# Data Efficiency of Segment Anything Model for Optic Disc and Cup Segmentation

**Citation for published version:**
Yii, F, MacGillivray, T & Bernabeu, MO 2024, Data Efficiency of Segment Anything Model for Optic Disc and Cup Segmentation. in J Woo, A Hering, W Silva, X Li, H Fu, X Liu, F Xing, S Purushotham, TS Mathai, P Mukherjee, M De Grauw, RB Tan, V Corbetta, E Kotter, M Reyes, CF Baumgartner, Q Li, R Leahy, B Dong, H Chen, Y Huo, J Lv, X Xu, X Li, D Mahapatra, L Cheng, C Petitjean & B Presles (eds), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops: MTSAIL 2023, LEAF 2023, AI4Treat 2023, MMMI 2023, REMIA 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8–12, 2023, Proceedings. vol. 14394, Lecture Notes in Computer Science, vol. 14394, Springer, Cham, pp. 336-346, 26th International Conference on Medical Image Computing and Computer Assisted Intervention, 2023, Vancouver, British Columbia, Canada, 8/10/23. https://doi.org/10.1007/978-3-031-47425-5_30

**Digital Object Identifier (DOI):**
10.1007/978-3-031-47425-5_30

**Link:**
Link to publication record in Edinburgh Research Explorer

**Document Version:**
Peer reviewed version

**Published In:**
Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops

# Data Efficiency of Segment Anything Model for Optic Disc and Cup Segmentation*

Fabian Yii [1,2] [0000-0002-4730-7363], Tom MacGillivray[1,2], Miguel O. Bernabeu[3,4]

[1] Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, UK
[2] Curle Ophthalmology Laboratory, The University of Edinburgh, Edinburgh, UK
[3] Centre for Medical Informatics, Usher Institute, The University of Edinburgh, Edinburgh, UK
[4] The Bayes Centre, The University of Edinburgh, UK
fabian.yii@ed.ac.uk

**Abstract.** We investigated the performance of Segment Anything Model (SAM) — the first promptable foundation model for image segmentation — for optic disc (OD) and optic cup (OC) segmentation when fine-tuned on progressively smaller number of fundus images. Three different implementations of SAM with an input prompt were considered: (1) SAM with an OD/OC-centred bounding box (*SAM GT*); (2) SAM with a noise-added (e.g. displacement, size variation) bounding box (*SAM Noise*); and (3) SAM with an automatically predicted (using Faster R-CNN) bounding box (*SAM Auto*). Two popular pre-trained semantic segmentation models, DeepLabV3 with a MobileNetV3-Large backbone and DeepLabV3 with a ResNet-50 backbone were used as baseline models. For OD segmentation, ResNet-50 exhibited comparable if not higher data efficiency (i.e. good performance despite limited training data) than even the most optimal implementation of SAM (*SAM GT*), although SAM was evidently more robust to small training set sizes, e.g. 25, than MobileNetV3-Large and in eyes with more challenging OD morphologies, e.g. significant peri-papillary atrophy. For OC segmentation, however, SAM GT and SAM Noise consistently demonstrated higher data efficiency, particularly in eyes with relatively small cup-to-disc ratio and ill-defined OC margin.

**Keywords:** Segment Anything Model, Optic Disc, Optic Cup

## 1 Background

Foundation models are changing the landscape of artificial intelligence [2]. Trained on enormous quantities of data on a scale hitherto unimaginable, these models, with such prominent examples as BERT and GPT-3, are designed to be adaptable/transferrable to a wide range of downstream tasks [2]. In computer vision, the recently introduced Segment Anything Model (SAM) has stirred interest as the first foundation model for

* Fabian Yii is supported by the Medical Research Council [grant number MR/N013166/1]. This research was conducted in part using data from UK Biobank under project ID 90655.

image segmentation [9]. SAM was trained on more than 1 billion masks and 11 million predominantly non-medical images. This, coupled with its ability to take input prompts including bounding boxes and foreground/background points, gives it tremendous zero-shot capabilities on diverse segmentation tasks involving natural images.

However, several preprints [6, 7, 15, 27] have shown that SAM's zero-shot performance is unsatisfactory when applied to medical imaging modalities including chest X-ray (lungs), MRI (hippocampus), ultrasound (breast tumour and nerve), colonoscopy (polyps) and whole slide imaging (renal tissue), although it performs well on hip X-ray and spleen CT [15]. Furthermore, SAM does not have good zero-shot performance for optic disc (OD), optic cup (OC) and retinal vessel segmentation when applied to fundus photographs [18, 19]. Despite these unfavourable results, simple fine-tuning has been shown to be highly effective in improving SAM's performance to levels comparable to, if not better than, the state-of-the-art medical image segmentation models [14, 25].

