Medical Image Analysis 100 (2025) 103404

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

# CLMS: Bridging domain gaps in medical imaging segmentation with source-free continual learning for robust knowledge transfer and adaptation

Weilu Li [1], Yun Zhang [1], Hao Zhou, Wenhan Yang, Zhi Xie [*], Yao He [*]

*State Key Laboratory of Ophthalmology, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China*

ARTICLE INFO

*Keywords:*
Medical segmentation
Continual learning
Source-free domain adaptation
Domain shift

ABSTRACT

Deep learning shows promise for medical image segmentation but suffers performance declines when applied to diverse healthcare sites due to data discrepancies among the different sites. Translating deep learning models to new clinical environments is challenging, especially when the original source data used for training is unavailable due to privacy restrictions. Source-free domain adaptation (SFDA) aims to adapt models to new unlabeled target domains without requiring access to the original source data. However, existing SFDA methods face challenges such as error propagation, misalignment of visual and structural features, and inability to preserve source knowledge. This paper introduces Continual Learning Multi-Scale domain adaptation (CLMS), an end-to-end SFDA framework integrating multi-scale reconstruction, continual learning, and style alignment to bridge domain gaps across medical sites using only unlabeled target data or publicly available data. Compared to the current state-of-the-art methods, CLMS consistently and significantly achieved top performance for different tasks, including prostate MRI segmentation (improved Dice of 10.87 %), colonoscopy polyp segmentation (improved Dice of 17.73 %), and plus disease classification from retinal images (improved AUC of 11.19 %). Crucially, CLMS preserved source knowledge for all the tasks, avoiding catastrophic forgetting. CLMS demonstrates a promising solution for translating deep learning models to new clinical imaging domains towards safe, reliable deployment across diverse healthcare settings.

## 1. Introduction

Deep learning has made remarkable progress in medical image analysis such as lesion detection, organ segmentation, and disease classification (Topol, 2019). However, the integration of deep learning models into clinical settings faces a major hurdle: the degradation in model performance when applied to medical imaging data across different healthcare sites (Wynants et al., 2020; Roberts et al., 2021; De Fauw et al., 2018). This arises due to variations in scanning protocols, imaging devices, patient populations, and technician proficiency across the sites. Consequently, discrepancies emerge between the source data, where the model was trained on, and the target data, where the model was applied to, which is known as domain shift (Zhang et al., 2020; Ju et al., 2021; Yasaka and Abe, 2018). Moreover, privacy restrictions often preclude access to the source data, exacerbating the challenge of reconciling domain differences (Dayan et al., 2021). One solution is to use transfer learning, which adapts models to the target data in the presence of labels for the target data (Kora et al., 2022). However, manual labeling is frequently infeasible, particularly for tasks such as segmentation whose labeling is labor-intensive. Therefore, domain shift persists as a central challenge in translating state-of-the-art deep learning models to varied clinical settings in practice.

Source-free domain adaptation (SFDA) adapts source models to target domains relying solely on unlabeled target data, without labor-intensive labelling and source data (Li et al., 2024). Common SFDA methods include image-level SFDA and feature-level SFDA (Li et al., 2024). Image-level SFDA generates virtual source domain images utilizing the domain-related features of the source model trained by source data to reconcile domain differences (Zhou et al., 2022; Hu et al., 2022; Wang et al., 2023). Feature-level SFDA involves image information

* Corresponding author.
*E-mail addresses:* xiezhi@gmail.com (Z. Xie), scheyao@hotmail.com (Y. He).
[1] Equally contributed and sorted alphabetically by last name

https://doi.org/10.1016/j.media.2024.103404
Received 8 May 2024; Received in revised form 1 October 2024; Accepted 19 November 2024
Available online 24 November 2024

alignment methods and self-training methods. Image information alignment methods encode target images to a source-consistent embedding to be recognized by the source model, usually via alignment to the batch normalization (BN) layers of the source model (Hong et al., 2022; Ye et al., 2022; Yu et al., 2023). Self-training methods fine-tune the source model by the target data through generating pseudo-labels, entropy minimization, or contrastive learning (Hong et al., 2022; Bateson et al., 2022; Liu et al., 2023; Kondo, 2022).

Although existing SFDA methods have made great progress in reducing domain shift, key challenges remain. Image-level SFDA may generate fake source images that resemble real source data but contain discrepancies, leading to errors that accumulate over training (Zhou et al., 2022). For feature-level SFDA, image information alignment methods primarily focus on low-level visual differences (Yu et al., 2023; Stan and Rostami, 2021). Consequently, these methods lack the selectivity required to map invariant morphological structures while simultaneously adapting the visual style, thereby risking misalignment issues (Hong et al., 2022; Yu et al., 2023). Hence, it often used in-conjunction with self-training methods to constrain the structural feature learning. Self-training methods can produce ambiguous outputs arising from the uncertainty of source model predictions for target domain data, such as incorrectly high-confidence samples (Li et al., 2024; Li et al., 2023). Although many studies focus on improving the confidence of source model predictions, avoiding errors introduced by incorrect predictions is infeasible due to the inherent uncertainty and potential inaccuracies of any source model applied to unlabeled target domain data (Wang et al., 2023; Yang et al., 2022; Cai et al., 2023). Moreover, existing methods typically involve a combination of approaches executed in multiple steps, leading to error propagation throughout the training process (Yu et al., 2023; Li et al., 2023; Yang et al., 2022). Adding to these challenges, retraining deep learning models on new target data can lead to catastrophic forgetting of source knowledge (Pianykh et al., 2020; Perkonigg et al., 2021), an effect often overlooked by the current SFDA methods. The risk of forgetting important morphological features during this process poses a risk to the performance of target applications. Therefore, there is a pressing need for more advanced SFDA approaches that can selectively reconcile domain gaps while retaining anatomical knowledge.

Here, we introduce CLMS, an end-to-end source-free domain adaptation solution by integrating multi-scale image reconstruction, continual learning, and style feature alignment. The key novelty of CLMS lies in its ability to map invariant morphological structures and preserve important anatomical knowledge, while simultaneously adapting the visual style by disentangling low-level visual and high-level structural features through the interplay of these components. Multi-scale reconstruction maps target images into a canonical form, establishing structural and visual feature mappings across scales; continual learning finetunes the source model through a replay-based approach (Wang et al., 2024), retaining high-level structural representations by transforming them into the canonical form using these mappings, while also augmenting target feature responsiveness; and style alignment constrains the canonical form's low-level visual representations to the source style. Together with joint end-to-end optimization, CLMS is positioned to effectively adapt models without risky pseudo-labels or error propagation issues. We compared CLMS to the state-of-the-art approaches using diverse multimodal datasets, including MRI, colonoscopy, and retinal images. CLMS consistently achieved the top performance and demonstrated robustness in adapting models to diverse medical imaging domains without catastrophic forgetting. Notably, CLMS demonstrated its ability to adapt segmentation models while retaining morphological features critical for subsequent clinical classification tasks like Plus disease diagnosis in retinal images. A comprehensive benchmark underscored CLMS's potential to transform model deployment in clinical settings by effectively bridging domain gaps and advancing deep learning in medical imaging and healthcare.

## 2. Related work

Domain shift problem can occur in medical segmentation across modalities, such as MRI (Guan et al., 2021), colonoscopy (Liu et al., 2021), fundus imaging (Wang et al., 2020), CT (Dong et al., 2022), X-ray (Sanchez et al., 2022) and among others. This issue may lead to diagnostic and analytical errors, thereby affecting patient treatment and diagnostic outcomes.

### 2.1. Unsupervised domain adaptation

Unsupervised domain adaptation (UDA) adapts knowledge from labeled source data to unlabeled target domains. Recent UDA approaches can be categorized into image-level and feature-level adaptation. Image-level UDA translates the image style between source and target domains to reduce style disparities (Zhu et al., 2017; Chen et al., 2019; Palladino et al., 2020), with cycle-consistent generative adversarial networks (CycleGAN) a common practice in this category (Palladino et al., 2020). Feature-level UDA focuses on learning domain-invariant feature representations based on the premise that domain gaps exist more in low-level characteristics (e.g., intensity values) than high-level traits (e.g., anatomical structures) (Dou et al., 2018; Tran et al., 2019; Tzeng et al., 2017). These methods align high-level features in later convolutional neural network (CNN) layers while fine-tuning earlier layers for low-level target adaptation (Yu et al., 2022). However, focusing solely on low or high-level features may lead to suboptimal adaptation as they are hierarchically composed in the image. Simultaneous image and feature adaptation has been proposed (Chen et al., 2019), but image-level prioritizes low-level features, while feature-level struggles to differentiate low/high-level features, potentially causing misalignment and error propagation (Kumari and Singh, 2024). Importantly, most UDA methods require access to both source and target data, which is often infeasible in medical domains due to data privacy restrictions preventing source data sharing.

### 2.2. Source free domain adaptation

The advantage of SFDA over UDA is the ability to adapt knowledge from labeled source data to unlabeled target domains without requiring access to the actual source data. Similar to UDA, existing SFDA methods can be broadly categorized into image-level and feature-level adaptation. Image-level SFDA aims to generate synthetic source-like image data, which can then be used with standard UDA techniques (Fang et al., 2024). A common approach is using Fourier transforms to stylize target domain images to resemble the source domain appearance (Wang et al., 2023; Yang et al., 2022). Feature-level SFDA focuses on learning domain-invariant features through techniques like feature alignment and self-training. Feature alignment synchronizes the feature distributions or prototypes between source and target models (Yu et al., 2023; Fang et al., 2024). Self-training constrains high-level semantics using pseudo-labeling (Wang et al., 2023; Yu et al., 2023; Yang et al., 2022), self-supervision (Li et al., 2023), or entropy minimization (Hong et al., 2022). SFDA faces challenges similar to UDA. Image-level SFDA struggles to adapt high-level semantics, often requiring combination with feature-level SFDA (Wang et al., 2023; Yang et al., 2022). Meanwhile, feature alignment methods has difficulty disentangling low and high-level features, necessitating self-training methods (Yu et al., 2023). But, self-training methods can produce incorrect pseudo-labels and error propagation when applying the source model to the unlabeled target data (Li et al., 2024; Fang et al., 2024; Luo et al., 2024; Cao et al., 2024). Several studies tried to improve prediction confidence through implementing false label filtering mechanism to reject unreliable pseudo labels (Yu et al., 2023; Yang et al., 2022) and introducing intra-class level threshold to select the voxels with intra-class confidence (Wang et al., 2023; Cai et al., 2023). Additionally, the inability to guarantee the retention of high-level semantic features from the source domain,

coupled with the possibility of crucial features from the source domain being discarded or overlooked during adaptation, can lead to catastrophic forgetting issues.

### 2.3. Continual learning

To address the catastrophic forgetting problem, continual learning can be categorized into five types: regularization-based, replay-based, optimization-based, representation-based, and architecture-based methods (Wang et al., 2024). The first four assume new and old tasks share high-level semantics, allowing fine-tuning while retaining old knowledge. For instance, regularization-based approaches constrain parameter changes based on importance to old tasks (Ritter et al., 2018; Schwarz et al., 2018; Rebuffi et al., 2017) or using knowledge distillation (Dhar et al., 2019). Replay-based approaches approximate old data distributions via buffered (Lopez-Paz and Ranzato, 2017) or generated (Shin et al., 2017; Wu et al., 2018) old training samples. Optimization-based approaches limit direction of gradient updates, typically at orthogonal to the previous input space (Chaudhry et al., 2018; Farajtabar et al., 2020). Representation-based approaches enhances generalizability through self-supervision (Cha et al., 2021) or pre-training (Mehta et al., 2023; Ramasesh et al., 2021). In contrast, architecture-based methods assume the new task introduces novel high-level semantics (Rusu et al., 2016; Mallya et al., 2018; Ebrahimi et al., 2020; Li and Hoiem, 2018). Hence, these methods construct new model components by reusing frozen features from the old task model and dividing the new model into shared and task-dedicated sections. While domain adaptation in medical imaging can safely assume consistent high-level anatomical structures cross domains, we utilize a replay-based continual learning approach to prevent catastrophic forgetting of the source domain when adapting to the target.

