathcal{L}_{\text{PL}}$). The length loss constrains only the total polyline length and cannot distinguish how the length is distributed along the polyline. We therefore introduce a part length loss that matches the length of each polyline segment:

$$\mathcal{L}_{\text{PL}}(P_i, P_i^*) = \sum_{k=1}^{K-1} \text{Smooth}_{\ell_1}\big(\|p_{i,k+1}^* - p_{i,k}^*\|_2 - \|p_{i,k+1} - p_{i,k}\|_2\big). \tag{5}$$

Segments follow the ground truth order chosen for $\mathcal{L}_{\text{loc}}$, so both polylines share the same orientation. For $K = 2$, $\mathcal{L}_{\text{PL}}$ reduces to $\mathcal{L}_{\text{L}}$ and $\mathcal{L}_{\text{PL}}$ is only effective for $K \geq 3$. The overall training objective combines the localization, classification, and measurement losses:

$$\mathcal{L}_{\text{final}} = \lambda_1 \mathcal{L}_{\text{loc}} + \lambda_2 \mathcal{L}_{\text{focal}} + \lambda_3 \mathcal{L}_{\text{EMD}} + \lambda_4 \mathcal{L}_{\text{L}} + \lambda_5 \mathcal{L}_{\text{PL}}. \tag{6}$$

The loss weights λ_i are set through a grid search on a validation set (Sec. 4). In summary, our proposal, MEASURENET, adapts a known bounding box object detector (DINO-DETR) for our task of detecting polylines, and adds additional loss terms, since our primary goal is object measurement.

4 IMPLEMENTATION DETAILS

Datasets: We evaluate MEASURENET on five datasets spanning five imaging modalities. Four are public; the fifth, CeDeM, is a duodenal biopsy dataset which we contribute; splits are described as train/validation/test images. KiTS19 ($K = 2$): Kidney lesion CT slice dataset (Heller et al., 2021) with a dataset split of $730/166/146$. We derive a two point polyline from the longest chord within each lesion mask and use its length as the tumor size. DeepBacs ($K = 2$): Live cell brightfield images of bacteria (Spahn et al., 2022) with dataset split $70/7/60$. The original segmentation masks are converted into two point polylines along each bacterium's principal axis. The bacteria are short, straight, and densely packed, making instances difficult to distinguish. Drosophila ($K = 3$): Optical coherence microscopy images of *Drosophila* larvae (Fishman et al., 2023) with dataset split $782/136/204$. Each larva is represented with a three point polyline along its principal axis. The larvae are small, curved, and frequently touch or overlap. BBBC010 ($K = 7$): Bright field and fluorescence images of coiled *C. elegans* worms (Wählby et al., 2012) with dataset split $66/16/16$. We convert the segmentation masks into seven point medial axis polylines. The worms exhibit high curvature and frequent overlap. CeDeM ($K = 3$; ours): A curated dataset of 752 H&E duodenal biopsy images from unique patients (see Fig. 5), with dataset split $612/70/70$. CeDeM is the first dataset for villi and crypt length measurement with polyline annotations. Following standard pathology practice, crypt length is defined as the distance between the villus shoulder (VS) and crypt border (CB). Accordingly, CeDeM provides polyline annotations for villi and crypts ($K = 3$), and VS and CB ($K = 5$). All annotations were performed by expert pathologists, verified by a senior pathologist, and disagreements were resolved by consensus. For each dataset, K is chosen to cover its most curved instances; straighter instances use K equidistant points (details in App. A).

Evaluation Metrics: We measure length accuracy with the Matched Mean Absolute Error, M-MAE $= \max(\text{MAE}_{G\to P}, \text{MAE}_{P\to G})$. $\text{MAE}_{G\to P}$ matches each ground truth (GT) polyline to its closest prediction under the Chamfer distance and averages the absolute length difference; $\text{MAE}_{P\to G}$ does the same in the other direction. A false positive therefore increases $\text{MAE}_{P\to G}$ and a false negative increases $\text{MAE}_{G\to P}$, while duplicate predictions of one GT leave M-MAE largely unaffected but affects precision. For localization, we report precision, recall, and mAP using a Chamfer-distance threshold (Barrow et al., 1977) of 50 squared pixels (≈ 7 pixels) for true positives, except BBBC010, for which we use 120 (set on a validation set).

MEASURENET Details: MEASURENET builds on DINO-DETR (Zhang et al., 2022) with a Swin-L backbone (Liu et al., 2022), six encoder layers, and six decoder layers with deformable attention. We use 900 object queries (100 for KiTS19, which has one lesion per image, and 200 for BBBC010). For CeDeM, two detection heads are used: one predicts villi and crypt polylines ($K = 3$), and the other predicts villus shoulder (VS) and crypt border (CB) polylines ($K = 5$). For BBBC010, both MEASURENET and the segmentation baselines are initialized from models trained on Drosophila owing to the limited training set (66 images). We train using AdamW with $\mathcal{L}_{\text{final}}$ (Eq. 6), a learning rate of 10^{-4} for the detection head and 10^{-5} for the backbone, batch size 8, 640×640 images, random rotations and flips, for up to 150 epochs. The Earth Mover loss $\mathcal{L}_{\text{EMD}}$ (Eq. 3) uses the GeomLoss (Feydy et al., 2019) Sinkhorn solver (blur 0.05, scaling 0.9), and the focal loss $\mathcal{L}_{\text{focal}}$ (Eq. 2) uses $\alpha = 0.25$ and $\gamma = 2$. A validation set grid search is used to select the loss weights $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$, which vary per dataset: CeDeM $(1, 1, 1, 1, 1)$, DeepBacs $(1, 1, 1, 1, 0)$, Drosophila $(1, 1, 1, 2, 2)$, KiTS19 $(1, 1, 2, 2, 0)$, and BBBC010 $(1, 1, 3, 5, 2)$. All detection baselines use the same backbone, data augmentation and adapted to predict same number of points (K). Segmentation baselines (SegFormer, BEiT, FRNet, and LWNet) are trained to predict a dilated ground truth polyline, from which length is recovered by skeletonization. For the annotation cost study (Sec. 5.1), the cost matched and oracle baselines are trained on full object masks. We use Omnipose for BBBC010 and DeepBacs, and SegFormer for the remaining datasets. In both cases, length is computed from the medial axis of the predicted segmentation. All length predictions are made in pixel space and can be converted to microscope units using a scaling factor. All experiments are conducted on an NVIDIA A100 GPU. Additional details are provided in Appendix C.

5 RESULTS

We organize our evaluation around four questions. (i) How does polyline annotation compare with dense segmentation for length measurement, when controlled for annotation effort (Fig. 3)? (ii) How does MEASURENET compare with other existing methods for polyline based length measurement across diverse imaging modalities (Tables 1 and 2, Fig. 4)? (iii) What is the contribution of each proposed loss term (Tables 3 and 4)? (iv) What are the failure modes of MEASURENET (Fig. 7)?

5.1 POLYLINE ANNOTATION IS EFFICIENT FOR LENGTH MEASUREMENT

Polyline annotation substantially reduces annotation effort compared with dense segmentation (see Fig. 3 (a)). Instead of outlining the entire object boundary, annotators place a small number of points along the object's principal axis (Fig. 1(b)). Across all datasets, two medical experts annotated polylines $2.5\times$–$5.8\times$ faster than dense masks. The largest gain is observed on CeDeM ($5.8\times$), where the irregular villi and crypt boundaries make dense mask annotation particularly challenging. Even for well defined straight objects ($K = 2$), polyline annotation is $2.8\times$–$4.5\times$ faster. For the highly curved worms in BBBC010 ($K = 7$), where both annotation types require greater effort, polyline annotation still provides a $2.5\times$ reduction in annotation time.