The ability to transfer SAM to the medical domain is a promising prospect, not least because its promptable nature allows for a "human in the loop", enabling greater trust to be fostered between clinicians and the system [17]. SAM's prompt engineering may also give it an extra edge over conventional pre-trained models in terms of data efficiency (i.e. robustness to limited training data), since its ability to query prior knowledge from domain experts may enable it to work with more limited or less diverse training data. To the best of our knowledge, though, no work has empirically investigated whether SAM offers any added value in terms of data efficiency in the medical domain. Here, we evaluated SAM's robustness to limited training data, with a specific focus on OD and OC segmentation due to the importance of these features for glaucoma (an eye disease associated with increased cup-to-disc ratio) assessment, not to mention that manual segmentation is labour intensive and annotated data from particular populations at risk of glaucoma, e.g. persons of African [1] and Latin American [23] descent, are extremely scarce.

Our main contributions are: (1) fine-tuning SAM's lightweight mask decoder is a highly effective transfer learning approach, as it yields competitive OD and OC segmentation performance; (2) fine-tuned SAM is *not* necessarily more data efficient than a pre-trained DeepLabV3 model with a ResNet-50 backbone insofar as OD segmentation is concerned, but it is evidently more data efficient when it comes to OC segmentation; and (3) SAM is more robust to challenging OD or OC morphologies.

## 2 Methods

### 2.1 Datasets

374 macula-centred colour fundus photographs (2048×1536 pixels) from the UK Biobank (UKB) were randomly sampled from a total of 117,175 participants (predominantly White British) who took part in a baseline ophthalmic assessment [5]. An annotator (F.Y) manually segmented the OD in each image using the Image Segmenter app in MATLAB, which allowed regions of interest to be drawn interactively using waypoints. OD margin was defined as the inner margin of the

Elschnig's scleral ring per the conventional ophthalmoscopic definition. A random subset (N=50) of images were re-annotated at least 1 week after the first annotation to assess the intra-rater agreement (Dice score: 96.4%). OC segmentation was not carried out due to time constraints.

Two other publicly available datasets, namely DRISHTI-GS [20] and PAPILA [10], were also used in this work. DRISHTI-GS contains 101 OD-centred colour fundus photographs (2896 × 1944 pixels) taken from patients at Aravind Eye Hospital in India. Most eyes (N=70) in this dataset have glaucoma. PAPILA, on the other hand, consists of 488 OD-centred colour fundus photographs (2576 × 1934 pixels) taken from predominantly non-glaucomatous patients (N=333) seen at an ophthalmology department in Spain. Ground-truth OD and OC masks were available from both datasets.

## 2.2 Experiments

UK Biobank images were first cropped to 1400 × 1400 pixels to remove the black border around each image, before being resized to 560 × 560 pixels. Images from DRISHTI-GS and PAPILA were resized to 523 × 613 pixels and 580 × 772 pixels, respectively. UK Biobank images were split into 299 training images and 75 test images, whereas 50 and 51 DRISHTI-GS images were used for training and testing per the original train/test split prescribed by the data provider [20]. The first 300 PAPILA images were used as training images, leaving 188 images for testing. In each experiment, different implementations of SAM and two baseline models were fine-tuned and tested using a pre-defined number of training images from one of the 3 datasets above. Training set size was progressively reduced after each experiment, i.e. each time choosing the first N images in the training set, down to a minimum of 25.

### Implementations

We considered three different implementations of SAM at inference:

(1) **SAM GT**: SAM with an OD/OC-centred, tight-fitting bounding box manually given as input prompt to simulate the most ideal (noise-free) user scenario.

(2) **SAM Noise**: SAM with noise added to the ground truth bounding box. Noise refers to random addition of up to 10 pixels to each point coordinate of the box, resulting in displacement, size variation and orientation variation.

(3) **SAM Auto**: SAM with an automatically predicted bounding box given as input prompt. The bounding box was predicted using Faster R-CNN (with a ResNet-50-FPN backbone) [11] pre-trained on the COCO dataset [12]. In each experiment, Faster R-CNN was fine-tuned for 10 epochs using a batch size of 10. ADAM optimizer (initial learning rate: 5e-4; weight decay: 1e-2) with a cosine annealing scheduler was used. Data augmentation including random changes of brightness/saturation, horizontal flip and rotation up to ±60° was applied on the fly. The model at the last epoch was used at inference. Note that similar training set size was used to fine-tune Faster R-CCN and SAM in each experiment.

Two popular semantic segmentation models, DeepLabV3 [3] with a MobileNetV3-Large backbone and DeepLabV3 with a ResNet-50 backbone (hereafter known as MobileNet and ResNet-50), both pre-trained on the COCO dataset [12], were used as baseline models. The official PyTorch implementations of these models were used in this work. All experiments were conducted using a NVIDIA RTX A5000 24GB GPU.

