Journal of Machine Learning for Biomedical Imaging. 2022:007. pp 1–27 Submitted 09/2021; Published 03/2022
Special Issue: Information Processing in Medical Imaging (IPMI) 2021
Guest Editors: Aasa Feragen, Stefan Sommer, Mads Nielsen, Julia Schnabel

Continual Active Learning Using Pseudo-Domains for Limited Labelling Resources and Changing Acquisition Characteristics

Matthias Perkonigg matthias.perkonigg@meduniwien.ac.at
Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab (CIR), Medical University of Vienna, Austria

Johannes Hofmanninger
Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab (CIR), Medical University of Vienna, Austria

Christian Herold
Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Austria

Helmut Prosch
Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Austria

Georg Langs georg.langs@meduniwien.ac.at
Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab (CIR), Medical University of Vienna, Austria

Abstract

Machine learning in medical imaging during clinical routine is impaired by changes in scanner protocols, hardware, or policies resulting in a heterogeneous set of acquisition settings. When training a deep learning model on an initial static training set, model performance and reliability suffer from changes of acquisition characteristics as data and targets may become inconsistent. Continual learning can help to adapt models to the changing environment by training on a continuous data stream. However, continual manual expert labelling of medical imaging requires substantial effort. Thus, ways to use labelling resources efficiently on a well chosen sub-set of new examples is necessary to render this strategy feasible. Here, we propose a method for continual active learning operating on a stream of medical images in a multi-scanner setting. The approach automatically recognizes shifts in image acquisition characteristics – new *domains* –, selects optimal examples for labelling and adapts training accordingly. Labelling is subject to a limited budget, resembling typical real world scenarios. In order to avoid catastrophic forgetting while learning on new domains the proposed method utilizes a rehearsal memory. To demonstrate generalizability, we evaluate the effectiveness of our method on three tasks: cardiac segmentation, lung nodule detection and brain age estimation. Results show that the proposed approach outperforms other active learning methods on a continuous data stream with domain shifts.

Keywords: Continual learning, Active learning, Domain adaptation.

1. Introduction

The performance of deep learning models in the clinical environment is hampered by frequent changes in scanner hardware, imaging protocols, and heterogeneous composition of acquisition routines. Ideally, models trained on a large data set should be continuously adapted to the changing characteristics of the data stream acquired in imaging depart-

https://www.melba-journal.org/papers/2022:007.html.

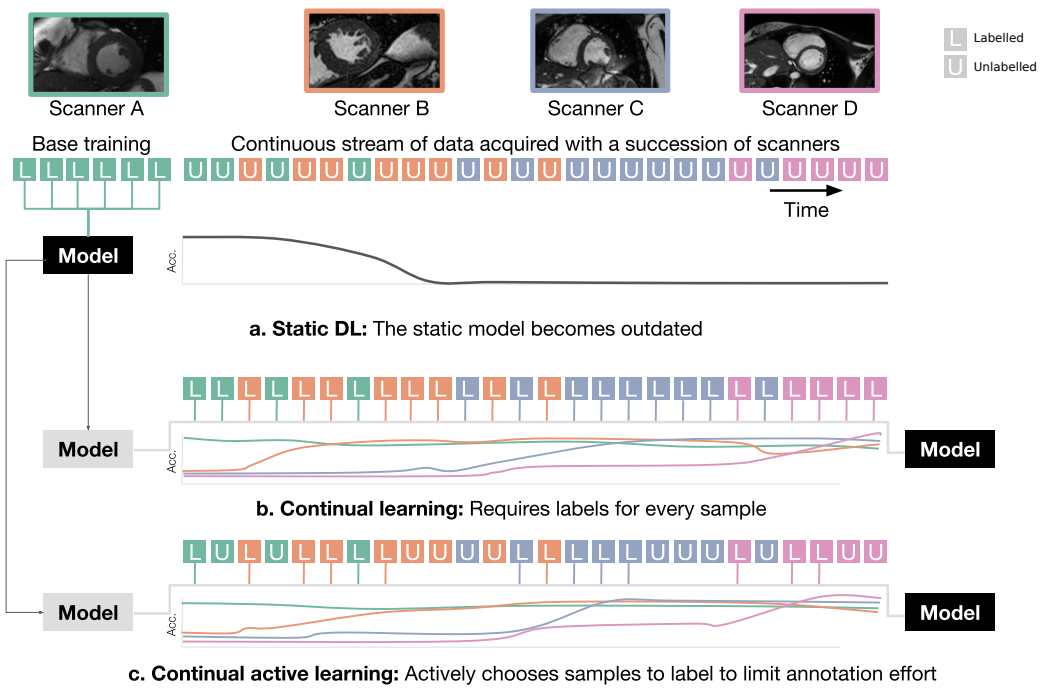

Figure 1: Experimental setup: A model is pre-trained on scanner A data (base training) and then subsequently updated on a data stream gradually including data of scanner B, C and D. (a) The accuracy of a model trained on a static data set of only scanner A drops as data from other scanner appear in the data stream. (b) Continual learning can incorporate new knowledge, but requires all samples in the data stream to be labelled. (c) Active continual learning actively chooses the labels to annotated from the stream and is able to keep the model up to date while limiting the annotation effort.

ments. However, training on a data stream of images acquired solely by recent acquisition technology can lead to *catastrophic forgetting* (McCloskey and Cohen, 1989), a deterioration of performance on preceding domains or tasks, see Figure 1 (a). Therefore, a continual learning strategy is required to counteract forgetting. Counteracting forgetting is important in medical imaging to ensure backward compatibility of the model, as well as to enable faster adaptation to possibly related domains in the future. Model training in a medical context requires expert labelling of data in new domains. This is often prohibitively expensive and time-consuming. Therefore, reducing the number of cases requiring labelling, while still providing training with the variability necessary to generalize well, is a key challenge in active learning on medical images (Budd et al., 2019). Here, we propose an active learning approach to make efficient use of annotation resources during continual machine learning. In a continual data stream of examples from an unlabelled distribution, it identifies those that are most informative if labelled next.

We focus on accounting for domain shifts occurring in a continual data stream, without knowledge about when those shifts occur. Figure 1 depicts the scenario our method is designed for. A deep learning model is trained on a base data set of labelled data of one

domain (Scanner A), afterwards it is exposed to the data stream in which scanners B, C and D occur. For each sample of the data stream, continual active learning has to take a decision on whether or not labelling is required for the given image. Labelled images are then used for continual learning with a rehearsal memory. Previously proposed continual active learning methods either disregard domain shifts in the training distribution or assume that the domain membership of images is known (Lenga et al., 2020; Özgün et al., 2020; Karani et al., 2018). However, this knowledge can not be assumed in clinical practice due to the variability in encoding the meta data (Gonzalez et al., 2020). Therefore, a technique to detect those domain shifts in a continuous data stream is needed. A combination of continual active learning with automatic detection of domain shifts is desirable to ensure that models can deal with a diverse and growing number of image acquisition settings, and at the same time minimizing the manual efforts and resources needed to keep the models up to date.

Contribution Here, we propose an continual active learning method. The approach operates without domain membership knowledge and learns by selecting informative samples to annotate from a continuous data stream. We first run a base training on data of a single scanner. Subsequently, the continuous data stream is observed and domain shifts in the stream are detected. This detection triggers the labelling of samples of the newly detected pseudo-domain and knowledge about the new samples is incorporated to the model. At the same time, the model should not forget knowledge about previous domains, thus we evaluate the final model on data from all observed domains. Our approach combines continual active learning with a novel domain detection method for continual learning. We refer to our approach as *Continual Active Learning for Scanner Adaptation (CASA)*. CASA uses a rehearsal method to alleviate catastrophic forgetting and an active labelling approach without prior domain knowledge. CASA is designed to learn on a continuous stream of medical images under the restriction of a labelling budget, to keep the required manual annotation effort low. The present paper expands on our prior work on continual active learning (Perkonigg et al., 2021b). In this prior work we introduced the novel setup on active and continual learning on a data stream of medical imaging and proposed the CASA method. Here, we expand the approach in several ways: (1) The pseudo-domain assignment is refined and simplified. While previous work used a method based on isolation forests (Liu et al., 2008), in this work we use a distance metric for pseudo-domain assignment. (2) Experiments with two additional machine learning tasks, cardiac segmentation in MR imaging and lung nodule detection in CT are included to demonstrate the generalizability of CASA. (3) Active learning with uncertainty is added as a reference method across all experiments to compare the performance of CASA. (4) A more detailed analysis of the composition of the rehearsal memory, and the influence of the sequential nature of the data stream are included.

2. Related Work

The performance of machine learning models can be severely hampered by changes in image acquisition settings (Castro et al., 2020; Glocker et al., 2019; Prayer et al., 2021). *Harmonization* can counter this in medical imaging (Fortin et al., 2018; Beer et al., 2020), but requires all data to be available at once, a condition not feasible in an environment where

data arrives continually. *Domain adaptation (DA)* addresses domain shifts, and in particular approaches dealing with continuously shifting domains are related to the proposed method. Wu et al. (2019) showed how to adapt a machine learning model for semantic segmentation of street scenes under different lightning conditions. Rehearsal methods for domain adaptation have been shown to perform well on benchmark data sets such as rotated MNIST (Bobu et al., 2018) or Office-31 (Lao et al., 2020). In the area of medical imaging, DA is used to adapt between different image acquisition settings Guan and Liu (2021). However, similar to harmonization, most DA methods require that source and target domains are accessible at the same time.

