Machine Vision and Applications (2022) 33:39
https://doi.org/10.1007/s00138-022-01276-z

ORIGINAL PAPER

Cross-validation of a semantic segmentation network for natural history collection specimens

Abraham Nieva de la Hidalga[1] · Paul L. Rosin[2] · Xianfang Sun[2] · Laurence Livermore[3] · James Durrant[4] · James Turner[5] · Mathias Dillen[6] · Alicia Musson[7] · Sarah Phillips[7] · Quentin Groom[6] · Alex Hardisty[2]

Received: 14 August 2020 / Revised: 28 October 2021 / Accepted: 25 December 2021 / Published online: 21 March 2022

Abstract

Semantic segmentation has been proposed as a tool to accelerate the processing of natural history collection images. However, developing a flexible and resilient segmentation network requires an approach for adaptation which allows processing different datasets with minimal training and validation. This paper presents a cross-validation approach designed to determine whether a semantic segmentation network possesses the flexibility required for application across different collections and institutions. Consequently, the specific objectives of cross-validating the semantic segmentation network are to (a) evaluate the effectiveness of the network for segmenting image sets derived from collections different from the one in which the network was initially trained on; and (b) test the adaptability of the segmentation network for use in other types of collections. The resilience to data variations from different institutions and the portability of the network across different types of collections are required to confirm its general applicability. The proposed validation method is tested on the Natural History Museum semantic segmentation network, designed to process entomological microscope slides. The proposed semantic segmentation network is evaluated through a series of cross-validation experiments designed to test using data from two types of collections: microscope slides (from three institutions) and herbarium sheets (from seven institutions). The main contribution of this work is the method, software and ground truth sets created for this cross-validation as they can be reused in testing similar segmentation proposals in the context of digitization of natural history collections. The cross-validation of segmentation methods should be a required step in the integration of such methods into image processing workflows for natural history collections.

Keywords Semantic segmentation · Cross-validation · Digitization workflow · Data extraction · Natural history collection specimens

✉ Abraham Nieva de la Hidalga
nievadelahidalgaa@cardiff.ac.uk

Paul L. Rosin
RosinPL@cardiff.ac.uk

Xianfang Sun
SunX2@cardiff.ac.uk

Laurence Livermore
l.livermore@nhm.ac.uk

James Durrant
jdurrant14@gmail.com

James Turner
James.Turner@museumwales.ac.uk

Mathias Dillen
mathias.dillen@plantentuinmeise.be

Alicia Musson
A.Musson@kew.org

Sarah Phillips
sarah.phillips@kew.org

Quentin Groom
quentin.groom@plantentuinmeise.be

Alex Hardisty
HardistyAR@cardiff.ac.uk

[1] School of Chemistry, Cardiff University, Cardiff, UK

[2] School of Computer Science and Informatics, Cardiff University, Cardiff, UK

[3] Natural History Museum, London, UK

[4] London, UK

[5] National Museum Wales, Cardiff, UK

[6] Meise Botanic Garden, Meise, Belgium

[7] Royal Botanic Gardens Kew, Richmond, UK

1 Introduction

The need to increase global accessibility to natural history collection specimens and to reduce handling and deterioration of valuable, and often fragile, physical samples has spurred the evolution of advanced digitization practices. Early online databases recording specimens' catalog data have morphed into modern online portals which allow browsing digital specimens including taxonomic data, location, specimen-specific traits and images, along with other types of media (videos, audio recordings, links to related specimens and scientific publications). Collections consisting of millions of diverse specimens facilitated the emergence of high-throughput digitization workflows which also prompted research into novel acquisition methods, image standardization, curation, preservation and publishing. In some areas, this has promoted the creation of successful processing workflows capable of processing high volumes of specimens. However, the advance has not been extended to all areas, resulting in various activities of the digitization workflows which still rely on manual processes, and therefore throttle the speed with which the images can be processed and published. Image quality control and information extraction from specimen labels are among the digitization workflow activities which can benefit from greater automation. In this context, semantic segmentation methods can support the automation and improvement of image quality management and information extraction from images of physical specimens [25]. The processing speed of human operators and the high cost of hiring and training personnel for these activities directly affect the throughput of the whole workflow, which in turn prevents the speedy publishing of digitization results.

The use of artificial intelligence has been proposed to speed up some processing steps of the workflows after image acquisition [13, 29, 32, 35, 36]. The adoption of these methods requires being able to determine whether they are fit for purpose which means that the method is flexible and resilient requiring minimal training and validation for processing different datasets from different collections and institution.

This paper reports on a study that aimed to (a) validate the effectiveness of a semantic segmentation method for use in image sets from collections different from the one in which the model was initially trained on; and (b) validate the adaptability of the segmentation method for use in other types of collections. The purpose of these experiments is to determine whether the segmentation model is robust enough for creating a segmentation service which can then be incorporated into automated workflows supporting to speed up image processing/curation. The resilience to variations in data from different institutions and the portability of the semantic segmentation models across different types of collections are required to assure the reliability of the model in operating conditions.

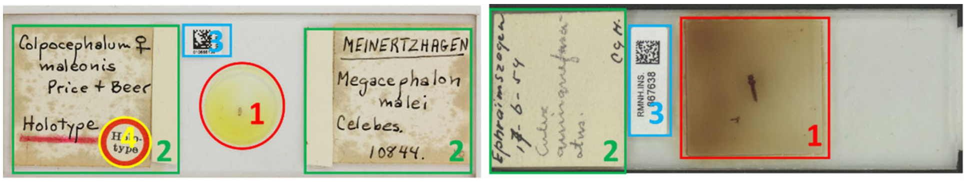
The target segmentation method is the Natural History Museum, London (NHM), semantic segmentation network created for the segmentation of entomological microscope slide images [10]. The NHM semantic segmentation network (NHM-SSN) has been openly published and used for processing NHM entomology slides. However, its application to collections from other institutions or other types of collections had not been addressed. This paper presents a group of cross-validation experiments designed to test the applicability of NHM-SSN using data from different institutions and two types of collections: microscope slides (from three institutions) and herbarium sheets (from seven institutions). The paper is structured as follows: Section two describes the digitization of microscope slides and herbarium sheets from the perspective of the main features of the images produced in each case and the requirements for further processing that may be improved with segmentation. Section three describes the NHM semantic segmentation network, its development and architecture. Section four describes the design of the cross-validation experiments, including the details of the collections used. Section five presents the results and analysis from the cross-validation experiments. Section six analyzes the results, providing insights into the strengths and weaknesses of the segmentation method. Finally, section seven provides a conclusion and suggestions for further work.

2 Digitization of microscope slides and herbarium sheets

Microscope slides and herbarium sheets collections contain specimens that are close to 2D representations, i.e., although the specimens exist physically three dimensional (length, width, vertical depth), they are pressed flat and the depth dimension is such that for imaging purposes the majority of slides and herbarium sheets can be treated as two-dimensional. The equipment used for imaging in each case can vary between institutions and collections; however, there is overall consistency of characteristics of the images to be produced for each type of these collections.

2.1 Microscope slides

Microscope slide digitization produces images of individual slides that can contain up to four kinds of elements: (1) specimen itself (coverslip area), (2) labels(s), (3) barcode and (4) nomenclatural type labels(s). All these elements are contained within the slide itself. In some slides, labels may be placed on both sides of the slide and require more than one pass through the image acquisition step, producing two images per slide. Figure 1 shows two examples of microscope slides, from the Natural History Museum (NHM) and Naturalis Biodiversity Center (Naturalis). After acquisition,



(a) Natural History Museum Specimen

(b) Naturalis Biodiversity Center Specimen

Fig. 1 Examples of Microscope Slide Images from NHM (NHM Data Portal [23]. Specimen: http://data.nhm.ac.uk/object/5b804af3-5e82-44f6-9861-ed13b4b13f26) and Naturalis (Naturalis Bioportal [24]. Specimen: http://data.biodiversitydata.nl/naturalis/specimen/RMNH.INS.867638). The image elements highlighted in both images are: (1) coverslip and specimen, (2) labels, (3) barcode and (4) type label. Notice that the type label is not always present

slide images may need further processing, which includes naming and linking the images to the corresponding specimen records, marking type specimens and extracting data from labels.

The goal of the semantic segmentation approach for these types of images is to correctly identify all the image elements and differentiate between the instances present on each image. That is, for semantic segmentation, pixels in an image are assigned to a class corresponding to one of the four image elements types listed above. For instance, segmentation pixels can be assigned the type label so that multiple instances from the same type can be identified (i.e., multiple labels). As the examples from Fig. 1 show, the colors, textures and shapes of elements can vary between collections, for instance, the barcode labels used are clearly different. Notice also that the specimen image in Fig. 1a has more labels than the specimen image Fig. 1b. Other potential issues are the overlapping of labels, for instance in Fig. 1a the type label is placed over one of the larger labels on the slide; this also can happen with barcodes. A further potential issue is text written directly on the slides. In such a case there are no clear borders for the label and the text area merges with the background.

The physical features of the microscope slides influence the resulting images, and these include the type of specimens being preserved, the mounting techniques and curation processes used, and the slides themselves [2]. In this study we consider slides having a standard size of 25 mm × 75 mm (approximately 3″ × 1″) [4]. The resolution of the specimen images used can vary from 900 pixels per inch (ppi)[1] to 28,500 ppi.[2]

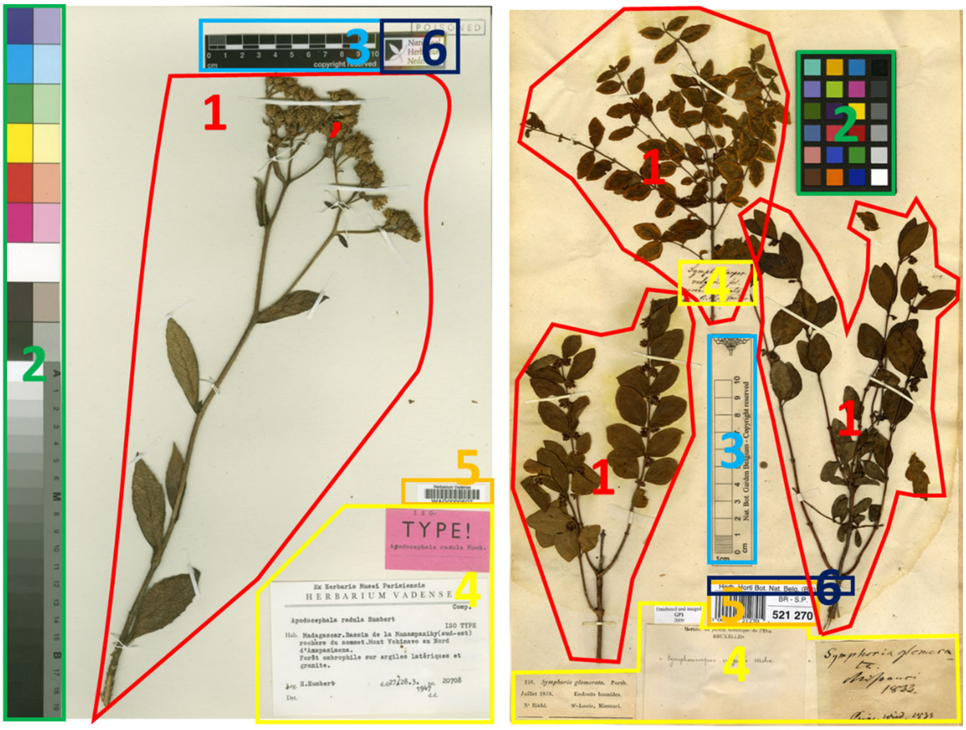
2.2 Herbarium sheets

The majority of botanical institutions follow the digitization guidelines of the Global Plants Initiative (GPI) [14, 15] which specifies the elements to include and the resolution for herbarium sheet images. Consequently, most of the images produced when digitizing herbarium sheets are homogeneous. According to these GPI guidelines each herbarium sheet must include: (1) specimen, (2) color chart, (3) scale bar, (4) labels, (5) barcode and (6) institution name. The GPI guidelines also consider the needs for multiple images per specimen because of drawings and letters attached to herbarium specimens, labels covered by specimen parts or information attached to the back of the sheet. Some specimens also have envelopes or capsules containing loose material associated with the specimen (such as seeds, flowers or sprouts). In some workflows these are left unopened. In others these are opened, emptied onto trays and imaged as well. For instance, at RGBK, the digitization process specifies opening the capsules. In this case they take one image of the sheet with the capsule closed. Then one with the capsule open and normally cut and paste the contents in the open capsule on top of the closed capsule image. However, if there is writing on the closed capsule or multiple capsules to open, then multiple images might be used, one with the capsules open and one with capsules shut. Consequently, digitized herbarium sheet specimens can include of more than one image. The GPI guidelines recommend scanning at 600 ppi for archival quality images. Herbarium sheets generally have a standard size of 29 × 43 cm (or 11.4 × 16.9 in), although there is some variation in this. Specimens also vary in size. The goal of applying semantic segmentation for processing images of herbarium sheets is to correctly identify all the image elements and differentiate between the instances of each type present in each image (as shown in Fig. 2).

Despite the existence of the standard guidelines from GPI, there can be several variations in the types of elements used, such as type of color chart, scales and barcodes. Some specimens from MBG (Fig. 2b), for example, have a transparent scale that is sometimes placed in between the specimen parts. The sizes and shapes of color charts are also variable. The Finnish Museum of Natural History, for example, uses two small color charts on the side of the sheet, while others use long color charts spanning the length of the sheet. Some color

[1] Approximated from the dimensions of specimen image obtained from Naturalis, Fig. 1b.

[2] Approximated from the dimensions of specimen image obtained from NHM, Fig. 1a.



(a) Naturalis Biodiversity Center Specimen **(b)** Meise Botanic Garden Specimen

Fig. 2 Examples of Herbarium Sheet Images from Naturalis (Meise Virtual Herbarium [22]. Specimen: https://www.botanicalcollections.be/specimen/BR0000005212705) and MBG (Naturalis Bioportal [24]. Specimen: https://data.biodiversitydata.nl/naturalis/specimen/WAG0000507). The image elements highlighted in the specimens are: (1) specimen, (2) color chart, (3) scale, (4) labels, (5) barcode and (6) institution name. Notice that institution name is included as part of the scale bar. Other institutions may include it on the barcode, color chart or as a separate stamp

charts also include a scale bar. Additionally, herbarium sheets can contain more than one specimen, and as a result they may contain more than one barcode. Some barcodes are simple and only include the specimen identifier, while others are printed in labels that also include the name of the institution.

2.3 Image processing and segmentation

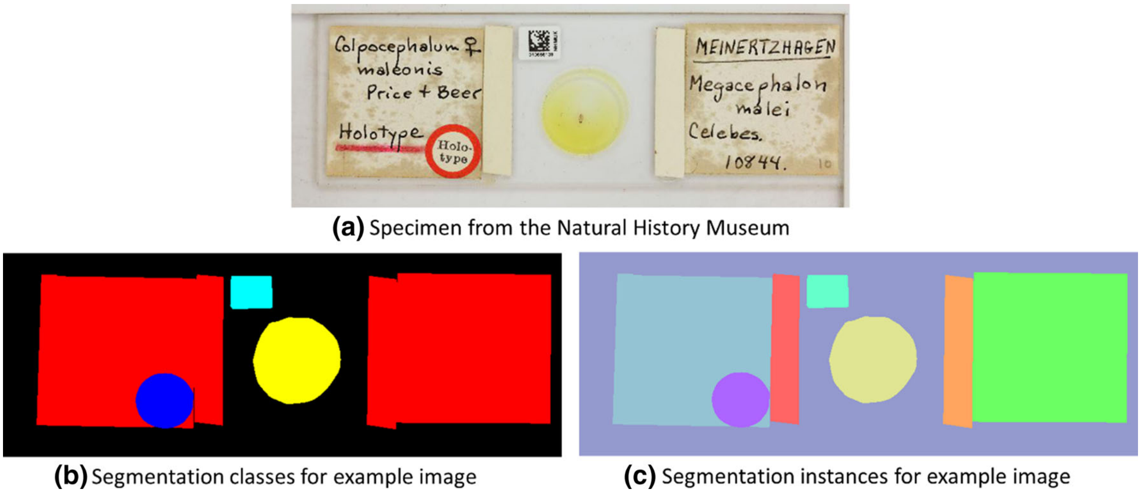
The further processing activities of microscope slides and herbarium sheets images that can benefit from segmentation include identification of image elements, identification of regions of interest, identification of nomenclatural type specimens (identification of type labels), verification of image names (reading barcodes) and image quality verification (color, sharpness, cropping).

The presence of the different elements is a basic requirement for both herbarium sheets and microscope slides. Verifying that large batches of images contain the minimal required elements can be delegated to the semantic segmentation process. This identification in turn can facilitate verification of file names and linking to physical specimen records in collection management systems through the barcodes. Similarly, breaking up the large image into smaller regions of interest can also benefit optical character recognition (OCR) processes, improving both speed and accuracy [27]. For a longer discussion of these aspects see the report on a pilot study performed to evaluate the suitability of segmentation comparing processing segmented images vs. processing full images [25].

In addition to the issues highlighted for microscope slides and herbarium sheets, it is important to acknowledge that digitization projects are time and resource limited. Consequently, projects tend to work on subsets of the collections and are separated from one another in time by months or even years. As a result, the quality of the images, the image elements used and the layout of the specimens may change between digitization projects targeting different portions of large collections. This can be the consequence of many factors, including changes in working procedures, standards and best practices applied and equipment and techniques used.