## 3. Methods

We introduced CLMS, a solution integrating multi-scale image reconstruction, continual learning, and style feature alignment (Fig. 1). This framework, utilizing solely unlabeled data from the target domain or publicly available data, not only enhanced the model's performance in the target domain but also ensured the preservation of its performance in the source domain.

CLMS enables end-to-end adaptation of a source model to target domain images through a multi-stage process that includes multi-scale reconstruction, continual learning, and style feature alignment. The multi-scale reconstruction module maps target domain images into a canonical form and then reconstructs them back to the target domain. This process establishes a transformation of both high-level structural and low-level visual features across different scales between the target domain and the canonical form (Fig. 1A). Importantly, the reconstructed images remain in the target domain, preserving detailed information while providing a bridge to the canonical representation. The continual learning module finetunes a clone of the source model (clone model) using a replay approach (Fig. 1B). It enforces consistency between the source model's responses to the target images in the original and reconstructed forms. This approach offers two key advantages: (a) since the images are in the same domain, finetuning the clone model doesn't introduce the incorrect pseudo-label problem that could occur if fine-tuning on the canonical form, (b) the consistency of responses helps retain high-level anatomical representations transformed into the canonical form via the target-canonical mapping, while also augmenting model responsiveness to target features. Furthermore, maintaining prediction consistency on an augmented public dataset further retains representations and reinforces responsiveness (Fig. 1C), expanding the model's response range without propagation errors. The style feature alignment module constrains the low-level visual representations of the

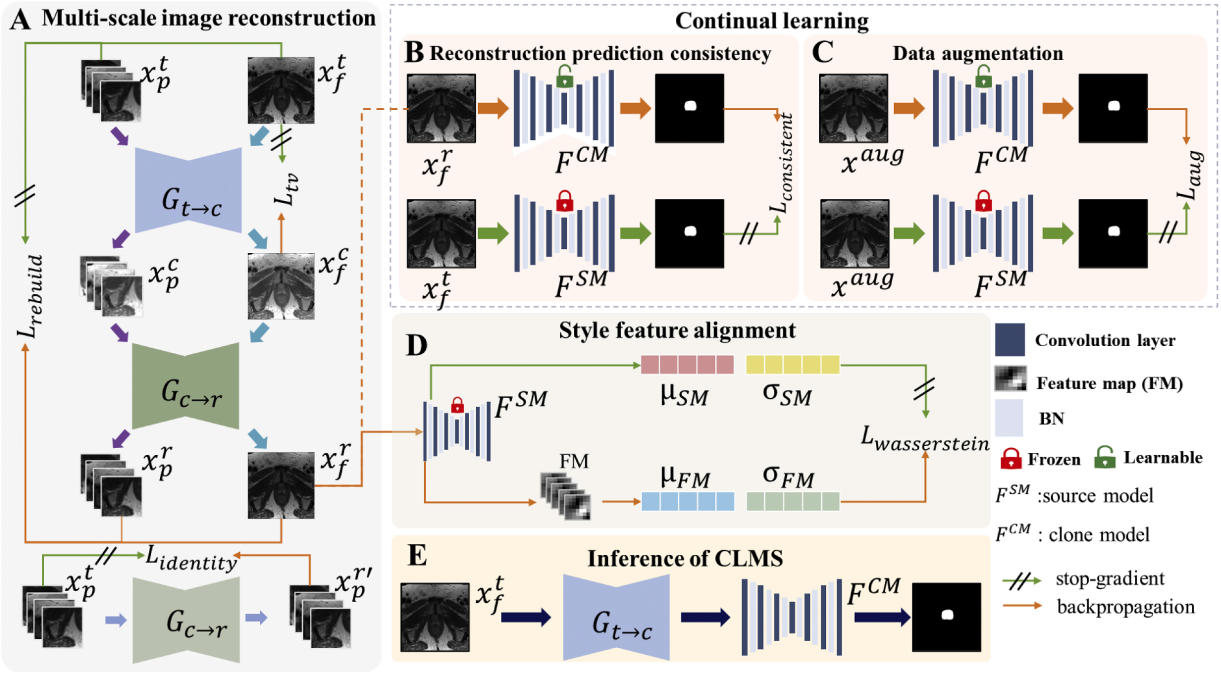


**Fig. 1.** | Overview of the architecture of CLMS. (A) Multi-scale image reconstruction module captures both global and local features of the target domain images, fusing information across scales to enhance local details in the reconstructed images. Continual learning module simulates source domain response via (B) the reconstructed images in the target domain, and (C) publicly available data for augmentation. This module allows the clone model to learn more accurate source-target difference while preserving the anatomical knowledge. (D) Style feature alignment module adjusts the style representations of canonical form images to match the style of the source image. (E) Inference of CLMS transforms the target data into the canonical form to be processed by the downstream models. Abbreviations: *t*: target domain; *c*: canonical form; *r*: reconstruction; *f*: whole-level; *p*: patch-level; *aug*: augmentation; $\mu_{FM}$ and $\sigma_{FM}$ represent the mean and standard deviation of the batch normalization layer of the feature maps computed using the canonical image $x_f^c$; $\mu_{SM}$ and $\sigma_{SM}$ denote the running mean and the running standard deviation of the batch normalization layer of the source model.

canonical form to be consistent with the source domain style (Fig. 1D). This is achieved by aligning the features to the batch normalization (BN) layers of the source model. Importantly, this alignment is performed on the canonical representation to prevent domain collapse that could occur if finetuning on the reconstructed images. During the inference, CLMS transforms the target data into the canonical form to be processed by the downstream models (Fig. 1E). This established a novel approach that interleaves multi-scale reconstruction, style feature alignment, and continual learning to disentangle low-level visual and high-level anatomical representations, thus ensuring model robustness and adaptability across different medical environments.

### 3.1. Multi-scale image reconstruction module

The primary goal of this module was to transform target domain images into a canonical form through an encoder-decoder generative model while retaining both global and local image information through whole-level and patch-level image reconstructions. Patch-level images were derived from the whole image through random cropping.

For patch-level reconstruction, a generator $G_{t\to c}$transforms a patched image $x_p^t$ into a canonical form image $x_p^c$. The generator $G_{t\to c}$ is formulated as follows:

$$x_p^c = \left(G_{t\to c}\left(x_p^t\right) + x_p^t\right)\Big/2.$$

Then, another generator $G_{c\to r}$ of the same type as $G_{t\to c}$ transforms $x_p^c$ into the reconstruction image $x_p^r$ and the generator $G_{c\to r}$ is formulated as follows:

$$x_p^r = \left(G_{c\to r}\left(x_p^c\right) + x_p^c\right)\Big/2.$$

The image $x_p^r$ is constrained by the reconstruction loss $L_{rebuild\_p}$ as follows:

$$L_{rebuild\_p} = \left|\left|x_p^r - x_p^t\right|\right|_1,$$

where $||\cdot||_1$ represents L1 norm.

In addition, the generator $G_{c\to r}$ processes $x_p^t$ into image $x_p^{r'}$, and the L1 norm between images $x_p^t$ and $x_p^{r'}$, denoted as the identity mapping loss $L_{identity}$, is formulated as:

$$L_{identity} = \left|x_p^{r'} - x_p^t\right|_1.$$

For whole-level reconstruction, the generator $G_{t\to c}$ transforms a target whole image $x_f^t$ to obtain the canonical form image $x_f^c$, and then the generator $G_{c\to r}$ transforms $x_f^c$ to obtain the reconstruction image $x_f^r$. Finally, the reconstruction loss at the whole-level, denoted as $L_{rebuild\_f}$, is computed as follows:

$$L_{rebuild\_f} = \left|x_f^r - x_f^t\right|_1.$$

In addition, we proposed an improved total variation loss (TV Loss) to constrain the quality of image $x_f^c$ for both spatial and channel aspects.

The spatial-level loss function $L_{sptial}$ is defined as follows:

$$L_{sptial} = \frac{sum\left(\left|\left(\left(x_f^c - x_f^t\right)_{h-1,w} - \left(x_f^c - x_f^t\right)_{h,w}\right) \times hMask_{h,w}\right|\right)}{sum\left(hMask_{h,w}\right)} + \frac{sum\left(\left|\left(\left(x_f^c - x_f^t\right)_{h,w-1} - \left(x_f^c - x_f^t\right)_{h,w}\right) \times wMask_{h,w}\right|\right)}{sum\left(wMask_{h,w}\right)},$$

$$hMask_{h,w} = \begin{cases} 1 & if\left(\left(x_f^t 0:h-1, w- x_f^t 1:h, w\right) \times \left(x_f^c 0:h-1, w- x_f^c 1:h, w\right)\right) < 0 \\ 0 & if\left(\left(x_f^t 0:h-1, w- x_f^t 1:h, w\right) \times \left(x_f^c 0:h-1, w- x_f^c 1:h, w\right)\right) \geq 0 \end{cases},$$

$$wMask_{h,w} = \begin{cases} 1 & if\left(\left(x_f^t h, 0:w-1- x_f^t h, 1:w\right) \times \left(x_f^c h, 0:w-1- x_f^c h, 1:w\right)\right) < 0 \\ 0 & if\left(\left(x_f^t h, 0:w-1- x_f^t h, 1:w\right) \times \left(x_f^c h, 0:w-1- x_f^c h, 1:w\right)\right) \geq 0 \end{cases},$$

where $h$ and $w$ represent the image coordinates along the y-axis and x-axis, respectively, where sum denotes the matrix summation function, and $|\cdot|$ denote the absolute value.

The channel-level loss function $L_{channel}$ is formulated as:

$$L_{channel} = \frac{sum\left(\left|\left(\left(x_f^c r - x_f^t r\right) - \left(x_f^c g - x_f^t g\right)\right) \times rgMask\right|\right)}{sum(rgMask)} + \frac{sum\left(\left|\left(\left(x_f^c g - x_f^t g\right) - \left(x_f^c b - x_f^t b\right)\right) \times gbMask\right|\right)}{sum(gbMask)} + \frac{sum\left(\left|\left(\left(x_f^c r - x_f^t r\right) - \left(x_f^c b - x_f^t b\right)\right) \times rbMask\right|\right)}{sum(rbMask)},$$

$$rgMask = \begin{cases} 1 & if\left(\left(x_f^t r - x_f^t g\right) \times \left(x_f^c r - x_f^c g\right)\right) < 0 \\ 0 & if\left(\left(x_f^t r - x_f^t g\right) \times \left(x_f^c r - x_f^c g\right)\right) \geq 0 \end{cases},$$

$$gbMask = \begin{cases} 1 & if\left(\left(x_f^t g - x_f^t b\right) \times \left(x_f^c g - x_f^c b\right)\right) < 0 \\ 0 & if\left(\left(x_f^t g - x_f^t b\right) \times \left(x_f^c g - x_f^c b\right)\right) \geq 0 \end{cases},$$

$$rbMask = \begin{cases} 1 & if\left(\left(x_f^t r - x_f^t b\right) \times \left(x_f^c r - x_f^c b\right)\right) < 0 \\ 0 & if\left(\left(x_f^t r - x_f^t b\right) \times \left(x_f^c r - x_f^c b\right)\right) \geq 0 \end{cases},$$

where $x_f^t r$, $x_f^t g$ and $x_f^t b$ represent the R, G and B channels of image $x_f^t$, respectively. Similarly, $x_f^c r$, $x_f^c g$ and $x_f^c b$ represent the R, G and B channels of image $x_f^c$ ,respectively. The loss function $L_{tv}$ is then:

$$L_{tv} = L_{sptial} \times \lambda_{sptial} + L_{channel} \times \lambda_{channel},$$

where $\lambda_{sptial}$ and $\lambda_{channel}$ are scalar weights used to balance the spatial and channel losses.