Continual learning (CL) is used to incorporate new knowledge into ML models without forgetting knowledge about previously seen data. For a detailed review on CL see (Parisi et al., 2019; Delange et al., 2021). An overview of the potential of CL combined with medical imaging combined is given in (Pianykh et al., 2020). Ozdemir et al. (2018) used continual learning to incrementally add new anatomical regions into a segmentation model. Related to this work, CL has been used for domain adaptation for chest X-ray classification (Lenga et al., 2020) and for brain MRI segmentation (Özgün et al., 2020). Karani et al. (2018) proposed a simple, yet effective approach for lifelong learning for brain MRI segmentation by using separate batch normalization layers for each protocol. The rehearsal memory approach of our work is closely related to *dynamic memory*, a continual learning method based on an image style-based rehearsal memory (Hofmanninger et al., 2020; Perkonigg et al., 2021a). However dynamic memory assumes a fully labelled data set, while this approach limits the annotation need by using an active stream-based selective sampling method.

Active Learning is an area of research where the goal is to identify samples to label next to minimize annotation effort, while maximizing training efficiency. A detailed review of active learning in medical imaging is given in (Budd et al., 2019). In context of this review our work is closest related to *stream-based selective sampling*. Also Pianykh et al. (2020) discuss human-in-the-loop concepts with continual learning, which is similar to the approach presented in this work. Active learning was used to classify fundus and histopathological images by Smailagic et al. (2020) in an incremental learning setting. Zhou et al. (2021) combine transfer learning and active learning to choose samples for labelling based on entropy and diversity. They show the benefits of their method on polyp detection and pulmonary embolism detection. Different from the proposed method, those approaches do not take data distribution shifts during training into account and do not perform selective sampling based on a continuous data stream.

3. Methods

The continual active learning method CASA uses a *rehearsal memory* and performs active training sample selection from an unlabelled, continuous data stream $\mathcal{S}$ to keep a task model up-to-date under the presence of domain shifts, while at the same time countering forgetting. For active sample labelling an oracle can be queried to return task annotations $y = \mathbf{o}(x) \mid x \in \mathcal{S}$. In a real world clinical setting this oracle can be thought of as a radiologist. Due to the cost of manual labelling, the queries to the oracle are limited by the labelling budget β. CASA aims at training a task network on a continuous data stream under the restriction of β, while at the same time alleviating catastrophic forgetting. CASA detects

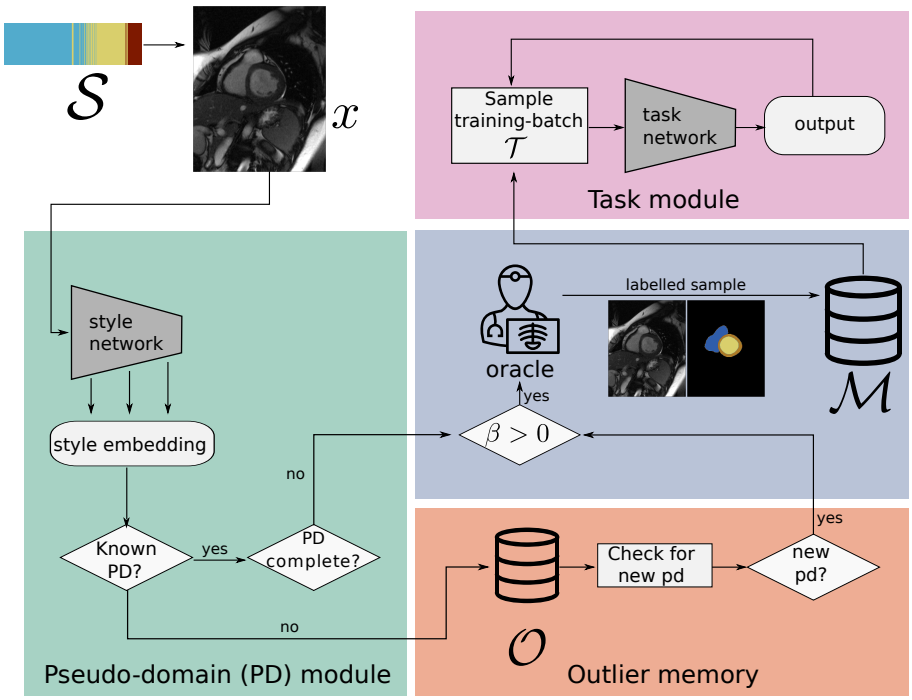

Figure 2: Overview of the *CASA* algorithm. Each sample from the data stream is processed by the pseudo-domain module to decide whether its routed to the oracle or to the outlier memory. Whenever a new item is added to the outlier memory it is evaluated if a new pseudo-domain (pd) should be created. The oracle labels a sample and stores it in the training memory, from which the task module trains a network. Binary decision alternatives resulting in discarding the sample are left out for clarity of the figure.

pseudo-domains to keep a diverse set of training samples in the rehearsal memory. Those pseudo-domains are formed by groups of examples with similar appearance. Similarity is measured as style difference of images. The proposed method consists of a pseudo-domain module, a task module and and two memories (outlier memory and training rehearsal memory), that are controlled by the CASA-Algorithm described in the following.

3.1 CASA Training Scheme

Before starting continual training, the task module is pre-trained on a labelled data set $\mathcal{L} = \{\langle i_1, l_1\rangle, \ldots, \langle i_L, l_L\rangle\}$ of image-label pairs $\langle i, l\rangle$ obtained on a particular scanner (*base training*). This base training is a conventional epoch based training procedure assuming a static, fully labelled data set. In this training phase samples can be revisited in a supervised training scheme without restriction of the labelling budget. After base training is finished, continual training is started from a model which performs well on data of a single scanner. Continual active training follows the scheme depicted in Figure 2 and outlined in Algorithm 1. First, an *input-mini-batch* $\mathcal{B} = \{x_1, \ldots, x_B\}$ is drawn from $\mathcal{S}$ and the *pseudo-domain module* evaluates the style embedding (see Section 3.2) of each image $x \in \mathcal{B}$. Based on this embedding, a decision is taken to store x in one of the memories ($\mathcal{O}$ or $\mathcal{M}$) or discard

x. The fixed sized *training memory* $\mathcal{M} = \{\langle m_1, n_1, d_1\rangle \ldots, \langle m_M, n_M, d_M\rangle\}$, where m is the image, n the corresponding label and d the assigned *pseudo-domain*, holds samples the task network can be trained on. Labels n can only be generated by querying the *oracle* $\mathbf{o}(x)$, and are subject to the limited labelling budget β. $\mathcal{M}$ is initialized with a random subset of $\mathcal{L}$ before starting continual training. Pseudo-domain detection is performed within the *outlier memory* $\mathcal{O} = \{\langle o_1, c_1\rangle \ldots, \langle o_n, c_n\rangle\}$, which holds a set of unlabelled images. o represents the image and c is a counter, how long the image is part of $\mathcal{O}$. Details about the outlier memory are given in Section 3.5. Given that training has not saturated on all pseudo-domains, a training step is performed by sampling *training-mini-batches* $\mathcal{T} = \{\langle t_1, u_1\rangle, \ldots, \langle t_T, u_T\rangle\}$ of size T from $\mathcal{M}$, and training the task module (Section 3.3) for one step. This process is continued by drawing the next mini-batch $\mathcal{B}$ from $\mathcal{S}$.

3.2 Pseudo-domain module

CASA does not assume direct knowledge about domains (e.g. scanner vendor, scanning protocol). Therefore, in the pseudo-domain module the *style* of each image x is evaluated and x is assigned to a pseudo-domain. Pseudo-domains represent groups of images which exhibit similar style. A set of pseudo-domains $\mathcal{D} = \{\langle c_1, d_1, \bar{p}_1\rangle \ldots \langle c_D, d_D, \bar{p}_D\rangle\}$ is defined by their *style embedding* center c_j and the maximum distance d_j from c_j that an image is considered belonging to pseudo-domain j. In addition, a running average $\bar{p}_j$ of the performance on j for each pseudo-domain $j \in \{1, \ldots, D\}$ is stored.

Style embedding A style embedding is calculated for an image x based on a *style network* pre-trained on a different dataset (not necessarily related to the task) and not updated during training. The choice of the style network is dependent on the dataset used for training, the specific style networks used in this paper are discussed in Section 4.2. From this network, we evaluate the style of an image based on the gram matrix $G^l \in \mathbb{R}^{N_l \times N_l}$, where N_l is the number of feature maps in layer l. Following (Gatys et al., 2016; Hofmanninger et al., 2020), $G^l_{ij}(x)$ is defined as the inner product between the vectorized activations $\mathbf{f}_{il}(x)$ and $\mathbf{f}_{jl}(x)$ of two feature maps i and j in a layer l given a sample image x:

$$G^l_{ij}(x) = \frac{1}{N_l M_l} \mathbf{f}_{il}(x)^\top \mathbf{f}_{jl}(x) \tag{1}$$

where M_l denotes the number of elements in the vectorized feature map. Based on the gram matrix a *style embedding* $\mathbf{e}(x)$ is defined: For a set of convolutional layers $\mathcal{C}$ of the style network, gram matrices $(G^l \mid l \in \mathcal{C})$ are calculated and Principle Component Analysis (PCA) is applied to reduce the dimensionality of the embedding to a fixed size of e. PCA is fitted on style embeddings of the base training set.