3 Semantic segmentation of natural history specimen images

Image processing in combination with artificial intelligence methods has been proposed to address different issues of natural history digitization projects, such as: identification of nomenclatural type specimens [33], morphological analysis [11, 37], specimen identification [6, 13], identification of the elements present in a specimen image [35, 36], automated information extraction [16], phenological research [38] and phenotype studies [19, 30]. Some methods are specifically designed to take advantage of the large quantities of image



(a) Specimen from the Natural History Museum

(b) Segmentation classes for example image

(c) Segmentation instances for example image

Fig. 3 Segmentation of microscope slide image (same as Fig. 1a, from NHM Data Portal [22]. Permanent URL: https://data.nhm.ac.uk/dataset/collection-specimens/resource/05ff2255-c38a-40c9-b657-4ccb55ab2feb/record/8023163. Retrieved: 15:05 18 Sep 2018 (GMT)). **a** Is the original image. **b** Shows the five classes in which the image pixels are grouped: specimen (yellow), labels (red), barcode (light blue), type label (dark blue) and background (black). **c** Shows the instances present on the image, with each instance region being indicated by a different color

data made available recently, while others focus on improving the quantity and quality of the data included in those datasets. The NHM semantic segmentation network (NHM-SSN) falls within the latter category as it is part of the continuing effort to automate and improve the museum's digitization workflows [10]. The NHM-SSN was developed in the context of the efforts for digitizing microscope slide collections [1, 2, 18] as a resource which could speed up and automate some portions of the image processing and curation steps. One of the envisaged advantages of combining artificial intelligence and image processing methods is the creation of services that can be seamlessly integrated within the image processing workflows of large digitization projects. For this to be possible, however, it is necessary to ensure that the methods are flexible and adaptable for use in the imaging workflows of different projects, targeting different collections and implemented by different institutions. This portability goal is one of the unexplored areas of the application of semantic segmentation.

3.1 The NHM semantic segmentation network

The NHM-SSN is a semi-supervised semantic and instance segmentation network developed originally for the segmentation of images of microscope slides. First, a semantic segmentation step breaks an image into smaller segments grouping pixels into different classes predefined to represent the element types of interest. In a second step, the separate instances of each element class are identified. This process is illustrated by the example in Fig. 3. Figure 3a shows the original microscope slide image. The elements present are classified as i) specimen (in the center), ii) labels (either side of the specimen), iii) type label (small label circled in red in the left side), iv) barcode label (small white label to upper left of specimen), while the rest of the image is classified as 'background.' The labeling of these different elements as colored classes in the first step is shown in Fig. 3b with the definition of classes linked to predefined colors (yellow, red, light blue, dark blue, black). The result of the second step identifying instances of each class is shown in Fig. 3c, with each instance being assigned a different color.

The training of the semantic segmentation network to support this process requires creating a large set of ground truth images. However, sets of ground truth images are expensive and hard to obtain, because ground truths for selected specimen images need to be manually generated and this requires training personnel and assigning resources for them to work in creating the ground truths. In this scenario, it is desirable to use methods that can perform well with small ground truth datasets for training. Semi-supervised learning covers several techniques that employ large datasets of images without ground truths (unlabeled data) to enhance the capability of models otherwise learned on small ground truth sets.

3.2 Network structure

The NHM-SSN developed by the NHM [10] is available online.[3] The github repository wiki explains the architecture of the NHM-SSN (shown in Fig. 4). The SSN consists of two branches: primary segmentation branch (right, in

[3] https://github.com/NaturalHistoryMuseum/semantic-segmentation.

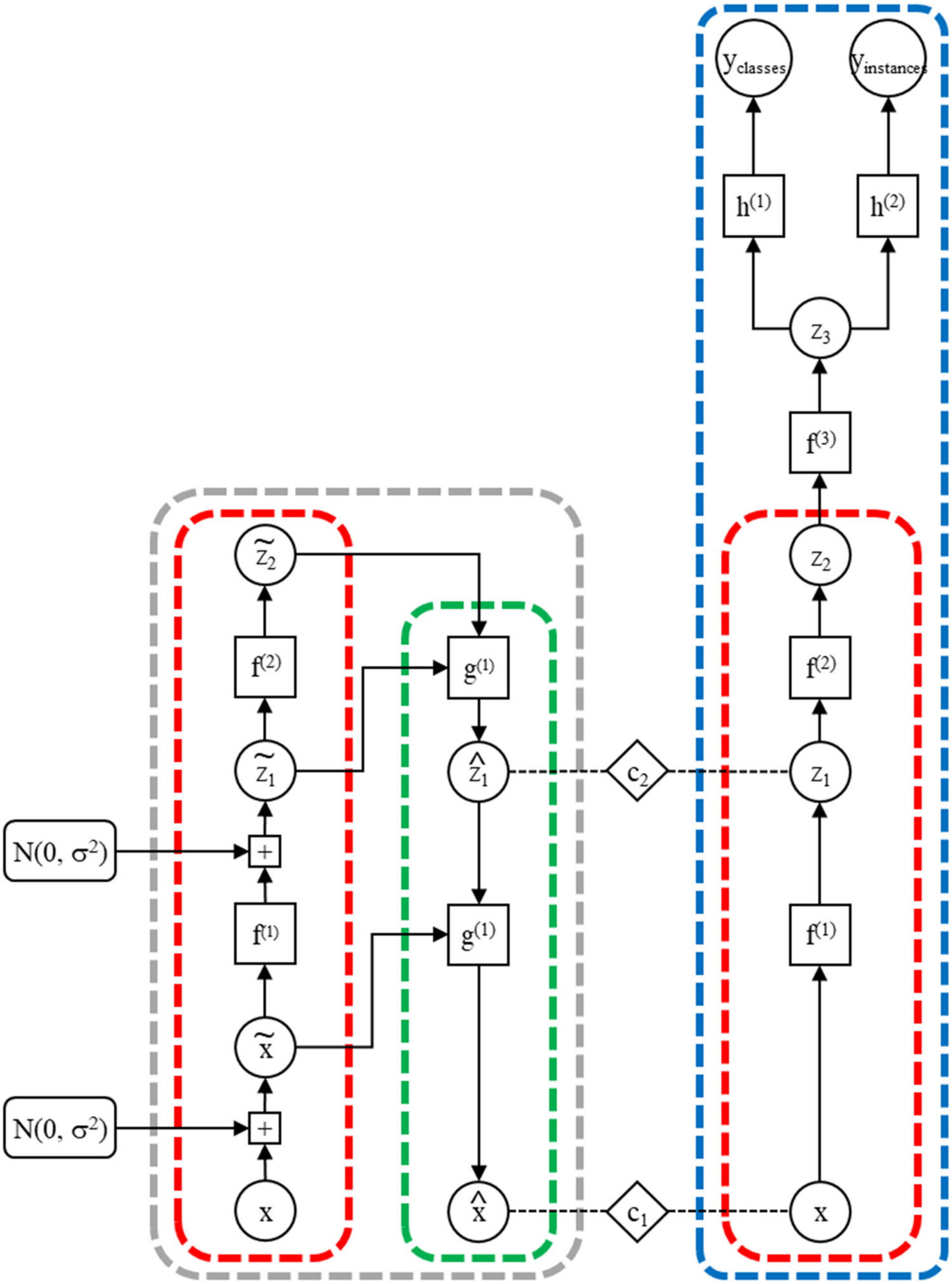


Fig. 4 Diagram of the Network Architecture and Flow. The primary segmentation branch is shown in blue (right) and the reconstruction branch for regularizing by semi-supervised learning is in gray (left). In the diagram, $f^{(1)}$ and $f^{(2)}$ represent smaller networks consisting of the highest layers of a ResNet-18 [12] model that has been pretrained on ImageNet; they are divided up to enable denoising losses [8]. $f^{(3)}$ is also a network, this time consisting of dilated convolutional layers (also known as atrous convolution layers) of decreasing sizes [7]. The branches $h^{(1)}$ and $h^{(2)}$ are simple two-layer CNNs. The feedforward path $x - z_1 - z_2$ shares the mappings $f^{(1)}$ and $f^{(2)}$ with the encoder path (corrupted path) $x - \tilde{x} - \tilde{z_1} - \tilde{z_2}$. The decoder path $\tilde{z_2} - \widehat{z_1} - \hat{x}$ includes two denoising functions $g^{(2)}$ and $g^{(1)}$. The cost functions c_1 and c_2 are intended to minimize the differences between $\widehat{z_l}$ and z_l. Noise with a $N\left(0, \sigma^2\right)$ distribution is introduced as a component-wise batch normalization function [28]

blue) and reconstruction branch for regularizing the network and enabling semi-supervised learning (left, in gray). Both branches contain an identical sub-network with shared weights (in red). The difference between them being that noise is injected before each layer in the reconstruction side. The up-sampling part of this branch then tries to reconstruct these inputs to be the same as they were before the noise was added (derived from the architecture of ladder networks [28]). The shared sub-network is the embedding network (red), and the other subnetwork (in green) is for reconstruction denoising.

3.3 Related work

Recently others have taken a similar deep learning approach to the identification of the elements present in specimen images [35, 36], using methods such as You Only Look Once (YOLO), Region-Based Convolutional Neural Networks (R-CNN), and Single-Shot Detector (SSD) [35], comparing the accuracy per element detected and the processing time per image, subsequently refining the results of the most successful segmentation method (YOLO [36]). While these methods show successful segmentation results, they have not carried out thorough cross-validation experiments as described in this paper.

Apart from the specific efforts for digitization of natural history specimens, the work of identifying elements on 2D images is closely related to document layout segmentation efforts which support the analysis of digitized printed documents [5, 17, 20, 21, 26, 31]. In these areas page segmentation aims at identifying distinct text regions, images, tables and other non-text objects. In this research area, the identification and differentiation of text orientation and grouping has been used as a step for supporting OCR and data extraction [5], document classification [21], differentiation of text and images [17], text block order [31] and text unwarping [20].

The main differences between the segmentation of specimens and documents are: (1) the amount of text, (2) the type of non-text entities, (3) the purpose of extraction. The amount of text on specimens varies greatly from some lines in microscope slides to large paragraphs or even booklets attached to herbarium sheets, however, but they are nowhere near the amount of text on books or printed articles. Non-text entities on documents are typically images and tables, the types of objects on a digital specimens (scanned herbarium sheets and microscope slides) can be simpler but harder to identify because of differences in position, sizes, and color, in some cases the placement of elements is maintained for a digitization campaign but can change for the next. For instance, the placement of elements added to herbarium sheets can vary, and sometimes elements are placed on top of the sheet, on the borders of the sheet or a mix of places. Purpose of extraction is also different. In this area, information extraction, in this case OCR follows segmentation, is closer to the targets of layout segmentation. Quality control is another case; this can be part of the digitization workflow to ensure that required elements are present in images, to ensure the lighting and visibility of elements is correct. Quality control using segmentation can also be used when receiving batches of digitized specimens from contractors, to ensure that the images are consistent with the specification for the type of specimens being digitized. Rapid cataloguing and classification are a third purpose which mainly focuses on the identification of barcodes, which can vary in placement, type, coloring and size.

4 Experiment design

The objective of this paper is to validate if the NHM-SSN is suitable for deployment as a generic component that could be integrated into automated image processing workflows. This demands a low training cost and portability across datasets. This means that some retraining is required when switching the type of images being segmented; however, the model should be robust and ready with relatively small training sets, requiring hundreds rather than thousands of examples. The portability requirement is in turn covered by the fact that the images to be segmented have sufficient visual similarity required for the NHM-SSN to generalize in new data. In practice this should mean that the model provides consistent classification results that are not affected by the origin of the training sets, making it project and institution independent.

4.1 Training and cross-validation process

A cross-validation process was designed to test whether these two requirements are fulfilled by NHM-SSN. The cross-validation process encompasses training, testing and validating the models with different datasets. The requires training/testing of the NHM-SSN is performed with images from a single institution (Fig. 5). The training/testing dataset consists of a training dataset (80% of the images with ground truths), an unlabeled dataset (100% of specimen images w/o ground truths) and a testing dataset (10% of the images with ground truths). The training process will produce a set of learned models which then are applied to the testing dataset to determine the accuracy of the model. This accuracy is then used to determine which learned model to use in the cross-validation step. This process is repeated independently for the images from each institution producing a learned model associated to each institution.

After training and selecting learning models, the evaluation datasets (remaining 10% of the images with ground truths) are used in the cross-validation of the models, to determine the actual accuracy of the learned models when applied to sets from other institutions (Fig. 6).

The experiments were performed in two batches, one for microscope slides and another for herbarium sheets. Each cross-validation batch consists of three image datasets from three different institutions. Each dataset is split accordingly and used in training following the process described above. The results of the cross-validation are then tabulated and plotted for comparison. The results from these experiments should indicate whether NHM-SSN fulfills the portability requirements.

Further analysis of the results provides insights into the strengths and weaknesses of the segmentation method and the effectiveness of the cross-validation process. The validation and evaluation of the semantic segmentation process classifies each pixel of the image. The results of the pixel classification are compared to the ground truth images to generate the accuracy scores reported above. These data can be further analyzed by looking at the actual results and trying to measure further performance indicators such as specificity, recall and precision. The process for analyzing the performance of the

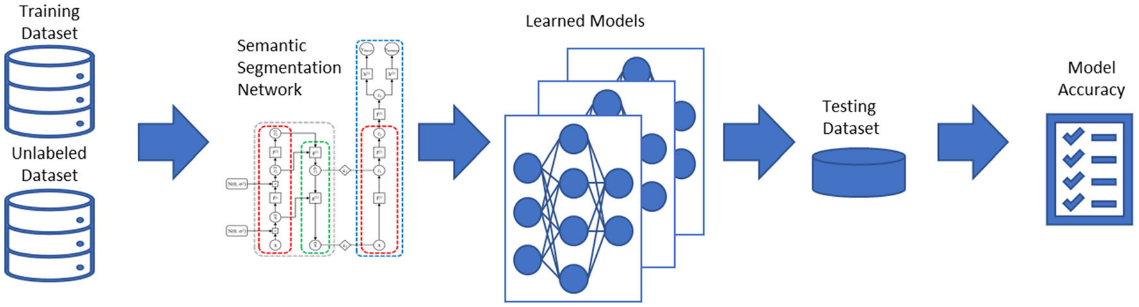


Fig. 5 Training: The training/testing dataset consists of 80% of the ground truth images, the unlabeled images and 10% of the ground truth images for testing. The training step is designed to assess the performance of the learning method and determine which of the learned models is the best for segmenting images from a given institution

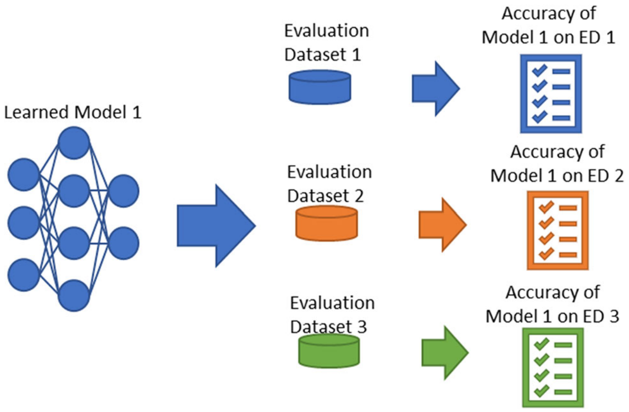

Learned Model 1
Evaluation Dataset 1
Accuracy of Model 1 on ED 1
Evaluation Dataset 2
Accuracy of Model 1 on ED 2
Evaluation Dataset 3
Accuracy of Model 1 on ED 3

Fig. 6 Cross-validation: The evaluation datasets consist of 10% of the images with ground truths. The first evaluation dataset contains images from the same institution as those used for training, while the remaining evaluation subsets contain images from other institutions

models involves measuring how well the predicted instance-class pairs match the instance-class pairs of the ground truth. This type of mapping will allow the identification of true positives, true negatives, false positives and false negatives. These values can then be used to create a confusion matrix and use the matrix = values to calculate additional evaluation measures (Table 1). Accuracy and error are complementary measures (ACC = 1 – ERR and ERR = 1-ACC).

Sensitivity and specificity measure the performance of a binary classification test.[4] Sensitivity (a.k.a. true positive rate, recall or probability of detection) measures the proportion of actual positives that are correctly identified as such. Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such. The false positive rate[5] (a.k.a. false alarm rate) usually refers to the expectancy of the false positive ratio (the probability of falsely rejecting the null hypothesis for a test). Precision (a.k.a. positive predictive value) is the fraction of relevant instances among the retrieved instances.[6] True positive rate, false positive rate and precision can be used to compare model performance side by side using receiver operating characteristics (ROC) plots and precision recall plots (PRC). These comparisons rely on using the evaluation sets to generate the corresponding confusion matrices for each model. The following subsections use these measures to perform those comparisons and give further insights on the performance of the models.

4.2 Data collection

The preparation of image data for cross-validation required obtaining image datasets for the planned experiments. Collecting image data was straight forward since most institutions have digitized large portions of their collections. However, the availability of microscope slides is limited in comparison with herbarium sheets.