### Training Details

The mask decoder of SAM with the smallest (91M parameters) backbone size (ViT-B image encoder) was fine-tuned due to its lightweight nature. Empirical evidence also indicates that the smallest model is more computationally efficient, as the largest backbone (ViT-H) is slowest at inference and does not have significantly better performance [8]. Noise-added bounding boxes were used as input prompts during training to improve SAM's robustness at inference. Unless otherwise stated, the following training details and hyperparameter settings were consistently applied. Five-fold cross validation was used for hyperparameter tuning and model selection, i.e. model with the lowest validation loss used at inference. Each model was trained for 10 epochs (for each fold) using a batch size of 15, except when fine-tuning ResNet-50 on PAPILA where a batch size of 12 was used (maximum possible size permitted by the GPU memory). ADAM optimizer (initial learning rate: 5e-4; weight decay: 1e-4) was used with a cosine annealing scheduler. Binary cross-entropy loss was used as the loss function, weighting the foreground (OD or OC) pixels by the median ratio of background to foreground pixels due to significant class imbalance. Data augmentation including random changes of brightness/saturation, horizontal flip and rotation up to ±60° was applied on the fly.

### Evaluation Metrics

Average Precision (AP), commonly used to summarise a precision-recall curve, was adopted as the primary performance metric because it is not contingent upon an arbitrary binary threshold:

$$AP = \sum_{n}(R_n - R_{n-1})P_n \tag{1}$$

where $R_n$ and $P_n$ denotes recall and precision at the n$^{th}$ threshold. We also computed the Dice score to facilitate comparison with other studies, using 0.95 as the binary threshold (based on the observation that this gave the most clinically accurate binary masks on average):

$$Dice = \frac{2\ x\ TP}{(TP+FP)+(TP+FN)} \tag{2}$$

where TP, FP and FN denote true positives, false positives and false negatives. Unless otherwise stated, all results were summarised as mean ± standard deviation.

# 3 Results

## 3.1 OD Segmentation

As shown in Table 1 and Figure 1, ResNet-50 almost invariably demonstrated higher data efficiency than even the most optimal implementation of SAM, i.e. SAM GT. For example, while the AP achieved by ResNet-50 remained largely unchanged on the UKB test set (around 90%) when the training set size was reduced from 299 to 50, an appreciable reduction in SAM GT's AP, i.e. from 90% to 86%, was observed with a similar reduction in training set size. The AP achieved by ResNet-50 was also noticeably higher than SAM GT, e.g. 93.0% vs 90.9% when the training set size was equal to 25, on the DRISHTI-GS test set. That said, a qualitative assessment of the predicted masks (some examples shown in the top panel of Figure 2) suggested that SAM was more robust to challenging (also more unusual) OD morphologies, such as when the margin was ill-defined due to significant peripapillary atrophy, which could be seen as irregular pigmentation/brightness in the area adjacent to OD in the presence of high myopia (second image of the UKB image pair in Figure 2). However, this was contingent upon good bounding box placement, a point best illustrated by the fact that the reduction in SAM Auto's performance when the training set size was decreased (irrespective of dataset) could be attributed to a drastic drop in Faster R-CNN's performance when there was limited training data (Supplementary S1)

## 3.2 OC Segmentation

SAM GT/Noise consistently yielded superior performance on the DRISHTI-GS test set, with the gain being most evident when the training set size was reduced to 25 (Table 2 and Figure 2). On the PAPILA test set, SAM (irrespective of implementation type) demonstrated disproportionately better performance than the baseline models. Of note, significant discrepancies in model performance were observed between datasets. On the DRISHTI-GS test set, for instance, where 50 training images would suffice to yield a high AP of 88.1% using SAM GT, a similar level of performance on the PAPILA test set was far from evident even with 300 training images (AP equal to 72.8%). In keeping with this, Moris et al. [16] also observed that their best model yielded significantly poorer performance on the PAPILA test set than DRISHTI-GS (more than 20% difference in median Dice score) and other test sets (Figure 2 in their paper). These discrepancies are attributable to the fact that a higher proportion of eyes in the DRISHTI-GS dataset have a large cup-to-disc ratio (large OC relative to OD) due to glaucoma, giving rise to a "deeper" cup appearance and more defined OC margin. A qualitative assessment of the predicted masks (some examples shown in the bottom panel of Figure 1) also revealed that small cup-to-disc ratio — commonly found in PAPILA — was generally detrimental to segmentation quality, although SAM (irrespective of implementation type) was more robust to such cases (i.e. PAPILA images) than the baseline models.

**Table 1.** Optic disc segmentation: mean test performance (standard deviation in parentheses) of different implementations of SAM and two DeepLabV3 baseline models (one with a MobileNet backbone, B1, and another with a ResNet-50 backbone, B2) across training set sizes (N). Best test performance is highlighted in maroon and bold (note that in the printed version this is only highlighted in bold).