Pseudo-domain assignment CASA uses pseudo-domains to assess if training for a specific style is needed and to diversify the memory $\mathcal{M}$. A new image $x \in \mathcal{B}$ is assigned to the pseudo-domain minimizing the distance between the center of the pseudo-domain and the style embedding $\mathbf{e}(x)$ according to the following equation:

$$\mathbf{p}(x) = \begin{cases} \arg\min_d |\mathbf{e}(x) - c_d| & \text{if } |\mathbf{e}(x) - c_d| < d_d \\ -1, & \text{otherwise} \end{cases} \mid d \in \mathcal{D} \tag{2}$$

Algorithm 1: CASA Training Algorithm

Input : Pre-trained task model t, continual data stream $\mathcal{S}$, limited budget β, task-memory $\mathcal{M}$, outlier memory $\mathcal{O}$, b training steps per batch

while $\mathcal{B} \leftarrow$ `nextBatch`$(\mathcal{S})$ **do**
  for $x \in \mathcal{B}$ **do**
    $e \leftarrow$ `styleembedding`(x)
    $pd \leftarrow$ `pd-assignment`(e)
    if $pd == -1$ **then**
      $\mathcal{O}$.`add`(x)
    end if
    else
      if `pd-complete` *(pd)* **then**
        if $\beta > 0$ **then**
          $\mathcal{M}$.`add`(x)
          $\beta \leftarrow \beta - 1$
        end if
      end if
    end if
  end for
  $\mathcal{N} \leftarrow$ `newPseudodomainCheck`$(\mathcal{O}, \beta)$; `/* elements of discovered pd */`
  if $\mathcal{N} \neq \emptyset$ **then**
    for $n \in \mathcal{N}$ **do**
      if $\beta > 0$ **then**
        $\mathcal{M}$.`add`(x)
        $\beta \leftarrow \beta - 1$
      end if
    end for
  end if
  for $i \leftarrow 0$ **to** b **do**
    $\mathcal{T} \leftarrow$ `sample`$(\mathcal{M})$
    `train`$(t, \mathcal{T})$
  end for
end while

If $\mathbf{p}(x) = -1$, the image is added to the outlier memory $\mathcal{O}$ from which new pseudo-domains are detected (see Section 3.5). For the threshold distance d_d, let $\mathcal{M}_d$ be the subset of samples assigned to domain d. Then d_d is calculated as two times the mean distance between the center of d and the style embedding of all samples in $\mathcal{M}_d$:

$$d_d = 2 \cdot \frac{\sum_{x \in \mathcal{M}_d} (c_d - \mathbf{e}(x))^2}{|\mathcal{M}_d|} \tag{3}$$

If the pseudo-domain $\mathbf{p}(x)$ is known and has completed training, we discard the image, otherwise it is added to $\mathcal{M}$ according to the strategy described in Section 3.4.

Average performance metric $\bar{p}_j$ is the running average of a performance metric of the target task calculated on the last P elements of pseudo-domain j that have been labelled by the *oracle*. The performance metric is measured before training on the sample. $\bar{p}_j$ is used to evaluate if the pseudo-domain completed training, that is $\bar{p}_j > k$ for classification tasks and $\bar{p}_j < k$ for regression tasks, where k is a fixed performance threshold. If that threshold is reached, subsequent samples assigned to the corresponding pseudo-domain do not require manual annotation. The specific choice of the performance metric depends on the learning task, see Section 4.2 for the metrics used in the experiments of this work.

3.3 Task module

The task module is responsible for learning the target task (e.g. cardiac segmentation), where the main component of this module is the *task network* ($\mathbf{t}(x) \mapsto y$), mapping from input image x to target label y. During base training, this module is trained on a labelled data set $\mathcal{L}$. During continual active training, the module is updated in every step by drawing n training-input-batches $\mathcal{T}$ from the memory $\mathcal{M}$ and performing a training step on each of the batches. The aim of CASA is to train a task module performing well on images of all image acquisition settings available in $\mathcal{S}$ without suffering catastrophic forgetting.

3.4 Training memory

The M sized training memory $\mathcal{M}$ is balanced between the pseudo-domains currently in $\mathcal{D}$. Each of the D pseudo-domains can occupy up to $\frac{M}{D}$ elements in the memory. If a new pseudo-domain is added to $\mathcal{D}$ (see Section 3.5) a random subset of elements of all previous domains is flagged for deletion, so that only $\frac{M}{D}$ elements are kept protected in $\mathcal{M}$. If a new element $e = \langle m_k, n_k, d_k \rangle$ is inserted to $\mathcal{M}$ and $\frac{M}{D}$ is not reached, an element currently flagged for deletion is replaced by e. Otherwise the element will replace the one in $\mathcal{M}$, which is of the same pseudo-domain and minimizes the distance between the style embeddings. Formally, the element with index ξ is replaced:

$$\xi(i) = \arg\min_j (\mathbf{e}(m_k) - \mathbf{e}(m_j))^2 |\ n_k = n_j,\ j \in \{1, \ldots, M\}. \tag{4}$$

3.5 Outlier memory and pseudo-domain detection

The outlier memory $\mathcal{O}$ holds candidate examples that do not fit an already identified pseudo-domain, and might form a new pseudo-domain by themselves. Whether they form a pseudo-domain is determined based on their proximity in the style embedding space. Examples

are stored until they are assigned a new pseudo-domain, or if a fixed number of training steps z is reached. If no new pseudo-domain is discovered for an image within z steps, it is considered a 'real' outlier and removed from the outlier memory. Within $\mathcal{O}$, new pseudo-domains are discovered, and subsequently added to $\mathcal{D}$. The discovery process is started when $|\mathcal{O}| = o$, where o is a fixed threshold of minimum elements in $\mathcal{O}$. To detect a dense region in the style embedding space of samples in the outlier memory, the pairwise euclidean distances of all elements in $\mathcal{O}$ are calculated. If there is a group of images for which all pair-wise distances are below a threshold t, a new pseudo-domain is established by these images. For all elements belonging to the new pseudo-domain, labels are queried from the oracle and they are transferred from $\mathcal{O}$ to $\mathcal{M}$.

4. Experimental Setup

We evaluate CASA on data streams containing imaging data sampled from different scanners. To demonstrate the generalizability of CASA to a range of different areas in medical imaging, three different tasks are evaluated:

- *Cardiac segmentation* on cardiovascular magnetic resonance (CMR) data
- *Lung nodule detection* in computed tomography (CT) images of the lung
- *Brain Age Estimation* on T1-weighted MRI data

For all tasks, the performance of CASA is compared to several baseline techniques (see Section 4.3).

4.1 Data set

Cardiac segmentation 2D cardiac segmentation experiments were performed on data from a multi-center, multi-vendor challenge data set (Campello et al., 2021). The data set included CMR data from four different scanner vendors (Siemens, General Electric, Philips and Canon), where we considered each vendor as a different domain. We split the data into base training, continual training, validation, and test set on a patient level. Table 1 (a) shows the number of slices for each domain in those splits. Manual annotations for left ventricle, right ventricle and left ventricular myocardium were provided. 2D images were center-cropped to 240×196px and normalized to a range of [0-1]. In the continual data set, the scanners appeared in the order Siemens, GE, Philips and Canon and are referred to Scanner C1-C4.

Lung nodule detection For lung nodule detection, we used two data sources: the LIDC-database (Armato et al., 2011), with the annotations as provided for the LUNA16-challenge (Setio et al., 2017) and the LNDb challenge data set (Pedrosa et al., 2019). Lung nodule detection was performed as 2D bounding box detection, therefore bounding boxes around annotated lesions were constructed for all available lung nodule annotations. From LIDC, the three most common domains, in terms of scanner vendor and reconstruction kernel, were used to collect a data set suitable for continual learning with shifting domains. Those domains were GE MEDICAL SYSTEMS with low frequency reconstruction algorithm (GE/L), GE MEDICAL SYSTEM with high frequency reconstruction algorithm

	Siemens (C1)	GE (C2)	Philips (C3)	Canon (C4)	Total
Base	1120	0	0	0	1120
Continual	614	720	2206	758	4298
Validation	234	248	220	258	960
Test	228	246	216	252	942

(a) Cardiac segmentation data set

	GE/L (L1)	GE/H (L2)	Siemens (L3)	LNDb (L4)	Total
Base	253	0	0	0	253
Continual	136	166	102	479	883
Validation	53	23	10	55	141
Test	85	26	18	91	220

(b) Lung nodule detection data set

	1.5T IXI (B1)	1.5T OASIS (B2)	3.0T IXI (B3)	3.0T OASIS (B4)	Total
Base	201	0	0	0	201
Continual	52	190	146	1504	1892
Validation	31	23	18	187	259
Test	31	23	18	187	259

(c) Brain age estimation data set

Table 1: Splitting of the data sets into a base training, continual training, validation, and test set. The number of cases in each split are shown.

(GE/H) and Siemens with B30f kernel (Siemens). In addition, data from LNDb was used as a forth domain which was comprised of data from multiple Siemens scanners. For LNDb, nodules with a diameter $< 3mm$ were excluded to match the definition in LIDC. Image intensities were cropped from -1024 to 1024 and normalized to [0-1]. From the images 2D slices were extracted and split into base training, continual training, validation and test data set according to Table 1 (b). For all continual learning experiments, the order of the domains was GE/L, GE/H, Siemens and LNDb, those are referred to as L1-L4.

Brain age estimation Data pooled from two different data sets containing three different scanners was used for brain age estimation. The IXI data set[1] and data from OASIS-3 (LaMontagne et al., 2019) was used to collect a continual learning data set. From IXI, we used data from a Philips Gyroscan Intera 1.5T and a Philips Intera 3.0T scanner, from OASIS-3 we used data from a Siemens Vision 1.5T and a Siemens TrioTim 3.0T scanner. Images were resized to 64x128x128px and normalized to a range between 0 and 1. Data was split into base base training, continual training, validation and test set (see Table 1 (c)). In continual training data occurred in the order: Philips 1.5T, Siemens 1.5T, Philips 3.0T and Siemens 3.0T, the scanner domains are referred to B1-B4 in the following.