### 3.2. Continual learning module

Continual learning was formulated through two constraints: source prediction consistency and data augmentation. Reconstruction prediction consistency between source responses simulated by target images and reconstructed images helped preserve key semantic features related to the downstream analytical task while also retaining source domain knowledge. Additionally, an augmented dataset leveraging diverse

public medical images provided regularization to maintain performance on the source domain. Together, reconstruction prediction consistency and data augmentation constrained the image translation so that visually transformed images still contained crucial semantic content and source knowledge needed for the analytical task.

#### 3.2.1. Reconstruction prediction consistency

This module leverages the source model $F^{SM}$ to supervise the training of a clone model $F^{CM}$ on the source domain prediction instead of target domain prediction commonly used by previous SFDA methods. By minimizing a consistency loss, $L_{consistent}$, $F^{CM}$ learns to make predictions consistent with the source knowledge contained in $F^{SM}$.

For a target image group, $x^t$, comprising whole-level image $x_f^t$ and patch-level image $x_p^t$, and the corresponding reconstructed image group is $x^r$. The model $F^{CM}$ is initialized with the weights of $F^{SM}$. The parameters of $F^{CM}$ are updated during adaptation while the parameters of $F^{SM}$ remain fixed. During adaptation, the steps are described as follows:

1. The soft pseudo-label of masks, $y^t$, for source domain are obtained by simulating the model $F^{SM}$ with the target image group $x^t$.
2. The prediction masks, $y^r$, are obtained by simulating the model $F^{CM}$ using the image group $x^r$.
3. $L_{consistent}$ enforces consistency between the outputs of $F^{CM}$ and $F^{SM}$. Generally, $L_{consistent}$ uses the binary cross-entropy loss function, $F_{BCE}$, but can employ the additional Dice loss function, $F_{DICE}$, for further improvement as given by:

$$\begin{aligned} L_{consistent} &= F_{BCE}\left(y^r, y^t\right) OR\, F_{BCE}\left(y^r, y^t\right) \\ &\quad + F_{DICE}\left(y^r, y^t\right),\ F_{BCE}\left(\mathrm{x}_1,\ \mathrm{x}_2\right) \\ &= -(\mathrm{x}_2 \times \log(\mathrm{x}_1) + (1-\mathrm{x}_2) \times \log(1-\mathrm{x}_1)), \end{aligned}$$

$$F_{DICE}\left(\mathrm{x}_1, \mathrm{x}_2\right) = 1 - \frac{2 \times sum(\mathrm{x}_1 \times \mathrm{x}_2)}{sum(\mathrm{x}_1) + sum(\mathrm{x}_2)},$$

where log represents the natural logarithm.

#### 3.2.2. Data augmentation

Given the limited target domain data and lack of diversity, this module was used to expand and diversify the dataset.

For an augmentation image group, $x^{aug}$, comprising whole-level image $x_f^{aug}$ and patch-level image $x_p^{aug}$. The data augmentation module shares the same model, $F^{CM}$, with the source prediction consistency module. During adaptation, the steps are followed:

1. The soft pseudo-label of masks, $y^{SM}$, for source domain are obtained by simulating the model $F^{SM}$ with the target image group $x^{aug}$.
2. The prediction masks, $y^{CM}$, are obtained by simulating the model $F^{CM}$ using the image group $x^{aug}$.
3. An augmentation loss $L_{aug}$ enforces consistency between the outputs of $F^{CM}$ and $F^{SM}$. Generally, $L_{aug}$ uses $F_{BCE}$ but can employ an additional $F_{DICE}$ for further improvement as given by:

$$L_{aug} = F_{BCE}\left(y^{CM}, y^{SM}\right) OR\, F_{BCE}\left(y^{CM}, y^{SM}\right) + F_{DICE}\left(y^{CM}, y^{SM}\right).$$

### 3.3. Style feature alignment module

The purpose of this module was to match the style of the canonical form to the style of the source domain. Prior work had shown that batch normalization layers in trained models capture statistics representing the source style (Yang et al., 2022). We leveraged this by constraining the batch normalization statistics of the canonical image, $x_f^c$, to match the statistics derived from the source model. Specifically, the Wasserstein distance $L_{wasserstein}$ between those statistics is defined as:

$$L_{wasserstein} = \sum_{n=1}^{K} \left|\left|\mu_{FM}^n(t) - \mu_{SM}^n\right|\right|_2 + \left|\left|\sigma_{FM}^n(t) - \sigma_{SM}^n\right|\right|_2,$$

$$\mu_{FM}^n(t) = \mu_{BN}^n \times \alpha + \mu_{FM}^n(t-1) \times (1-\alpha),$$

$$\sigma_{FM}^n(t) = \sigma_{BN}^n \times \alpha + \sigma_{FM}^n(t-1) \times (1-\alpha),$$

where $\mu_{BN}^n$ and $\sigma_{BN}^n$ represent the mean and standard deviation of the $n^{th}$ batch normalization layer of the feature maps computed using the canonical image $x_f^c$. $t$ is current training iteration, $t$-1 is the previous training iteration and $\alpha$ is a scalar weight to balance statistical values between the iterations to calculate the running average, $\mu_{FM}^n(t)$ and $\sigma_{FM}^n(t)$. $\mu_{SM}^n$ and $\sigma_{SM}^n$ denote the running mean and the running standard deviation of the $n^{th}$ batch normalization layer of the source model, and $||\cdot||_2$ represent the L2 norm. The summation by $L_{wasserstein}$ is over the first K layers most likely to capture low-level style information. By minimizing this Wasserstein loss, the image generator was encouraged to output canonical images with a style aligned to the source domain while preserving target content.

### 3.4. Optimization summary

In each adaptation iteration, the target domain images are firstly augmented by random flipping and the following sequential steps are repeated until the pre-set number of training iterations are met:

1. Patch level image reconstruction to minimize the loss function $L_{patch}$, given by:

$$L_{patch} = L_{rebuild\_p} \times \lambda_{rebuild\_p} + L_{identity} \times \lambda_{identity},$$

Where $\lambda_{rebuild\_p}$ and $\lambda_{identity}$ are scalar weights balancing the two loss terms.

2. Whole level image reconstruction and image quality preservation to minimize the loss function $L_{whole}$, given by:

$L_{whole} = L_{wassertein} + L_{rebuild\_f} \times \lambda_{rebuild\_f} + L_{tv}$, where $\lambda_{rebuild\_f}$ is a scalar weight controlling the whole level image reconstruction.

3. Continual learning module to minimize the loss function $L_{CM}$, given by:

$$L_{CM} = L_{consistent} + L_{aug} \times \lambda_{aug},$$

where $\lambda_{aug}$ is a scalar weight balancing the two loss terms.

### 3.5. Experimental settings

#### 3.5.1. Datasets

**Prostate MRI dataset:** three publicly available prostate MRI datasets were utilized in this study: NCI-ISBI13 dataset (Roberts et al., 2021), I2CVB dataset (Yu et al., 2023) and PROMISE12 dataset (Stan and Rostami, 2021) (Table S1). NCI-ISBI13 dataset contained 2 sites, I2CVB dataset contained 1 site and PROMISE12 dataset contained 3 sites, for a total of 6 distinct sites. The NCI-ISBI13 and I2CVB datasets served as the source domain, comprising 3 sites total. The PROMISE12 dataset served as the target domain, with each of its 3 sites adapted individually. For each target domain site adaptation, the other 2 PROMISE12 sites were used for augmentation. For all datasets, 80 % of cases were assigned to training and 20 % to testing. All images were resized to 384 × 384 resolution.

**Colonoscopy Image dataset:** three publicly available colonoscopy video datasets were utilized in this study: CVC-ClinicDB (Vázquez et al., 2017), ETIS-Larib (Silva et al., 2014; Chen et al., 2020) and HyperKvasir (English 2020) (Table S2). The CVC-ClinicDB, ETIS-Larib and HyperKvasir served as source domain, target domain and augmentation dataset, respectively. Videos were examined to ensure images from the

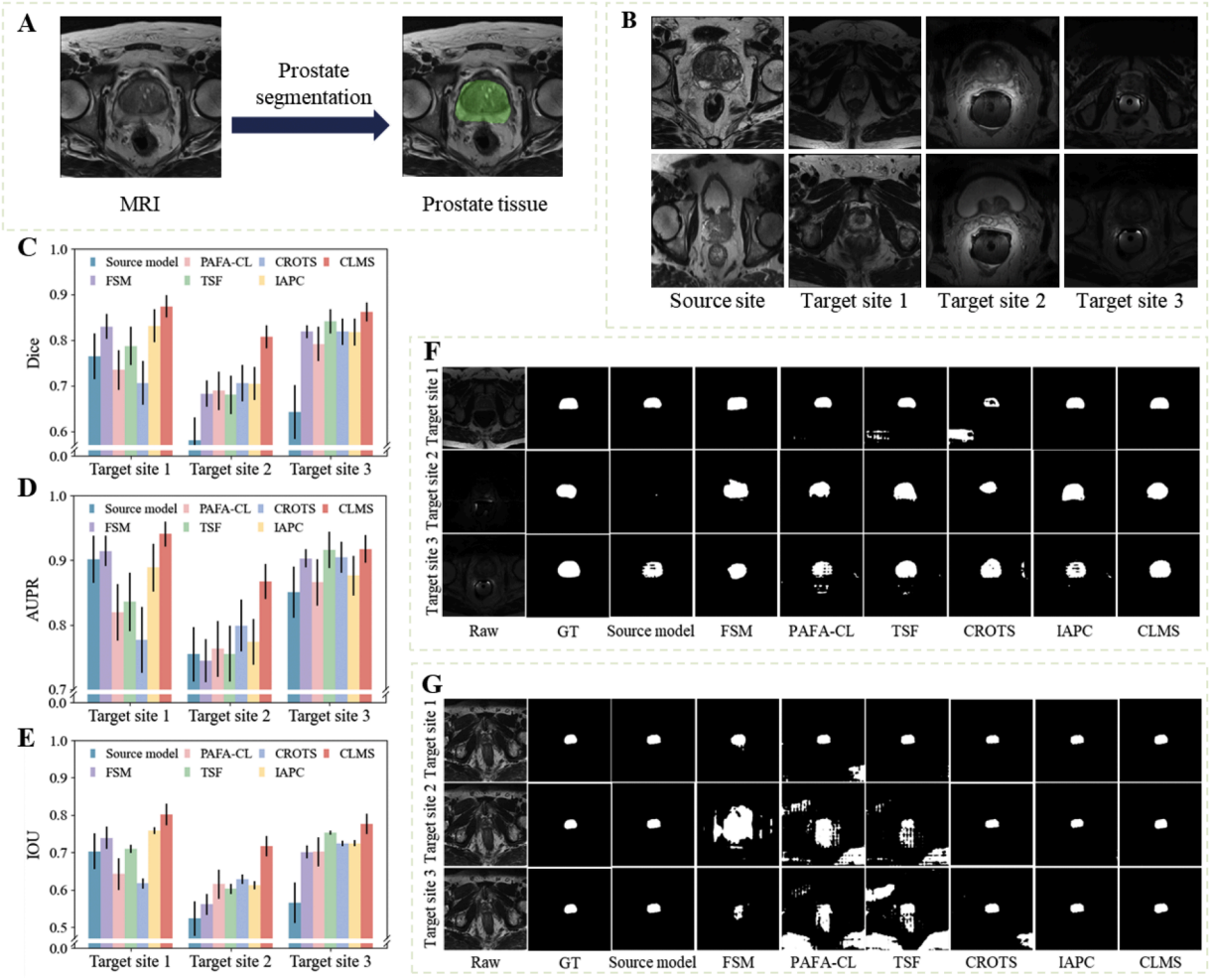


**Fig. 2.** | Experimental results on the prostate MRI datasets. (A) Demonstration of prostate segmentation from MRI that is critical for computer-aided diagnosis and treatment planning in prostate cancer. (B) The prostate MRI datasets (see Table S1 for the details) exhibit clear appearance disparities. The comparison of the performance of source-free domain adaptation methods on three target sites using three common segmentation metrics: Dice (C), AUPR (D), and IOU (E); the error bar represents the standard error mean (SEM). Visualization of prostate segmentation on the target domain (F) and source domain (G).

same segment were designated for either training or testing. 80 % of the colonoscopy images from each segment were used to comprise the training set, while the remaining 20 % were held out for testing. All images were resized to 384 × 384 resolution.