To evaluate the effectiveness of polyline annotation, we compare it with full instance segmentation under two settings. (1) **Same annotation budget:** both polyline and dense segmentation are allocated the same annotation time. Using the annotation times in Fig. 3(a), we determine how many dense masks can be annotated within the time required to annotate the complete polyline dataset. (2) **Same instance count:** both methods are trained on the same number of instances, irrespective of annotation time. In both settings, object length is recovered from the predicted segmentation (mask $\rightarrow$ skeleton $\rightarrow$ length). For the segmentation baselines, we use the best performing segmentation model for each dataset: Omnipose (Cutler et al., 2022) on BBBC010 and DeepBacs, and SegFormer (Xie et al., 2021) on the remaining datasets.

Under the same annotation budget, MEASURENET achieves lower measurement error than dense segmentation on four of the five datasets, with the largest improvements on CeDeM (56.73 to 12.92) and BBBC010 (41.03 to 10.59), corresponding to $4.4\times$ and $3.9\times$ lower measurement error, respectively, followed by DeepBacs (15.62 to 10.17) and Drosophila (2.00 to 1.51). Even when segmentation is trained using every available dense mask (same instance count), MEASURENET still matches or outperforms the full-mask setting on these datasets (Drosophila: 1.72 to 1.51; DeepBacs: 13.33 to 10.17; BBBC010: 15.46 to 10.59), despite requiring only a fraction of the annotation effort (Fig. 6 (b)). Only on KiTS19 does the full mask retain an advantage (6.75 vs. 9.53), as the target is sin-

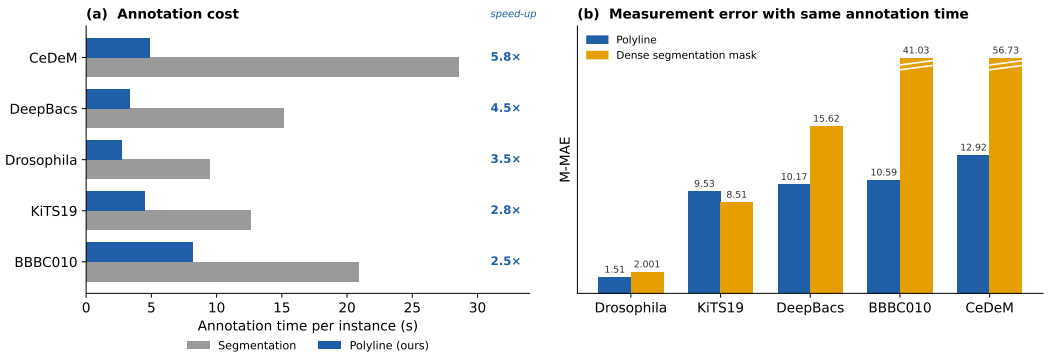

Figure 3: **Polyline annotation is cheaper while matching the measurement accuracy of dense segmentation. (a)** Average per instance annotation time measured by two medical experts. Polyline annotations are 2.5–5.8× faster than dense masks across all five datasets. **(b)** Measurement error (M-MAE, lower is better) of polyline (MEASURENET) and a same annotation budget segmentation model. We use Omnipose for DeepBacs and BBBC010, and SegFormer for the remaining datasets. MEASURENET matches or outperforms on CeDeM, Drosophila, DeepBacs, and BBBC010; only on KiTS19, where the target is a single well defined structure, do dense masks retain a slight advantage.

Table 1: Measurement and localization performance on four datasets (KiTS19, DeepBacs, Drosophila, and BBBC010). Measurement accuracy is reported as mean absolute error (M-MAE, ↓) and localization performance as mAP (↑). Best results in each column are shown in **bold**; closest baseline shown with underline; MEASURENET is highlighted (shaded).

	KiTS19 (K=2)		DeepBacs (K=2)		Drosophila (K=3)		BBBC010 (K=7)	
Model	M-MAE↓	mAP↑	M-MAE↓	mAP↑	M-MAE↓	mAP↑	M-MAE↓	mAP↑
BEiT (2021)	11.21 ±0.23	59.77 ±3.73	25.83 ±0.60	52.74 ±0.30	2.73 ±0.04	88.96 ±0.19	41.78 ±0.26	56.65 ±0.46
FRNet (2024)	14.02 ±1.08	53.67 ±2.47	24.77 ±0.44	51.62 ±0.36	3.76 ±0.15	89.37 ±0.38	41.47 ±0.32	55.59 ±1.31
LWNet (2022)	19.57 ±0.74	25.17 ±2.52	27.09 ±0.29	50.43 ±0.37	3.25 ±0.08	89.51 ±0.36	43.17 ±1.68	59.58 ±0.49
SegFormer (2021)	15.30 ±0.34	63.26 ±1.94	25.69 ±0.43	51.59 ±3.79	2.89 ±0.08	89.47 ±0.69	44.11 ±0.59	56.02 ±0.81
LETR (2021)	15.13 ±0.65	63.12 ±0.29	13.82 ±0.43	61.66 ±0.61	2.42 ±0.15	97.44 ±0.48	28.07 ±5.86	57.40 ±2.30
MapTR (2025)	12.46 ±0.36	64.41 ±1.28	13.35 ±0.25	57.27 ±0.46	5.95 ±0.82	95.78 ±0.29	23.46 ±2.44	70.67 ±2.64
RANK-LETR (2025)	10.54 ±0.07	71.68 ±0.15	14.54 ±0.33	42.30 ±1.37	5.76 ±0.67	91.24 ±0.08	29.42 ±2.21	54.39 ±0.11
MEASURENET (ours)	**9.53** ±0.32	**75.73** ±0.27	**10.17** ±0.02	**64.33** ±0.06	**1.51** ±0.12	**97.81** ±0.07	**10.59** ±0.18	**82.23** ±0.04

gle, circular and well defined. Overall, for curved or densely packed structures, dense segmentation provides no measurement benefit, and requires several times more annotation. This supports our hypothesis that polyline annotation is an efficient representation for object length measurement.

5.2 MEASURENET OUTPERFORMS OTHER POLYLINE DETECTORS

MEASURENET achieves the best measurement performance across diverse datasets compared to other polyline detection methods. Tables 1 and 2 compare MEASURENET against segmentation for polylines (BEiT (Bao et al., 2021), FRNet (Ning et al., 2024), LWNet (Galdran et al., 2022), and SegFormer (Xie et al., 2021)), lane and map detection methods (Yolino (Meyer et al., 2021) and MapTR (Liao et al., 2025)), and line detection methods (LETR (Xu et al., 2021) and RANK-LETR (Tong et al., 2025)). Across the four single class datasets (Table 1), MEASURENET consistently attains the lowest M-MAE and the highest mAP. Compared with the strongest baseline, MEASURENET improves M-MAE from 10.54 to 9.53 on KiTS19, where the target is a well defined lesion represented by a straight two point chord. The gains are larger on more challenging datasets with densely packed or highly curved objects, reducing M-MAE from 13.35 to 10.17 on DeepBacs, from 2.42 to 1.51 on Drosophila, and by over 50% on BBBC010 (23.46 to 10.59), while simultaneously improving mAP by over 10 points (70.67 to 82.23) (for all metrics see Table 6). On the CeDeM dataset (see Table 2), MEASURENET achieves the lowest measurement error for villi (12.92), crypts (8.80), and the clinically relevant villi to crypt ratio (0.572), while also obtaining the highest recall (76.11) and mAP (53.01). For all datasets, compared with the strongest baseline, the gains are statistically significant (two-sided Student's t-test, $p < 0.01$).