4.2 Experimental setup

Hyperparameters Multiple hyperparamters are used in CASA. In the experiments the focus is to analyze those parameters with the most influence on the methods performance are the memory size M and the labelling budget β. These two parameters are extensively evaluated and analyzed in Section 5.3 and 5.2. Besides that the dimensions of the style embedding after PCA is fixed to $e = 30$ and the minimal number of elements in $\mathcal{O}$ to discover new pseudo-domains is fixed to $o = 10$ after preliminary experiments showed little influence on model performance of those parameters. Another hyperparameter is the choice of the style network that is fixed during training, here preliminary experiments showed that the choice of the style network is not critical for the performance of CASA. The performance threshold k depends on the task and performance metric used and is set to be approximately as high as the average performance of domain specific models (see Section 4.3). The threshold used t is dependent on the data set used, therefore the threshold was empirically set by analyzing mean distances of the style embeddings of the base training data set. Details on k and t are given in the following.

Cardiac segmentation For segmentation, a 2D-UNet (Ronneberger et al., 2015) was used as task network. The style network was a ResNet-50 (Ren et al., 2017), pretrained on ImageNet and provided in the torchvision package. For segmentation, the performance metric used in all experiments was the mean dice score (DSC) over the three annotated anatomical regions (left ventricle, right ventricle and left ventricular myocardium). The performance threshold k is fixed to a mean DSC of 0.75 based on domain specific models. The distance threshold was fixed to $t = 0.025$.

1. https://brain-development.org/ixi-dataset/

Lung nodule detection As a task network, Faster R-CNN with a ResNet-50 backbone was used (Ren et al., 2017). For evaluating the style, we used a ResNet-50 pretrained on ImageNet. For lung nodule detection, we used average precision (AP) as the performance metric to evaluate the performance of the models with a single metric. We followed the AP definition by Everingham et al. (2010). For lung nodule detection we set k to an AP of 0.5 and $t = 0.025$ for all experiments.

Brain age estimation As a task network, a simple 3D feed-forward network was used (Dinsdale et al., 2020). The style network used in the pseudo-domain module was a 3D-ModelGenesis model, pre-trained on computed tomography images of the lung (Zhou et al., 2020). For brain age estimation a 3D data set was used, thus we used a different style model as for cardiac segmentation and lung nodule detection. The main performance measure for brain age estimation we used was the mean absolute error (MAE) between predicted and true age. For brain age estimation we set k to a MAE to 5.0 and $t = 0.025$ for all experiments.

4.3 Methods compared

Throughout the experiments, five methods were evaluated and compared:

1. *Joint model (JM)*: a model trained in a standard, epoch-based approach on samples from all scanners in the experiment jointly.

2. *Domain specific models (DSM)*: a separate model is trained for each domain in the experiment with standard epoch-based training. The evaluation for a domain is done for each domain setting separately.

3. *Naive AL (NAL)*: a naive continuously trained, active learning approach of labelling every n-th label from the data stream, where n depends on the labelling budget β.

4. *Uncertainty AL (UAL)*: Is a common type of active learning which labels samples where the task network is uncertain about the output (Budd et al., 2019). Here, uncertainty is calculated using dropout at inference as an approximation for Bayesian inference (Gal and Ghahramani, 2016).

5. *CASA (proposed method)*: The method described in this work.

Joint models and DSM require the whole training data set to be labelled, and thus are an upper limit to which the continual learning methods are compared to. CASA, UAL and NAL use an oracle to label specific samples only. The comparison to NAL and UAL evaluates if the detection of pseudo-domains and labelling based on them is beneficial in an active learning setting. Note, that the aim of our experiments is to show the gains of CASA compared to other active learning methods, not to develop new state-of-the-art methods for either of the three tasks evaluated.

4.4 Experimental evaluation

We evaluate different aspects of CASA:

1. **Performance across domains:** For all tasks, we evaluate the performance across domains at the end of training, and highlight specific properties of CASA in comparison to the baseline methods. Furthermore, we evaluate the ability of continual learning to improve accuracy on existing domains by adding new domains *backward transfer* (BWT), and the contribution of previous domains in the training data to improving the accuracy on new domains *forward transfer* (FWT) (Lopez-Paz and Ranzato, 2017). BWT measure how learning a new domain influences the performance on previous tasks, FWT quantifies the influence on future tasks. Negative BWT values indicate catastrophic forgetting, thus avoiding negative BWT is especially important for continual learning.

2. **Influence of labelling budget β:** For cardiac segmentation, the influence of the β is studied. The labelling budget is an important parameter in clinical practice, since labelling new samples is expensive. We express β as a fraction of the continual data set. Different settings of β are analysed $\beta = \frac{1}{5}$, $\beta = \frac{1}{8}$, $\beta = \frac{1}{10}$ and $\beta = \frac{1}{20}$. To solely study the influence of β, the memory size in this experiments is fixed $M = 128$ for all settings.

3. **Influence of memory size M:** For cardiac segmentation different settings for the memory size M are evaluated. The memory size is the number of samples that are stored for rehearsal, and might be limited due to privacy concerns and/or storage space. Here, M is evaluated for $[64, 128, 256, 512, 1024]$ and a fixed $\beta = \frac{1}{10}$. We assume that the diversity of the data addressed by CASA is a result of the set of scanners, not the number of images in the dataset, therefore M is fixed to specific numbers rather than a fraction of the dataset.

4. **Memory composition and pseudo-domains:** We study if our proposed method of detecting pseudo-domains is keeping samples in memory that are representative of the whole training set for cardiac segmentation. In addition, we evaluate how the detected pseudo-domains are connected to the real domains determined by the scanner types.

5. **Learning on a random stream:** We study how CASA is performing on a random stream of data, where images of different acquisition settings are appearing randomly in the data stream. In contrast to the standard setting, where these acquisition settings appear subsequently with a phase of transition in between.

5. Results

5.1 Performance across domains

Here, the quantitative results at the end of the continual training are compared for a memory size $M = 128$ and a labelling budget of $\beta = \frac{1}{10}$. Different settings for M and β are evaluated in Section 5.2 and 5.3 respectively.

Cardiac segmentation Performance for cardiac segmentation was measured using the mean dice score. Continual learning with CASA applied to cardiac segmentation outperformed UAL and NAL for scanners C2, C3 and C4 (Table 2). For scanner C1, the performance of the model trained with CASA was slightly below UAL and NAL. This was

due to the distribution in the rehearsal memory, where CASA balanced between all four scanner domains, while for UAL and NAL, a majority of the rehearsal memory was filled with C1 images (further details are discussed in Section 5.4). Compared to the base model, which corresponds to the model performance prior to continual training the performance of CASA remained constant for C1 and at the same time rose significantly for C2 (+0.041), C3 (+0.085) and C4 (+0.336), showing that CASA was able to perform continual learning without forgetting the knowledge acquired in base training. UAL and NAL were also able to learn without forgetting during continual learning, this is also reflected in a BWT of around 0 for all compared methods. However, UAL and NAL performed worse in terms of FWT and overall dice for C2 to C4. As expected, JModel outperformed all other training strategies since it has access to the fully labelled training set at once and thus can perform epoch-based deep learning.

Meth.	C1	C2	C3	C4	BWT	FWT
CASA	0.812 ± 0.017	0.731 ± 0.025	0.803 ± 0.015	0.676 ± 0.158	-0.006 ± 0.009	0.086 ± 0.046
UAL	0.816 ± 0.006	0.700 ± 0.009	0.764 ± 0.023	0.652 ± 0.078	-0.003 ± 0.013	0.067 ± 0.031
NAL	0.819 ± 0.003	0.707 ± 0.005	0.761 ± 0.013	0.564 ± 0.064	-0.004 ± 0.003	0.060 ± 0.026
DSM	0.835 ± 0.047	0.718 ± 0.018	0.773 ± 0.016	0.833 ± 0.003		
JModel	0.828 ± 0.009	0.758 ± 0.020	0.818 ± 0.023	0.825 ± 0.016		
Base	0.814	0.690	0.718	0.340		

Table 2: Cardiac segmentation: Quantitative results for $M = 128$, $\beta = \frac{1}{10}$ measured in mean dice score. $\pm$ marks the standard deviations over $n = 5$ independent training runs. Comparison between CASA (proposed method), Uncertainty AL (UAL), Naive AL (NAL), Domain specific models (DSM), Joint Model (JModel) and the model after base training. C1-C4 denote the scanners occuring in the continuous data stream. BWT and FWT mark backward and forward transfer respectively.

Lung nodule detection In Table 3, results for lung nodule detection measured as average precision are compared. CASA performed significantly better than NAL and UAL for all scanners. For L4, which were the images extracted from LNDb, the distribution of nodules was different. For scanners L1-L3, the mean lesion diameter was 8.29mm, while for L4, lesion diameter was 5.99mm on average. This lead to a worse performance on L4 for all approaches. Nevertheless, CASA was the only active learning approach able to label a large enough amount of images for L4 such that it can significantly outperform the base model, as well as NAL and UAL.

Brain age estimation Table 4 (c) shows the results for brain age estimation in terms of MAE. CASA was able to perform continual learning without forgetting, and outperformed UAL and NAL for all scanners (B1-B4) at the end of the continuous data stream. Comparing MAE for B1 data for UAL (7.01) and NAL (11.91) with the base model (6.44) shows that forgetting has occurred for UAL and NAL. For CASA, MAE for B2 and B3 was notably higher than for B1 and B4 respectively. Due to the composition of the continual training set, B2 (n=190) and B3 (n=146) occurred less than B4 (n=1504) in the data stream, consequently leading to fewer B2 and B3 images seen during training, and consequently a worse performance. Nevertheless, CASA was able to handle this data set composition better than UAL and NAL.