**Fundus image dataset:** three private fundus image datasets were collected from the following sites: Zhongshan Ophthalmic Center at Sun Yat-sen University (ZOC), Guangdong Women and Children Hospital Panyu Branch (PY), and Zhuhai Center for Maternal and Child Healthcare (OA) (Table S3). The PY dataset served as the source domain, the ZOC dataset served as the target domain, and the OA dataset was used for augmentation. All images were resized to a resolution of 512 × 512 pixels. 146 ROP fundus images from ZOC dataset served as the test set; the rest formed the training set. The test set were annotated with the presence of Plus disease by two experienced ophthalmologists (with over 10 years of experience) and reviewed by a clinical professor serving as the reference standard diagnosis (RSD).

### 3.5.2. Implementation details

*3.5.2.1. Prostate segmentation.* Firstly, we trained a source domain model $F^{SM}$ using the labeled source domain data from the prostate MRI dataset. The source model was formulated based on DeepLab-v2 using a ResNet-101 backbone. The parameters of source model were initialized on ImageNet, and the training batch size was 24 for 150 epochs, using the Adam optimizer with a learning rate of 1e-3.

In the adaptation stage, the weights of the $F^{CM}$ model was initialized using the weights of the $F^{SM}$ model and the adaptation using CLMS was based on the target domain dataset. The adaptation batch size was 4 for 50 epochs, using the Adam optimizer with a learning rate of 1e-4. Other hyperparameters are shown in Table S4.

*3.5.2.2. Polyp segmentation.* Firstly, we trained a source domain model $F^{SM}$ using labeled source domain data from the colonoscopy image dataset. The source model was formulated based on DeepLab-v2 using a ResNet-101 backbone. The parameters of source model were initialized on ImageNet, and the training batch size was 24 for 150 epochs, using the Adam optimizer with a learning rate of 1e-3.

In the adaptation stage, the weights of the $F^{CM}$ model was initialized using the weights of the $F^{SM}$ model, and the adaptation using CLMS was based on the target domain dataset and augmentation dataset. The adaptation batch size was 4 for 100 epochs, using the Adam optimizer with a learning rate of 1e-4. Other hyperparameters are shown in Table S4.

*3.5.2.3. Plus disease classification.* Firstly, we trained a source domain model $F^{SM}$ using pseudo-labeled source domain data from the fundus image dataset. The source model was formulated based on a UNet-type Retinal Segmentation Network ResUNet. The training batch size was 18

**Table 1**
| Benchmark results on prostate segmentation for target site 1.

| Method | Train dataset | | Target site 1 | | | | | Source site |
|---|---|---|---|---|---|---|---|---|
| | Source image | Target label | Dice (95%CI) | AUPR (95%CI) | IOU (95%CI) | Sensitivity (95%CI) | Specificity (95%CI) | Dice (95%CI) |
| Source model | √ | × | 0.7651 (0.666-0.864) | 0.9016 (0.830-0.974) | 0.7027 (0.607-0.798) | 0.7434 (0.642-0.844) | 0.9987 (0.998-0.999) | 0.9085 (0.900-0.917) |
| FSM[1] | × | × | 0.8299 (0.775-0.884) | 0.9143 (0.868-0.961) | 0.7393 (0.679-0.800) | 0.8548 (0.799-0.911) | 0.9965 (0.996-0.997) | 0.7949 (0.777-0.813) |
| PAFA-CL[1] | × | × | 0.7347 (0.650-0.820) | 0.8194 (0.733-0.906) | 0.6421 (0.557-0.727) | 0.7901 (0.705-0.875) | 0.9889 (0.983-0.994) | 0.8488 (0.834-0.864) |
| TSF[1] | × | × | 0.7872 (0.705-0.870) | 0.8356 (0.747-0.925) | 0.7097 (0.627-0.792) | 0.8046 (0.727-0.882) | 0.9942 (0.992-0.997) | 0.8917 (0.882-0.902) |
| CROTS[2] | × | × | 0.7063 (0.612-0.801) | 0.7765 (0.676-0.877) | 0.6172 (0.527-0.708) | 0.6783 (0.584-0.773) | 0.9923 (0.988-0.996) | 0.8840 (0.872-0.896) |
| IAPC[2] | × | × | 0.8313 (0.759-0.904) | 0.8888 (0.816-0.962) | 0.7584 (0.688-0.8290 | 0.8244 (0.749-0.900) | 0.9979 (0.997-0.998) | 0.8986 (0.890-0.907) |
| **CLMS** | × | × | **0.8738 (0.826-0.922)** | **0.9404 (0.902-0.979)** | **0.8017 (0.746-0.858)** | **0.8580 (0.803-0.913)** | **0.9984 (0.998-0.999)** | **0.9018 (0.894-0.910)** |
| LwF | × | √ | 0.8453 (0.779-0.912) | 0.9177 (0.859-0.977) | 0.7740 (0.707-0.841) | 0.8507 (0.779-0.923) | 0.9975 (0.997-0.998) | 0.8969 (0.887-0.907) |
| iCaRL | √ | √ | 0.8232 (0.750-0.897) | 0.9270 (0.873-0.981) | 0.7487 (0.667-0.817) | 0.8484 (0.773-0.924) | 0.9975 (0.997-0.998) | 0.9033 (0.894-0.913) |
| Target model | × | √ | 0.8323 (0.784-0.880) | 0.9139 (0.865-0.963) | 0.7387 (0.680-0.797) | 0.8548 (0.813-0.897) | 0.9965 (0.995-0.998) | 0.5828 (0.552-0.614) |

* 1: SFDA methods in medical imaging; 2: SFDA methods in natural imaging.
The best and second-best performing results are highlighted with underlines below.

**Table 2**
| Benchmark results on prostate segmentation for target site 2.

| Method | Train dataset | | Target site 2 | | | | | Source site |
|---|---|---|---|---|---|---|---|---|
| | Source image | Target label | Dice (95%CI) | AUPR (95%CI) | IOU (95%CI) | Sensitivity (95%CI) | Specificity (95%CI) | Dice (95%CI) |
| Source model | √ | × | 0.5813 (0.483-0.679) | 0.7546 (0.672-0.837) | 0.5237 (0.433-0.615) | 0.5689 (0.470-0.668) | 0.9967 (0.996-0.998) | 0.9085 (0.900-0.917) |
| FSM[1] | × | × | 0.683 (0.626-0.739) | 0.7446 (0.679-0.810) | 0.5608 (0.506-0.616) | 0.8058 (0.751-0.861) | 0.9829 (0.979-0.987) | 0.4016 (0.381-0.423) |
| PAFA-CL[1] | × | × | 0.6892 (0.606-0.772) | 0.7629 (0.678-0.848) | 0.6149 (0.537-0.693) | 0.7055 (0.617-0.794) | **0.9924 (0.991-0.994)** | 0.7366 (0.711-0.762) |
| TSF[1] | × | × | 0.6803 (0.598-0.762) | 0.7551 (0.670-0.840) | 0.6024 (0.525-0.680) | 0.7057 (0.622-0.789) | 0.9887 (0.985-0.992) | 0.6081 (0.578-0.639) |
| CROTS[2] | × | × | 0.7056 (0.627-0.784) | 0.7990 (0.720-0.878) | 0.6284 (0.552-0.704) | 0.7179 (0.636-0.800) | 0.9920 (0.989-0.995) | 0.8484 (0.831-0.866) |
| IAPC[2] | × | × | 0.7046 (0.633-0.776) | 0.7740 (0.704-0.844) | 0.6121 (0.544-0.681) | 0.7321 (0.661-0.803) | 0.9920 (0.989-0.995) | 0.8444 (0.830-0.859) |
| **CLMS** | × | × | **0.8078 (0.759-0.857)** | **0.8669 (0.814-0.920)** | **0.7172 (0.664-0.771)** | **0.8492 (0.801-0.897)** | **0.9924 (0.991-0.994)** | **0.8675 (0.856-0.879)** |
| LwF | × | √ | 0.7059 (0.625-0.787) | 0.8063 (0.731-0.881) | 0.6327 (0.556-0.709) | 0.7119 (0.625-0.798) | 0.9949 (0.994-0.996) | 0.6172 (0.587-0.647) |
| iCaRL | √ | √ | 0.8083 (0.752-0.865) | 0.9149 (0.872-0.958) | 0.7268 (0.670-0.784) | 0.8247 (0.766-0.884) | 0.9952 (0.994-0.996) | 0.8997 (0.890-0.909) |
| Target model | × | √ | 0.8559 (0.820-0.892) | 0.9166 (0.885-0.949) | 0.7727 (0.729-0.816) | 0.9007 (0.871-0.930) | 0.9952 (0.994-0.996) | 0.1673 (0.139-0.196) |

* 1: SFDA methods in medical imaging; 2: SFDA methods in natural imaging.
The best and second-best performing results are highlighted with underlines below.