| Dataset | N | Average Precision (%) | | | | | Dice (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **SAM GT** | **SAM noise** | **SAM Auto** | **B1** | **B2** | **SAM GT** | **SAM noise** | **SAM Auto** | **B1** | **B2** |
| UKB | 299 | 90.2<br>(3.5) | 82.3<br>(6.0) | 88.1<br>(6.1) | **91.1**<br>**(4.1)** | 90.0<br>(4.4) | 94.8<br>(1.9) | 90.4<br>(3.5) | 93.6<br>(3.7) | **95.3**<br>**(2.3)** | 94.7<br>(2.5) |
| | 200 | 89.3<br>(3.8) | 82.1<br>(6.8) | 87.3<br>(5.5) | 89.4<br>(4.6) | **90.7**<br>**(4.2)** | 94.3<br>(2.1) | 90.3<br>(3.9) | 93.2<br>(3.2) | 94.4<br>(2.6) | **95.1**<br>**(2.3)** |
| | 100 | 88.2<br>(4.3) | 80.2<br>(6.7) | 83.9<br>(11.8) | 85.1<br>(7.5) | **90.0**<br>**(4.4)** | 93.7<br>(2.4) | 89.1<br>(4.1) | 90.6<br>(11.4) | 91.8<br>(4.7) | **94.7**<br>**(2.5)** |
| | 50 | 86.2<br>(4.7) | 80.3<br>(5.9) | 84.5<br>(7.7) | 78.4<br>(15.6) | **89.7**<br>**(6.1)** | 92.6<br>(2.7) | 89.2<br>(3.6) | 91.5<br>(4.8) | 86.6<br>(15.6) | **94.5**<br>**(3.6)** |
| | 25 | 82.5<br>(4.6) | 77.1<br>(7.2) | 79.3<br>(7.4) | 68.2<br>(14.9) | **86.7**<br>**(7.7)** | 90.6<br>(2.7) | 87.3<br>(4.3) | 88.5<br>(4.6) | 79.9<br>(15.0) | **92.8**<br>**(4.7)** |
| DRISHTI-GS | 50 | 92.3<br>(3.3) | 91.3<br>(3.5) | 91.4<br>(4.3) | 89.0<br>(7.0) | **95.3**<br>**(2.4)** | 95.9<br>(1.9) | 95.3<br>(2.0) | 95.4<br>(2.5) | 93.9<br>(4.4) | **97.5**<br>**(1.3)** |
| | 25 | 90.9<br>(9.1) | 88.7<br>(9.3) | 88.3<br>(14.7) | 87.0<br>(5.8) | **93.0**<br>**(3.3)** | 94.8<br>(7.5) | 93.5<br>(7.8) | 92.5<br>(14.6) | 92.9<br>(3.4) | **96.3**<br>**(1.8)** |
| PAPILA | 300 | 93.2<br>(2.4) | 89.9<br>(3.4) | 89.8<br>(3.3) | 92.3<br>(3.6) | **93.1**<br>**(3.5)** | **96.4**<br>**(1.3)** | 94.6<br>(1.9) | 94.6<br>(1.8) | 95.9<br>(2.0) | **96.4**<br>**(1.9)** |
| | 150 | 92.2<br>(2.6) | 89.1<br>(3.4) | 88.8<br>(3.6) | 90.6<br>(4.6) | **93.1**<br>**(3.2)** | 95.8<br>(1.4) | 94.2<br>(1.9) | 94.0<br>(2.0) | 95.0<br>(2.6) | **96.3**<br>**(1.8)** |
| | 50 | **91.8**<br>**(3.1)** | 87.9<br>(4.2) | 87.7<br>(4.2) | 85.0<br>(11.3) | 90.5<br>(4.7) | **95.7**<br>**(1.7)** | 93.5<br>(2.4) | 93.4<br>(2.4) | 91.3<br>(10.5) | 94.9<br>(2.6) |
| | 25 | **90.9**<br>**(3.3)** | 88.0<br>(3.8) | 88.1<br>(3.9) | 76.4<br>(10.3) | 89.9<br>(5.2) | **95.2**<br>**(1.8)** | 93.6<br>(2.1) | 93.6<br>(2.2) | 86.2<br>(7.8) | 94.6<br>(3.0) |

**Table 2.** OC segmentation: mean test performance (standard deviation in parentheses) of different implementations of SAM and two DeepLabV3 baseline models (one with a MobileNet backbone, B1, and another with a ResNet-50 backbone, B2) across training set sizes (N). Best test performance is highlighted in maroon and bold (note that in the printed version this is only highlighted in bold).

| Dataset | N | Average Precision (%) | | | | | Dice (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **SAM GT** | **SAM noise** | **SAM Auto** | **B1** | **B2** | **SAM GT** | **SAM noise** | **SAM Auto** | **B1** | **B2** |
| PAPILA | 300 | **72.8**<br>**(10.7)** | 60.4<br>(18.3) | 61.4<br>(17.9) | 56.2<br>(21.1) | 56.2<br>(22.1) | **84.0**<br>**(7.3)** | 74.7<br>(16.4) | 75.7<br>(15.1) | 70.3<br>(19.0) | 69.9<br>(20.1) |
| | 150 | **72.8**<br>**(13.1)** | 58.0<br>(20.0) | 58.7<br>(20.2) | 51.0<br>(24.2) | 54.1<br>(23.8) | **83.5**<br>**(12.5)** | 72.5<br>(19.1) | 72.6<br>(20.7) | 64.9<br>(23.7) | 67.8<br>(22.3) |
| | 50 | **69.7**<br>**(13.3)** | 57.4<br>(19.7) | 57.3<br>(19.3) | 31.4<br>(20.5) | 40.1<br>(26.1) | **81.5**<br>**(12.0)** | 71.7<br>(19.7) | 72.1<br>(18.7) | 44.8<br>(23.3) | 52.6<br>(28.8) |
| | 25 | **60.7**<br>**(17.2)** | 51.9<br>(21.0) | 52.9<br>(20.6) | 22.2<br>(16.6) | 29.6<br>(20.2) | **74.0**<br>**(17.2)** | 66.6<br>(21.9) | 67.5<br>(21.6) | 33.9<br>(20.0) | 42.2<br>(23.7) |
| DRISHTI-GS | 50 | **88.1**<br>**(5.7)** | 82.4<br>(9.1) | 78.2<br>(16.1) | 75.9<br>(15.4) | 82.5<br>(12.1) | **93.6**<br>**(3.4)** | 90.3<br>(5.9) | 86.7<br>(12.2) | 85.5<br>(11.1) | 90.1<br>(7.7) |
| | 25 | **84.4**<br>**(9.7)** | 82.0<br>(10.1) | 77.1<br>(16.8) | 72.3<br>(17.0) | 79.3<br>(15.3) | **91.2**<br>**(6.4)** | 89.9<br>(6.4) | 85.9<br>(12.9) | 82.7<br>(13.1) | 87.5<br>(11.5) |