Meth.	L1	L2	L3	L4	BWT	FWT
CASA	0.664 ± 0.016	0.543 ± 0.080	0.816 ± 0.005	0.229 ± 0.026	0.023 ± 0.037	0.025 ± 0.036
UAL	0.650 ± 0.011	0.394 ± 0.058	0.738 ± 0.038	0.180 ± 0.021	-0.003 ± 0.048	-0.015 ± 0.023
NAL	0.619 ± 0.025	0.472 ± 0.057	0.765 ± 0.041	0.184 ± 0.019	-0.019 ± 0.025	0.004 ± 0.009
DSM	0.644 ± 0.036	0.440 ± 0.060	0.488 ± 0.102	0.365 ± 0.062		
JModel	0.728 ± 0.033	0.649 ± 0.037	0.793 ± 0.017	0.454 ± 0.024		
Base	0.644	0.458	0.807	0.159		

Table 3: Lung nodule detection: Quantitative results for $M = 128$, $\beta = \frac{1}{10}$ measured in average precision. $\pm$ marks the standard deviations over $n = 5$ independent training runs. Comparison between CASA (proposed method), Uncertainty AL (UAL), Naive AL (NAL), Domain specific models (DSM), Joint Model (JModel) and the model after base training. L1-L4 denote the scanners occuring in the continuous data stream. BWT and FWT mark backward and forward transfer respectively.

Meth.	B1	B2	B3	B4	BWT	FWT
CASA	6.40 ± 0.35	8.96 ± 1.16	8.56 ± 1.14	6.54 ± 0.70	0.45 ± 0.59	4.97 ± 0.23
UAL	7.01 ± 0.68	12.22 ± 1.78	9.75 ± 0.50	12.92 ± 1.47	1.16 ± 0.39	1.66 ± 0.52
NAL	11.91 ± 2.31	17.67 ± 2.90	14.16 ± 3.52	15.54 ± 2.20	1.87 ± 2.36	2.82 ± 3.44
DSM	6.28 ± 0.37	4.77 ± 0.57	7.16 ± 0.53	4.42 ± 1.13		
JModel	6.51 ± 1.07	6.63 ± 2.09	4.38 ± 1.31	5.99 ± 0.53		
Base	6.44	18.26	11.43	15.86		

Table 4: Brain Age Estimation: Quantitative results for $M = 128$, $\beta = \frac{1}{10}$ measured in mean absolute error. $\pm$ marks the standard deviations over $n = 5$ independent training runs. Comparison between CASA (proposed method), Uncertainty AL (UAL), Naive AL (NAL), Domain specific models (DSM), Joint Model (JModel) and the model after base training. B1-B4 denote the scanners occuring in the continuous data stream. BWT and FWT mark backward and forward transfer respectively.

5.2 Influence of labelling budget β

The influence of β on cardiac segmentation performance is shown in Figure 3. For the first scanners C1 and C2, a similar performance can be observed for all methods and values of β. This was due to the fact that all methods have a sufficient amount of budget to adapt from C1 to C2. For C3, the performance for CASA was slightly higher compared to UAL and NAL. The most striking difference between the methods can be seen for scanner C4. There, CASA performed significantly better for $\beta = \frac{1}{5}$ and $\beta = \frac{1}{8}$. For $\beta = \frac{1}{10}$ CASA outperformed UAL and NAL on average however, a large deviation between the individual five runs is observable. Investigating this further revealed that CASA ran out of budget before C4 data appeared in the stream for one of the random seeds. Thus it was not able to adapt to C4 for this random seed. For $\beta = \frac{1}{20}$, CASA consumed the whole labelling budget before C4 data occured in the stream. Thus, it was not able to adapt to C4 data properly. UAL only had little budget left when C4 data came in and runs out afterwards thus, the performance was significantly worse to the results with more labelling budget. NAL performed the best for the setting with the lowest budget $\beta = \frac{1}{20}$, due to the fact that NAL labels every 20th step and thus did not run out of budget until the end of the stream is reached.

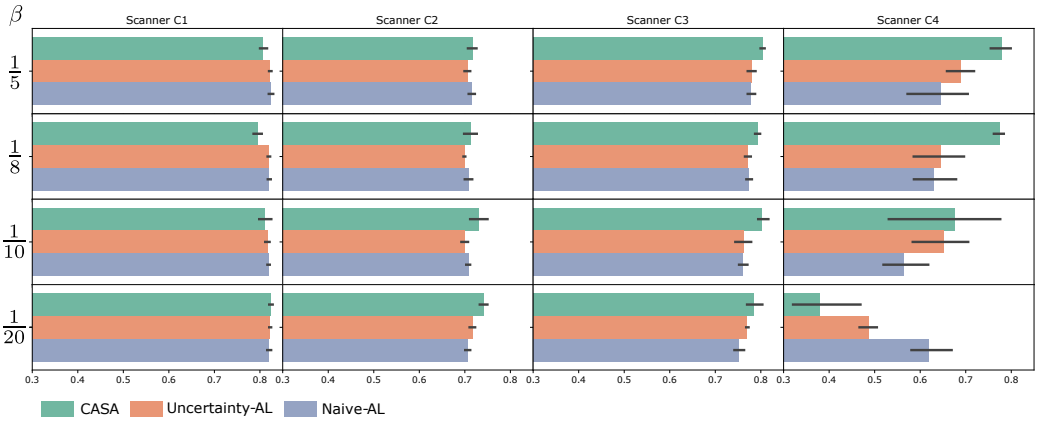

Figure 3: Influence of labelling budget β for cardiac segmentation with $M = 128$ comparing CASA, uncertainty AL and naive AL. Performance was measured in mean DSC.

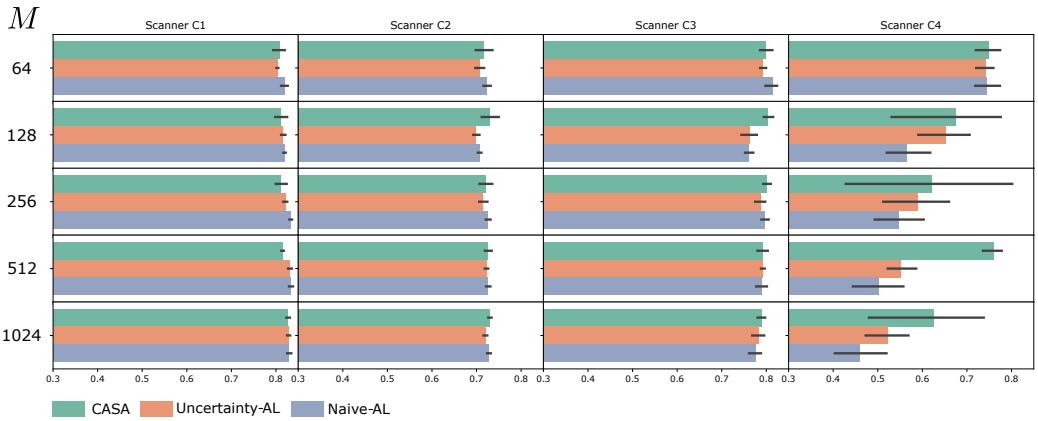

Figure 4: Influence of memory size M for cardiac segmentation with labelling budget $\beta = \frac{1}{10}$ comparing CASA, uncertainty AL and naive AL. Performance was measured as mean DSC.

5.3 Influence of memory size M

For cardiac segmentation, we investigate the influence of the training memory size M. The memory size M influences the the adaption to new domains the the continuous data stream. In Figure 4, CASA, UAL and NAL are compared for $M = \langle 64, 128, 256, 512, 1024 \rangle$. For the first scanners C1 and C2 in the stream, the performance was similar across AL methods and settings of M. For $M = 64$, CASA could not gain any improvement in comparison to UAL and NAL, meaning that the detection of pseudo-domains and balancing based on them is more useful for reasonable large memory sizes. All methods performed best for $M = 64$ however, the memory on the end of the stream was primarily filled with C3 and C4 data, which would lead to forgetting effects if training continues. In addition, a performance drop for $M >= 128$ compared to $M = 64$ for C4 data can be observed. This is a sign that the higher the memory size, the longer it takes for all methods to adapt to new domains. At the end of the data stream training for C4 data has not saturated for a rehearsal memory of size 128 and larger. The large variation in performance of CASA on C4 data across different

sizes of M might be due to the early consumption of the whole labelling budget. For each independent test run, the ordering of the continuous data stream is randomly changed, some of those orderings led to CASA running out of budget before the stream ended, thus the adaptation to C4 data was not completed.