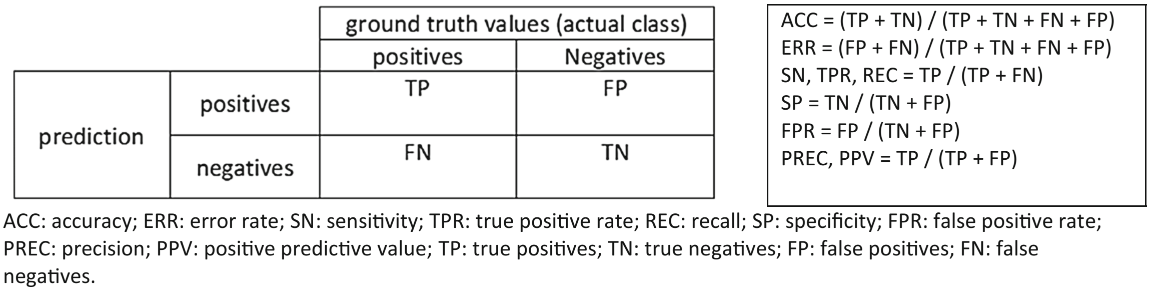
Microscope slides were provided by Natural History Museum (NHM, UK), Royal Botanic Gardens, Kew (RBGK, UK) and Naturalis Biodiversity Center (Naturalis, Netherlands). The microscope slides datasets from NHM and Naturalis are similar as both are sets of slides of invertebrates and contain the same element types (specimen, labels, type labels, barcode). The shapes of the different types of specimens, labels and barcodes are similar. The main difference is with the type labels. The Naturalis set contains less type specimens and the shape and color of the type labels on these varies. The RBGK set is quite homogeneous as the variations between slides are minimal. However, the images in this set are different from the NHM and Naturalis sets. In this set, the specimens are wood cut tissue samples (mounted in Euparal), clearly distinct from the preparations of invertebrate specimens, which are typically mounted using some resin (e.g., Canada balsam) which gives them a yellow pigment. The expectation prior to the experiments was that the cross-validation of the learned models derived from the NHM

[4] https://en.wikipedia.org/wiki/Precision_and_recall.

[5] https://en.wikipedia.org/wiki/Sensitivity_and_specificity.

[6] https://en.wikipedia.org/wiki/False_positive_rate.

Table 1 Confusion matrix and basic evaluation measures [34]



		ground truth values (actual class)	
		positives	Negatives
prediction	positives	TP	FP
	negatives	FN	TN

ACC = (TP + TN) / (TP + TN + FN + FP)
ERR = (FP + FN) / (TP + TN + FN + FP)
SN, TPR, REC = TP / (TP + FN)
SP = TN / (TN + FP)
FPR = FP / (TN + FP)
PREC, PPV = TP / (TP + FP)

ACC: accuracy; ERR: error rate; SN: sensitivity; TPR: true positive rate; REC: recall; SP: specificity; FPR: false positive rate; PREC: precision; PPV: positive predictive value; TP: true positives; TN: true negatives; FP: false positives; FN: false negatives.

Table 2 Microscope slides datasets

Dataset	Ground Truth	Unlabeled
National History Museum (NHM)	500	500
Naturalis Biodiversity Center (Naturalis)	500	500
Royal Botanic Gardens, Kew (RBGK)	500	500

Table 3 Herbarium sheet datasets

Dataset	Ground Truth	Unlabeled
National Museum Wales (NMW)	500	500
Muséum National d'Histoire Naturelle—MNHN	500	500
Mix	1000	1000
MFN	200	200
LUOMUS	200	200
MBG	200	200
NHM	200	200
Naturalis	200	200

and Naturalis sets would yield better results than the learned model derived from the RBGK set. Table 2 shows the composition of the datasets, and each dataset is composed of two large subsets of ground truth and unlabeled data (columns 2 and 3).

Herbarium sheets were provided by National Museum Wales (NMW, UK), Muséum national d'histoire naturelle (MNHN, France), Museum für Naturkunde (MfN, Berlin), Finnish Museum of Natural History (LUOMUS, Finland), Meise Botanic Garden (MBG, Belgium), Natural History Museum (NHM, UK) and Naturalis Biodiversity Center (Naturalis, Netherlands). Each set of herbarium sheets contained 500 images. The NMW dataset is a homogeneous dataset, since all the images provided were originally produced as part of a single digitization project. By homogeneous we mean that the images use the same type of control elements (barcodes, scales and color charts) and the positioning within the scanned image is regular. The MNHN dataset presents more variation as the images were produced during different projects. The control elements (barcodes, scales and color charts) in this set vary, and the quality of the images (illumination, cropping, naming conventions) also varies. The Naturalis image set includes not only images produced in different projects, but also images from collections acquired by Naturalis from other institutions. This means that even the textures of the sheets and the types of identifiers (barcodes) used vary within that single image set.

The three datasets used to train, test and validate the neural network transfer learning were those from National Museum Wales (NMW, Cardiff, UK), National Museum of Natural History (Muséum national d'histoire naturelle—MNHN, Paris, France) and the mixed set containing data from five institutions (Museum für Naturkunde—MFN, Berlin, Germany, Finnish Museum of Natural History—LUOMUS, Helsinki, Finland, Meise Botanic Garden—MBG, Meise, Belgium, Natural History Museum—NHM, London, UK, and Naturalis Biodiversity Center—Naturalis, Leiden, Netherlands). Table 3 shows the composition of the datasets, and each dataset is composed of two large subsets of ground truth and unlabeled data (columns 2 and 3).

4.3 Data preparation

Data preparation consisted of ground truthing and harmonization. Ground truthing is the process of labeling classes and instances for each of the images to be used in training, testing and validation. Harmonization is the process of homogenizing data to ensure that the results from the cross-validation experiments can be compared to each other.

Ground truthing is a resource-intensive task requiring hardware, software and trained operators. This part of the data preparation process was performed by colleagues from MBG, NMW and Cardiff University. This was the part that consumed most of the time due to the nature of the tasks and resource constraints. Only the NMW and MNHN image sets were labeled completely (500 images with ground truths for each). The remaining five image sets from MfN, LUOMUS, MBG, NHM and Naturalis were used to produce a mixed set of 1,000 labeled images (200 from each institution), while the remainder was used as the unlabeled images for training.

MBG contributed a set of 300 images from different institutions with ground truths, which were part of a larger dataset published as part of a pilot study [9]. In addition to this, MBG created a labeling protocol that helped to speed up the creation of ground truth images. The protocol consisted of a set of steps to create ground truth images that were easy to follow by volunteers with little training. The protocol was generic enough to allow using either GIMP[7] or Photoshop[8] software tools. NMW collaborated with a set of 800 images with ground truths from their collection as well as 500 images from different collections for the mixed set. Additionally, NMW improved on the MBG protocol, creating a set of Photoshop scripts that also helped speed up ground truthing of images. NMW were able to quickly train two volunteers who produced the labeled sets in less than a week. The rest of the images were labeled by colleagues from Cardiff University, using the MBG protocol and NMW scripts.

Harmonization was required because of the origin of the image sets and the ground truthing process. In relation to their origin, the image sets were provided at different resolutions and with varying image sizes. In relation to ground truthing, the labeled sets were produced using different software and hardware, which required verifying that image sets were consistent. Thus, harmonization included: ensuring homogeneity of image sizes, verifying correctness of the color scheme used for specifying the ground truth of classes and verifying that class and instance ground truths match. Harmonizing the image sets was intended to give greater confidence that the cross-validation results would be dependable and unbiased (i.e., not affected by image origin and/or ground truth process).

5 Training and cross-validation experiments

Two sets of experiments were performed, one with the microscope slides image sets and another with the herbarium sheets image sets. In both experimental sets, the SSN-NHM architecture was the same, but required independent training for each institutional set.

5.1 Microscope slides training

Each of the training datasets was used to train the segmentation network producing three models. The results from the training are shown in Table 4. As explained above, each of the ground truth sets was divided for training and validation in as 80%/10%/10% split, using 80% of the images for training, 10% of the images for testing and 10% of the images for validation. The testing accuracy (column 6) is the mean of the accuracy of the model when applied to the corresponding testing subset. The selection of the learned models to validate for each training experiment is depicted in Fig. 7. In each case, the learned model selected for cross-validation corresponds to the last epoch in which the test results are above the corresponding learning trend. The testing accuracies in the epochs after the selected one are lower and do not surpass the accuracy reported while training.

Table 4 Training–testing results for the microscope slides datasets

Dataset	Images	Training	Epoch	Testing	
				Images	Accuracy (%)
NHM	500	400	8	50	95.68
Naturalis	500	400	4	50	90.31
RBGK	500	400	2	50	93.66

5.2 Microscope slides cross-validation

The cross-validation of the models involves measuring how well the predicted instance-class pairs match the instance-class pairs of the ground truth. Comparing the ground truths to the segmentation results allows the identification of true positives, true negatives, false positives and false negatives. These indicators are needed to calculate accuracy, true positive rate, false positive rate and precision for each dataset. Comparing predicted instance-class pairs to ground truth instance-class pairs is illustrated in Fig. 8.

Table 5 shows the number of elements for each of the three evaluation datasets. The results presented in Tables 6, 7 and 8 allow comparing the results of segmenting the specimen images from the evaluation datasets using the NHM, Naturalis and RBGK learned models. The indicators for each type of image element are calculated to show how the performance of learned models varies depending on the type of image element. The total for each model allows the rapid comparison of the learned models.

The results in Table 6 show that the best model for segmenting the NHM dataset is, as expected, the NHM model. The NHM model lowest performance is the detection specimen elements, while the most successful is barcodes. The worst model for segmenting the NHM dataset is the Naturalis model, affected by the identification of false positives for type labels (detecting 724 false positives). The performance of the RGBK model is better than that of the Naturalis model; however, it performs poorly in the detection of labels missing

[7] GNU Image Manipulation Program: https://www.gimp.org/.

[8] Adobe Photoshop: https://www.adobe.com/products/photoshop.html.

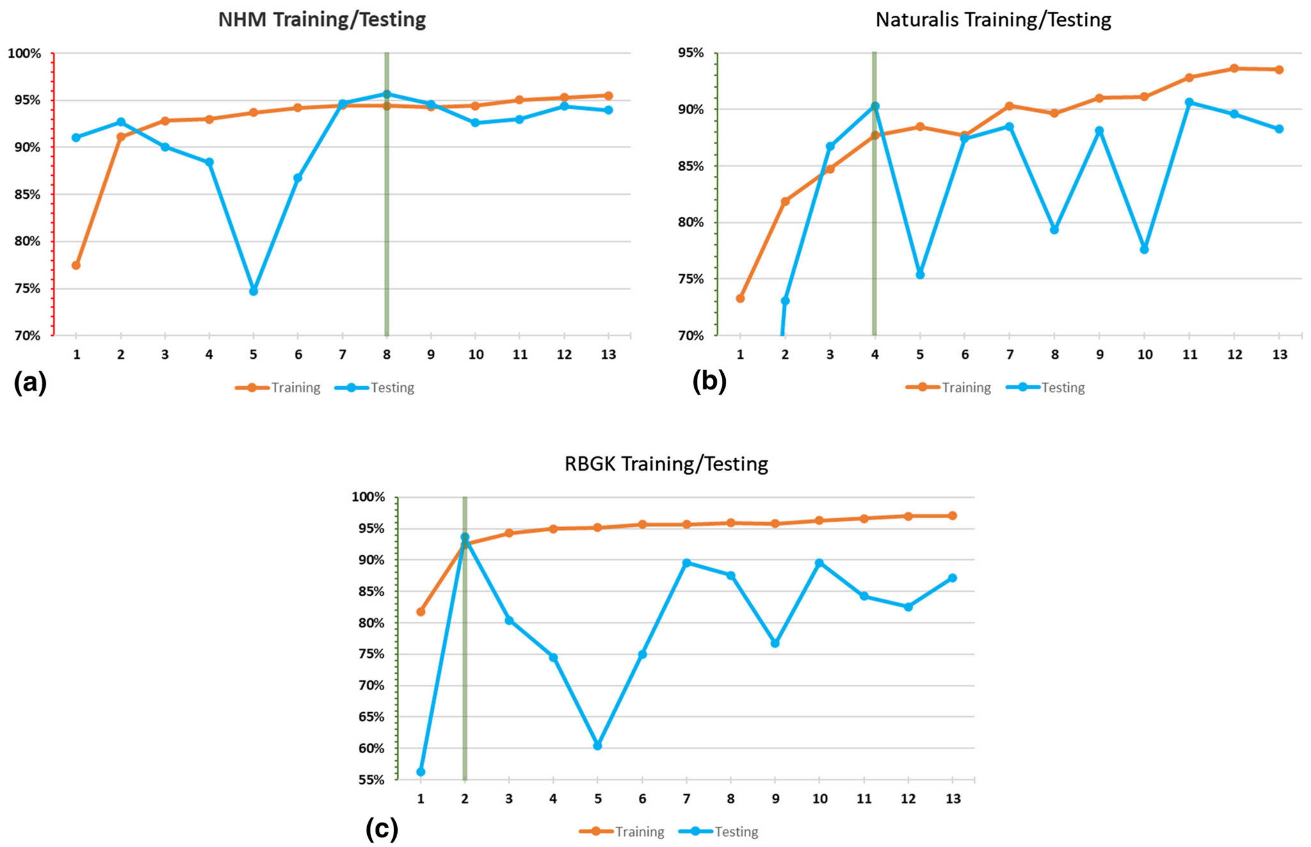


Fig. 7 Graphical view of the identification of the peak learned models for the microscope slides datasets: **a** NHM, **b** Naturalis and **c** RBGK. The orange line indicates the testing accuracies, which improves constantly as learning progresses, while the blue line marks the results of segmenting the test set. The green vertical indicates the epoch corresponding to the model selected for cross-validation

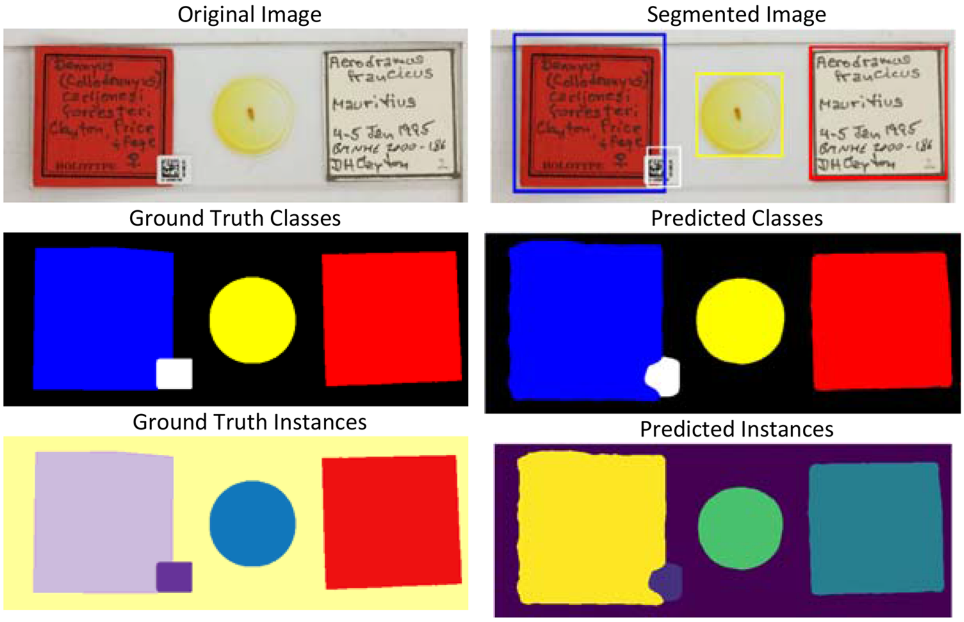


Fig. 8 Mapping of ground truth to predicted classes-instances. The images on the left show the specimen (from evaluation dataset, original [23]: https://data.nhm.ac.uk/object/e3898745-65ff-4e2a-a27c-879460df6e04/1586390400000) and its ground truths. The images on the right show the predictions and segmented image(top)

or 120 out of 156 (76.9% of the total). The detection of barcodes is the only element for which the tree models perform well, being element identified with the highest score by the three learned models.

The results in Table 7 indicate that the best model for segmenting the Naturalis dataset is the Naturalis model, followed closely by the NHM model. Both models perform poorly in the detection of type labels; however, Naturalis is significantly worse as it detects more false positives (61), while the NHM model does a better work detecting less false positives and more true negatives. The worst model for segmenting the Naturalis dataset is, as expected, the RBGK model. The RBGK model performs poorly for all elements except barcodes. This is consistent with the results for the NHM dataset above, in which the three models perform well in the detection of barcodes.

The cross-validation results (Table 8) indicate that the best model for segmenting the RBGK dataset is the RBGK model, as expected. The worst model for segmenting the RBGK dataset is Naturalis model, affected by the identification of false positives, especially for type labels (detecting 4,743 false positives). The performance of the NHM model is better than that of the Naturalis model; however, it performs poorly in the detection of type labels detecting an excess of 56 instances. As with the NHM and Naturalis datasets, the three models perform well in the detection of barcodes.

5.3 Herbarium Sheets Training

The three herbarium image datasets (Table 3) were used to train and validate the segmentation network, which resulted in the creation of three models. The results from the training are shown in Table 9. For training and validation, the ground truth datasets were split in the same way as the microscope slides sets above (80/10/10). The testing accuracy (column 6) is the mean of the accuracy of the model when applied to the corresponding testing subset. The validation accuracy (column 8) is the mean of the accuracy of the model when applied to the corresponding validation subset. The selection of the learned models for each training experiment is depicted in Fig. 9. In each case, the learned model selected for cross-validation corresponds to the last epoch in which the test results are above the corresponding learning trend. The testing accuracies in the epochs after the selected one are lower and do not surpass the accuracy reported while training.

5.4 Herbarium sheets cross-validation

The cross-validation of the models involves measuring how well the predicted instance-class pairs match the instance-class pairs of the ground truth. Comparing the ground truths to the segmentation results allows the identification of true positives, true negatives, false positives and false negatives. These indicators are needed to calculate true positive rate, false positive rate and precision for each dataset. Comparing predicted instance-class pairs to ground truth instance-class pairs is illustrated in Fig. 10.

Table 10 shows the number of elements for each of the three evaluation datasets. The results presented in Tables 12, 11 and 13 allow comparing the results of segmenting the specimen images from the evaluation datasets using the NMW, MNHN and MIXED learned models. The indicators for each type of image element are calculated to show how the performance of learned models varies depending on the type of image element. The total for each model allows the rapid comparison of the learned models.