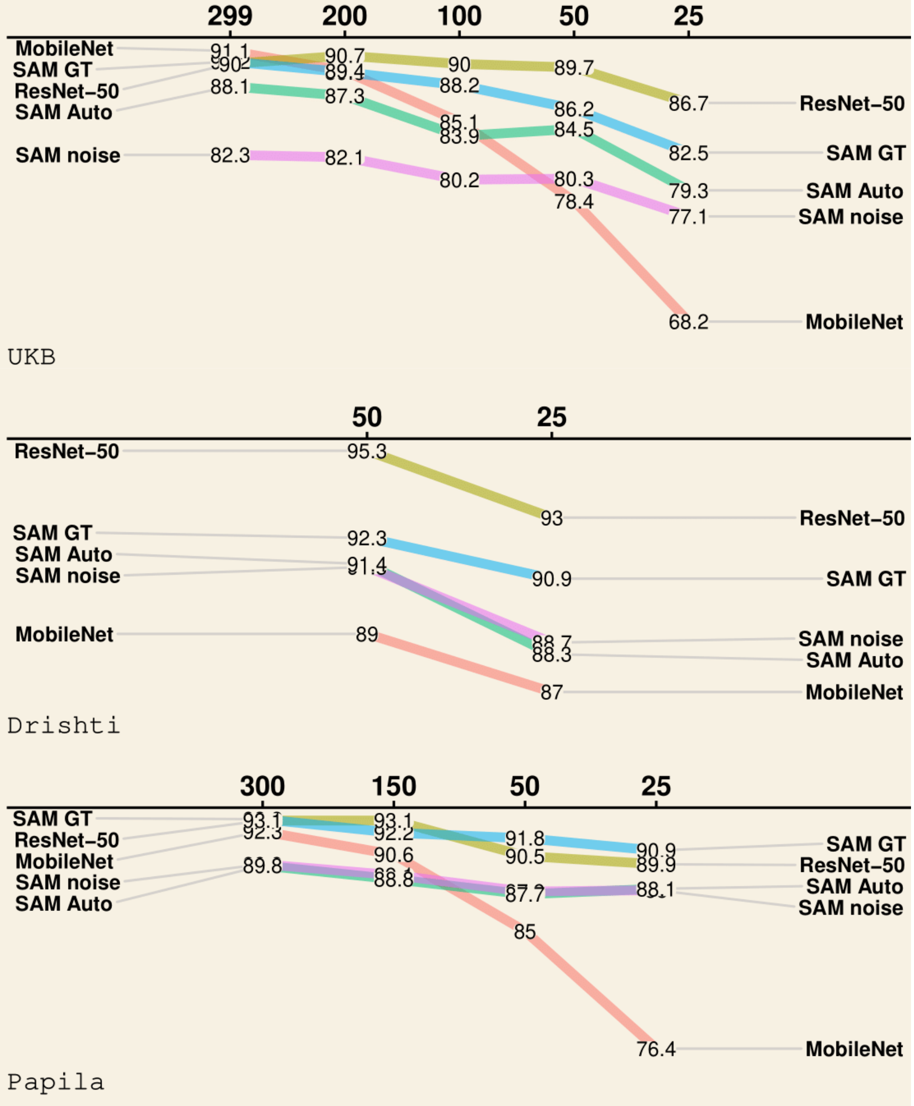

**Fig. 1.** Optic disc segmentation: mean average precision (%) by training set size and dataset.

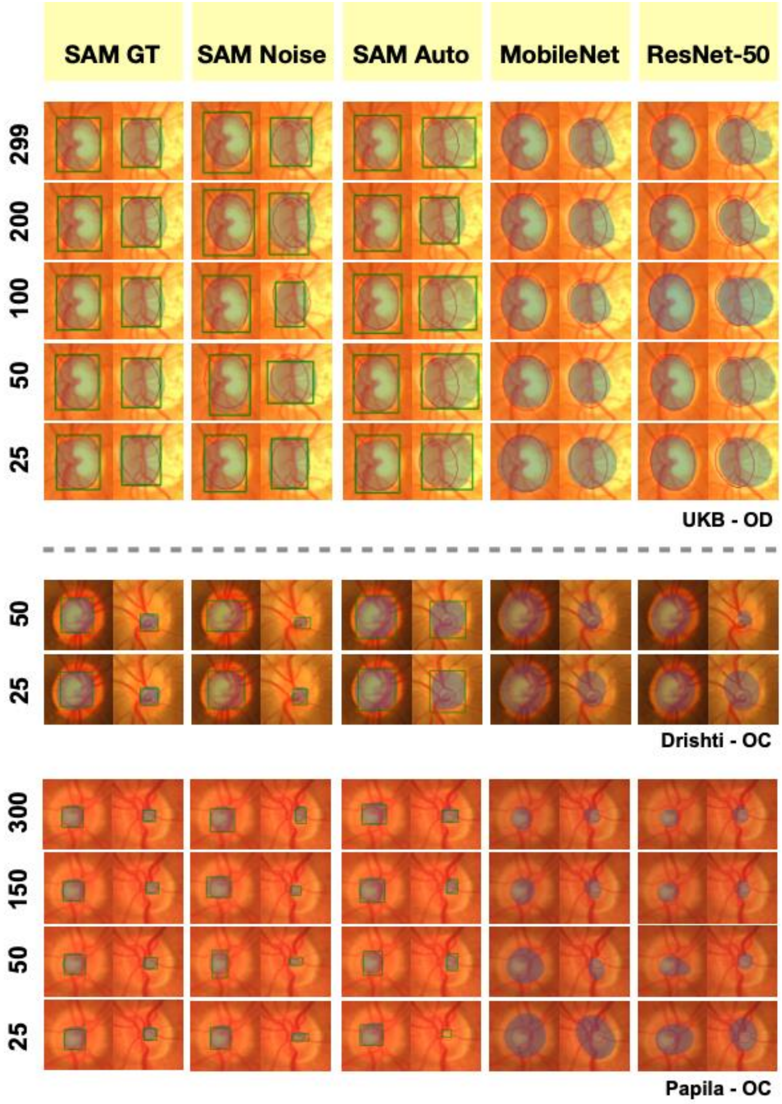

**Fig. 2.** Top panel (above dotted line) displays some examples of OD segmentation masks (UK Biobank) predicted by different models. The first image in each image pair is more typical, while the second image with significant peri-papillary atrophy (bright area adjacent to OD margin) is more unusual. Bottom panel displays some examples of OC segmentation masks (DRISHTI-GS and PAPILA) predicted by different models. The first image in each DRISHTI-GS image pair with a large cup-to-disc ratio ("deeper" cup and more well-defined margin) is more typical of DRISHTI-GS due to more eyes being glaucomatous. Conversely, most eyes in PAPILA have a normal (small) cup-to-disc ratio, and thus more ill-defined OC margin, as represented by the second image in the PAPILA image pair. Predicted masks are represented by blue overlays; ground truth (OD or OC) margin is circled in red; bounding box (predicted or otherwise) is displayed as appropriate.