5.4 Evaluation of the Memory and Pseudo-Domains

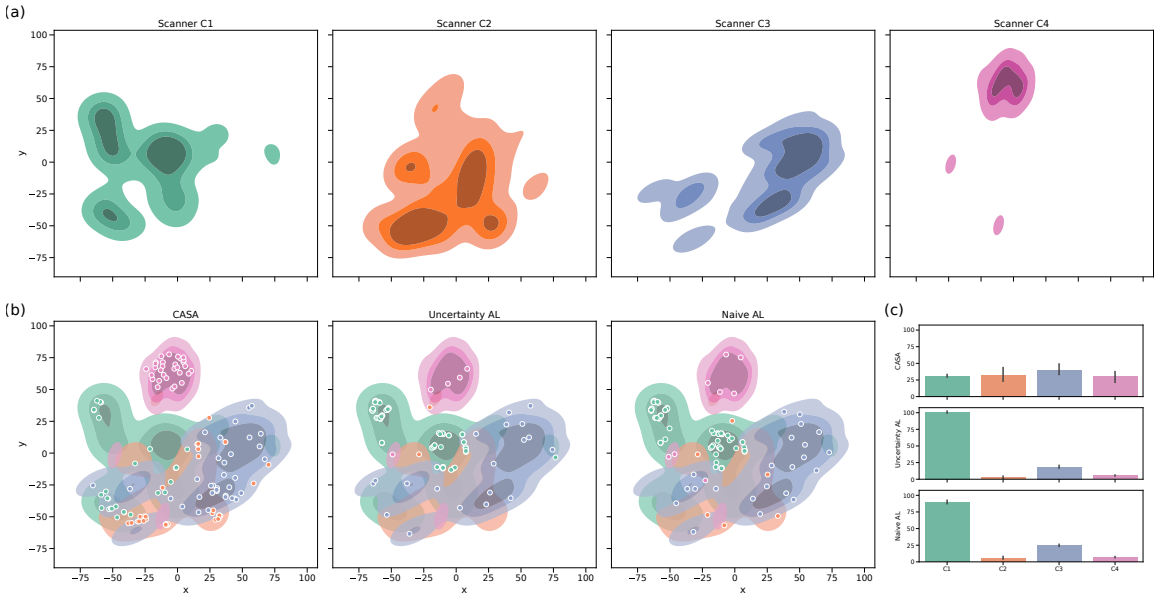

Figure 5: t-SNE visualization of style embeddings for cardiac segmentation. (a) shows the distribution of the different domains C1-C4 in the embedding space. (b) For CASA, UAL and NAL the memory elements at the end of continual training are marked in the embedding space, showing a balanced distribution for CASA. (c) Counts of elements in the rehearsal memory at the end of training for CASA, UAL and NAL.

We analyzed the balancing of our memory at the end of training (with $M = 128, \beta = \frac{1}{10}$) and the detection of different pseudo-domains by extracting the *style embedding* for all samples in the training set (combined base and continual set). Those embeddings were mapped to two dimensions using t-SNE (Maaten and Hinton, 2008) for plotting. In Figure 5 (a), it is observable that different scanners are located in different areas of the embedding space. Especially scanner C4 forms a compact cluster separated from the other scanners. Furthermore, a comparison of the distribution of the samples in the rehearsal memory at the end of training (Figure 5(b)) shows that for CASA, the samples distributed over the whole embedding including scanner C4. For UAL and NAL, most samples focused on scanner C1 samples (base training scanner), and a lower number of images of later scanners were kept in memory. Note, that this does not mean that UAL and NAL labelled primarily scanner C1 samples but that those methods did not balance the rehearsal memory. So, labelled images from C2-C4 might be lost in the process of choosing what samples to keep. As shown in Appendix Figure 8, those observations were stable over differnt test runs. Figure 5 (c) confirms the finding. Here, the memory distribution for the compared methods over

five independent runs (with different random seeds) demonstrated the capability of CASA to balance the memory across scanner, although the real domains are not known during training.

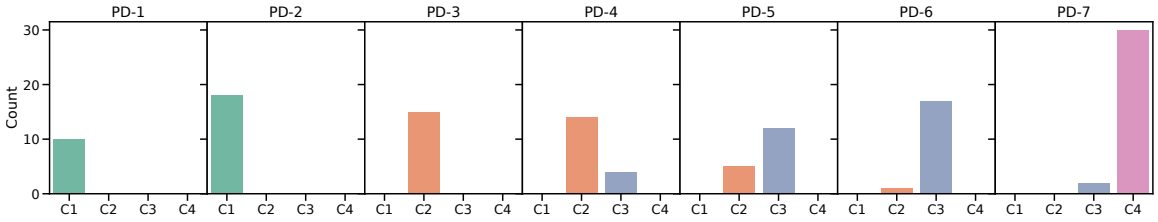

Figure 6: Distribution of images of specific scanners (C1-C4) to the discovered pseudo-domains for one run of CASA training with $M = 128$, $\beta = \frac{1}{10}$.

Pseudo-domain discovery in CASA (with $M = 128$, $\beta = \frac{1}{10}$) resulted in 6-8 pseudo-domains, showing that the definition of pseudo-domains does not exactly capture the definitions of domains in terms of scanners. In addition, the detection was influenced by the order of the continuous data stream. Figure 6 shows the distribution of images from certain scanners in the pseudo-domains for one training run of CASA (results for all $n = 5$ independent runs are shown in Appendix Figure 9). The first two pseudo-domains were dominated by samples from scanner C1, while the last pseudo-domain consisted mainly of scanner C4. The pseudo-domains 3-6 represented a mix of C2 and C3 data. This is consistent with Figure 5 where we see that the distributions of C2 and C3 overlap while C1 and especially C4 data is more separated.

5.5 Learning on a random stream of data

To analyze the influence of the sequential nature of the stream, we show how CASA performs on a random stream of data, with no sequential order of the scanners used. Results are given in Table 5. Note, that in this comparison adding randomness to the stream results in eliminating the domain shifts between scanners and making the samples within the stream approximately independent and identically distributed (i.i.d.). CASA performed well on a random stream however, for scanners with few samples in the training set (Scanner C2, C4), a drop in performance was observed. This is due to the fact that the pseudo-domain detection based on the outlier memory was not as effective as on a continuous stream, where we expect a accumulation of outliers as new domains start to occur in the stream.

NAL performed well on a random stream, outperforming NAL on a ordered stream, as well as CASA. Due to the randomness in the stream and the sampling strategy of NAL (taking every n-th sample), NAL learned on a diverse set of samples and managed to reach a more balanced training. Similar observations hold for uncertainty based active learning. Mixing up the order of samples in the stream leads to an earlier occurrence of data from C3 and C4, thus a better rehearsal set can be constructed with UAL. This highlights the ability of CASA to learn under the influence of domain shifts. If no domain shifts occur in the data the specific design of the approach does not provide a benefit.

Meth.	C1	C2	C3	C4
CASA - Random	0.816 ± 0.009	0.719 ± 0.011	0.778 ± 0.013	0.652 ± 0.078
UAL - Random	0.820 ± 0.006	0.717 ± 0.010	0.791 ± 0.006	0.740 ± 0.024
NAL - Random	0.826 ± 0.008	0.728 ± 0.007	0.802 ± 0.006	0.745 ± 0.021
CASA	0.812 ± 0.017	0.731 ± 0.025	0.803 ± 0.015	0.676 ± 0.158
UAL	0.816 ± 0.006	0.700 ± 0.009	0.764 ± 0.023	0.652 ± 0.078
NAL	0.819 ± 0.003	0.707 ± 0.005	0.761 ± 0.013	0.564 ± 0.064

Table 5: Dice scores for cardiac segmentation on a random stream. CASA with $M = 128$, $\beta = \frac{1}{10}$ is compared to naive active learning. $\pm$ marks the standard deviations over $n = 5$ independent training runs.

6. Discussion and Conclusion

We propose a continual active learning method to adapt deep learning models to changes of medical imaging acquisition settings. By detecting novel pseudo-domains occurring in the data stream, our method is able to keep the number of required annotations low, while improving the diversity of the training set. Pseudo-domains represent groups of images with similar but new imaging characteristics. Balancing the rehearsal memory based on pseudo-domains ensures a diverse set of samples is kept for retraining on both new and preceding domains.

Experiments showed that the proposed approach improves model accuracy in a range of different medical imaging tasks ranging from segmentation, to detection and regression. Performance of the models is improved across all domains for each task, while effectively counteracting catastrophic forgetting. Extensive experiments to gain insights on the effect of the composition of the rehearsal memory showed that CASA successfully balances training data between the real, but unknown domains.

A question that could be addressed by future research is that CASA needs to store samples that are part of the memory from preceding domains until the end of training, which could lead to data privacy concerns. A possible direction of further research is how to combine the concepts presented in this work with pseudo-rehearsal methods that do not store samples directly but rather a privacy conserving representation of the previously seen samples. In our experiments, the memory size and the labelling budget β was fixed before training while in real world applications running for possibly unlimited time, strategies how to expand the memory to include a sufficient amount of samples to cover the whole data distribution are needed. The labelling budget β might be increased based on the samples already observed, for example a simple strategy would be to add budget for every $n - th$ sample observed. Another aspect relevant for the practical implementation is that when deploying active learning approaches to real world applications, there is the need for interaction with a human annotator during the annotation. While the trained model can be used for clinical application if accuracy requirements are met, in practice active learning would indicate the necessity of annotating further cases to update the model. CASA is limiting this necessity of annotating and can be used to speed up the manual annotation process by suggesting a possible annotation for human annotators. For example when manual segmentation is required to derive a diagnosis, the proposed approach can be easily extended to enable the generation of a suggestion for a segmentation that the human

annotator only has to correct, leading to time-savings compared to manual annotation from scratch.

In experimental validation we assumed a data stream to which data from new scanners are added sequentially, i.e., one by one. However, in clinical practice, commonly multiple scanners are used in parallel, leading to possible simultaneous updates and newly entering scanners. Note that such a data stream with simultaneous additions would still be different from the randomly mixed stream analyzed in Section 5.5 as multiple acquisition shifts would still appear at specific time points in clinical practice. In addition to multiple scanners used in parallel, it is desirable not only to include data from one hospital, but to include data from different sites. Future work is needed to explore such a multi-site and multi-stream setting. A possible approach might combine the presented approach with federated learning.

Acknowledgments

This work was partially supported by the Austrian Science Fund (FWF): P 35189, by the Vienna Science and Technology Fund (WWTF): LS20-065, and by Novartis Pharmaceuticals Corporation. Part of the computations for research were performed on GPUs donated by NVIDIA.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

M.P. and J.H. declare no conflicts of interests. C.H.: Research Consultant for Siemens Healthineers and Bayer Healthcare, Stock holder at Hologic Inc. H.P.: Speakers Honoraria for Boehringer Ingelheim and Roche. Received a research grant by Boehringer Ingelheim. G.L.: Co-founder and stock holder at contextflow GmbH. Received research funding by Novartis Pharmaceuticals Corporation.