The cross-validation results on Table 12 indicate that the best model for segmenting the NMW dataset is the MNHN model. The worst model for segmenting the NMW dataset is NMW model, which is not the expected outcome. The main issue of the model is the high detection of false positives across all element types (detecting 438 false positives). The performance of the MIXED model is close to the performance of the MNHN model, outperforming it in the detection of barcodes. None of the models perform above 0.5 in general, and labels are the only elements which are consistently detected above this threshold.

The results in Table 11 indicate that the best model for segmenting the MNHN dataset is the MNHN model, as expected. However, the total accuracy of the model is below the one reported in the training and testing of the model, having a high incidence of false positives for all elements, while still outperforming the other two models by 0.2. The worst model for segmenting the MNHN dataset is the NMW model, affected by the identification of 394 false positives for all element types, performing bad in the detection of color charts and barcodes. The performance of the Mixed model is better than that of the NMW model, while still reporting a high number of false positives, detecting an excess of 365 elements.

The results in Table 13 show that the best model for segmenting the Mixed dataset is the Mixed model, which has the lowest misidentifications of the three models being used, failing to detect 217 of the 777 total (27.9%). The worst model for segmenting the Mixed dataset is the NMW model, affected by the identification of 591 false positives for all element types while also missing large significant numbers of labels and barcodes (218 and 39, respectively). The performance of the MNHN model is close to that of the NMW model; however, it performs poorly in the detection of color charts (detecting an excess of 162). As with the NMW dataset, none of the models performs above 0.5.

The cross-validation of models using the datasets from the other institutions indicates that the best performing model is the one generated when training on the MNHN set, given that the mean difference with the results from the other sets

Table 5 Ground truths of evaluation datasets

Element	Dataset		
	NHM	Naturalis	RBGK
Label	156	88	52
Type Label	12	0	19
Specimen	57	52	153
Barcode	50	50	50
Total	275	190	274

Table 6 Predictions on the NHM evaluation set using the three learned models

Model	Element	TP	TN	FP	FN	ACC	TPR	FPR	PREC
NHM Model	Label	131	0	2	25	0.829	0.840	1.000	0.985
	Type Label	7	40	6	5	0.810	0.583	0.130	0.538
	Specimen	55	0	18	2	0.733	0.965	1.000	0.753
	Barcode	50	0	2	0	0.962	1.000	1.000	0.962
	Total	243	40	28	32	0.825	0.884	0.412	0.897
Naturalis Model	Label	10	0	0	146	0.064	0.064	0.000	1.000
	Type Label	12	10	724	0	0.029	1.000	0.986	0.016
	Specimen	47	0	14	10	0.662	0.825	1.000	0.770
	Barcode	45	0	7	5	0.789	0.900	1.000	0.865
	Total	114	10	745	161	0.120	0.415	0.987	0.133
RBGK Model	Label	36	0	0	120	0.231	0.231	0.000	1.000
	Type Label	6	16	26	6	0.407	0.500	0.619	0.188
	Specimen	28	0	8	29	0.431	0.491	1.000	0.778
	Barcode	50	0	7	0	0.877	1.000	1.000	0.877
	Total	120	16	41	155	0.410	0.436	0.719	0.745

TP true positives, *TN* true negatives, *FP* false positives, *FN* false negatives, *ACC* accuracy, *TPR* true positive rate, *FPR* false positive rate, *PREC* precision (see Table 1 for calculation formulas)

Table 7 Predictions on the Naturalis evaluation set using the three learned models

Model	Element	TP	TN	FP	FN	ACC	TPR	FPR	PREC
NHM Model	Labels	73	0	22	15	0.664	0.830	1.000	0.768
	Type Labels	0	23	29	0	0.442	0.000	0.558	0.000
	Specimen	48	0	37	4	0.539	0.923	1.000	0.565
	Barcode	40	0	2	10	0.769	0.800	1.000	0.952
	Total	161	23	90	29	0.607	0.847	0.796	0.641
Naturalis Model	Labels	70	0	12	18	0.700	0.795	1.000	0.854
	Type Labels	0	18	61	0	0.228	0.000	0.772	0.000
	Specimen	46	0	9	6	0.754	0.885	1.000	0.836
	Barcode	48	0	4	2	0.889	0.960	1.000	0.923
	Total	164	18	86	26	0.619	0.863	0.827	0.656
RBGK Model	Labels	41	0	0	47	0.466	0.466	0.000	1.000
	Type Labels	0	16	35	0	0.314	0.000	0.686	0.000
	Specimen	13	0	0	39	0.250	0.250	0.000	1.000
	Barcode	41	0	1	9	0.804	0.820	1.000	0.976
	Total	95	16	36	95	0.459	0.500	0.692	0.725

TP true positives, *TN* true negatives, *FP* false positives, *FN* false negatives, *ACC* accuracy, *TPR* true positive rate, *FPR* false positive rate, *PREC* precision (see Table 1 for calculation formulas)

Table 8 Predictions on the RBGK evaluation set using the three learned models

Model	Element	TP	TN	FP	FN	ACC	TPR	FPR	PREC
NHM Model	Labels	47	0	22	5	0.635	0.904	1.000	0.681
	Type Labels	17	4	56	2	0.266	0.895	0.933	0.233
	Specimen	74	0	0	79	0.484	0.484	0.000	1.000
	Barcode	47	0	11	3	0.770	0.940	1.000	0.810
	Total	185	4	89	89	0.515	0.675	0.957	0.675
Naturalis Model	Labels	4	0	6	48	0.069	0.077	1.000	0.400
	Type Labels	18	0	4743	1	0.004	0.947	1.000	0.004
	Specimen	42	0	6	111	0.264	0.275	1.000	0.875
	Barcode	42	0	21	8	0.592	0.840	1.000	0.667
	Total	106	0	4776	168	0.021	0.387	1.000	0.022
RBGK Model	Labels	40	0	12	12	0.625	0.769	1.000	0.769
	Type Labels	18	5	33	1	0.404	0.947	0.868	0.353
	Specimen	133	0	25	20	0.747	0.869	1.000	0.842
	Barcode	50	0	3	0	0.943	1.000	1.000	0.943
	Total	241	5	73	33	0.699	0.880	0.936	0.768

TP true positives, *TN* true negatives, *FP* false positives, *FN* false negatives, *ACC* accuracy, *TPR* true positive rate, *FPR* false positive rate, *PREC* precision (see Table 1 for calculation formulas)

Table 9 Validation and testing results for herbarium datasets

Dataset	Training Images	Epoch	Testing	
			Images	Accuracy (%)
NMW	400	11	50	91.59
MNHN	400	9	50	93.11
MIX	800	9	100	92.62

Table 10 Ground truths of evaluation datasets

Element	Herbarium Sheets Dataset		
	NMW	MNHN	MIXED
Label	149	186	457
Scale	52	28	104
Color Chart	50	27	95
Barcode	50	87	121
Total	301	328	777

is 4.93%.[9] The second-best performing model is the one from the NMW dataset, having a mean difference of 18.39%. The worst performing model is the one derived from the MIX dataset, having a mean difference of 23.65%.

6 Analysis of results

The cross-validation confirmed that the learned models perform better for segmenting images from the same origin in all cases, however with lower accuracy than that reported in the training–testing phases. The main issues highlighted by the cross-validation experiments that need closer inspection are lower performance in segmenting data from the same origin, fragmentation of instances and detection of false positives. The following sections will analyze these issues for microscope slides and herbarium sheets.

[9] Mean difference calculated as $(|M_0 - M_1| + |M_0 - M_2|)/2$ where M_0 is the mean validation accuracy for using the validation subset from the model, and M_1 and M_2 are the mean validation accuracies when applying the model to the other two validation subsets from.

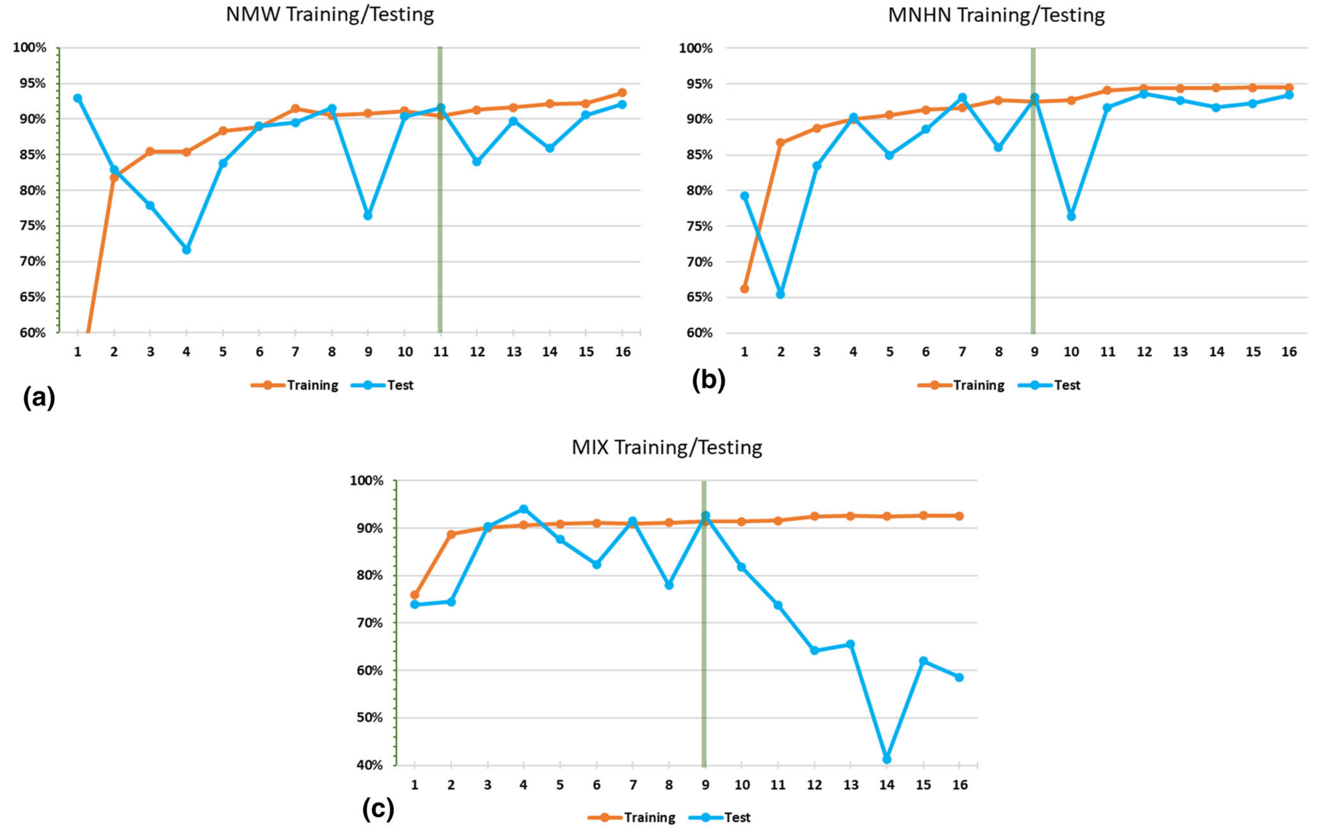


Fig. 9 Graphical view of the identification of the peak learned models when training with the herbarium datasets: **a** NMW, **b** MNHN and **c** MIXED set. The orange line indicates the testing accuracies, which improve constantly as learning progresses, while the blue line marks the results of segmenting the test set. The green vertical indicates the epoch selected as the ideal model for segmenting the images in the test sets

6.1 Microscope slides segmentation issues

The analysis of the results of cross-validation can be visually assessed by looking at the actual segmentation results using the learned models. This visual comparison is provided in Figs. 11, 12, 13. In each of these figures, the vertical order is determined by the success of the learned model created using the same original image set as the one from which the evaluation dataset is derived. The first two columns correspond to the image and the ground truths for the evaluation images, and the third column presents the segmentation results from the learned model which provides the best results, followed by the second and third best models for the dataset (as shown in Tables 6, 7 and 8).

The results from the third column on each of the figures coincide with the prediction which states that the best model for each dataset matches the learned model produced by using the same original image set as the one from which the evaluation dataset is derived. The results from the last two columns show the shortcomings of each model in segmenting the same images.

For the NHM dataset, Fig. 11, the RGBK model was evaluated as performing better than the Naturalis model, this is confirmed by the fact that the RGBK model is more effective in the identification of labels and barcodes, while the Naturalis model seems to be skewed toward identifying more type labels and specimens.

For the Naturalis dataset, Fig. 12, NHM model was evaluated as performing better than the RBGK model, this is confirmed by the fact that the NHM model is more effective in the identification of labels, specimens and barcodes, while the RBGK model seems to struggle with the identification of specimens, except for the specimen which Naturalis segmented poorly. The interesting thing is that in this case the Naturalis specimen is a wood cut tissue sample, which is the same type of the specimens used to train the RBGK model.

For the RBGK dataset, Fig. 13, NHM model was evaluated as performing better than the Naturalis model, this is confirmed by the fact that the NHM model is more effective in the identification of labels and barcodes, while the Naturalis model seems to struggle with the identification of specimens and labels.

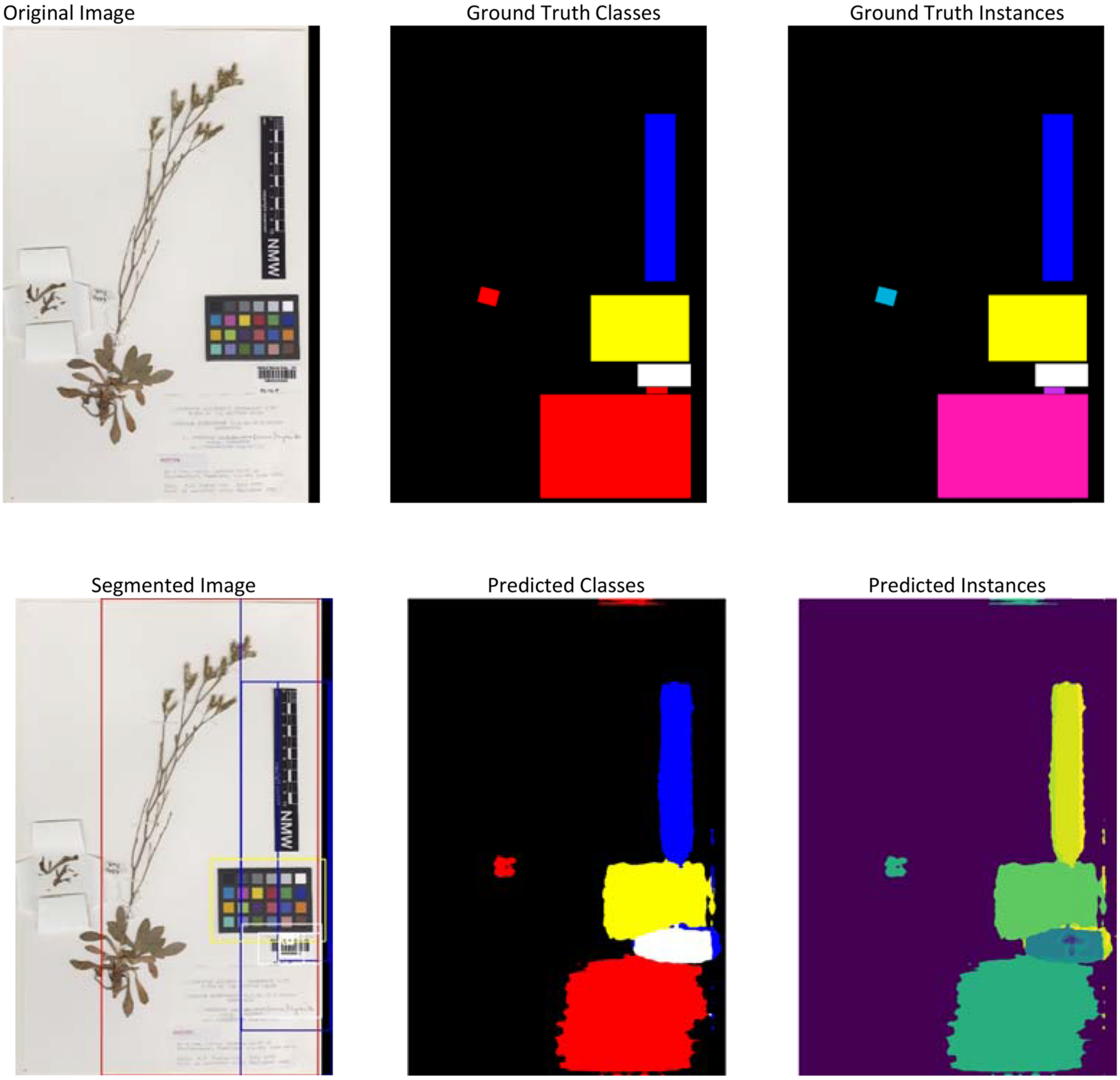


Fig. 10 Mapping of ground truth to predicted classes-instances. The images on the top show the specimen image (the images are part of the evaluation dataset, the specimen image is derived from the specimen provided by NMW, from the GPI digitization project. (https://plants.jstor.org/stable/history/10.5555/al.ap.specimen.nmw0000050)) and its corresponding ground truth classes and instances. The images on the bottom show the predictions, the rightmost image showing the segmentation of the original

6.2 Herbarium sheet segmentation issues

The visual comparison of the results of cross-validation can also be performed for the herbarium sheet experiments (Figs. 14, 15, 16). In the figures, the vertical order is determined by the success of the learned model segmenting the evaluation dataset, while the vertical order from column three is according to the predicted success of the learned models. The results for the NMW dataset indicated that the best model would be the MIX model, followed by the NMW and MNH models. As Fig. 14 shows, the results on the NMW model (fourth column) are close to those of the most MIX model; however, it is possible to see that the label and barcode instances are more segmented, creating more false negatives. The lowest performance of the MNHN model is confirmed by the fact that it tends to overestimate the size of labels.