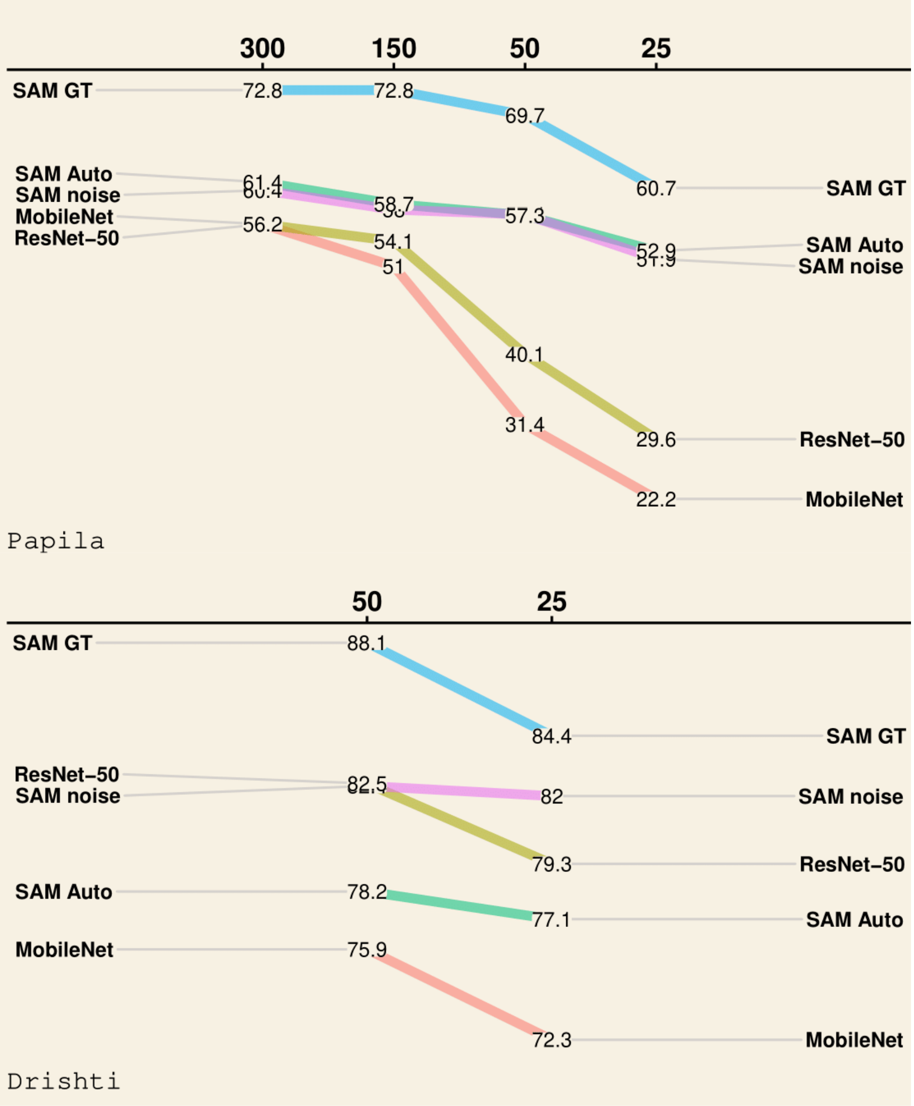

**Fig. 3.** Optic cup segmentation: mean average precision (%) by training set size and dataset.

## 4 Discussion and conclusions

In this work, we demonstrated that fine-tuning SAM's lightweight mask decoder was a simple yet highly effective transfer learning approach. On the DRISHTI-GS test set, for instance, we improved SAM's out-of-the-box Dice score from 55.6% (OD) and 57.1% (OC) [19] to 95.3-95.9% and 86.7-93.6% (depending on the implementation type) after fine-tuning it on just 50 images.

However, insofar as data efficiency for OD segmentation is concerned, SAM offers little added value because an existing model like DeepLabV3 (with a ResNet-50 backbone), coupled with appropriate data augmentation and transfer learning techniques, is already optimised and robust to limited training data. To illustrate, the Dice score achieved by our ResNet-50 model on the benchmark DRISHTI-GS test set using only 50 (97.5%) and 25 (96.3%) training images is comparable to the current state-of-the-art result, i.e. 97.8% [13, 24], and to studies working with significantly more training data. Yu et al. [26], for example, reported a Dice score of 97.4% on the DRISHTI-GS test set by fine-tuning a U-Net pre-trained on more than 600 fundus images (designated for OD segmentation task). Similarly, Sun et al. [21] achieved 97.0% on the same test set using a fine-tuned U-Net pre-trained on 400 fundus images. Moreover, using only 25 images from the PAPILA train set, our ResNet-50 yielded a similar Dice score (94.6%) to that recently reported by Moris et al. (94.5%) [16], who used a combined total of 1169 training images from different datasets. Having said that, our qualitative assessment of the predicted masks indicated that SAM's ability to tap into domain-specific prior knowledge (mask prediction from bounding box), rendered it more robust to images with challenging OD morphologies (ill-defined margin).

For OC segmentation, though, SAM (irrespective of implementation type) is evidently more data efficient than the baseline models across datasets and training set sizes. In addition, SAM GT and SAM Noise (fine-tuned on 50 DRISHTI-GS images) yielded superior/comparable Dice score (93.6% and 90.3%) to the state of the art (92.4%) reported by a study that used a larger (internal) training set (70/30 train/test split) [22] and studies that tapped into more (external) training data, e.g. Sun et al. (87.7%) [21] and Yu et al. (88.8%) [26]. The retinal ganglion cell axons exit the eye via the OD and *descend* (as seen en face) into the central pale/whitish excavation that forms the OC [4]. On a colour fundus photograph, therefore, the contrast between OC and its surrounding tissue is intrinsically more attenuated (as some depth information is inevitably lost because of 2D projection) compared to that between OD and its surrounding tissue [4]. As such, the uncertainty in ground truth is invariably higher for OC than that for OD segmentation. For example, in the PAPILA dataset, the intra-observer agreement ranges from 82.7% to 83.4% (Dice score) for OC segmentation, similar to what our fine-tuned SAM GT achieved, compared to 95.5 to 95.8% for OD segmentation. The fact that SAM offers a clear added value when used for OC but not OD segmentation suggests that it only has an extra edge when the boundary of the region of interest is not well defined (also tying in with the observation that SAM is more robust to challenging OD morphologies). Indeed, as discussed earlier, SAM's added value is most evident when applied to PAPILA — where most eyes have a normal (small) cup-to-disc ratio, and therefore (naturally) ill-defined OC margin — which stands in contrast to DRISHTI-GS in which the majority of eyes have a "deeper" and more well-defined OC appearance due to glaucomatous retinal ganglion cell loss.