References

Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J.R. Van Beek, David Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, David P.Y. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batra, Philip Caligiuri, Ali Farooqi, Gregory W. Gladish, C. Matilda Jude, Reginald F. Munden, Iva Petkovska, Leslie E. Quint,

Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallam, Michael D. Heath, Michael H. Kuhn, Ekta Dharaiya, Richard Burns, David S. Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, Barbara Y. Croft, and Laurence P. Clarke. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 2011. ISSN 00942405. doi: 10.1118/1.3528204.

Joanne C. Beer, Nicholas J. Tustison, Philip A. Cook, Christos Davatzikos, Yvette I. Sheline, Russell T. Shinohara, and Kristin A. Linn. Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage*, 220, 10 2020. ISSN 10959572. doi: 10.1016/j.neuroimage.2020.117129.

Andreea Bobu, Eric Tzeng, Judy Hoffman, and Trevor Darrell. Adapting to continously shifting domains. In *ICLR Workshop*, 2018.

Samuel Budd, Emma C Robinson, and Bernhard Kainz. A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis. 2019. URL `http://arxiv.org/abs/1910.02923`.

Victor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreno, Alberto Albiol, Fanwei Kong, Shawn C. Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, Bjoern Menze, Firas Khader, Christoph Haarburger, Cian M. Scannell, Mitko Veta, Adam Carscadden, Kumaradevan Punithakumar, Xiao Liu, Sotirios A. Tsaftaris, Xiaoqiong Huang, Xin Yang, Lei Li, Xiahai Zhuang, David Vilades, Martin L. Descalzo, Andrea Guala, Lucia La Mura, Matthias G. Friedrich, Ria Garg, Julie Lebel, Filipe Henriques, Mahir Karakas, Ersin Cavus, Steffen E. Petersen, Sergio Escalera, Santi Segui, Jose F. Rodriguez-Palomares, and Karim Lekadir. Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. *IEEE Transactions on Medical Imaging*, pages 1–1, 6 2021. doi: 10.1109/TMI.2021.3090082.

Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020. ISSN 20411723. doi: 10.1038/s41467-020-17478-w. URL `http://dx.doi.org/10.1038/s41467-020-17478-w`.

Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. ISSN 0162-8828. doi: 10.1109/TPAMI.2021.3057446. URL `https://ieeexplore.ieee.org/document/9349197/`.

Nicola K. Dinsdale, Mark Jenkinson, and Ana I.L. Namburete. Unlearning Scanner Bias for MRI Harmonisation in Medical Image Segmentation. *Communications in Computer and Information Science*, 1248 CCIS:15–25, 2020. ISSN 18650937. doi: 10.1007/978-3-030-52791-4{_}2.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 6 2010. ISSN 0920-5691. doi: 10.1007/s11263-009-0275-4. URL http://link.springer.com/10.1007/s11263-009-0275-4.

Jean Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L. Phillips, Madhukar H. Trivedi, Myrna M. Weissman, and Russell T. Shinohara. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120, 2 2018. ISSN 10959572. doi: 10.1016/j.neuroimage.2017.11.024.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning Zoubin Ghahramani. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059, 2016.

Leon Gatys, Alexander Ecker, and Matthias Bethge. A Neural Algorithm of Artistic Style. *Journal of Vision*, 16(12):326, 2016. ISSN 1534-7362. doi: 10.1167/16.12.326.

Ben Glocker, Robert Robinson, Daniel C. Castro, Qi Dou, and Ender Konukoglu. Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects. *arXiv Preprint*, 2019. URL http://arxiv.org/abs/1910.04597.

Camila Gonzalez, Georgios Sakas, and Anirban Mukhopadhyay. What is Wrong with Continual Learning in Medical Image Segmentation? *arXiv Preprint*, 10 2020. URL http://arxiv.org/abs/2010.11008.

Hao Guan and Mingxia Liu. Domain Adaptation for Medical Image Analysis: A Survey. 2 2021. URL http://arxiv.org/abs/2102.09508.

Johannes Hofmanninger, Matthias Perkonigg, James A. Brink, Oleg Pianykh, Christian Herold, and Georg Langs. Dynamic Memory to Alleviate Catastrophic Forgetting in Continuous Learning Settings. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12262 LNCS: 359–368, 2020. ISSN 16113349. doi: 10.1007/978-3-030-59713-9{_}35.

Neerav Karani, Krishna Chaitanya, Christian Baumgartner, and Ender Konukoglu. A lifelong learning approach to brain MR segmentation across scanners and protocols. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11070 LNCS, pages 476–484. Springer, Cham, 2018. ISBN 9783030009274. doi: 10.1007/978-3-030-00928-1{_}54.

Pamela J LaMontagne, Tammie L S Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, Marcus E Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *medRxiv*, page 2019.12.13.19014902, 2019. doi: 10.1101/2019.12.13.19014902. URL http://medrxiv.org/content/early/2019/12/15/2019.12.13.19014902.abstract.

Qicheng Lao, Xiang Jiang, Mohammad Havaei, and Yoshua Bengio. Continuous Domain Adaptation with Variational Domain-Agnostic Feature Replay. 2020. URL http://arxiv.org/abs/2003.04382.

Matthias Lenga, Heinrich Schulz, and Axel Saalbach. Continual Learning for Domain Adaptation in Chest X-ray Classification. In *Conference on Medical Imaging with Deep Learning (MIDL)*, 2020. URL http://arxiv.org/abs/2001.05922.

Tony Fei Liu, Ming Kai Ting, and Zhi-Hua Zhou. Isolation Forest. In *International Conference on Data Mining*, 2008.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, pages 6468–6477, 2017. ISSN 10495258.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Michael McCloskey and Neal J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, 24(C):109–165, 1989. ISSN 00797421. doi: 10.1016/S0079-7421(08)60536-8.

Firat Ozdemir, Philipp Fuernstahl, and Orcun Goksel. Learn the New, Keep the Old: Extending Pretrained Models with New Anatomy and Images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11073 LNCS:361–369, 2018. ISSN 16113349. doi: 10.1007/978-3-030-00937-3{_}42.

Sinan Özgün, Anne-Marie Rickmann, Abhijit Guha Roy, and Christian Wachinger. Importance Driven Continual Learning for Segmentation Across Domains. Number Cl, pages 423–433. 2020. doi: 10.1007/978-3-030-59861-7{_}43. URL https://link.springer.com/10.1007/978-3-030-59861-7_43.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 5 2019. ISSN 08936080. doi: 10.1016/j.neunet.2019.01.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608019300231.

João Pedrosa, Guilherme Aresta, Carlos Ferreira, Márcio Rodrigues, Patrícia Leitão, André Silva Carvalho, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, and Aurélio Campilho. LNDb: A lung nodule database on computed tomography. *arXiv*, pages 1–12, 2019. ISSN 23318422.

Matthias Perkonigg, Johannes Hofmanninger, Christian J. Herold, James A. Brink, Oleg Pianykh, Helmut Prosch, and Georg Langs. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. *Nature Communications*, 12(1): 5678, 12 2021a. ISSN 2041-1723. doi: 10.1038/s41467-021-25858-z. URL https://www.nature.com/articles/s41467-021-25858-z.

Matthias Perkonigg, Johannes Hofmanninger, and Georg Langs. Continual Active Learning for Efficient Adaptation of Machine Learning Models to Changing Image Acquisition. In *Advances in Information Processing in Medical Imaging, IPMI*, 2021b.

Oleg S. Pianykh, Georg Langs, Marc Dewey, Dieter R. Enzmann, Christian J. Herold, Stefan O. Schoenberg, and James A. Brink. Continuous learning AI in radiology: Implementation principles and early applications. *Radiology*, 297(1):6–14, 2020. ISSN 15271315. doi: 10.1148/radiol.2020200038.

Florian Prayer, Johannes Hofmanninger, Michael Weber, Daria Kifjak, Alexander Willenpart, Jeanny Pan, Sebastian Röhrich, Georg Langs, and Helmut Prosch. Variability of computed tomography radiomics features of fibrosing interstitial lung disease: A test-retest study. *Methods*, 188:98–104, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. ISSN 01628828. doi: 10.1109/TPAMI.2016.2577031.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. pages 1–8, 2015. ISSN 16113349. doi: 10.1007/978-3-319-24574-4{_}28. URL `http://arxiv.org/abs/1505.04597`.

Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas de Bel, Moira S.N. Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, Robbert van der Gugten, Pheng Ann Heng, Bart Jansen, Michael M.J. de Kaste, Valentin Kotov, Jack Yu-Hung Lin, Jeroen T.M.C. Manders, Alexander Sóñora-Mengana, Juan Carlos García-Naranjo, Evgenia Papavasileiou, Mathias Prokop, Marco Saletta, Cornelia M Schaefer-Prokop, Ernst T. Scholten, Luuk Scholten, Miranda M. Snoeren, Ernesto Lopez Torres, Jef Vandemeulebroucke, Nicole Walasek, Guido C.A. Zuidhof, Bram van Ginneken, and Colin Jacobs. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Medical Image Analysis*, 42:1–13, 12 2017. ISSN 13618415. doi: 10.1016/j.media.2017.06.015. URL `https://linkinghub.elsevier.com/retrieve/pii/S1361841517301020`.

Asim Smailagic, Pedro Costa, Alex Gaudio, Kartik Khandelwal, Mostafa Mirshekari, Jonathon Fagert, Devesh Walawalkar, Susu Xu, Adrian Galdran, Pei Zhang, Aurélio Campilho, and Hae Young Noh. O-MedAL: Online active deep learning for medical image analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(4):1–15, 2020. ISSN 19424795. doi: 10.1002/widm.1353.