The results for the MNHN dataset indicated that the best model for segmenting was the MNHN model, followed by the MIX and NMW models (Fig. 15). The results of the MNHN model are higher than those of the MIX model because it does not detect as many false positives, even when it can overestimate the size of labels. The MIX model in turn is better because it does not excessively fragment instances, whereas the NMW model fragments and reports false positives.

The results for the MIX dataset indicated that the best model for segmenting was the MIX model, followed by the

Table 11 Predictions on the MNHN evaluation set using the three learned models

Model	Element	TP	TN	FP	FN	ACC	TPR	FPR	PREC
NMW Model	Label	132	0	39	54	0.587	0.710	1.000	0.772
	Scale	27	6	88	1	0.270	0.964	0.936	0.235
	Color Chart	27	0	158	0	0.146	1.000	1.000	0.146
	Barcode	81	0	109	6	0.413	0.931	1.000	0.426
	Total	267	6	394	61	0.375	0.814	0.985	0.404
MNHN Model	Label	112	0	12	74	0.566	0.602	1.000	0.903
	Scale	28	19	23	0	0.671	1.000	0.548	0.549
	Color Chart	27	2	64	0	0.312	1.000	0.970	0.297
	Barcode	85	0	23	2	0.773	0.977	1.000	0.787
	Total	252	21	122	76	0.580	0.768	0.853	0.674
Mixed Model	Label	125	0	42	61	0.548	0.672	1.000	0.749
	Scale	28	0	176	0	0.137	1.000	1.000	0.137
	Color Chart	27	2	97	0	0.230	1.000	0.980	0.218
	Barcode	84	0	50	3	0.613	0.966	1.000	0.627
	Total	264	2	365	64	0.383	0.805	0.995	0.420

TP true positives, *TN* true negatives, *FP* false positives, *FN* false negatives, *ACC* accuracy, *TPR* true positive rate, *FPR* false positive rate, *PREC* precision (see Table 1 for calculation formulas)

Table 12 Predictions on the NMW evaluation set using the three learned models

Model	Element	TP	TN	FP	FN	ACC	TPR	FPR	PREC
NMW Model	Label	93	4	19	56	0.564	0.624	0.826	0.830
	Scale	52	0	200	0	0.206	1.000	1.000	0.206
	Color Chart	50	0	151	0	0.249	1.000	1.000	0.249
	Barcode	50	0	68	0	0.424	1.000	1.000	0.424
	Total	245	4	438	56	0.335	0.814	0.991	0.359
MNHN Model	Label	121	0	56	28	0.590	0.812	1.000	0.684
	Scale	52	0	52	0	0.500	1.000	1.000	0.500
	Color Chart	50	0	106	0	0.321	1.000	1.000	0.321
	Barcode	50	0	69	0	0.420	1.000	1.000	0.420
	Total	273	0	283	28	0.467	0.907	1.000	0.491
Mixed Model	Label	96	0	27	53	0.545	0.644	1.000	0.780
	Scale	52	0	134	0	0.280	1.000	1.000	0.280
	Color Chart	50	0	43	0	0.538	1.000	1.000	0.538
	Barcode	50	0	47	0	0.515	1.000	1.000	0.515
	Total	248	0	251	53	0.449	0.824	1.000	0.497

TP true positives, *TN* true negatives, *FP* false positives, *FN* false negatives, *ACC* accuracy, *TPR* true positive rate, *FPR* false positive rate, *PREC* precision (see Table 1 for calculation formulas)

MNHN and NMW models (Fig. 16). The results of the MIX model are higher than those of the MNHN model because it does not detect as many false positives. Additionally, the MNHN model is again faulty in overestimating the size of the labels. The NMW model fragments and reports more false positives than the other two models, and for this it is scored as the lowest performing model.

6.3 Discussion

The results obtained for microscope slides and for herbarium sheets point to the need for retraining the segmentation network each time a new source dataset is to be processed. To assess this, we performed an additional experiment for microscope slides to compare the results from a model generated when training against all the datasets, using all the training–testing datasets. Fig. 17 shows the results of the

Table 13 Predictions on the Mixed evaluation set using the three learned models

Model	Element	TP	TN	FP	FN	ACC	TPR	FPR	PREC
NMW Model	Label	239	0	32	218	0.489	0.523	1.000	0.882
	Scale	103	1	149	1	0.409	0.990	0.993	0.409
	Color Chart	95	0	302	0	0.239	1.000	1.000	0.239
	Barcode	82	7	108	39	0.377	0.678	0.939	0.432
	Total	519	8	591	258	0.383	0.668	0.987	0.468
MNHN Model	Label	258	0	24	199	0.536	0.565	1.000	0.915
	Scale	78	17	70	26	0.497	0.750	0.805	0.527
	Color Chart	90	1	162	5	0.353	0.947	0.994	0.357
	Barcode	105	4	120	16	0.445	0.868	0.968	0.467
	Total	531	22	376	246	0.471	0.683	0.945	0.585
Mixed Model	Label	263	0	24	194	0.547	0.575	1.000	0.916
	Scale	97	3	175	7	0.355	0.933	0.983	0.357
	Color Chart	94	3	111	1	0.464	0.989	0.974	0.459
	Barcode	106	3	69	15	0.565	0.876	0.958	0.606
	Total	560	9	379	217	0.488	0.721	0.977	0.596

TP true positives, *TN* true negatives, *FP* false positives, *FN* false negatives, *ACC* accuracy, *TPR* true positive rate, *FPR* false positive rate, *PREC* precision (see Table 1 for calculation formulas)

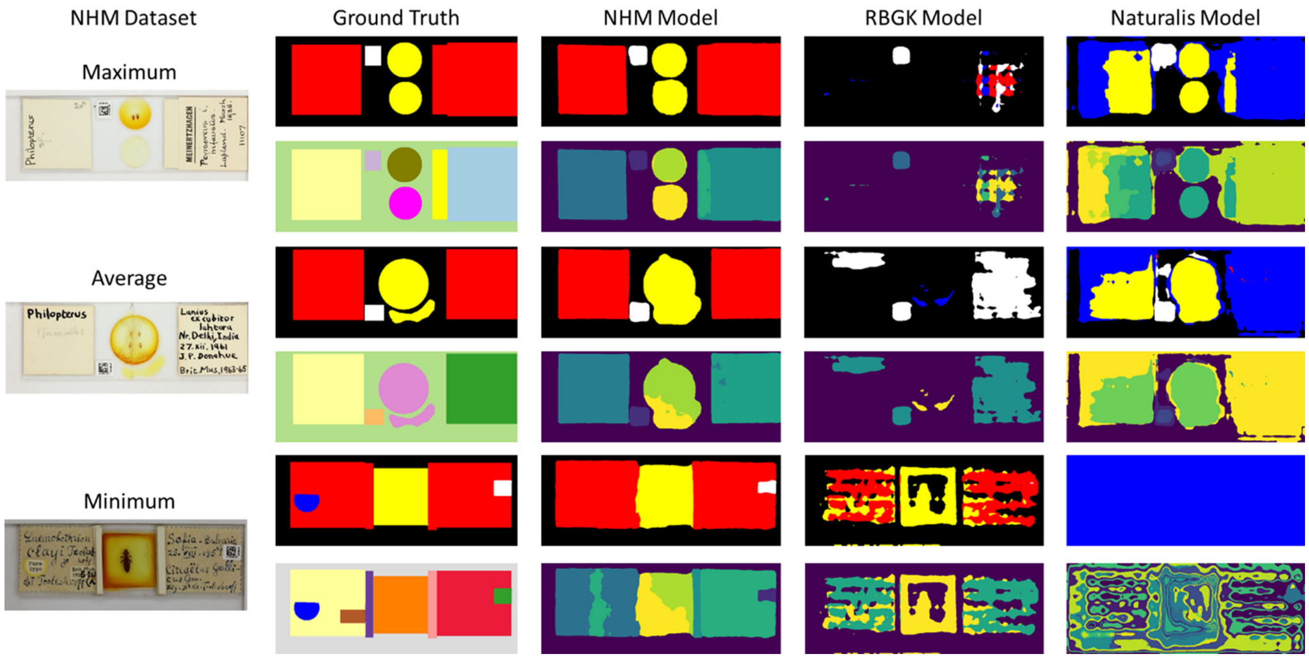


Fig. 11 NHM dataset segmentation results: specimen images (first column), ground truth images for each specimen (classes and instances, second column), results of segmentation with the NHM model (third column), results from segmenting with the RBGK (fourth column) and Naturalis models (fifth columns)

training–testing of the segmentation network and indicates the model selected for segmenting the evaluation sets.

After training and selecting a combined learned model for segmentation, the same cross-validation process applied before was used to evaluate segmentation results, i.e., measuring how well the predicted instance-class pairs match the instance-class pairs of the ground truths (Fig. 8).

Table 14 shows the evaluation of the results of the combined learned model when used for segmenting the combined evaluation set. The overall accuracy is of the model 0.807 which is above the average than that of the individual cross-validation experiments performed earlier. Individually, when separating the results of NHM, Naturalis and RBGK, the

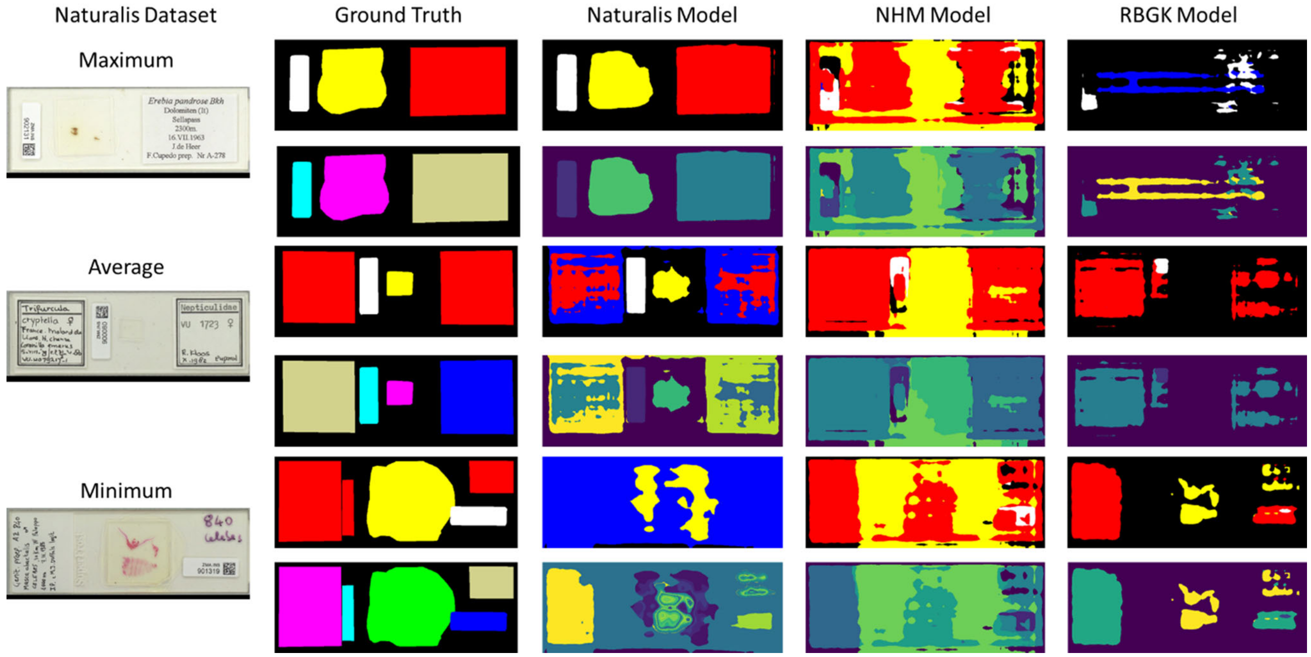


Fig. 12 Naturalis dataset segmentation results: specimen images (first column), ground truth images for each specimen (classes and instances, second column), results of segmentation with the Naturalis model (third column), results from segmenting with the NHM (fourth column) and RBGK models (fifth columns)

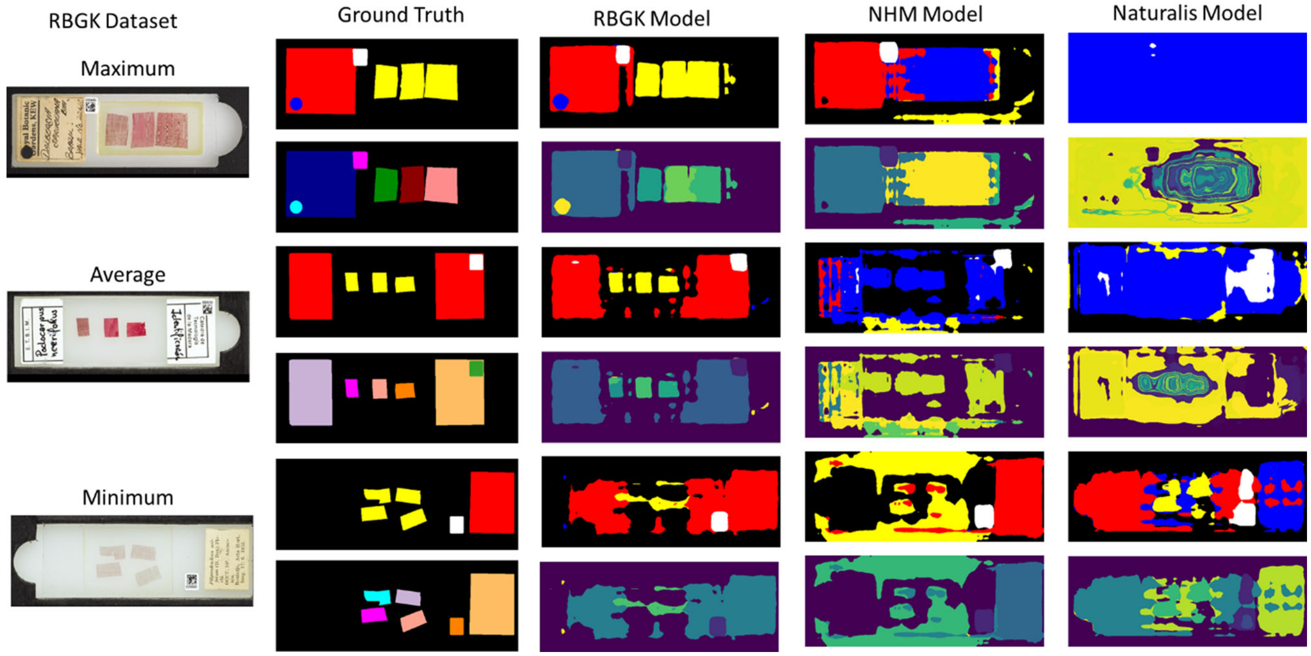


Fig. 13 RBGK dataset segmentation results: specimen images (first column), ground truth images for each specimen (classes and instances, second column), results of segmentation with the RBGK model (third column), results from segmenting with the NHM (fourth column) and Naturalis models (fifth columns)

model performs slightly worse than the NHM model, but significantly better than the Naturalis and RBGK models.