Table 2: Measurement performance on CeDeM ($K = 3$). MAE-V and MAE-C are M-MAE for villi and crypts; MAE-Ratio is the M-MAE of the villi to crypt length ratio. Values are mean ± s.d. over 5 repeated runs; best per column in **bold**, our method (shaded) is best on all measurement columns.

Model	M-MAE-V ↓	M-MAE-C ↓	MAE-Ratio ↓	Precision ↑	Recall ↑	mAP ↑
FRNet (2024)	68.07 ±5.1	34.86 ±2.9	1.110 ±0.12	31.83 ±2.7	27.16 ±3.1	36.52 ±2.8
SegFormer (2021)	33.05 ±1.75	19.09 ±1.27	0.858 ±0.07	36.61 ±2.3	55.72 ±3.6	40.91 ±2.9
BEiT (2021)	62.22 ±4.6	38.82 ±3.8	1.100 ±0.15	37.91 ±2.4	26.88 ±2.7	47.71 ±3.2
Yolino (2021)	62.98 ±4.2	21.14 ±2.5	1.650 ±0.20	14.30 ±1.9	54.18 ±3.3	20.31 ±1.6
MapTR (2025)	21.41 ±3.3	19.32 ±1.1	1.440 ±0.08	**54.61** ±0.7	31.27 ±0.7	42.93 ±2.5
LETR (2021)	16.94 ±0.9	11.81 ±0.6	0.704 ±0.03	39.52 ±2.1	70.85 ±2.7	40.39 ±2.4
RANK-LETR (2025)	62.04 ±4.5	23.51 ±2.6	1.190 ±0.16	17.20 ±2.1	49.80 ±3.2	20.23 ±1.7
MEASURENET (ours)	**12.92** ±0.6	**8.80** ±0.2	**0.572** ±0.02	53.01 ±1.7	**76.11** ±1.8	**53.01** ±1.5

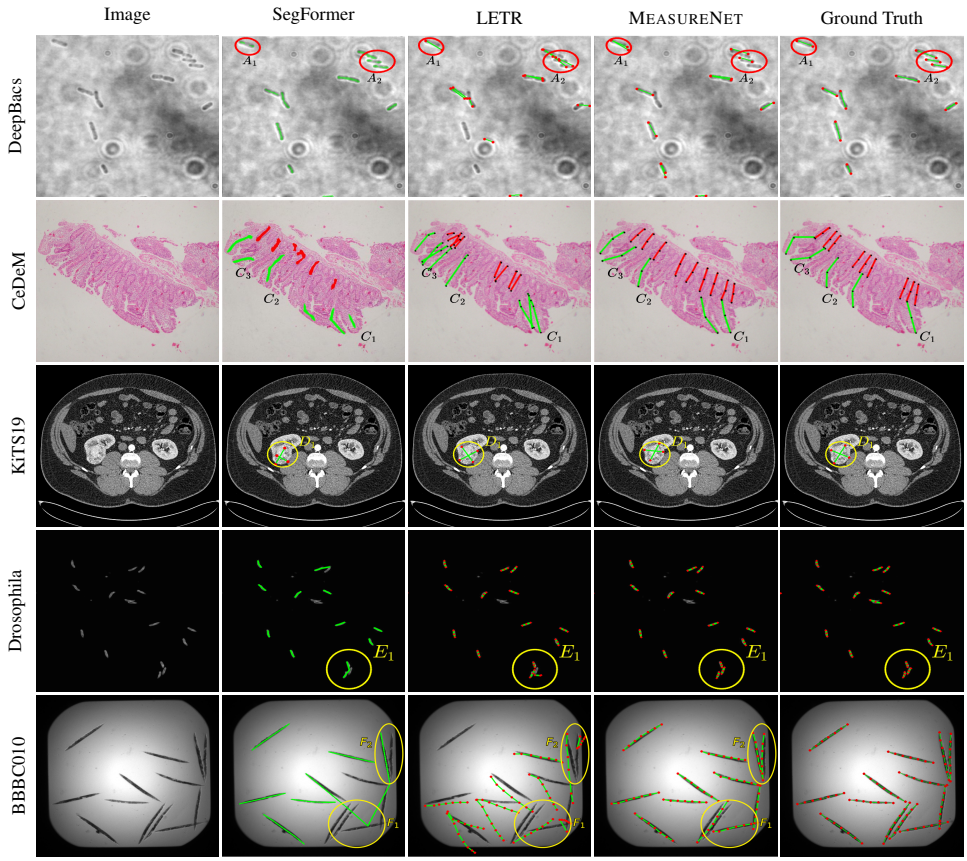

Figure 4: Qualitative comparisons. Columns: input image, SegFormer (segmentation for polylines), LETR (line detection), MEASURENET, and ground truth. Rows: different datasets. SegFormer yields fragmented predictions and LETR misaligns endpoints, while MEASURENET produces continuous, well localized polylines that better match the ground truth.

MEASURENET produces accurate polylines across diverse imaging modalities. Fig. 4 compares MEASURENET with the segmentation baseline (SegFormer) and the polyline baseline (LETR) across all five datasets. SegFormer often produces fragmented or discontinuous masks (Fig. 4, A_1, A_2, C_1, and C_3), while LETR frequently misplaces endpoints or fails to follow object curvature (C_1, D_1, and E_1). In contrast, MEASURENET predicts a continuous polyline that closely follows the principal axis of each object (C_1, C_3, F_1 and F_2), from straight kidney lesions and bacteria to curved Drosophila larvae, coiled *C. elegans* worms, and villi and crypts in H&E histopathology. The predicted endpoints also align well with the true object boundaries (C_1, D_1, and E_1), resulting in

more accurate length measurements. In case of BBBC010 (F_1 and F_2), where objects are curved, SegFormer merges neighboring worms into a single continuous line and LETR collapses multiple points onto the same location, whereas MEASURENET keeps the disjoint instances separated owing to its measurement specific losses.

Failure modes. Errors arise mainly from partially visible structures, whose length MEASURENET underestimates on CeDeM, and from the rare strongly curved larvae in Drosophila, which it sometimes represents with a nearly straight polyline (App. F, Fig. 7).

5.3 ABLATION STUDIES OF MEASURENET

We perform an ablation study on the CeDeM dataset to quantify the contribution of each loss component introduced in Section 3.2 (Table 3). Starting from the DINO-DETR baseline (row 1), adding the Earth Mover loss ($\mathcal{L}_{\text{EMD}}$) improves geometric alignment, reducing MAE-V from 16.94 to 15.47 and MAE-C from 11.81 to 10.89 (row 2). Incorporating the length loss ($\mathcal{L}_{\text{L}}$) further improves measurement accuracy, lowering MAE-V to 14.36 and MAE-C to 9.62 (row 3). Finally, adding the part length loss ($\mathcal{L}_{\text{PL}}$) yields the best overall performance, reducing MAE-V, MAE-C, and MAE-Ratio to 12.92, 8.80, and 0.572, respectively (row 4). Together, these results show that each proposed loss improves the measurement accuracy.