**Table 3**
| Benchmark results on prostate segmentation for target site 3.

| Method | Train dataset | | Target site 3 | | | | | Source site |
|---|---|---|---|---|---|---|---|---|
| | Source image | Target label | Dice (95%CI) | AUPR (95%CI) | IOU (95%CI) | Sensitivity (95%CI) | Specificity (95%CI) | Dice (95%CI) |
| Source model | √ | × | 0.6424 (0.527-0.758) | 0.8503 (0.772-0.928) | 0.5652 (0.458-0.672) | 0.5856 (0.475-0.696) | 0.9991 (0.999-1.000) | 0.9085 (0.900-0.917) |
| FSM[1] | × | × | 0.8187 (0.791-0.846) | 0.9027 (0.875-0.931) | 0.7013 (0.666-0.737) | 0.8020 (0.773-0.831) | 0.9973 (0.996-0.998) | 0.4289 (0.396-0.461) |
| PAFA-CL[1] | × | × | 0.7917 (0.718-0.865) | 0.8657 (0.794-0.937) | 0.7016 (0.626-0.778) | 0.8119 (0.744-0.880) | 0.9926 (0.989-0.996) | 0.6367 (0.610-0.663) |
| TSF[1] | × | × | 0.8417 (0.789-0.894) | 0.9159 (0.861-0.971) | 0.7524 (0.695-0.810) | 0.7958 (0.742-0.850) | 0.9976 (0.996-0.999) | 0.5184 (0.488-0.549) |
| CROTS[2] | × | × | 0.8184 (0.762-0.875) | 0.9048 (0.857-0.952) | 0.7238 (0.659-0.789) | 0.7977 (0.730-0.866) | 0.9971 (0.996-0.998) | 0.8664 (0.854-0.879) |
| IAPC[2] | × | × | 0.8169 (0.759-0.875) | 0.8761 (0.816-0.937) | 0.7247 (0.656-0.794) | **0.8503 (0.808-0.892)** | 0.9945 (0.992-0.997) | 0.8804 (0.870-0.891) |
| **CLMS** | × | × | **0.8613 (0.821-0.902)** | **0.9173 (0.876-0.959)** | **0.7759 (0.722-0.830)** | 0.8035 (0.750-0.857) | **0.9991 (0.999-0.999)** | **0.8814 (0.872-0.891)** |
| LwF | × | √ | 0.8718 (0.844-0.900) | 0.9332 (0.903-0.964) | 0.7828 (0.743-0.823) | 0.8841 (0.851-0.917) | 0.9964 (0.995-0.997) | 0.7184 (0.900-0.917) |
| iCaRL | √ | √ | 0.8976 (0.872-0.924) | 0.9680 (0.953-0.983) | 0.8232 (0.786-0.861) | 0.9270 (0.909-0.945) | 0.9975 (0.997-0.998) | 0.8973 (0.692-0.744) |
| Target model | × | √ | 0.9033 (0.882-0.925) | 0.9753 (0.968-0.983) | 0.8303 (0.798-0.863) | 0.9189 (0.901-0.937) | 0.9979 (0.997-0.999) | 0.4045 (0.372-0.437) |

* 1: SFDA methods in medical imaging; 2: SFDA methods in natural imaging.
The best and second-best performing results are highlighted with underlines below.

for 30 epochs, using the Adam optimizer with a learning rate of 1e-3.

In the adaptation stage, the weights of the $F^{CM}$ model was initialized using the weights of the $F^{SM}$ model and the adaptation using CLMS was based on the target domain dataset and augmentation dataset. The adaptation batch size was 4 for 150 epochs, using the Adam optimizer with a learning rate of 1e-4. Other hyperparameters are shown in Table S4.

### 3.5.3. Benchmarking methods

We compared CLMS to several state-of-the-art SFDA methods, encompassing approaches applied in both medical and natural imaging domains. In the medical imaging category, we evaluated Fourier Style Mining (FSM), a two-stage framework involving image-level and self-training methods (Yang et al., 2022); Prototype-Anchored Feature Alignment and Contrastive Learning (PAFA-CL), which incorporates image information alignment and self-training methods (Yu et al., 2023); and Target-Specific Fine-tuning (TSF), a self-supervised based method (Li et al., 2023). In the natural imaging category, we evaluated Importance-Aware and Prototype-Contrast (IAPC) (Cao et al., 2024) and CROss domain Teacher-Student learning framework (CROTS) (Luo et al., 2024), both of which are based on self-supervised approaches.

Furthermore, to explore the effectiveness of continual learning in the SFDA scenario, we compared CLMS with established continual learning methods, specifically Learning without Forgetting (LwF) (Li and Hoiem, 2018) and Incremental Classifier and Representation Learning (iCaRL) (Rebuffi et al., 2017). Unlike SFDA approaches, these continual learning methods typically require more than just target domain data. Due to this requirement, we limited our comparison to MR and colonoscopy datasets, excluding the fundus image dataset which lacks target domain segmentation labels. We included these comparisons to provide a broader perspective on adaptation strategies and to illustrate the performance of continual learning paradigms in this context.

To evaluate performance, we used various metrics for segmentation and classification tasks. For segmentation, we calculated Dice score, Area Under the Precision-Recall curve (AUPR), Intersection over Union (IOU), sensitivity, and specificity. For classification, we computed Area Under the Receiver Operating Characteristic curve (AUC), weighted F1 score, type I error, type II error, sensitivity, and specificity. Detailed explanations of these metrics can be found in the supplementary section S2 Metrics.

### 3.5.4. Statistics analysis

All statistics analysis was using Wilcoxon-test by python package Scipy (v.1.7.3) and SPSS Statistics (R26.0.0.0). All plots were generated using python package brokenaxes (v.0.5.0) and matplotlib (v.3.5.3).

### 3.5.5. Ablation analysis

We conducted ablation analysis on the prostate MRI dataset and colonoscopy image dataset. For each of the target sites, we assessed the impact of removing modules on the performance of the CLMS framework. Each time, one of the modules, including multi-scale image reconstruction, style feature alignment, reconstruction prediction consistency, data augmentation and continual learning, was removed and the same optimization process was applied as previous demonstrated.

### 3.5.6. Architecture analysis

We conducted an architectural analysis of CLMS using prostate MRI and colonoscopy image datasets to investigate the crucial interplay

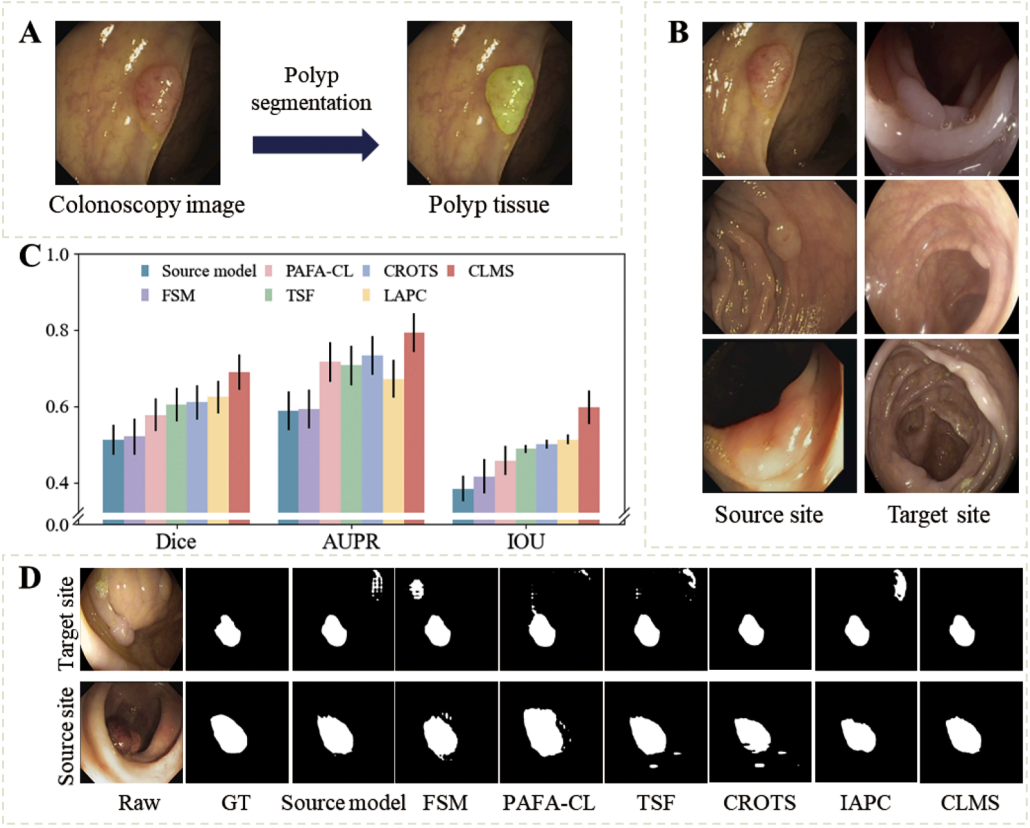


**Fig. 3.** | Experimental results on the colonoscopy dataset. (A) Demonstration of polyp segmentation from a colonoscopy image that is critical for early diagnosis and treatment of colonoscopy cancer. (B) The colonoscopy image datasets (see Table S2 for the details) exhibit clear appearance disparities. (C) The comparison of the performance of source-free domain adaptation methods on target sites using average dice, AUPR, and IOU; the error bar represents SEM. (D) Visualization of polyp segmentation on the target domain and source domain.

between its key components. Our analysis focused on two critical modifications:

(1) In the continual learning module, we changed the input to the $F^{CM}$ model from the reconstructed target image to the canonical form (Fig. S1). This alteration tests the importance of maintaining target domain characteristics during finetuning, as the reconstructed target image preserves detailed target information while the canonical form represents a more source-like representation.
(2) In the style feature alignment module, we switched the input from the canonical form to the reconstructed target image (Fig. S2). This change examines the impact of aligning low-level visual features directly on target-like images rather than on the source-like canonical form.

## 4. Results

### 4.1. Evaluation of CLMS in prostate segmentation on MRI

Prostate cancer ranked among the most prevalent male malignancies, with MRI serving as a primary diagnostic tool renowned for its heightened detectability (Vente et al., 2021). The accurate segmentation of the prostate from MRI images facilitates the detection, treatment planning, and therapeutic evaluation of this form of cancer (Khan et al., 2021) (Fig. 2A). However, considerable variations in MRI intensities across different medical centers, stemming from diverse scanning protocols, posed a notable obstacle. These disparities rendered the existing models less adaptable to domain shift that alter the image appearance of prostate boundaries and tissue heterogeneity, further compounded by the absence of labeled data for fine-tuning.

We employed prostate T2-weighted MRI datasets sourced from six different medical sites, each characterized by distinct scanning protocols and devices (Table S1) (Liu et al., 2020). We used images from three sites as source sites and the other sites as target sites. Images from these different sites displayed significant visual disparities, highlighting substantial style variations across domains (Fig. 2B). Compared to the images from the target domain, the source domain images had brighter grayscale, larger visible tissue coverage, and higher contrast between the prostate and surrounding tissues, accompanied by sporadic bright speckles. Overall, these substantial discrepancies in quality and visuals posed considerable challenges for model generalization.

We combined the source sites 1, 2, and 3, characterized by relatively consistent data profiles, as the source domains for training a model according to a previous study (Liu et al., 2020), a prostate segmentation source model achieved average Dice score (Dice) of 90.85 % (95 % CI = 0.900–0.917). When applied to the target sites 1, 2 and 3, Dice decreased to 76.51 % (95 % CI = 0.666–0.864; $n$ = 43), 58.13 % (95 % CI = 0.483–0.679; $n$ = 71), and 64.24 % (95 % CI = 0.527–0.758; $n$ = 40), respectively, showing a significant performance drop (Table 1–3). Training models on datasets from three source sites, either individually or collectively, also led to a noteworthy performance reduction on the unseen target sites, despite the enhancement across the seen source sites (Table S5). This observation suggested that while collective supervised learning, such as federated learning (Dayan et al., 2021), could yield excellent generalization across the domains within the training data distributions, it encounters considerable performance degradation when encountering with new, out-of-distribution data. Furthermore, bimodal performances were observed, where the higher peak represented good generalization to target data similar to source, while the lower peak reflected poor generalization to dissimilar target data (Fig. S3A).