Taken together, SAM does not offer higher data efficiency when it comes to OD segmentation, although with good bounding box placement it is more robust to challenging (and more unusual) OD morphologies. When it comes to OC segmentation, though, SAM is evidently more data efficient. SAM holds potential as a *clinician-in-*

*the-loop* tool to facilitate accurate OD and OC segmentation in the presence of ill-defined boundaries without necessarily requiring huge amounts of training data.

## References

1. Boland, M.V., Quigley, H.A.: Risk factors and open-angle glaucoma: classification and application. J Glaucoma 16, 406-418 (2007)
2. Bommasani, R., et al.: On the Opportunities and Risks of Foundation Models. arXiv e-prints arXiv 2108.07258 (2021)
3. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv e-prints arXiv:1706.05587 (2017)
4. Cheng, J., Li, Z., Gu, Z., Fu, H., Wong, D.W.K., Liu, J.: Chapter 11 - Structure-preserving guided retinal image filtering for optic disc analysis. In: Trucco, E., MacGillivray, T., Xu, Y. (eds.) Computational Retinal Image Analysis, pp. 199-221. Academic Press (2019)
5. Chua, S.Y.L., et al.: Cohort profile: design and methods in the eye and vision consortium of UK Biobank. BMJ Open 9, e025077 (2019)
6. Deng, R., et al.: Segment Anything Model (SAM) for Digital Pathology: Assess Zero-shot Segmentation on Whole Slide Imaging. arXiv e-prints arXiv 2304.04155 (2023)
7. He, S., et al.: Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets. arXiv e-prints arXiv 2304.09324 (2023)
8. Huang, Y., et al.: Segment Anything Model for Medical Images? arXiv e-prints arXiv:2304.14660 (2023)
9. Kirillov, A., et al.: Segment Anything. arXiv e-prints arXiv 2304.02643 (2023)
10. Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., Sancho-Gómez, J.L.: PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. Sci Data 9, 291 (2022)
11. Li, Y., Xie, S., Chen, X., Dollar, P., He, K., Girshick, R.: Benchmarking Detection Transfer Learning with Vision Transformers. arXiv e-prints arXiv 2111.11429 (2021)
12. Lin, T.-Y., et al.: Microsoft COCO: Common Objects in Context. arXiv e-prints arXiv 1405.0312 (2014)
13. Liu, B., Pan, D., Song, H.: Joint optic disc and cup segmentation based on densely connected depthwise separable convolution deep network. BMC Med Imaging 21, 14 (2021)
14. Ma, J., Wang, B.: Segment Anything in Medical Images. arXiv e-prints arXiv 2304.12306 (2023)
15. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment Anything Model for Medical Image Analysis: an Experimental Study. arXiv e-prints arXiv 2304.10517 (2023)
16. Moris, E., et al.: Assessing coarse-to-fine deep learning models for optic disc and cup segmentation in fundus images. pp. 125670R. eprint: arXiv:2209.14383 (Year)
17. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, Á.: Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review 56, 3005-3054 (2023)
18. Qiu, Z., Hu, Y., Li, H., Liu, J.: Learnable Ophthalmology SAM. arXiv e-prints arXiv 2304.13425 (2023)
19. Shi, P., Qiu, J., Dalike Abaxi, S.M., Wei, H., Lo, F.P.W., Yuan, W.: Generalist Vision Foundation Models for Medical Imaging: A Case Study of Segment Anything Model on Zero-Shot Medical Segmentation. arXiv e-prints arXiv 2304.12637 (2023)
20. Sivaswamy, J., R., K.S., Gopal, D.J., Madhulika, J., Ujjwaft, S.T.A.: Drishti-GS: Retinal image dataset for optic nerve head(ONH) segmentation. pp. 53-56 (2014)

21. Sun, J.D., Yao, C., Liu, J., Liu, W., Yu, Z.K.: GNAS-U2Net: A New Optic Cup and Optic Disc Segmentation Architecture With Genetic Neural Architecture Search. IEEE Signal Processing Letters 29, 697-701 (2022)

22. Tabassum, M., et al.: CDED-Net: Joint Segmentation of Optic Disc and Optic Cup for Glaucoma Screening. IEEE Access 8, 102733-102747 (2020)

23. Vajaranant, T.S., Wu, S., Torres, M., Varma, R.: The changing face of primary open-angle glaucoma in the United States: demographic and geographic changes from 2011 to 2050. Am J Ophthalmol 154, 303-314.e303 (2012)

24. Wei, Z., et al.: RMSDSC-Net: A robust multiscale feature extraction with depthwise separable convolution network for optic disc and cup segmentation. International Journal of Intelligent Systems 37, 11482-11505 (2022)

25. Wu, J., et al.: Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. arXiv e-prints arXiv 2304.12620 (2023)

26. Yu, S., Xiao, D., Frost, S., Kanagasingam, Y.: Robust optic disc and cup segmentation with deep learning for glaucoma detection. Comput Med Imaging Graph 74, 61-71 (2019)

27. Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., Gong, C.: Can SAM Segment Polyps? arXiv e-prints arXiv 2304.07583 (2023)