Table 3: Ablation study evaluating the contribution of the Earth Mover loss ($\mathcal{L}_{\text{EMD}}$), length loss ($\mathcal{L}_{\text{L}}$), and part length loss ($\mathcal{L}_{\text{PL}}$) on the CeDeM dataset. Measurement error is reported for villi (MAE-V), crypts (MAE-C), and the villus to crypt length ratio (MAE-R).

Base	$\mathcal{L}_{\text{EMD}}$	$\mathcal{L}_{\text{L}}$	$\mathcal{L}_{\text{PL}}$	MAE-V↓	MAE-C↓	MAE-R↓
✓				16.94	11.81	0.704
✓	✓			15.47	10.89	0.618
✓	✓	✓		14.36	9.62	0.584
✓	✓	✓	✓	**12.92**	**8.80**	**0.572**

Choice of distance metric: Earth Mover best. Table 4 compares three distance metrics for the polyline loss while keeping all other components unchanged on CeDeM. Earth Mover Distance (EMD) consistently achieves the lowest measurement error across all three evaluation metrics, obtaining the best MAE for villi (12.92), crypts (8.80), and the villi to crypt ratio (0.572). Chamfer distance performs worst because its nearest neighbor matching is non bijective, allowing multiple predicted points to match the same ground truth point. Manhattan (ℓ_1) distance enforces index wise correspondence but cannot account for point permutations. In contrast, EMD computes a bijective optimal transport plan, preventing point collapse.

Table 4: Comparison of distance metrics on the CeDeM dataset. Earth Mover Distance (EMD) achieves the lowest measurement error across all evaluation metrics.

Distance	MAE-V ↓	MAE-C ↓	MAE-R ↓
Chamfer	13.00	8.99	0.634
Manhattan (ℓ_1)	13.81	8.81	0.575
Earth Mover	**12.92**	**8.80**	**0.572**

6 CONCLUSIONS & DISCUSSION

We show that length measurement does not require dense segmentation. A polyline of K points along the principal axis is 2.5 to $5.8\times$ faster to annotate than a dense mask and carries sufficient information needed to measure object lengths. MEASURENET predicts these polylines directly, supervised by Earth Mover, length, and part length losses that target the measurement itself. Across five datasets spanning microscopy, computed tomography, and histopathology (including CeDeM, which we release as the first polyline benchmark for villi and crypt measurement), MEASURENET attains the lowest measurement error of all polyline detection baselines, and it outperforms dense segmentation trained under the same annotation budget on four of the five datasets.

Discussion: In a clinical setting, if a predicted measurement is inaccurate, a clinician can simply adjust points to obtain an updated length immediately, rather than correcting an entire segmentation mask and repeating the measurement pipeline. This combination of direct measurement, efficient editing, and human verification makes polylines a practical choice for computer assisted clinical measurement. However, currently, MEASURENET uses a fixed number K of polyline points per dataset. Extending it to adaptive (per instance) polyline representations and volumetric centerline estimation, such as vessels or neurons in 3D imaging, are promising directions for future work.

AI USE STATEMENT

In this work we have used generative AI tools for correcting grammatical errors and restructuring text. Additionally, generative AI tools were used to clean and organize the code for easy reproducibility. We have verified the code and take full responsibility for the work.

ETHICS STATEMENT

CeDeM was collected after obtaining ethical clearance (IEC640/05.08.2022). All images were de-identified before annotation, and the dataset will be publicly released upon acceptance. The other four datasets are publicly available. MEASURENET is intended to assist clinicians. We have discussed the performance of MEASURENET for various tasks along with error analysis (App. F) and limitations (Sec. 6).

REPRODUCIBILITY STATEMENT

Code for MEASURENET is available at `https://anonymous.4open.science/r/MeasureNet_Polyline-2163`. The problem formulation and all losses are defined in Sec. 3 (Eqs. 1–6), and the full Hungarian matching cost with its weights is given in App. B.2. Sec. 4 describes each dataset, its train/validation/test split, and how segmentation masks are converted to K point polylines, and lists the evaluation metrics, architecture, optimizer, learning rates, batch size, input resolution, augmentation, number of queries, loss weights per dataset, and hardware. App. C gives the hyperparameter search and the training settings of every baseline, and Appendix A documents CeDeM acquisition, annotation, and quality control. All reported results are the mean and standard deviation over five repeated runs, and significance is tested with a two-sided Student's t-test. CeDeM, including images, polyline annotations, and splits, will be released upon acceptance.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report, 1977.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Kevin J. Cutler, Carsen Stringer, Teresa W. Lo, Luca Rappez, Nicholas Stroustrup, S. Brook Peterson, Paul A. Wiggins, and Joseph D. Mougous. Omnipose: a high-precision morphology-independent solution for bacterial cell segmentation. *Nature Methods*, 19(11):1438–1448, 2022.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Prasenjit Das, Gaurav PS Gahlot, Alka Singh, Vandana Baloda, Ramakant Rawat, Anil K Verma, Gaurav Khanna, Maitrayee Roy, Archana George, Ashok Singh, et al. Quantitative histology-based classification system for assessment of the intestinal mucosal histological changes in patients with celiac disease. *Intestinal research*, 17(3):387–397, 2019.

Marawan Elbatel, Anbang Wang, Keyuan Liu, Kaouther Mouheb, Enrique Almar-Munoz, Lizhuo Lin, Yanqi Yang, Karim Lekadir, and Xiaomeng Li. Medsapiens: Taking a pose to rethink medical imaging landmark detection. *Medical Image Analysis*, pp. 104015, 2026.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

Matthew Fishman, Abigail Matt, Fei Wang, Elena Gracheva, Jiantao Zhu, Xiangping Ouyang, Andrey Komarov, Yuxuan Wang, Hongwu Liang, and Chao Zhou. A drosophila heart optical coherence microscopy dataset for automatic video segmentation. *Scientific data*, 10(1):886, 2023.

Adrian Galdran, André Anjos, José Dolz, Hadi Chakor, Hervé Lombaert, and Ismail Ben Ayed. State-of-the-art retinal vessel segmentation with minimalistic models. *Scientific Reports*, 12(1): 6174, 2022.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the KiTS19 challenge. *Medical Image Analysis*, 67:101821, 2021.

Konstantinos Koukoutegos, Richard 's Heeren, Liesbeth De Wever, Frederik De Keyzer, Frederik Maes, and Hilde Bosmans. Segmentation-based quantitative measurements in renal ct imaging using deep learning. *European Radiology Experimental*, 8(1):110, 2024.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *International Journal of Computer Vision*, 133(3):1352–1374, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. doi: 10.1109/TPAMI.2018.2858826.

Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *International conference on machine learning*, pp. 22352–22369. PMLR, 2023.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019, 2022.

Annika Meyer, Philipp Skudlik, Jan-Hendrik Pauls, and Christoph Stiller. Yolino: Generic single shot polyline detection in real time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2916–2925, 2021.

Ki-Ryum Moon, Byoung-Dai Lee, and Mu Sook Lee. A deep learning approach for fully automated measurements of lower extremity alignment in radiographic images. *Scientific Reports*, 13(1): 14692, 2023.

Haojian Ning, Chengliang Wang, Xinrun Chen, and Shiying Li. An accurate and efficient neural network for octa vessel segmentation and a new dataset. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1966–1970. IEEE, 2024.

Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Rémi Pautrat, Daniel Barath, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. Deeplsd: Line segment detection and refinement with deep image gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17327–17336, 2023.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019.