**Table 4**
| Benchmark results on polyp segmentation.

| Method | Train dataset | | Target site | | | | | Source site |
|---|---|---|---|---|---|---|---|---|
| | Source image | Target label | Dice (95%CI) | AUPR (95%CI) | IOU (95%CI) | Sensitivity (95%CI) | Specificity (95%CI) | Dice (95%CI) |
| Source model | √ | × | 0.5122 (0.436-0.588) | 0.5882 (0.487-0.690) | 0.3839 (0.316-0.451) | 0.7108 (0.605-0.817) | 0.9569 (0.947-0.967) | 0.7779 (0.746-0.810) |
| FSM[1] | × | × | 0.5200 (0.427-0.613) | 0.5928 (0.493-0.692) | 0.4161 (0.327-0.506) | 0.5522 (0.442-0.662) | 0.9851 (0.981-0.989) | **0.7772 (0.745-0.809)** |
| PAFA-CL[1] | × | × | 0.5773 (0.493-0.661) | 0.7156 (0.613-0.818) | 0.4573 (0.382-0.533) | **0.7824 (0.684-0.880)** | 0.9527 (0.941-0.965) | 0.7156 (0.676-0.755) |
| TSF[1] | × | × | 0.6034 (0.518-0.689) | 0.7066 (0.603-0.810) | 0.4881 (0.409-0.567) | 0.7616 (0.662-0.861) | 0.9629 (0.952-0.973) | 0.7179 (0.678-0.758) |
| CROTS[2] | × | × | 0.6097 (0.521-0.699) | 0.7323 (0.632-0.832) | 0.5005 (0.416-0.585) | 0.6669 (0.558-0.776) | **0.9859 (0.980-0.991)** | 0.6401 (0.589-0.691) |
| IAPC[2] | × | × | 0.6238 (0.538-0.709) | 0.6715 (0.574-0.769) | 0.5126 (0.430-0.596) | 0.6950 (0.596-0.794) | 0.9798 (0.972-0.987) | 0.6943 (0.647-0.741) |
| **CLMS** | × | × | **0.6895 (0.598-0.781)** | **0.7932 (0.693-0.894)** | **0.5966 (0.510-0.684)** | 0.7633 (0.667-0.859) | 0.9813 (0.974-0.989) | 0.7638 (0.729-0.799) |
| LwF | × | √ | 0.6197 (0.538-0.702) | 0.6653 (0.571-0.760) | 0.5027 (0.424-0.581) | 0.7296 (0.633-0.827) | 0.9789 (0.973-0.985) | 0.7604 (0.724-0.797) |
| iCaRL | √ | √ | 0.6006 (0.517-0.684) | 0.6631 (0.563-0.763) | 0.4823 (0.405-0.560) | 0.7329 (0.633-0.832) | 0.9717 (0.965-0.979) | 0.7653 (0.733-0.798) |
| Target model | × | √ | 0.6281 (0.528-0.729) | 0.7288 (0.622-0.836) | 0.5380 (0.443-0.633) | 0.7253 (0.628-0.823) | 0.9579 (0.941-0.975) | 0.6898 (0.651-0.728) |

* 1: SFDA methods in medical imaging; 2: SFDA methods in natural imaging.
The best and second-best performing results are highlighted with underlines below.

CLMS demonstrated superior segmentation performance compared to its counterparts (Fig. 2C - 2E). Specifically, for the target site 1, CLMS achieved the top performance in Dice, AUPR, IOU, sensitivity and specificity (Fig. 2C and S4A, Table 1). Notably, CLMS attained the highest Dice of 87.38 % (95 % CI = 0.826–0.922), representing a significant improvement of 10.87 % over the source model ($P < 0.001$). Furthermore, CLMS significantly outperformed FSM, PAFA-CL, TSF, CROTS and IAPC by 4.39 % ($P < 0.0001$), 13.91 % ($P < 0.0001$), 8.66 % ($P < 0.0001$), 16.75 % ($P < 0.0001$) and 4.25 % ($P < 0.01$), respectively (Fig. S4A and Table 1). Additionally, CLMS demonstrated a higher median and a narrower interquartile range of Dice, indicative of a more stable and robust overall segmentation performance (Fig. S4A). For the target site 2, CLMS achieved the top position for all the evaluation metrics except specificity, securing the highest Dice of 80.78 % (95 % CI = 0.759–0.857), surpassing the source model by 22.65 % ($P < 0.0001$), and FSM, PAFA-CL, TSF, CROTS and IAPC by 12.48 % ($P < 0.0001$), 11.86 % ($P < 0.0001$), 12.75 % ($P < 0.0001$), 10.22 % ($P < 0.0001$) and 10.32 % ($P < 0.0001$), respectively (Fig. 2D and S4B, Table 2). For the target site 3, CLMS achieved the top position for all evaluation metrics except sensitivity, and secured the highest Dice of 86.13 % (95 % CI = 0.821–0.902), surpassing the source model by 21.89 % ($P < 0.0001$), and FSM, PAFA-CL, TSF, CROTS and IAPC by 4.26 % ($P < 0.01$), 6.96 % ($P < 0.05$), 1.96 %, 4.29 % and 4.44 %, respectively (Fig. 2E and S4C, Table 3).

Moreover, CLMS effectively preserved model performance on the source domain during adaptation, as shown in Tables 1-3. Specifically, CLMS achieved Dice of 90.18 %, 86.75 % and 88.14 % on the target sites 1, 2 and 3, respectively. These results surpassed FSM, TSF, PAFA-CL, CROTS and IAPC for all the target sites, except a tie with TSF on the target site 1 and with IAPC on target sites 1 and 3. In addition, CLMS was the only approach that prevented median Dice reduction across the target sites (Fig. S4). Further analysis uncovered while all the methods improved on low dice images, which represented target images dissimilar to source (Dice <0.75), CLMS was the top across all the target sites. Notably, CLMS was the only method that maintained segmentation accuracy on high Dice images, which represented target images similar to source (Dice $\geq$ 0.75) (Fig. S5). Ideally, SFDA should improve both peaks (Fig. S3B). However, the results demonstrated that existing SFDA techniques boosted the lower peak by catastrophically interfering with the higher peak during adaptation (Fig. S3C). This highlighted CLMS's dual capacity for preserving knowledge on source-consistent data while adapting to novel target characteristics.

Unlike established SFDA methods discussed earlier, established continual learning approaches successfully maintained performance on the source domain during the adaptation, achieving Dice scores comparable to the source domain model (Tables 1–3). Notably, these methods not only preserved but also improved performance on images similar to the source domain (Dice $\geq$ 0.75) (Fig. S6), highlighting the effectiveness of continual learning approaches in preventing catastrophic forgetting and promoting positive transfer.

While both CLMS and established continual learning approaches demonstrated exceptional performance across multiple target domains, CLMS achieved comparable results with significantly less information. In the comparison across three target domains (Fig. S7, Table 1–3), iCaRL outperformed CLMS in two sites, while CLMS excelled in one. Specifically, for target domain 1, CLMS surpassed both LwF and iCaRL across all metrics, achieving a Dice score 2.85 % higher than LwF and 5.06 % higher than iCaRL. In target domain 2, CLMS again outperformed LwF across all metrics, with a Dice score 10.19 % higher, and nearly matched iCaRL's performance with a Dice score only 0.05 % lower. In target domain 3, CLMS's Dice score was 1.05 % and 3.63 % lower than LwF and iCaRL, respectively. Notably, CLMS attained these comparable results using only unlabeled target images, whereas iCaRL required both source images and target labels, and LwF utilized target labels. This stark contrast in data requirements underscores CLMS's superior effectiveness in leveraging limited information compared to existing continual learning methods.

The segmentation masks further demonstrated that CLMS generated

more accurate and clearer prostate segmentation against the ground truth (Fig. 2F). In contrast, the other methods performed worse on all the target sites, producing many false positive segmentations. On the other hand, after adapting to the target sites, CLMS also precisely locate the position of prostate on the source site while other methods misidentified a significant amount of tissue around the intestinal prostate as prostate (Fig. 2G). CLMS showcased superior and more robust segmentation performance across all the three target domains (Fig. 2F) and the source domain (Fig. 2G). These qualitative results further affirm the efficacy and robustness of CLMS.

### 4.2. Evaluation of CLMS in polyp segmentation on colonoscopy images

Colorectal cancer (CRC) stands out as one of the most prevalent adenocarcinomas, primarily manifesting in the colon or rectum, with 80–95 % of cases originating from adenomatous polyps (Alzahrani et al., 2021). The most effective screening method for the early diagnosis and treatment of CRC in clinical practice is colonoscopy examination, and previous studies showed that the segmentation of colon polyps from colonoscopy images using deep learning can assist clinicians in the early detection of polyps (Biffi et al., 2022) (Fig. 3A). However, differences across medical centers introduced significant variations in the color distribution of colonoscopy images. These variations hindered the ability of existing models to adapt to domain shift characterized by markedly distinct image appearances.

We employed three colonoscopy video datasets, namely CVC-ClinicDB (Vázquez et al., 2017), ETIS-Larib (Silva et al., 2014; Chen et al., 2020), and HyperKvasir (English 2020) (Table S2). In this evaluation, the CVC-ClinicDB dataset, characterized by higher image quality and a larger volume, was designated as the source domain, the ETIS-Larib dataset as the target domain, and the HyperKvasir dataset was employed for data augmentation because of its arbitrary order and inability to be manually categorized according to video sequence. Notable discrepancies were evident upon visual inspection of colonoscopy images between the source and target domains (Fig. 3B). The source images exhibited uneven illumination, containing overexposed and underexposed areas. Overall color tone skewed yellow with lower contrast, obscuring certain vascular details. In comparison, the target images displayed more consistent lighting, a redder hue, and clearer vessel and structural definition. Furthermore, while the polyp segmentation model achieved Dice of 77.79 % on the source domain, Dice declined markedly to 51.22 % (95 % CI = 0.436–0.588; $n = 46$) when applied to the target data (Table 4). In summary, substantial domain divergences existed in image quality and visual characteristics, presenting challenges for direct generalization of models across diverse colonoscopy image datasets.

Again, CLMS demonstrated superior colon polyp segmentation performance on the target domain compared to its counterparts. Specifically, CLMS achieved the highest scores across key metrics, including Dice, AUPR, and IOU, and the second in sensitivity (Fig. 3C and Table 4). CLMS attained top Dice of 68.95 % (95 % CI = 0.598–0.781; $n = 46$), significantly improving 17.73 % ($P < 0.0001$) over the source model. Furthermore, it surpassed FSM, PAFA-CL, TSF, CROTS, and IAPC by significant improvements of 16.95 % ($P < 0.0001$), 11.22 % ($P < 0.0001$), 8.61 % ($P < 0.001$), 7.98 % ($P < 0.001$), and 6.57 % ($P < 0.05$), respectively (Fig. 3C and S8, Table 4). Notably, CLMS even exceeded supervised training on the target data alone (the target model) for Dice, sensitivity, and specificity. This validated CLMS's capacity to fuse cross-domain knowledge and learn invariant feature representations for reliable generalization.