Figure 18 compares the results presented as the most successful model for segmenting the NHM evaluation dataset (three first columns from Fig. 11) to the results of segmenting the same images with the combined model (fourth column). The results from the two models are roughly equivalent. Figures 19 and 20 show the corresponding comparison for the

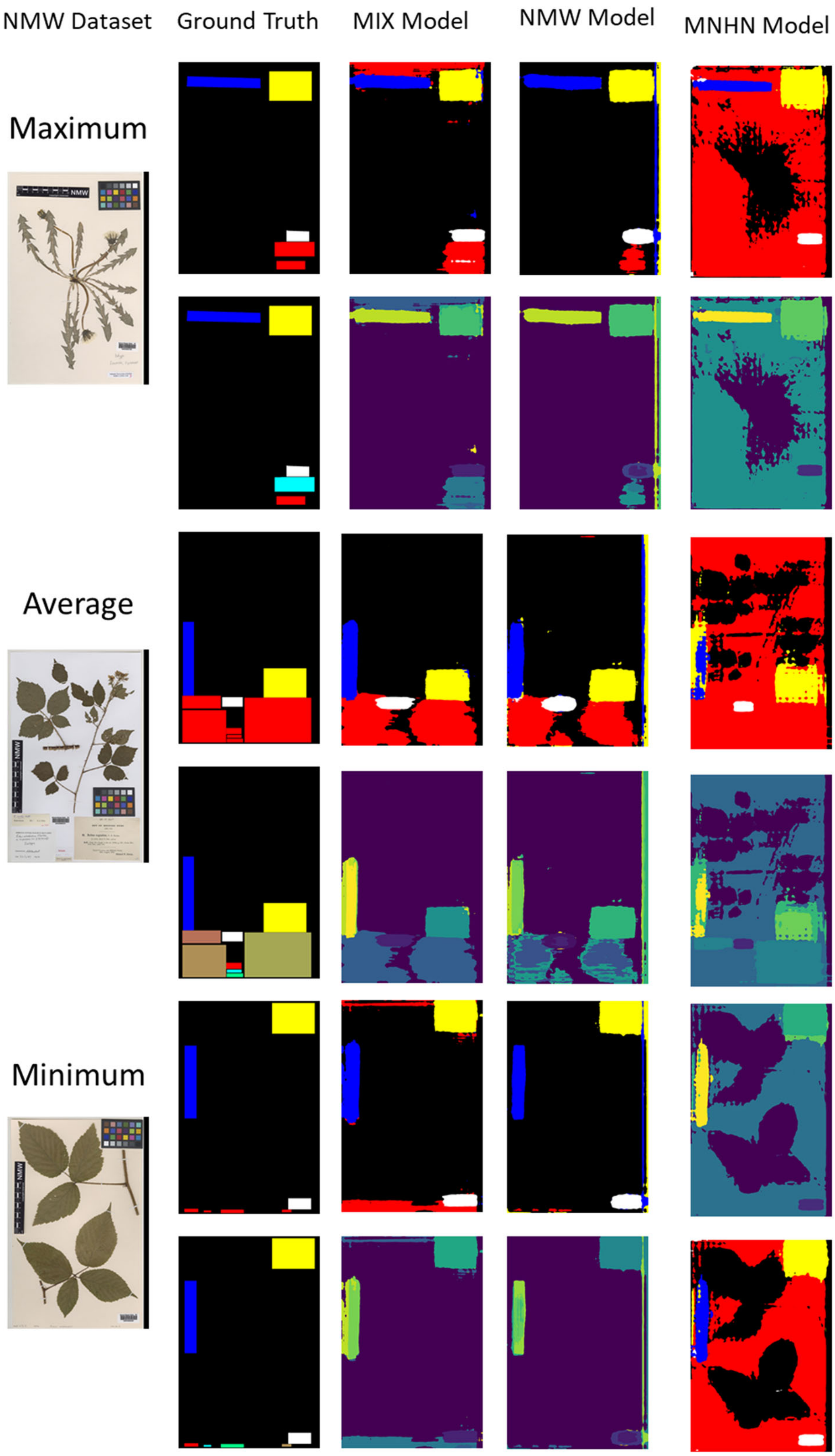


Fig. 14 Samples of segmentation results for the NMW dataset. The first column shows the specimen images, the second column shows the ground truths for each image (classes and instances), the third column shows the results of segmentation with the NMW model, this column shows a sample of the best segmented images at the top, a sample from an average result in the middle and a sample from the worst results from the model at the bottom. The remaining two columns show the corresponding results from segmenting with the MIX and MNHN models

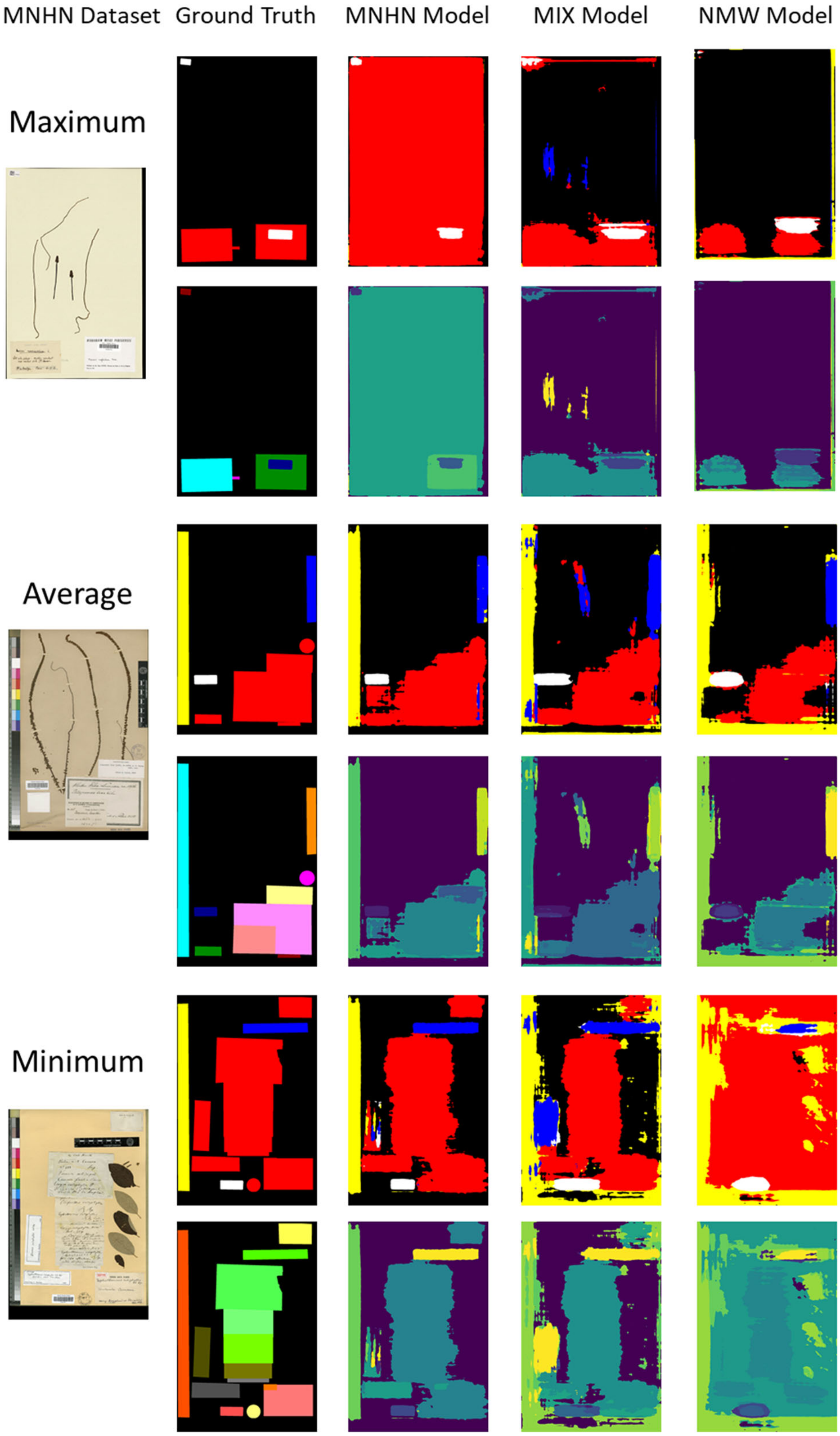
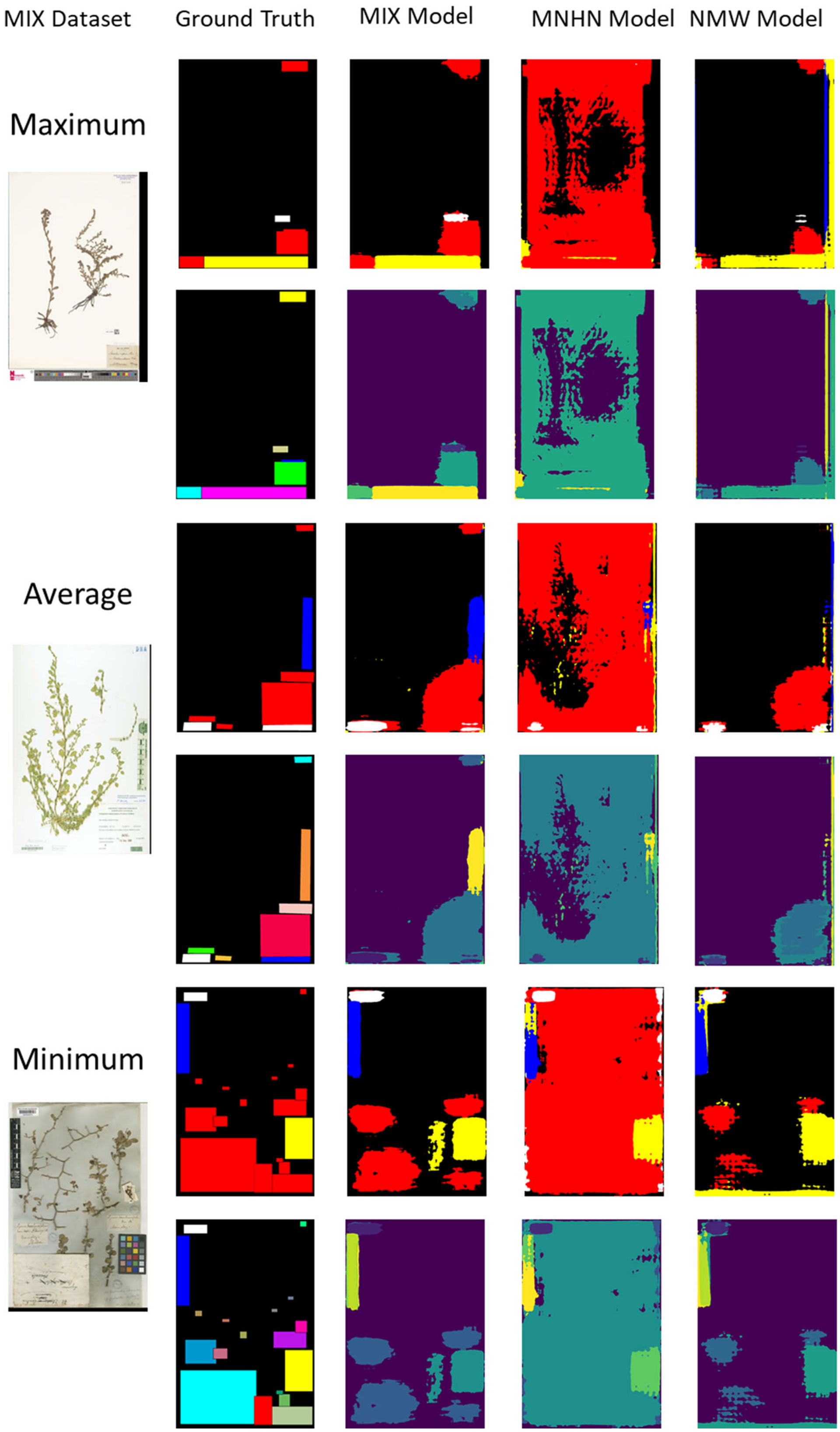


Fig. 15 Samples of segmentation results for the MNHN dataset. The first column shows the specimen images, the second column shows the ground truths for each image (classes and instances), the third column shows the results of segmentation with the MNHN model, this column shows a sample of the best segmented images at the top, a sample from an average result in the middle and a sample from the worst results from the model at the bottom. The remaining two columns show the corresponding results from segmenting with the MIX and NMW models

Fig. 16 Samples of segmentation results for the MIX dataset. The first column shows the specimen images, the second column shows the ground truths for each image (classes and instances), the third column shows the results of segmentation with the MIX model, this column shows a sample of the best segmented images at the top, a sample from an average result in the middle and a sample from the worst results from the model at the bottom. The remaining two columns show the corresponding results from segmenting with the MNHN and NMW models



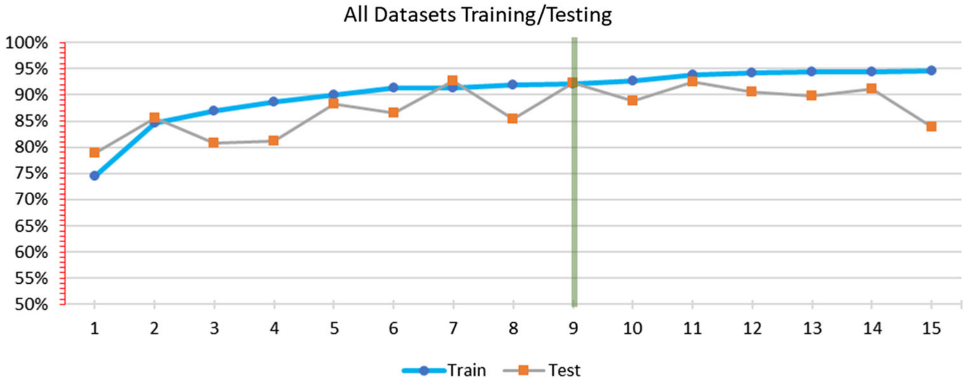


Fig. 17 Graphical view of the identification of the peak learned models when training against all datasets combined (NHM, Naturalis and RBGK). The blue line represents the training accuracy, which improves constantly as learning progresses, while the orange line represents the accuracy of segmenting the combined test set (NHM, Naturalis and RBGK). The green vertical indicates the epoch selected as the one producing the best learned model for segmenting the images in the evaluation sets

Table 14 Predictions on the combined evaluation set using the combined learned model

Dataset	Element	TP	TN	FP	FN	ACC	TPR	FPR	PREC
Combined	Label	263	0	28	33	0.812	0.889	1.000	0.904
	Scale	26	110	14	5	0.877	0.839	0.113	0.650
	Color Chart	233	0	42	29	0.766	0.889	1.000	0.847
	Barcode	144	0	29	6	0.804	0.960	1.000	0.832
	Total	666	110	113	73	0.807	0.901	0.507	0.855
NHM	Label	133	0	4	23	0.831	0.853	1.000	0.971
	Scale	9	40	3	3	0.891	0.750	0.070	0.750
	Color Chart	52	0	25	5	0.634	0.912	1.000	0.675
	Barcode	50	0	7	0	0.877	1.000	1.000	0.877
	Total	244	40	39	31	0.802	0.887	0.494	0.862
Naturalis	Label	82	0	5	6	0.882	0.932	1.000	0.943
	Scale	0	48	2	0	0.960	0.000	0.040	0.000
	Color Chart	49	0	14	3	0.742	0.942	1.000	0.778
	Barcode	46	0	13	4	0.730	0.920	1.000	0.780
	Total	177	48	34	13	0.827	0.932	0.415	0.839
RBGK	Label	48	0	19	4	0.676	0.923	1.000	0.716
	Scale	17	22	9	2	0.780	0.895	0.290	0.654
	Color Chart	132	0	3	21	0.846	0.863	1.000	0.978
	Barcode	48	0	9	2	0.814	0.960	1.000	0.842
	Total	245	22	40	29	0.795	0.894	0.645	0.860

TP: true positives; TN: true negatives; FP: false positives; FN: false negatives; ACC: accuracy; TPR: true positive rate; FPR: false positive rate; PREC: precision (see Table 1 for calculation formulas)

Naturalis and RBGK evaluation datasets. In both cases, the improved performance of the combined model when compared to the single-trained models is clearly visible.

This experiment demonstrates the possibility of creating learned models for segmentation which can be used to segment images from different origins (i.e., different collections and/or different institutions). These results support proposing the NHM-SSN as a resilient and portable segmentation network with general applicability beyond the domain for which it was originally developed.

6.4 Possible issues in model selection

There is a question about the cutoff point and the models selected for cross-validation. The use of accuracy for the selection of the best model appears to be supported by the

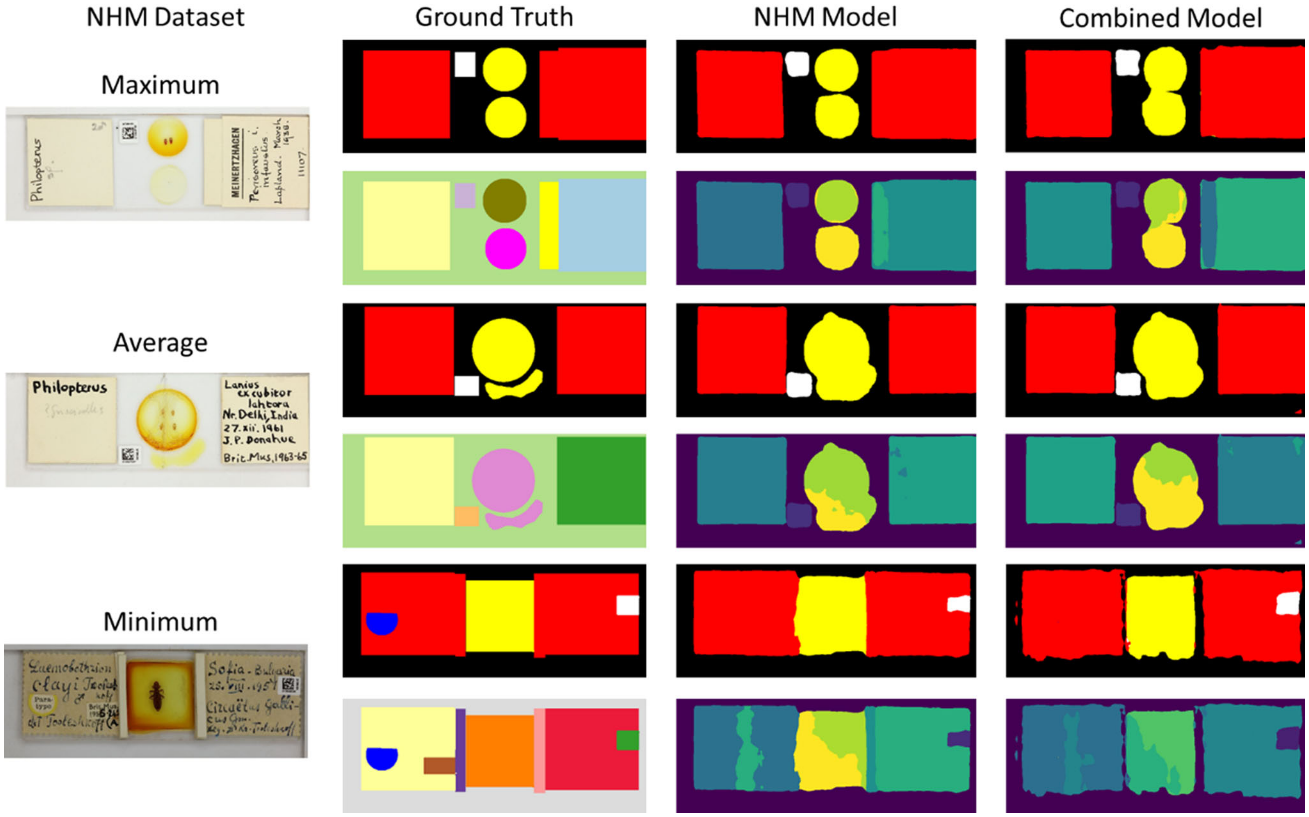


Fig. 18 Comparing maximum, average and minimum results from the NHM dataset: specimen images (first column), ground truth images for each specimen (classes and instances, second column), results of segmentation with the NHM model (third column), results from segmenting with the Combined model (fourth column). The three first columns are the same as those presented in Fig. 11

analysis of the corresponding ROC and PR plots for the models. For instance, in Fig. 21a, the ROC curve shows that Model 08 has the lowest false positive rate and a high true positive rate. This is confirmed by the PR plot (Fig. 21b) which shows that Model 08 has the highest recall and the highest precision compared to the other high accuracy models. However, the compared models are suboptimal since the precision and recall of all models are low, having precision in below 0.15 and recall between 0.30 and 0.55.