Se-Yeol Rhyou, Yongjin Cho, Jaechern Yoo, Sanghoon Hong, Sunghoon Bae, Hyunjae Bae, and Minyung Yu. Automatic lower-limb length measurement network (a3lmnet): a hybrid framework for automated lower-limb length measurement in orthopedic diagnostics. *Electronics*, 14(1):160, 2025.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, November 2000. ISSN 0920-5691. doi: 10.1023/A:1026543900054. URL https://doi.org/10.1023/A:1026543900054.

Christoph Spahn, Estibaliz Gómez-de Mariscal, Romain F Laine, Pedro M Pereira, Lucas von Chamier, Mia Conduit, Mariana G Pinho, Guillaume Jacquemet, Séamus Holden, Mike Heilemann, et al. Deepbacs for multi-task bacterial image analysis using open-source deep learning approaches. *Communications Biology*, 5(1):688, 2022.

Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Keep your eyes on the lane: Real-time attention-guided lane detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 294–302, 2021a.

Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Polylanenet: Lane estimation via deep polynomial regression. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6150–6156. IEEE, 2021b.

Yusuke Taniyama, Kentaro Murakami, Naoya Yoshida, Kozue Takahashi, Hisahiro Matsubara, Hideo Baba, and Takashi Kamei. Evaluating the effect of neoadjuvant chemotherapy for esophageal cancer using the recist system with shorter-axis measurements: a retrospective multicenter study. *BMC cancer*, 21:1–10, 2021.

Xin Tong, Shi Peng, Baojie Tian, Yufei Guo, Xuhui Huang, and Zhe Ma. Improving transformer based line segment detection with matched predicting and re-ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7428–7436, 2025.

Rafael Grompone Von Gioi, Jeremie Jakubowicz, Jean-Michel Morel, and Gregory Randall. Lsd: A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):722–732, 2008.

Carolina Wählby, Lee Kamentsky, Zihan H Liu, Tammy Riklin-Raviv, Annie L Conery, Eyleen J O'rourke, Katherine L Sokolnicki, Orane Visvikis, Vebjorn Ljosa, Javier E Irazoqui, et al. An image analysis toolbox for high-throughput c. elegans assays. *Nature methods*, 9(7):714–716, 2012.

Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Density-aware chamfer distance as a comprehensive metric for point cloud completion. *arXiv preprint arXiv:2111.12702*, 2021.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Yifan Xu, Weijian Xu, David Cheung, and Zhuowen Tu. Line segment detection using transformers without edges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4257–4266, 2021.

Nan Xue, Tianfu Wu, Song Bai, Fudong Wang, Gui-Song Xia, Liangpei Zhang, and Philip HS Torr. Holistically-attracted wireframe parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2788–2797, 2020.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

A Dataset annotation and Processing

CeDeM acquisition. CeDeM consists of 752 H&E-stained duodenal biopsy images collected from unique patients. Images were acquired using an Olympus BX50 microscope at $4\times$ magnification with a DP26 camera, producing 2448×1920 pixel images. The dataset contains sparse polyline annotations for 2,617 villi and 4,211 crypts, with average lengths of 104 ± 25 and 47 ± 15 pixels, for 640×640 resized images, respectively. It also contains 872 villus shoulder and 872 crypt border polylines. The dataset is split into $612/70/70$ training, validation, and test images.

Annotation and quality control. All annotations were performed by expert pathologists and independently verified by a senior pathologist. Following standard pathology practice, annotators selected interpretable regions with well-defined villi and crypts, excluding ambiguous tissue. Structures were annotated along their principal axis using two to four control points, depending on curvature, and standardized to $K = 3$ by interpolating or subsampling points. Annotations were created using LabelMe, with disagreements resolved through consensus.

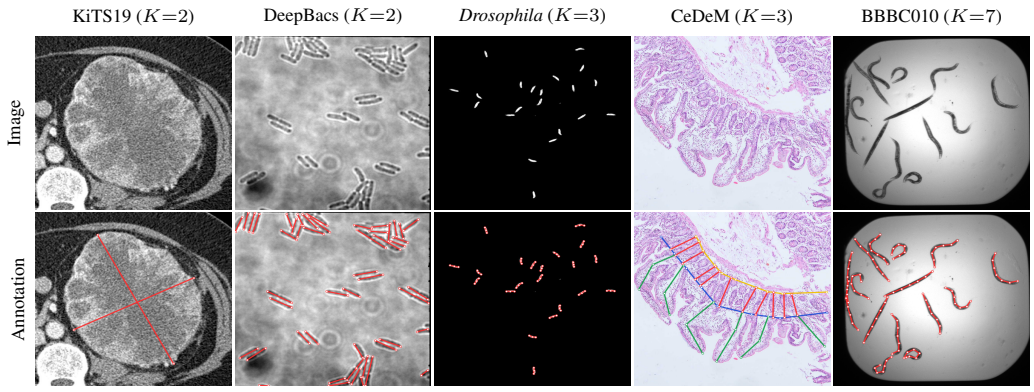

Figure 5: Image and annotation example: The first row shows the input image, and the second row shows the corresponding annotations. For CeDeM, villi (green), crypts (red), villus shoulders (blue), and crypt borders (yellow) are shown. All other datasets contain a single class, marked in red.

Public datasets. For the four public datasets, polyline annotations were derived from the segmentation masks. Images were resized to 640×640 using aspect-ratio-preserving padding. For KiTS19, we extracted the largest lesion in each annotated slice and defined a $K = 2$ polyline as its longest chord. For DeepBacs, a $K = 2$ polyline was obtained from the principal axis of each bacterial mask. For Drosophila, a $K = 3$ polyline was sampled from the medial axis at the head, midpoint, and tail. For BBBC010, masks were skeletonized, pruned to a single centerline, and resampled into a $K = 7$ polyline. For Drosophila and BBBC010, the ordered skeleton pixels are sampled at K points of equal arc length, and K is chosen by fitting $K = 2, \ldots, 8$ points to the skeletons and selecting the value with the minimum length error with respect to the full skeleton length. For each dataset, K is selected to ensure maximum curvature coverage. For straight objects, such as the bacteria in DeepBacs, $K = 2$ is sufficient, whereas complex curved structures, such as the worms in BBBC010, require $K = 7$. For CeDeM, we tried representing villi with 4 points instead of 3 but did not observe any additional improvement over 3 points. For CeDeM, in instances that required fewer than 3 points to represent, we placed a point midway to keep the number of points consistent within the class. We also tried duplicating the start or end point when a third point was not required; however, adding a middle point worked better for MeasureNet. Fig. 5 provides sample images along with their annotations.

B MeasureNet components

B.1 Detection backbone

DINO-DETR Zhang et al. (2022) is a DETR style detector with three training stability innovations relative to vanilla DETR Carion et al. (2020): (i) contrastive denoising, (ii) mixed query selection

that combines encoder positional queries with learnable content queries, and (iii) the look forward twice mechanism that propagates gradients from late layers into earlier ones. MEASURENET reuses these components and replaces the per query box head with a per query K point polyline head.