Similarly, CLMS effectively preserved source domain performance during adaptation (Table 4). Specifically, compared to SFDA methods, CLMS achieved a Dice of 76.38 %, close to the source model. This surpassed PAFA-CL, TSF, CROTS, and IAPC methods by significant

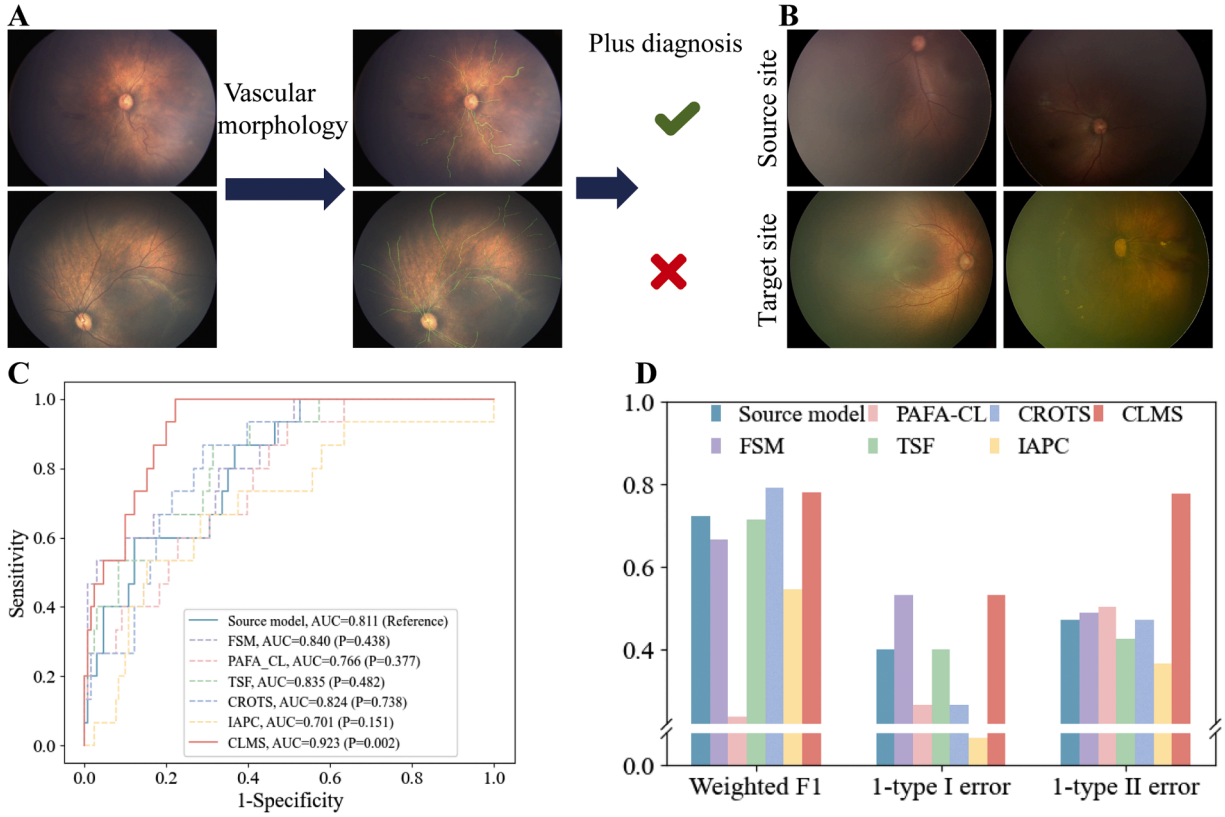


**Fig. 4.** | Experimental results on the retinal images dataset. (A) Demonstration of Plus disease diagnosis pipeline, involving exam retinal vascular morphology and diagnosis Plus disease. (B) The retinal image dataset (see Table S3 for the details) exhibits clear appearance color disparities. The comparison of the performance of source-free domain adaptation methods on the target site using ROC-area (C), weighted F1, type I error (0.05), and type II error (0.05) (D).

**Table 5**
| Comparison of the SFDA methods on plus classification.

| Method | AUC | weighted F1 | 1-Type I Error | 1-Type II Error | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Source model | 81.07 | 72.47 | 40.00 | 47.31 | 80.00 | 64.12 |
| FSM[1] | <u>84.02</u> | 66.70 | **53.33** | 48.84 | 86.67 | 55.73 |
| PAFA-CL[1] | 76.64 | 23.83 | 26.67 | 50.37 | **100.00** | 13.74 |
| TSF[1] | 83.51 | 71.35 | 40.00 | 42.73 | 86.67 | 61.83 |
| CROTS[2] | 82.44 | **79.31** | 26.67 | 47.32 | 73.33 | **74.81** |
| IAPC[2] | 70.07 | 54.62 | 6.67 | 36.62 | 80.00 | 41.98 |
| **CLMS** | **92.26** | <u>77.96</u> | **53.33** | **77.86** | **100.00** | 69.47 |

* 1: SFDA methods in medical imaging; 2: SFDA methods in natural imaging.
The best and second-best performing results are highlighted with underlines below.

**Table 6**
| CLMS architecture analysis results.

| Experiment | | | | | Prostate segmentation | | | Polyp segmentation |
|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | Target site 1 | Target site 2 | Target site 3 | |
| Ablation analysis | × | ● | ● | ● | 0.8545 | 0.8065 | 0.8367 | 0.5842 |
| | ● | × | ● | ● | 0.8351 | 0.7814 | 0.8543 | 0.5708 |
| | ● | ● | × | ● | 0.8675 | 0.7786 | 0.8466 | 0.6508 |
| | ● | ● | ● | × | 0.8652 | 0.7149 | 0.8291 | 0.5783 |
| | ● | × | × | ● | 0.7639 | 0.6999 | 0.8175 | 0.5767 |
| Modules B: $x^r \rightarrow x^c$ | | | | | 0.8600 | 0.7694 | 0.8232 | 0.5923 |
| Modules D: $x^c \rightarrow x^r$ | | | | | 0.8427 | 0.6876 | 0.8463 | 0.5484 |
| **CLMS** | | | | | **0.8738** | **0.8078** | **0.8613** | **0.6895** |

* $x^c$: the target image; $x^r$: the reconstructed target image; $x^c$: the canonical form.
A: Multi-scale image reconstruction; B: Reconstruction prediction consistency; C: Data augmentation; D: Style feature alignment.
●: Including this module; ×: Excluding this module.

improvements of 4.82 %, 4.59 %, 12.37 %, and 6.95 %, respectively, and matched FSM. While all the methods, including CLMS, improved on dissimilar, low Dice cases (Dice $< 0.75$), CLMS demonstrated the highest gains (Fig. S9). Further analysis revealed CLMS as the sole approach maintaining accuracy on high Dice images (Dice $\geq 0.75$) (Fig. S9). This again highlighted CLMS's unique capacity to prevent catastrophic forgetting of source knowledge, enabling robust cross-domain generalization.

As expected, established continual learning methods all preserved source domain performance during adaptation. Further analysis revealed these methods maintained accuracy on images similar to the source domain (Dice $\geq$ 0.75) (Fig. S10). These findings highlight continual learning methods unique capacity to prevent catastrophic forgetting of source knowledge while learning new information, enabling robust cross-domain generalization.

CLMS exhibited exceptional performance across target domains while requiring significantly less information than other established continual learning methods. CLMS outperformed both LwF and iCaRL across all metrics, with Dice scores improving by 6.98 % ($P < 0.001$) and 8.89 % ($P < 0.0001$), respectively (Fig. S11, Table 4). Notably, CLMS achieved these superior results using only unlabeled target images, while other methods required additional labeled data. These outcomes align with the findings from the MRI experiment, further confirming CLMS's remarkable effectiveness in leveraging limited information.

Visual inspection revealed key differences in segmentation quality between CLMS and the other methods. On the target domain, CLMS produced markedly more accurate and defined colon polyp boundaries, while the other methods displayed errors, particularly false positives (Fig. 3D). This adaptation performance affirmed CLMS's ability to mitigate domain gaps and learn transferable representations. On the source domain, CLMS maintained strong segmentation fidelity, generating clearer predictions (Fig. 3D). Competitors exhibited degraded output, indicating difficulty preserving source knowledge. These visual validations spotlighted the advantage of CLMS for sensitive medical image analysis, preserving detailed tissue patterns without target overfitting.

### 4.3. Evaluation of CLMS in retinopathy of prematurity classification

Retinal diseases stand out as a leading cause of visual impairment and blindness in children worldwide, with retinopathy of prematurity (ROP) being a major contributor (Yildiz et al., 2020; Blencowe et al., 2013). Symptoms include distorted peripheral vessels, fibrovascular growth, and neovascularization. Plus disease represents widening and tortuosity of retinal vessels, necessitating urgent treatment when observed (Gopal et al., 1995). Thus, early Plus disease identification can prevent ROP progression and vision loss. Clinically, diagnosis relies on ophthalmologists manually examining wide-field retinal images

collected from a wide-angle fundus camera (Roth et al., 2001), looking for Plus signs like peripheral dilation, tortuosity, and iris engorgement. Ideally, automated retinal image analysis using deep learning can enhance the efficiency and accuracy of Plus disease diagnosis (Fig. 4A).

We used three fundus datasets from Guangdong Women and Children Hospital at Panyu (PY), Zhongshan Ophthalmic Center, Sun Yat-sen University (ZOC) and ZhuHai Center for Maternal and Child healthcare (ZH) (Table S3). We used PY as the source, ZOC as the target, and ZH as the augmentation. We observed distinct differences between the source and the target fundus images (Fig. 4B). The source images had an overall reddish tone with lower contrast and clarity, mainly showing venous vessels. The target images were greenish, with higher contrast, and showed both arteries and veins. Such domain gaps present a challenge for clinical translation.

We adapted a deep learning system for Plus disease diagnosis, which involved two modules: a segmentation module for retinal vessel morphology, and a classification module for analyzing vessel morphology to detect Plus disease (Fig. 4A) (Yildiz et al., 2020). The model achieved an AUC of 81.1 % on the target data for Plus disease classification. We adapted only the segmentation module on target data while keeping the classification module fixed to validate if adapting just the segmentation module alone effectively transfers vascular morphology knowledge and integrates properly with the original classifier. Our analysis showed that CLMS was the only method that significantly improved AUC by 11.2 % to 92.3 % ($P < 0.01$). Other SFDA methods showed no significant changes from source performance (Fig. 4C). Moreover, CLMS achieved the highest metrics including sensitivity and Type I/II error rates at 0.05 significance level (Fig. 4D and Table 5). This affirmed that CLMS works not only by removing the discrepancies across domains but also retains the morphological structures necessary for clinical analysis.

Due to the severity of ROP and the urgency for early treatment, the screening method should minimize Type II error as much as possible to avoid missing disease detection and consequently missing the optimal treatment window. Compared with SFDA methods, it turned out that CLMS notably reduced the Type II error rate (22.14 %), whereas source model, FSM, PAFA-CL, TSF, CROTS, and IAPC demonstrate Type II error rates of 52.69 %, 51.16 %, 49.63 %, 57.27 %, 52.68 %, and 63.38 %, respectively (Table 5). Furthermore, while CLMS and PAFA-CL both achieved no missing disease detection (100 % sensitivity) in detecting Plus lesions, PAFA-CL (13.74 %) exhibits significantly lower specificity compared to CLMS (69.47 %). At the same time, in these methods, including the source model, all except PAFA-CL must modify the original Plus disease classifier to achieve no missing disease detection, complicating the adaptation.

In summary, CLMS minimizes missed diagnosis rates through lower Type II error while also reducing misdiagnoses through higher specificity, allowing it to retain morphological features critical for clinical diagnosis. Compared to other methods, CLMS adapted segmentation to new domains while excelling at Plus lesion detection, critical for this vision-threatening disease requiring urgent treatment.

### 4.4. CLMS Architecture Analysis

#### 4.4.1. Ablation analysis

We conducted comprehensive ablation experiments on both prostate and polyp segmentation tasks to assess the impact of each module within the CLMS framework (Table 6). The results consistently demonstrated that the removal of any individual module led to a decrease in performance compared to the full framework, thus confirming each component's significant contribution to the overall performance gain.

In both segmentation scenarios, the removal of the continual learning module resulted in the most significant decline in Dice scores. For prostate segmentation, this led to decreases of 10.99 %, 10.79 %, and 4.38 % for target sites 1, 2, and 3, respectively. Similarly, for polyp segmentation, it caused an 11.28 % Dice decline. These findings underscore the module's crucial role in facilitating adaptation.