The issues are more evident in the models derived from the Naturalis and RBGK datasets, as Figs. 22 and 23 show. The high false positive rates and low precisions affect all the models, regardless of the dataset used for building them.

6.5 Alternative segmentation methods

Rather than building a segmentation framework from scratch we looked at the existing segmentation proposals. The only one published and available at the time was the NHM-SSN. Rather than accepting and using this model as it was, we devised the set of tests presented in the article to validate that it had the characteristics required by the problem at hand. Nevertheless, we did try to see if other techniques could be suitable for the segmentation of natural history specimens and performed two experiments using YOLO V3. In both cases the networks were trained using the RBGK dataset consisting of 500 microscope slides using a 70–15-15 (70% training, 15% testing and 15% validation). The dataset was annotated using a script which extracted YOLO V3 coordinates into individual text files from the ground truths used for NHM-SSN. The training of YOLO V3 and testing of the prediction were performed using the Google Colaboratory [3] virtual environment. The first experiment targeted training YOLO3 with pretrained Darknet74 weights on a single element (barcodes). The results for identifying barcodes where encouraging as it gave close to 60% success in detecting the instances. The second experiment attempted to identify four elements (Barcode, Specimen, Label and Type Label). These were less efficient at detecting instances, resulting in 52.6% success for barcode and 25.0% for specimens while not detecting either type labels or labels (Table 15). A visual inspection of results additionally showed that sometimes elements were misidentified, for instance, labels as specimens and specimens as barcodes.

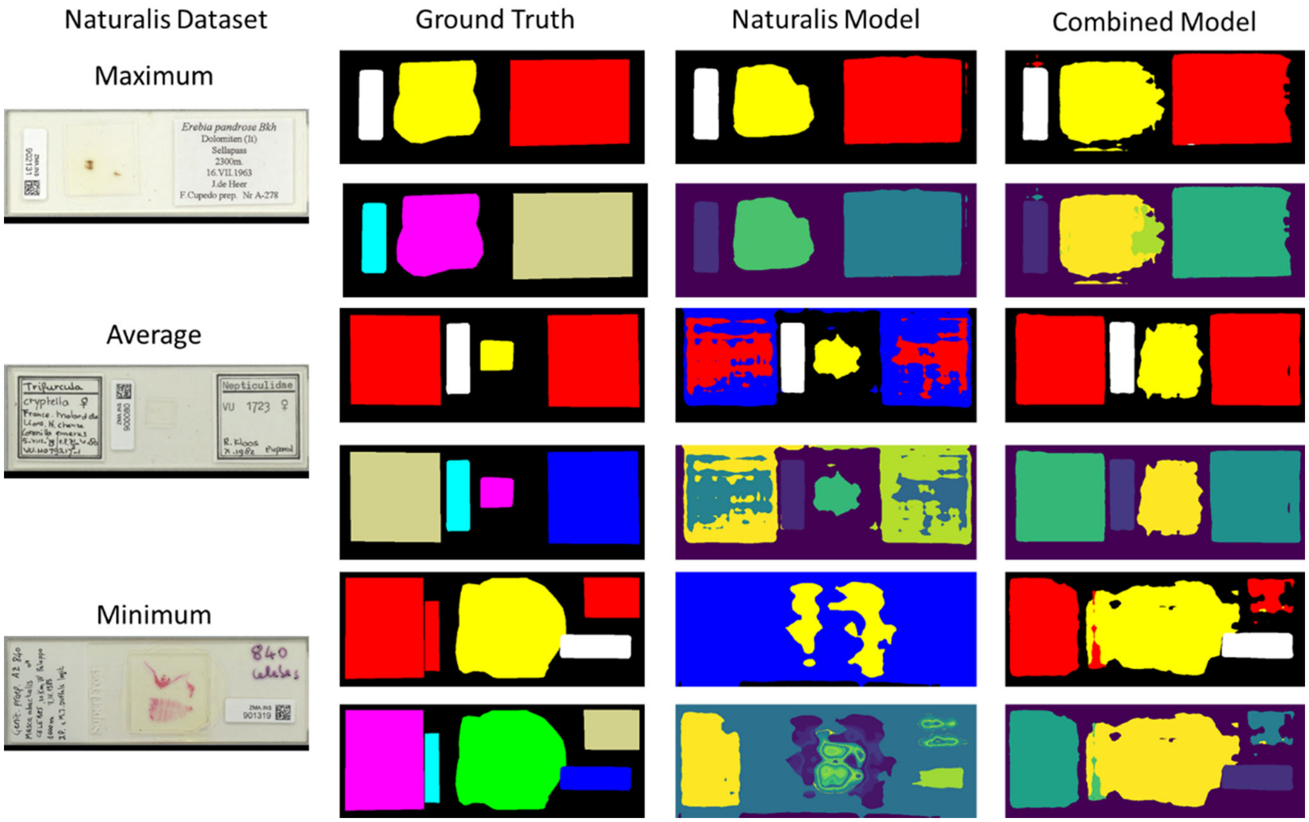


Fig. 19 Comparing maximum, average and minimum results from the Naturalis dataset: specimen images (first column), ground truth images for each specimen (classes and instances, second column), results of segmentation with the Naturalis model (third column), results from segmenting with the Combined model (fourth column). The three first columns are the same as those presented in Fig. 12

Figure 24 shows examples of the best, average and worst results obtained by applying YOLO. The average result was from images where two elements were detected (barcode and at least one specimen instance), the worst result where those case in which no element was identified, this happened in 11 of the 75 cases on the testing set (~ 15% of the test set).

These results are not mean to be compared side by side to the results presented for NHM-SSN as we did not attempt to perform larger experiments or alternative modifications to improve results. Instead, the results can be seen as the seed for further work using the same set of tests suggested in NHM-SSN to assess the suitability of YOLO or other techniques. These would require not only testing the predictive power of the approach but also the cost in terms of adapting, training and deploying.

7 Further work and conclusions

Rather than building a segmentation framework from scratch or accepting existing models as proposed, this paper proposes an evaluation strategy which may guide the selection of the most adequate method. The NHM-SSN model was designed specifically to address the segmentation of natural history specimen images. We devised and run a set of tests that have validated that it had the characteristics required to address this problem, which can also work on a wider range of datasets than those initially tested. Although the results indicate that the NHM-SSN model can be easily adapted for processing data of different collections and institutions, the results indicate that there is room for improvement and other models should be considered.

7.1 Further work

The results show that it is possible to test the flexibility of segmentation models to fit the requirements for wider use in natural history collections digitalization. Further work would require testing with other types of images and testing alternatives to training with a fully combined dataset, such as staged training. Apart from the evaluation method presented here, the datasets and ground truths produced are a valuable resource that can be used in future to evaluate improved versions of the NHM-SSN, as well as other segmentation

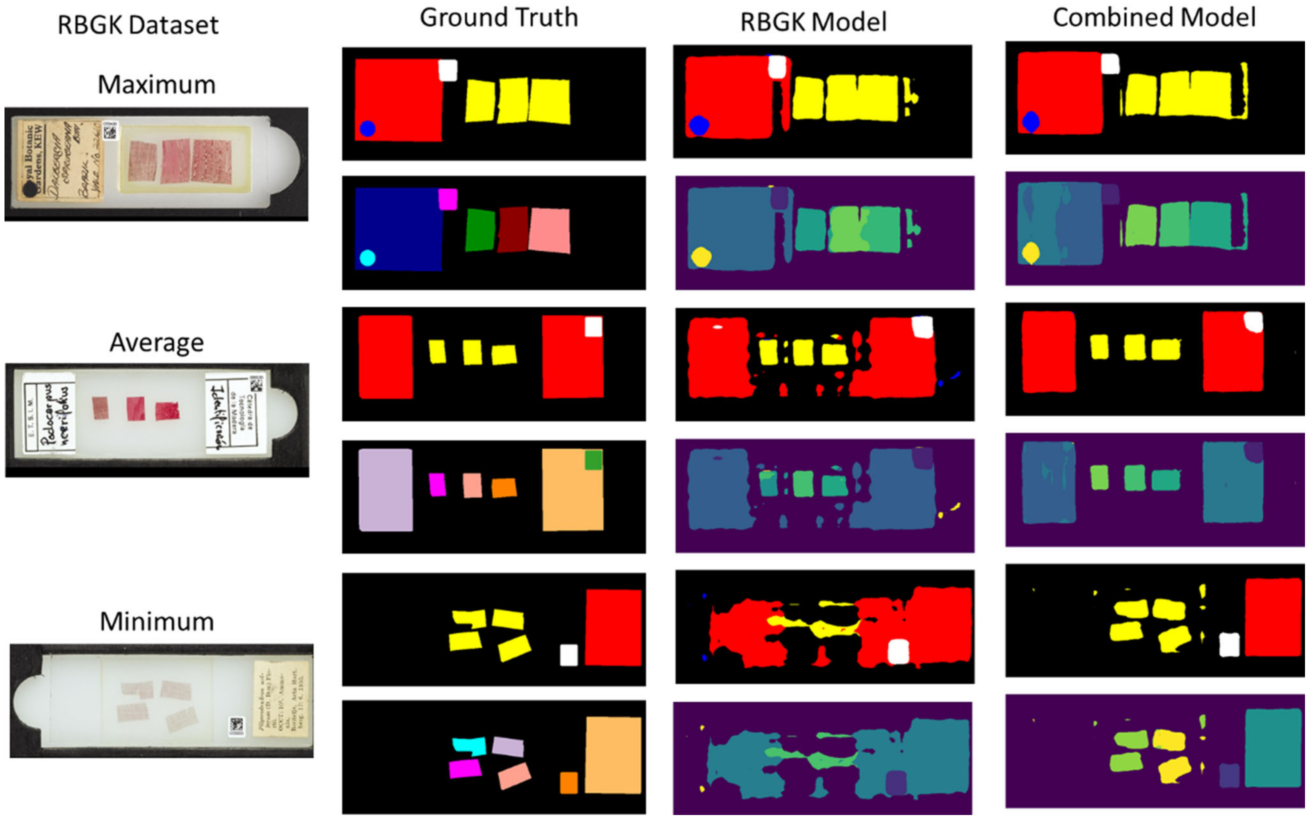


Fig. 20 Comparing maximum, average and minimum results from the RBGK dataset: specimen images (first column), ground truth images for each specimen (classes and instances, second column), results of segmentation with the RBGK model (third column), results from segmenting with the Combined model (fourth column). The first three columns are the same as those presented in Fig. 13

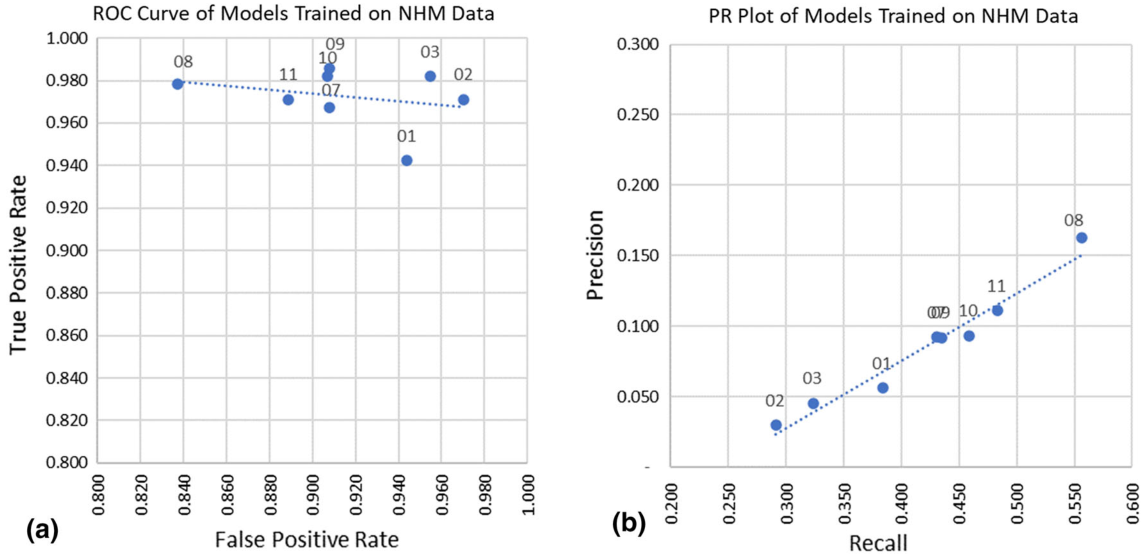


Fig. 21 ROC and PR plot of the models produced when training with the NHM dataset

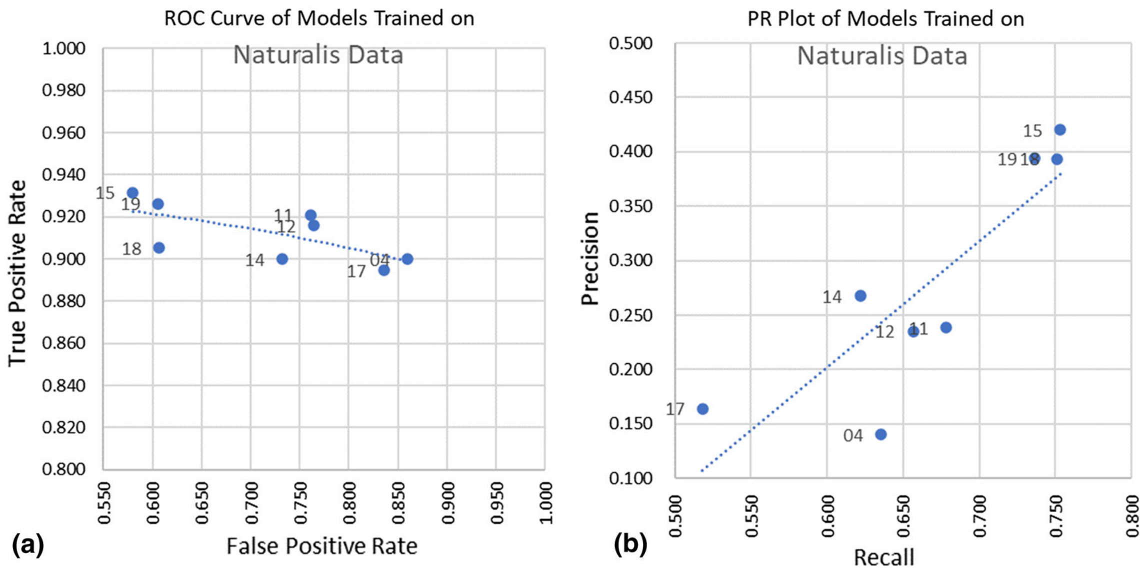


Fig. 22 ROC and PR plot of the models produced when training with the Naturalis dataset

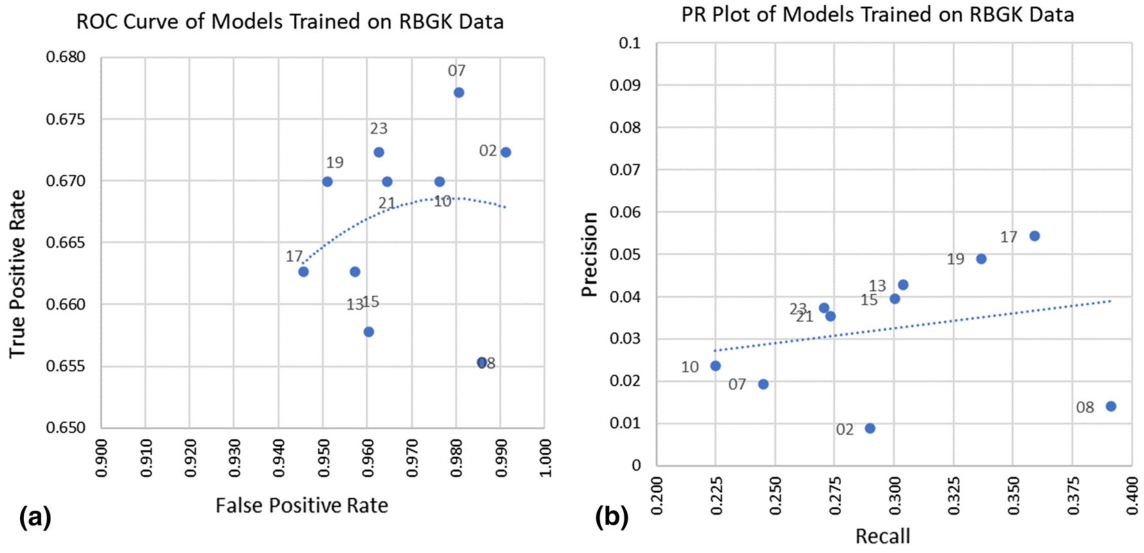


Fig. 23 ROC and PR plots of the models produced when training with the RBGK dataset

Table 15 Prediction on the RBGK set using YOLO V3 trained with four classes

Element	TP	FP	TN	FN	ACC	TPR	FPR	PREC
Label	0	0	0	81	0.000	0.000	0.000	0.000
Type Label	0	0	72	3	0.960	0.000	0.000	0.000
Specimen	60	2	0	178	0.250	0.252	1.000	0.968
Barcode	41	3	0	34	0.526	0.547	1.000	0.932
Total	101	5	72	296	0.365	0.254	0.065	0.953

proposals (such as the one proposed by [35, 36]). To facilitate this, each of the published datasets includes a csv file which contains the ground truths as coordinates. These can, for example, be used for evaluating other segmentation methods such as YOLO or a R-CNN (as shown in section 0).