B.2 POLYLINE MATCHING COST

Prediction i is matched to ground truth j by the Hungarian algorithm (Kuhn, 1955) on the cost

$$C_{ij} = \lambda_{\text{cls}}\, C_{\text{focal}}(i,j) + \lambda_{\ell_1}\, d_{\ell_1}(P_i, P_j^*) + \lambda_{\text{ch}}\, d_{\text{ch}}(P_i, P_j^*). \tag{7}$$

With s_{ij} the predicted probability of class c_j^* for query i, the focal classification cost of DINO-DETR is

$$C_{\text{focal}}(i,j) = -\alpha\,(1-s_{ij})^{\gamma} \log s_{ij} + (1-\alpha)\, s_{ij}^{\gamma} \log(1-s_{ij}). \tag{8}$$

The ℓ_1 term compares the $2K$ coordinates in both point orders, so a polyline annotated in the opposite direction is not penalized:

$$d_{\ell_1}(P_i, P_j^*) = \min\Big(\sum_{k=1}^{K} \|p_{i,k} - p_{j,k}^*\|_1,\ \sum_{k=1}^{K} \|p_{i,k} - p_{j,K+1-k}^*\|_1\Big). \tag{9}$$

The symmetric Chamfer distance (Barrow et al., 1977) depends only on the geometry, which helps separate touching and crossing instances:

$$d_{\text{ch}}(P_i, P_j^*) = \frac{1}{K}\sum_{k=1}^{K} \min_{l} \|p_{i,k} - p_{j,l}^*\|_2 + \frac{1}{K}\sum_{l=1}^{K} \min_{k} \|p_{j,l}^* - p_{i,k}\|_2. \tag{10}$$

The matching cost (Eq. 7) uses $\lambda_{\text{cls}} = 2$ and $\lambda_{\ell_1} = 5$ ($\lambda_{\ell_1} = 7$ on KiTS19); the Chamfer term is used only on BBBC010, whose worms frequently touch and cross ($\lambda_{\text{ch}} = 1.5$), and $\lambda_{\text{ch}} = 0$ otherwise.

C IMPLEMENTATION DETAILS

Grid search was performed on the development set over learning rate $\{10^{-4}, 10^{-5}, 10^{-6}\}$ and batch size $\{2, 4, 8\}$; $10^{-4}/8$ was best in all cases. LETR and MapTR baselines use the same DINO-DETR + Swin backbone and the same augmentations as MEASURENET for fair comparison; they are trained for 150 epochs with batch size 2, learning rate 10^{-4}, query dimension 4 and 900 queries. Yolino is trained for 80 epochs with Adam at learning rate 10^{-4}, using two separate single class models on a 32×32 grid. BEiT uses the `beit-base-finetuned-ade-640-640` backbone, batch size 4, learning rate 10^{-6} and 200 epochs. For mAP, we evaluate thresholds from 30 to 130 in increments of 20 for all datasets, and from 100 to 200 for BBBC010. At inference, MEASURENET takes 90 ms per image on a single NVIDIA A100 (80 GB) GPU, making it useful for pathological use.

LETR and RANK-LETR. Both baselines are adapted on the same DINO-DETR architecture and Swin backbone as MEASURENET, with the same K point polyline head, so they predict the same number of points per instance. LETR (Xu et al., 2021) is trained with the focal classification loss and an index wise ℓ_1 loss on the ordered points, and predictions are matched using classification, ordered ℓ_1. RANK-LETR (Tong et al., 2025) adds its two re-ranking components to this recipe: (i) a pairwise ranking loss (weight 0.2) that encourages predictions with a lower polyline error to receive a higher confidence, and (ii) a geometric re-ranking head, two convolutional branches on the highest resolution encoder feature map that predict endpoint and edge probability maps, supervised with an ℓ_2 loss against maps rendered from the ground truth (weight 1 each). At inference, the confidence of each prediction is offset by $\delta_e s_e + \delta_d s_d + \delta_l \log(1 + \text{len}(P_i))$, where s_e and s_d are the endpoint and edge map responses sampled along the polyline and $(\delta_e, \delta_d, \delta_l) = (1, 1, 0.1)$

Segmentation baselines. Segmentation models are used in two different settings. In the comparison of Sec. 5.2 (Tables 1 and 2), SegFormer, BEiT, FRNet, and LWNet are trained on the same polyline annotations as MEASURENET: each instance polyline is rasterized as a 2 pixel wide line and dilated with a 3×3 kernel, the models segment this medial axis mask of each instance, and length is recovered by skeletonizing each predicted instance. In the annotation cost study of Sec. 5.1, the segmentation model (SegFormer, or Omnipose on DeepBacs and BBBC010) is instead trained on

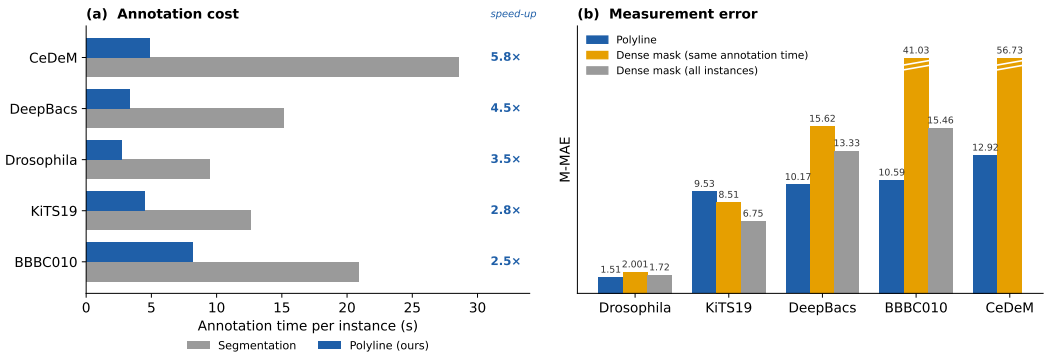

Figure 6: **Annotation cost versus measurement accuracy, all settings. (a)** Per instance annotation time, as in Fig. 3a. **(b)** Measurement error (M-MAE, lower is better) for the polyline model (MEASURENET) and dense segmentation trained under two budgets: at the same annotation time as the polylines, and on all instances. Dense segmentation uses Omnipose on DeepBacs and BBBC010 and SegFormer elsewhere. Bars above the axis limit are clipped, with the true value printed on top

dense masks of the complete objects. For the public datasets, these masks are the original instance segmentations. CeDeM has only polyline annotations, so for this study we additionally annotated dense masks of every complete villus and crypt on 120 images (110 training and 10 validation), sampled to balance the villus to crypt ratio grades; evaluation uses the polyline annotations of the CeDeM test set.

Omnipose oracle. For the annotation cost study (Sec. 5.1), the full mask oracle on DeepBacs and BBBC010 is Omnipose Cutler et al. (2022), a specialized instance segmentation model for elongated cells. On DeepBacs we fine tune the `bact_phase_omni` pretrained model for 200 epochs; on BBBC010 we trained Omnipose (`worm_omni`) is used to provided worm masks.

D ANNOTATION COST VERSUS MEASUREMENT ERROR

Table 5 reports the annotation time (mean $\pm$ standard deviation) for dense segmentation and polyline annotation across datasets. Polyline annotation is at least 2.54$\times$ faster than dense segmentation. Figure 6 extends the main-text comparison (Fig. 3) using the *same instance count* setting, where the dense segmentation model is trained on every instance, regardless of the additional annotation time required. Even under this unrestricted labeling budget, MEASURENET matches or outperforms dense segmentation on Drosophila (1.72 vs. 1.51), DeepBacs (13.33 vs. 10.17), and BBBC010 (15.46 vs. 10.59). Only on KiTS19 (6.75 vs. 9.53), a single well-defined structure whose measurement axis is a two-point chord, does the same-instance setting favor dense segmentation.