Zuxuan Wu, Xin Wang, Joseph Gonzalez, Tom Goldstein, and Larry Davis. ACE: Adapting to changing environments for semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2121–2130, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00221.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. Models Genesis. *Medical Image Analysis*, page 101840, 2020. ISSN 13618415. doi: 10.1016/j.media.2020.101840.

Zongwei Zhou, Jae Y. Shin, Suryakanth R. Gurudu, Michael B. Gotway, and Jianming Liang. Active, continual fine tuning of convolutional neural networks for reducing annotation efforts. *Medical Image Analysis*, 71, 7 2021. ISSN 13618423. doi: 10.1016/j.media.2021.101997.

Appendix A.

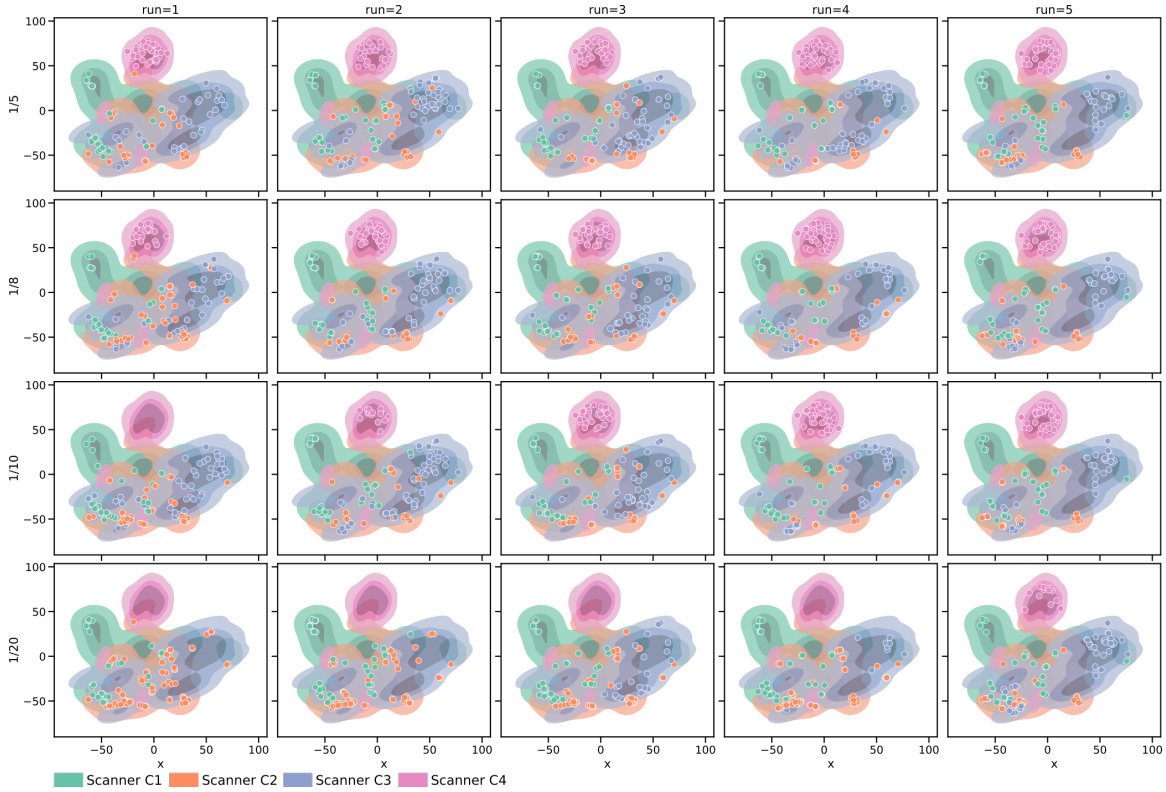

Figure 7: Style embeddings of CASA training memories for different runs with different labelling budgets β.

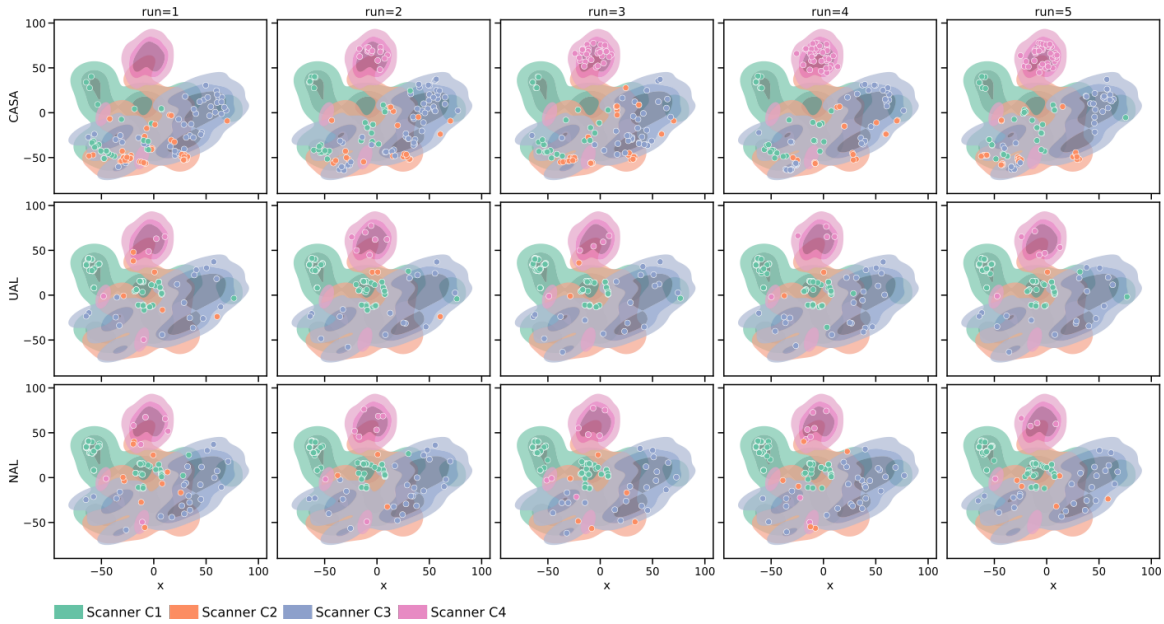

Figure 8: Style embeddings of training memories for different runs of CASA, UAL and NAL.

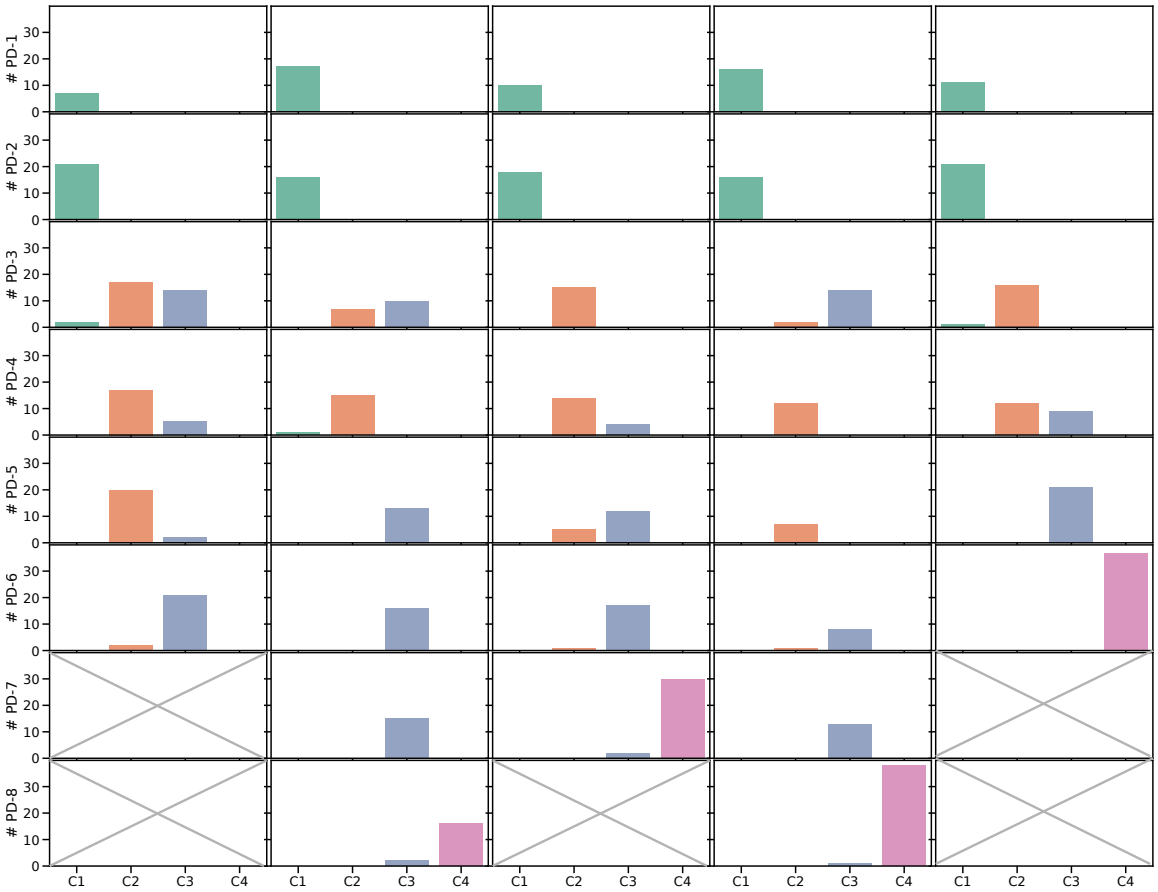

Figure 9: Distribution of images acquired with a specific scanner (C1-C4) to the discovered pseudo-domains for five runs of CASA training with $M = 128$, $\beta = \frac{1}{10}$.