7.2 Conclusions

The evaluation of the NHM-SSN segmentation network illustrates a viable proposal for determining whether a segmentation service API could be integrated into larger image

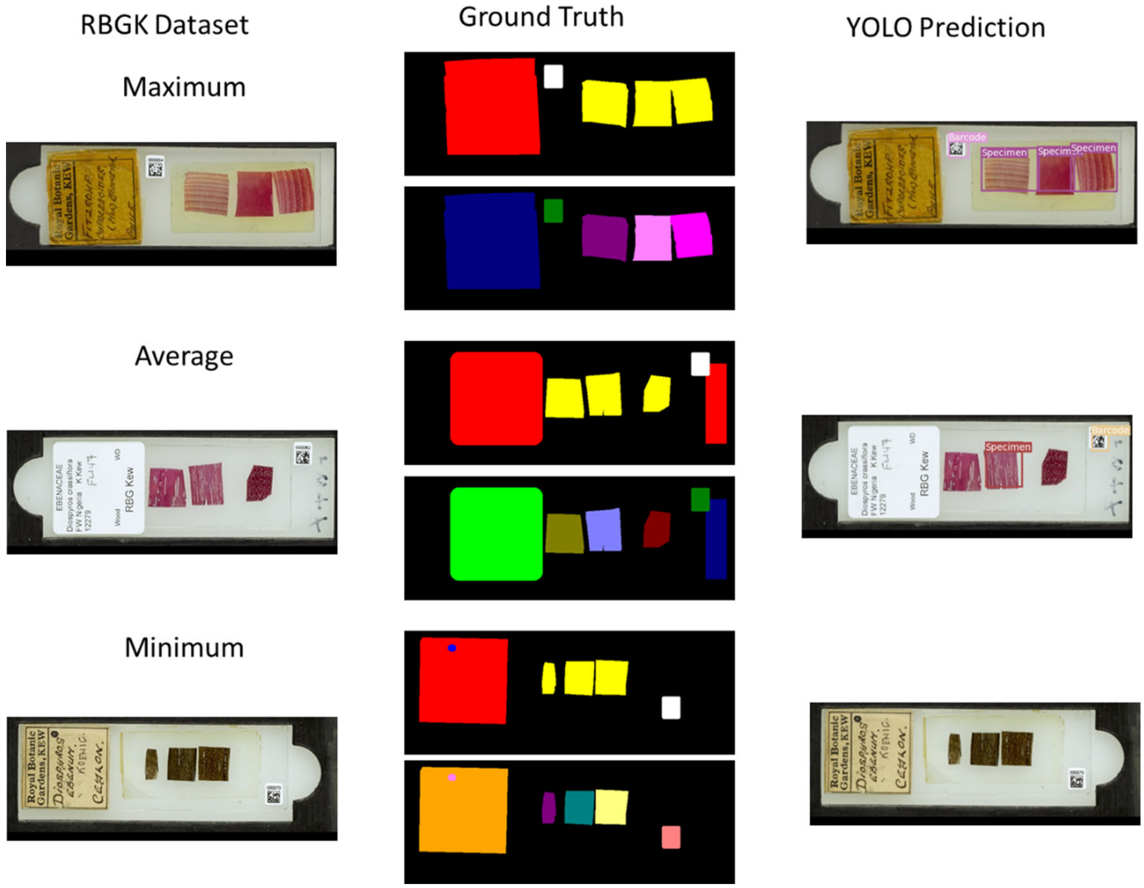


Fig. 24 RBGK dataset segmentation results: specimen images (first column), ground truth images for each specimen (classes and instances, second column), results of segmentation with the YOLO V3 (third column)

processing workflows for natural history collections. The results from the application of the methodology to two different types of collections (herbarium sheets and microscope slides) can be interpreted as the validation of the portability of the segmentation network and its potential for use in this context.

The initial results obtained pointed to the need for retraining the segmentation network each time a new source dataset is to be used. We carried a further experiment to compare individual training against training on a combined (larger) dataset, using the microscope slide datasets. The results are encouraging, showing that the NHM-SSN is resilient and adaptable for different collections.

Finally, the method and ground truth sets created for this work can be reused in testing other segmentation methods for other types of images, helping in the improvement of workflows for image processing in the context of digitization of natural history collections.

Acknowledgements We would like to thank our colleagues who declined participating as coauthors but helped us in obtaining datasets for the experiments particularly: Neil Raes and Wouter Addink from Naturalis, Vincent Smith and Matt Woodburn from the Natural History Museum, Pierre-Yves Gagnier from the Muséum National d'Histoire Naturelle and Hannu Saarenmaa from the Finnish Museum of Natural History. We also thank the reviewers for their thoughtful comments and efforts that helped us to improve this paper. We acknowledge the support of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government.

Funding ICEDIG Project was funded by the Horizon 2020 Framework Programme of the European Union. H2020-INFRADEV-2016–2017—Grant Agreement No. 777483.

Data availability The image datasets including ground truths, as well as the processed data and intermediate results are published as single package at: 10.5281/zenodo.4448186.

Code availability The code of the semantic segmentation network is published at: https://github.com/NaturalHistoryMuseum/semantic-segmentation. Additional processing code is available at: 10.5281/zenodo.4448207

Declarations

Conflict of interest All authors declare no conflicts of interest.

References

1. Allan, E., Dupont, S., Hardy, H., Livermore, L., Price, B., Smith, V.: High-throughput digitization of natural history specimens. Biodiversity Information Science and Standards **3**, e37337 (2019). https://doi.org/10.3897/biss.3.37337
2. Allan, E., Livermore, L., Price, B., Shchedrina, O., Smith, V.: A Novel Automated Mass Digitization Workflow for Natural History Microscope Slides. Biodiversity Data Journal **7**, e32342 (2019). https://doi.org/10.3897/BDJ.7.e32342
3. Bisong E. (2019) Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_7
4. Brown, P. A. (1997). A Review of Techniques Used in the Preparation, Curation and Conservation of Microscope Slides at the Natural History Museum, London. The Biology Curator, 10 - Supplement, 1 - 4. URL: http://www.natsca.org/article/455
5. Can, Y.S., Kabadayı, M.E.: CNN-Based Page Segmentation and Object Classification for Counting Population in Ottoman Archival Documentation. Journal of Imaging **6**(5), 32 (2020)
6. Carranza-Rojas, J., Goeau, H., Bonnet, P., Mata-Montero, E., & Joly, A. (2017). Going deeper in the automated identification of Herbarium specimens. BMC Evolutionary Biology, 17(1), 181. https://bmcevolbiol.biomedcentral.com/articles/https://doi.org/10.1186/s12862-017-1014-z
7. Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.
8. De Brabandere, B., Neven, D., & Van Gool, L. (2017). Semantic instance segmentation with a discriminative loss function. arXiv preprint arXiv:1708.02551.
9. Dillen, M., Groom, Q., Chagnoux, S., Güntsch, A., Hardisty, A., Haston, E., Livermore, L., Runnel, V., Schulman, L., Willemse, L., Wu, Z., Phillips, S.: A benchmark dataset of herbarium specimen images with label data. Biodiversity Data Journal **7**, e31817 (2019). https://doi.org/10.3897/BDJ.7.e31817
10. Durrant, J., Livermore, L. (2018) Semi-supervised semantic and instance segmentation, Labelling Training Data. Retrieved on 2018–02–13, from: https://github.com/NaturalHistoryMuseum/semantic-segmentation/wiki/Labelling-training-data
11. Gaikwad, J., Triki, A., Bouaziz, B.: Measuring Morphological Functional Leaf Traits From Digitized Herbarium Specimens Using TraitEx Software. Biodiversity Information Science and Standards **3**, e37091 (2019). https://doi.org/10.3897/biss.3.37091
12. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778). https://arxiv.org/abs/1512.03385
13. Hudson, L. N., Blagoderov, V., Heaton, A., Holtzhausen, P., Livermore, L., Price, B. W., ... & Smith, V. S. (2015). Inselect: automating the digitization of natural history collections. PLoS one, 10(11), e0143402.
14. JSTOR (2018). JSTOR Global Plants: Guidelines for Scanning Specimens. From: https://guides.jstor.org/ld.php?content_id=31764146
15. JSTOR (2018). JSTOR Plants Handbook. From http://www.snsb.info/SNSBInfoOpenWiki/attach/Attachments/JSTOR-Plants-Handbook.pdf
16. Kirchhoff, A., et al.: Toward a service-based workflow for automated information extraction from herbarium specimens. Database **2018**, 1–11 (2018). https://doi.org/10.1093/database/bay103
17. Kumar, S.S., Rajendran, P., Prabaharan, P., Soman, K.P.: Text/Image Region Separation for Document Layout Detection of Old Document Images Using Non-linear Diffusion and Level Set. Procedia Computer Science **93**, 469–477 (2016)
18. Livermore, L., et.al. (2017) Digitising Louse Slides. NERC / Natural History Museum pilot project. http://www.nhm.ac.uk/our-science/our-work/digital-museum/digital-collections-programme/digitising-slide-collections.html
19. Lorieul, T., Pearson, K.D., Ellwood, E.R., Goëau, H., Molino, J.F., Sweeney, P.W., Soltis, P.S.: Toward a large-scale and deep phenological stage annotation of herbarium specimens: Case studies from temperate, tropical, and equatorial floras. Applications in plant sciences **7**(3), e01233 (2019). https://doi.org/10.1002/aps3.1233
20. Ma, K., Shu, Z., Bai, X., Wang, J., & Samaras, D. (2018). DocUNet: document image unwarping via a stacked U-Net. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700–4709).
21. Mehri, M., Gomez-Krämer, P., Héroux, P., & Mullot, R. (2013). Old document image segmentation using the autocorrelation function and multiresolution analysis. In Document Recognition and Retrieval XX (Vol. 8658, p. 86580K). International Society for Optics and Photonics.
22. Meise Botanic Garden (2018) Botanical Collections Virtual Herbarium digital specimens cited where used.
23. Natural History Museum (2018) The Natural History Museum Data Portal, digital specimens cited where used.
24. Naturalis Biodiversity Center (2018) BioPortal, the Data Portal of the Naturalis Biodiversity Center, digital specimens cited where used.
25. Nieva de la Hidalga, A., Owen, D., Spasic, I., Rosin, P., Sun, X.: Use of Semantic Segmentation for Increasing the Throughput of Digitization Workflows for Natural History Collections. Biodiversity Information Science and Standards **3**, e37161 (2019). https://doi.org/10.3897/biss.3.37161
26. Okun, O., Dœrmann, D., & Pietikainen, M. (1999). Page segmentation and zone classification: the state of the art. OULU UNIV (FINLAND) DEPT OF ELECTRICAL ENGINEERING.
27. Owen, D., Groom, Q., Hardisty, A., Leegwater, T., van Walsum, M., Wijkamp, N., Spasić, I.: Methods for Automated Text Digitisation. Zenodo (2019). https://doi.org/10.5281/zenodo.3364501
28. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., & Raiko, T. (2015). Semi-supervised learning with ladder networks. In Advances in neural information processing systems (NIPS 2015). pp. 3546–3554. http://papers.nips.cc/paper/5947-semi-supervised-learning-with-ladder-networks.pdf

29. Rouhan, G., Chagnoux, S., Dennetière, B., Shchäfer, V., & Pignal, M. (2016). The herbonauts website: Recruiting the general public to acquire the data from herbarium labels. Botanists of the Twenty-First Century: Roles, Challenges and Opportunities. United Nations Educational, Scientific and Cultural Organisation, 143–148.
30. Scharr, H., Minervini, M., French, A.P., Klukas, C., Kramer, D.M., Liu, X., Yin, X.: Leaf segmentation in plant phenotyping: a collation study. Mach. Vis. Appl. **27**(4), 585–606 (2016). https://doi.org/10.1007/s00138-015-0737-3
31. Shafait, F. (2008). Geometric Layout Analysis of scanned documents.Doctoral Thesis, Technical University of Kaiserslautern.
32. Smith, V.S., Gorman, K., Addink, W., Arvanitidis, C., Casino, A., Dixey, K., Dröge, G., Groom, Q., Haston, E.M., Hobern, D., Knapp, S., Koureas, D., Livermore, L., Seberg, O.: SYNTHESYS+ Abridged Grant Proposal. Research Ideas and Outcomes **5**, e46404 (2019). https://doi.org/10.3897/rio.5.e46404
33. Soltis, P.S., Nelson, G., James, S.A.: Green digitization: Online botanical collections data answering real-world questions. Applications in Plant Sciences **6**(2), e1028 (2018). https://doi.org/10.1002/aps3.1028
34. Saito, T., Rehmsmeier, M.: (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE **10**(3), e0118432 (2015). https://doi.org/10.1371/journal.pone.0118432
35. Triki, A., Bouaziz, B., & Gaikwad, J. (2018) Refined Methodology for Accurately Detecting Objects from Digitized Herbarium Specimens. In ICEI 2018: 10th International Conference on Ecological Informatics-Translating Ecological Data into Knowledge and Decisions in a Rapidly Changing World. https://www.db-thueringen.de/receive/dbt_mods_00037908
36. Triki, A.; Bouaziz, B.; Mahdi, W. and Gaikwad, J. (2020). Objects Detection from Digitized Herbarium Specimen based on Improved YOLO V3.In Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, ISBN 978–989–758–402–2, pages 523–529. DOI: https://doi.org/10.5220/0009170005230529
37. White, A., Trizna, M., Frandsen, P., Dorr, L., Dikow, R., Schuettpelz, E.: Evaluating Geographic Patterns of Morphological Diversity in Ferns and Lycophytes Using Deep Neural Networks. Biodiversity Information Science and Standards **3**, e37559 (2019). https://doi.org/10.3897/biss.3.37559
38. Willis, C.G., Ellwood, E.R., Primack, R.B., Davis, C.C., Pearson, K.D., Gallinat, A.S., Yost, J.M., et al.: Old plants, new tricks: Phenological research using herbarium specimens. Trends Ecol. Evol. **32**, 531–546 (2017)

Abraham Nieva de la Hidalga is a postdoctoral research associate in data management and software development at the UK Catalysis Hub, contributing to the design of the Catalysis Data Infrastructure and the Catalysis Research Workbench. His work in the ICEDIG project included criteria for evaluation of image quality for image acquisition workflows, criteria for evaluation 3D modelling of specimens and use of semantic segmentation for processing specimen images.

Paul L. Rosin is a Professor at the School of Computer Science & Informatics, Cardiff University. He has more than 30 years of experience working on computer vision, covering areas such as low-level image processing, performance evaluation, shape analysis, facial analysis, medical image analysis, surveillance, 3D mesh processing, cellular automata and non-photorealistic rendering, as well as multidisciplinary collaborations (medical, social, geology, and cultural heritage).

Xianfang Sun received a PhD degree in control theory and its applications from the Institute of Automation, Chinese Academy of Sciences. He is a senior lecturer at Cardiff University. His research interests include computer vision and graphics, pattern recognition and artificial intelligence, system identification, and control. He is on the editorial board of Acta Aeronautica et Astronautica Sinica.

Laurence Livermore is the digital project manager at the Natural History Museum, specialized in digitization, biodiversity informatics and data. He has a background in entomology and a subject editor (Hemiptera: Heteroptera) for the Biodiversity data Journal and Zootaxa. Currently, he works on the development of the museum's standards and policy for mass digitization, delivery of rapid data entry web applications, and crowdsourcing technology for natural and cultural collections.

James Durrant is the lead design engineer at Synthesia. Previously, he worked at the Natural History Museum where he collaborated in different collection digitisation projects, including the development of the natural history museum semantic segmentation network.

James Turner is the manager of the Natural Sciences imaging laboratory at the National Museum Wales. He is interested in scientific imaging using advanced digital imaging techniques including photomicroscopy, scanning electron microscopy (SEM) and 3D scanning and modelling and also involved in design and development of web-based applications for Natural Sciences research and curation. He is interested in the application of new imaging techniques for natural science research collections, and the implementation of large-scale digitisation workflows.

Mathias Dillen is part of the Biodiversity Informatics team at Meise Botanic Garden in Belgium. He has worked on research projects dealing with the digitisation, standardisation and publication of biodiversity collection data, in particular in the context of the European DiSSCo research infrastructure and on topics concerning semantic interoperability and enrichment.

Alicia Musson oversees the digitising microscope slide collection at Royal Botanic Gardens, Kew. Included in the work so far are specimens of wood, leaf, flower, and pollen from a diverse range of plant taxa.

Sarah Phillips is the research leader for Digital Collections, leading the team developing and improving access to Kew's Herbarium Collections through digitisation. Her aim is to develop innovative methods to accelerate the rate of specimen imaging and label data capture by trialling different workflows. Her team also curates the herbarium specimen catalogue and manages digitisation projects.

Quentin Groom leads the Biodiversity Informatics team at Meise Botanic Garden in Belgium. He works at the interface of botany and information technology, particularly in relation to invasive species and the digitisation of biodiversity data. He often focuses on issues related to the interoperability of data, including biodiversity data standards and repeatable workflows.

Alex Hardisty recently retired Director of Informatics Projects at School of Computer Science and Informatics, Cardiff University. He is interested in bio-/geodiversity informatics, engineering large-scale distributed data management and processing systems, virtual research environments and socio-technical issues of technology adoption. Before retiring, he led the work on open Digital Specimens (openDS), Minimum Information about Digital Specimens/Collections (MIDS/MICS) and FAIR Digital Objects for DiSSCo.