Table 5: Annotation time (seconds, mean $\pm$ standard deviation) for dense masks and polylines.

Dataset	Dense mask (s)	Polyline (s)	Speed-up
KiTS19	12.63 $\pm$1.99	4.43 $\pm$0.60	2.85$\times$
DeepBacs	15.14 $\pm$3.06	3.36 $\pm$0.92	4.51$\times$
Drosophila	9.44 $\pm$1.47	2.72 $\pm$0.66	3.47$\times$
CeDeM	28.55 $\pm$8.82	4.90 $\pm$1.16	5.83$\times$
BBBC010	20.86 $\pm$2.89	8.20 $\pm$1.23	2.54$\times$

E PER DATASET DETECTION METRICS

Table 6 reports the complete measurement and detection metrics (M-MAE, precision, recall and mAP) for the four single structure datasets summarized in the main paper (Table 1). Datasets are ordered by polyline length K; the best value per column within each dataset block is shown in **bold**

Table 6: Complete per dataset metrics for the four single structure datasets, ordered by K. M-MAE↓; precision, recall, mAP↑. Mean ± standard deviation over five seeds. Best per column within each block in **bold**; second best underlined; MEASURENET rows shaded; "–" marks runs not available. Params: number of parameters in millions; Time: inference time per image in milliseconds on one NVIDIA A100 (80 GB) GPU

Model	Params (M)	Time (ms)	M-MAE ↓	Precision ↑	Recall ↑	mAP ↑
KiTS19 (CT, K=2)						
BEiT	163.3	84.1	11.21 ±0.23	59.62 ±1.90	68.40 ±4.71	59.77 ±3.73
FRNet	32.9	9.7	14.02 ±1.08	64.61 ±0.43	61.86 ±2.35	53.67 ±2.47
LWNet	0.3	7.4	19.57 ±0.74	43.03 ±3.56	33.45 ±1.51	25.17 ±2.52
SegFormer	47.2	25.8	15.30 ±0.34	69.79 ±2.46	66.84 ±3.32	63.26 ±1.94
LETR	217.3	88.5	15.13 ±0.65	62.57 ±0.14	75.37 ±0.53	63.12 ±0.29
MapTR	217.3	88.5	12.46 ±0.36	61.12 ±1.06	79.07 ±1.69	64.41 ±1.28
RANK-LETR	217.6	89.2	10.54 ±0.07	**75.65** ±0.16	79.41 ±0.21	71.68 ±0.15
MEASURENET	217.3	88.6	**9.53** ±0.32	71.93 ±0.58	**80.53** ±0.32	**75.73** ±0.27
DeepBacs (live cell, K=2)						
BEiT	163.3	126.2	25.83 ±0.60	55.45 ±0.44	96.52 ±0.46	52.74 ±0.30
FRNet	32.9	53.0	24.77 ±0.44	53.68 ±0.25	**97.75** ±0.19	51.62 ±0.36
LWNet	0.3	55.1	27.09 ±0.29	52.56 ±0.39	96.25 ±0.14	50.43 ±0.37
SegFormer	47.2	68.6	25.69 ±0.43	51.36 ±5.94	93.30 ±4.60	51.59 ±3.79
LETR	217.5	89.1	13.82 ±0.43	62.57 ±0.34	94.39 ±0.33	61.66 ±0.61
MapTR	217.5	89.1	13.35 ±0.25	44.83 ±0.73	91.49 ±1.18	57.27 ±0.46
RANK-LETR	217.8	89.9	14.54 ±0.33	32.11 ±1.59	49.38 ±1.77	42.30 ±1.37
MEASURENET	217.5	89.2	**10.17** ±0.02	**65.82** ±0.16	93.59 ±0.35	**64.33** ±0.06
Drosophila larvae (optical coherence microscopy, K=3)						
BEiT	163.3	130.7	2.73 ±0.04	93.43 ±0.25	93.29 ±0.30	88.96 ±0.19
FRNet	32.9	58.7	3.76 ±0.15	93.46 ±0.77	92.56 ±0.55	89.37 ±0.38
LWNet	0.3	58.4	3.25 ±0.08	91.88 ±0.09	93.56 ±0.08	89.51 ±0.36
SegFormer	47.2	75.7	2.89 ±0.08	**94.06** ±0.30	92.89 ±0.54	89.47 ±0.69
LETR	217.5	89.1	2.42 ±0.15	91.14 ±1.10	99.37 ±0.29	97.44 ±0.48
MapTR	217.5	89.1	5.95 ±0.82	90.32 ±0.04	98.39 ±0.27	95.78 ±0.29
RANK-LETR	217.8	89.9	5.76 ±0.67	90.27 ±0.23	92.43 ±0.41	91.24 ±0.08
MEASURENET	217.5	89.0	**1.51** ±0.12	91.97 ±0.36	**99.39** ±0.10	**97.81** ±0.07
BBBC010 C. elegans *(bright field+fluorescence, K=7)*						
BEiT	163.3	111.9	41.78 ±0.26	91.06 ±7.43	58.22 ±1.10	56.65 ±0.46
FRNet	32.9	71.5	41.47 ±0.32	91.87 ±5.30	55.09 ±0.94	55.59 ±1.31
LWNet	0.3	42.7	43.17 ±1.68	**92.81** ±3.64	61.12 ±0.97	59.58 ±0.49
SegFormer	47.2	60.7	44.11 ±0.59	79.41 ±2.07	49.95 ±2.33	56.02 ±0.81
LETR	217.5	89.3	28.07 ±5.86	37.68 ±5.95	93.47 ±4.84	57.40 ±2.30
MapTR	217.5	89.3	23.46 ±2.44	31.46 ±5.76	98.80 ±1.13	70.67 ±2.64
RANK-LETR	217.8	90.0	29.42 ±2.21	44.37 ±0.38	94.42 ±0.40	54.39 ±0.11
MEASURENET	217.3	90.2	**10.59** ±0.18	53.95 ±0.40	**100.00** ±0.00	**82.23** ±0.04

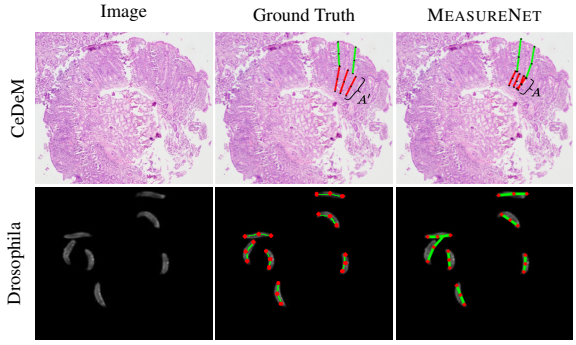

Figure 7: Failure modes of MEASURENET. Columns: input image, ground truth polyline, and MEASURENET prediction. (Top) On cluttered histopathology (CeDeM), length is underestimated for a partially visible structure. (Bottom) On Drosophila, where most larvae are near straight, MEASURENET occasionally represents strongly curved larvae with a nearly straight polyline and underestimates their arc length.

and the MEASURENET row is shaded. Every entry is the mean over five training runs that differ only

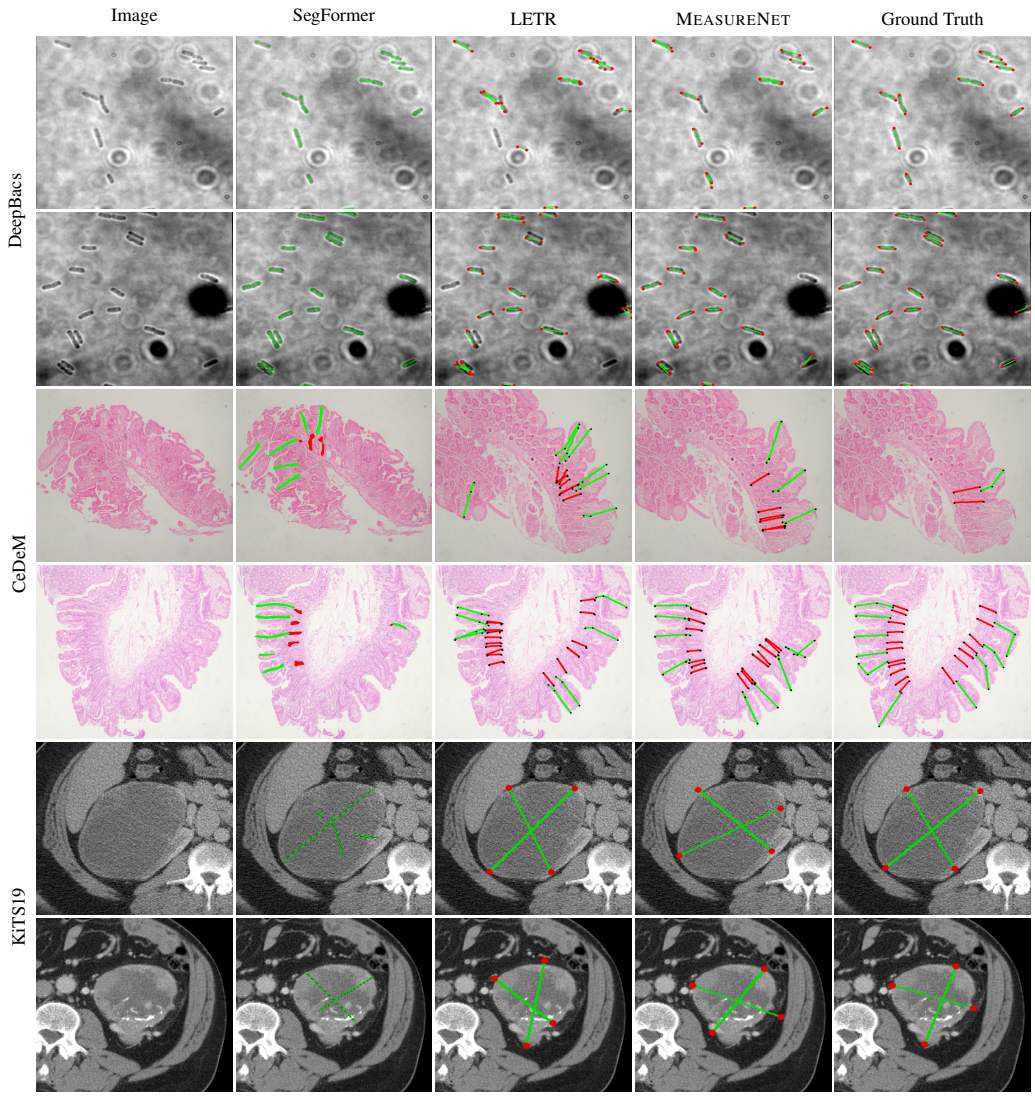

Figure 8: Additional qualitative results (part 1/2): two examples each for DeepBacs, CeDeM, and KiTS19. Columns: input image, SegFormer, LETR, MEASURENET, and ground truth. For KiTS19 the lesion region is shown, and predictions are the polylines matched to the ground truth long and short axes.

in the random seed, with the standard deviation after $\pm$. MEASURENET attains the lowest M-MAE and mAP on all datasets and on precision and recall it is competitive but not always best.

F ERROR ANALYSIS

Failures come from partially visible and highly curved structures. We observe two recurring failure modes (Fig. 7). First, on the duodenal biopsy (CeDeM) benchmark, MEASURENET underestimates length when a structure is only partially visible. Second, on datasets where most objects are well approximated by a straight line, such as Drosophila, MEASURENET sometimes misses the curvature of the few curved objects and represents them with a nearly straight polyline, underestimating their length.

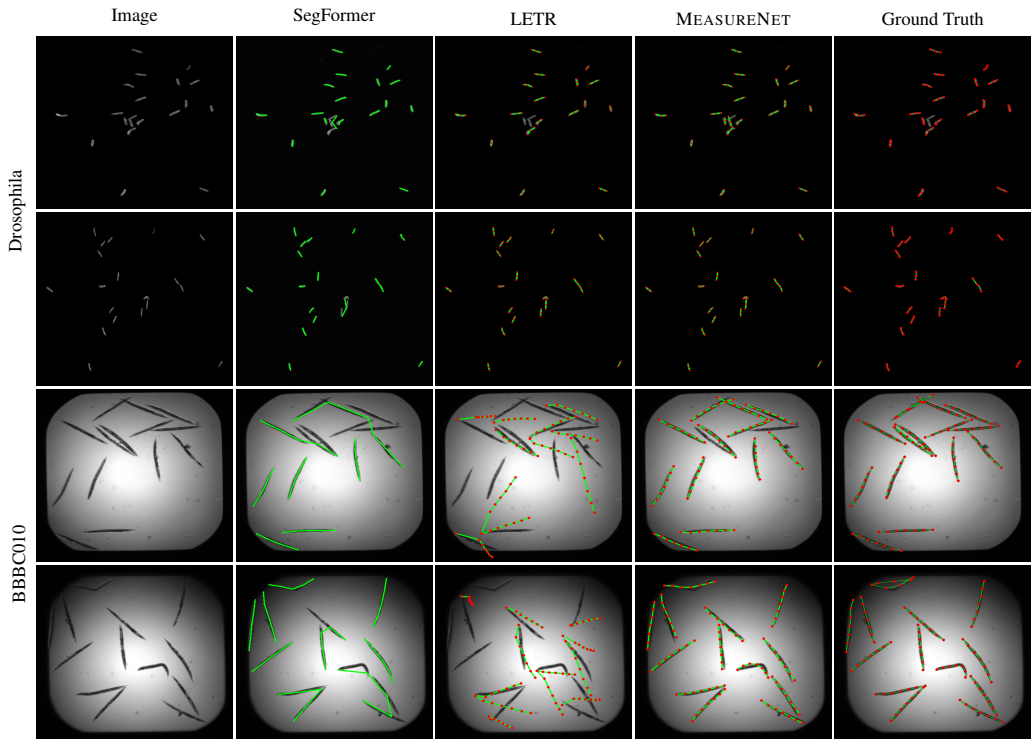

Figure 9: Additional qualitative results (part 2/2): two examples each for *Drosophila* and BBBC010. Columns: input image, SegFormer, LETR, MEASURENET, and ground truth. MEASURENET separates neighboring instances and follows object curvature, whereas SegFormer merges or fragments instances and LETR misplaces endpoints.

G ADDITIONAL QUALITATIVE RESULTS

Figures 8 and 9 provide two additional qualitative examples for each dataset (DeepBacs, CeDeM, KiTS19, *Drosophila*, and BBBC010), complementing the comparison in the main paper (Fig. 4). Across all examples, MEASURENET produces continuous, well localized polylines that separate neighboring instances and follow object curvature, whereas SegFormer tends to merge or fragment closely spaced instances and LETR frequently misplaces endpoints or collapses points.