To further elucidate the efficacy of the continual learning module, we disaggregated it into two components: the reconstruction prediction consistency module and the data augmentation module. For prostate segmentation, the individual absence of these components led to declines of 3.87 % and 0.63 % for target site 1, 2.64 % and 2.92 % for target site 2, and 0.7 % and 1.47 % for target site 3, respectively. In polyp segmentation, their individual absence resulted in declines of 11.87 % and 3.87 %, respectively. This proved that both target and external unlabeled data facilitate better adaptation. The improvements aligning with incorporated modules demonstrated the enhanced domain adaptation capability of our full framework.

#### 4.4.2. Architecture analysis

We conducted an architectural analysis of CLMS using prostate MRI and colonoscopy image datasets to investigate the crucial interplay between its key components. This analysis aimed to elucidate the roles of various connection routes within the CLMS framework and their effects on domain adaptation.

In one set of experiments, we modified continual learning module by substituting the reconstructed target image with the canonical form, which led to performance drops across all tests (Fig. S1). For prostate segmentation, the three target domain centers showed performance decreases of 1.38 %, 3.84 %, and 3.81 %. The polyp segmentation task experienced a more pronounced drop of 9.72 %. This performance decline likely stems from the canonical form and the target image belonging to different domains, potentially leading to label inconsistencies, disrupting the target-to-canonical mapping, and hindering the retention of high-level anatomical features.

A similar trend was observed when modifying style feature alignment module to use the reconstructed target image instead of the canonical form (Fig. S2). For prostate segmentation, performance declines of 3.11 %, 12.02 %, and 1.5 % were noted for the three target domain centers, while polyp segmentation performance dropped more dramatically by 14.11 %. This decrease in performance may be attributed to the stylistic differences between the reconstructed target image and the source domain features, leading to misalignment in feature mapping. Additionally, the reconstructed target image receives conflated style features from both source and target domains, making it challenging to separate them effectively.

## 5. Discussion

This study proposed CLMS, a novel source-free domain adaptation framework leveraging multi-scale image reconstruction, continual learning, and style feature alignment. CLMS demonstrates several key strengths. First, it effectively retains source domain performance while adapting to new target data, allowing it to generalize medical knowledge from the source without overfitting to the target. This dual learning capability is vital for maintaining robustness across domains. Second, unlike self-training approaches, CLMS avoids potentially inaccurate pseudo-labels that can compromise the learning process. Moreover, existing medical SFDA solutions often employ multi-step cascaded training, which can propagate errors. CLMS instead uses end-to-end joint optimization in a single model, reducing error accumulation and enabling more focused learning. This study demonstrates the efficacy of the proposed CLMS framework for source-free domain adaptation across three important medical image analysis tasks: prostate segmentation from MRI, colon polyp segmentation from colonoscopy images, and Plus disease classification from retinal images. CLMS consistently achieved state-of-the-art performance across all target domains, significantly outperforming existing methods like FSM, PAFA-CL, TSF, CROTS, and IAPC.

The continual learning module is essential for effective adaptation in CLMS. Its absence led to the most significant performance declines in ablation studies, underscoring the necessity of utilizing both target and

additional public data to reduce domain gaps while preserving source knowledge. One of the key benefits of the continual learning approach is CLMS's ability to adapt models to new target domains without sacrificing performance on the original source domain. In contrast, other SFDA methods may perform well on dissimilar target images but often lose accuracy on source-consistent data due to catastrophic forgetting. Established continual learning methods, however, not only maintain but can also enhance performance on images similar to the source domain. While inheriting this valuable characteristic of continual learning, CLMS demonstrated outstanding performance across target domains while requiring significantly less information than other continual learning methods. This capability highlights CLMS's strength in learning invariant representations, such as morphological structures, that transfer effectively across domains without overfitting to target distributions.

CLMS demonstrated efficacy across multiple medical imaging modalities and analysis tasks. For prostate and colon polyp segmentation, it directly adapted the segmentation model to the target domain images. Moreover, it improved integrated diagnosis in the Plus disease classification task by adapting only the retinal vessel segmentation module. This highlighted its ability to retain morphological features critical for subsequent clinical classification, beyond just removing domain gaps.

While CLMS showed reliable performance across modalities and datasets, evaluating CLMS on larger-scale target datasets with higher resolution images could better characterize its computational efficiency and scalability for real-time clinical analysis. Additionally, deploying CLMS in actual diagnostic settings with clinical evaluation would be essential to fully validating its utility and generalizability for enhancing real-world imaging pipelines. Finally, expanding the evaluation to more extensive manual labeling tasks beyond segmentation and classification would further verify its adaptable performance across diverse medical applications.

## 6. Conclusion

The study introduced CLMS, a novel end-to-end source-free domain adaptation framework that effectively bridges domain gaps across diverse medical imaging modalities. CLMS demonstrated superior performance compared to existing methods in segmentation tasks on prostate MRI, colonoscopy, and retinal image datasets. Crucially, CLMS retained crucial morphological features during adaptation, enabling seamless integration with downstream clinical analysis pipelines. Comprehensive evaluations validated CLMS's ability to adapt models to new target domains while preventing catastrophic forgetting of source knowledge. Overall, CLMS presents a promising solution for translating deep learning models to varied clinical settings, advancing medical imaging and healthcare applications.

## CRediT authorship contribution statement

**Weilu Li:** Writing – original draft, Data curation, Conceptualization, Formal analysis, Methodology, Visualization. **Yun Zhang:** Methodology, Writing – original draft, Conceptualization, Data curation, Writing – review & editing. **Hao Zhou:** Methodology. **Wenhan Yang:** Data curation. **Zhi Xie:** Writing – review & editing, Funding acquisition. **Yao He:** Writing – review & editing, Supervision, Methodology, Conceptualization, Project administration, Validation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### *Code availability*

The codes and algorithms developed for this study are available at GitHub (https://github.com/xie-lab/CLMS).

### *Benchmarking methods availability*

The compared SFDA methods include: FSM (open-sourced at https://github.com/CityU-AIM-Group/SFDA-FSM), PAFA-CL (open-sourced at https://github.com/CSCYQJ/MICCAI23-ProtoContra-SFDA), TSF (open-sourced at https://github.com/listiral/Source-Free-Cross-Tissues-Histopathological-Cell-Segmentation_Unet), CROTS (open-sourced at https://github.com/luoxin13/CROTS), and IAPC (open-sourced at https://github.com/yihong-97/Source-free-IAPC). We used the default parameters provided by these opensource methods.

The compared continual learning methods include LwF (open-sourced at https://github.com/MasLiang/Learning-without-Forgetting-using-Pytorch) and iCaRL (open-sourced at https://github.com/NiccoloCavagnero/IncrementalLearning).

### *Data availability*

The prostate and polyp images used in this study are shared publicly. The prostate image is open-sourced at https://liuquande.github.io/SAML/, and the polyp images include CVC-ClinicB (open-sourced at https://www.kaggle.com/datasets/balraj98/cvcclinicdb), ETIS-Larib (open-sourced at https://www.kaggle.com/datasets/nguyenvoquocduong/etis-laribpolypdb) and HyperKvasir (open-sourced at https://datasets.simula.no/hyper-kvasir/). Fundus images may be available for research purposes from the corresponding authors upon reasonable request.

### *Authors' contributions*

All authors read and approved the manuscript

### *Acknowledgments*

We would like to thank all the data and software contributors who made this research possible. We also thank the Center for Precision Medicine at Sun Yat-sen University for the long-term support.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2024.103404.

## Data availability

The data are publicly available. The code is available at https://github.com/xie-lab/CLMS

## Reference

Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25 (1), 44–56.
Wynants, L., et al., 2020. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 369, m1328.
Roberts, M., et al., 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat. Mach. Intell. 3 (3), 199–217.
De Fauw, J., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24 (9), 1342–1350.
Zhang, L., et al., 2020. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEe Trans. Med. ImAging 39 (7), 2531–2540.
Ju, L., et al., 2021. Leveraging regular fundus images for training UWF fundus diagnosis models via adversarial learning and pseudo-labeling. IEEe Trans. Med. ImAging 40 (10), 2911–2925.
Yasaka, K., Abe, O., 2018. Deep learning and artificial intelligence in radiology: Current applications and future directions. PLoS. Med. 15 (11), e1002707.
Dayan, I., et al., 2021. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat. Med. 27 (10), 1735–1743.

Kora, P., et al., 2022. Transfer learning techniques for medical image analysis: a review. Biocybern. Biomed. Eng. 42 (1), 79–107.
Li, J., et al., 2024. A comprehensive survey on source-free domain adaptation. IEEe Trans. Pattern. Anal. Mach. Intell. 46 (8), 5743–5762.
Zhou, C., et al., 2022. Domain adaptation for medical image classification without source data. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).
Hu, S., Liao, Z., Xia, Y., 2022. arXiv preprint.
Wang, Y., et al., 2023. FVP: fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. IEEe Trans. Med. ImAging 42 (12), 3738–3751.
Hong, J., Zhang, Y.D., Chen, W.T., 2022. Source-free unsupervised domain adaptation for cross-modality abdominal multi-organ segmentation. Knowl. Based. Syst. 250, 109155.
Ye, Y., et al., 2022. Alleviating style sensitivity then adapting: source-free domain adaptation for medical image segmentation. In: Proceedings of the 30th ACM International Conference on Multimedia. 2022, Association for Computing Machinery. Lisboa, Portugal, pp. 1935–1944.
Yu, Q., et al., 2023. Source-free domain adaptation for medical image segmentation via prototype-anchored feature alignment and contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer.
Bateson, M., et al., 2022. Source-free domain adaptation for image segmentation. Med. Image Anal. 82, 102617.
Liu, X., et al., 2023. Memory consistent unsupervised off-the-shelf model adaptation for source-relaxed medical image segmentation. Med. Image Anal. 83, 102641.
Kondo, S., 2022. arXiv preprint.
Stan, S., Rostami, M., 2021. Unsupervised model adaptation for continual semantic segmentation. In: Proceedings of the AAAI conference on artificial intelligence.
Li, Z., et al., 2023. Toward source-free cross tissues histopathological cell segmentation via target-specific finetuning. IEEe Trans. Med. ImAging 42 (9), 2666–2677.
Yang, C., et al., 2022. Source free domain adaptation for medical image segmentation with fourier style mining. Med. Image Anal. 79, 102457.
Cai, B.K., Ma, L.Y., Sun, Y., 2023. Dual consistent pseudo label generation for multi-source domain adaptation without source data for medical image segmentation. Front. Neurosci. 17.
Pianykh, O.S., et al., 2020. Continuous learning AI in radiology: implementation principles and early applications. Radiology. 297 (1), 6–14.
Perkonigg, M., et al., 2021. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nat. Commun. 12 (1), 5678.
Wang, L.Y., et al., 2024. A comprehensive survey of continual learning: theory, method and application. IEEe Trans. Pattern. Anal. Mach. Intell. 46 (8), 5362–5383.
Guan, H., et al., 2021. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. Med. Image Anal. 71, 102076.
Liu, X.Y., et al., 2021. Consolidated domain adaptive detection and localization framework for cross-device colonoscopic images. Med. Image Anal. 71, 102052.
Wang, S., et al., 2020. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. IEEe Trans. Med. ImAging 39 (12), 4237–4248.
Dong, D.Q., et al., 2022. An unsupervised domain adaptation brain CT segmentation method across image modalities and diseases. Expert. Syst. Appl. 207, 118016.
Sanchez, K., et al., 2022. CX-DaGAN: domain adaptation for pneumonia diagnosis on a small chest X-Ray dataset. IEEe Trans. Med. ImAging 41 (11), 3278–3288.
Zhu, J.-Y., et al., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision.
Chen, Y., et al., 2019. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Palladino, J.A., Slezak, D.F., Ferrante, E., 2020. Unsupervised domain adaptation via CycleGAN for white matter hyperintensity segmentation in multicenter MR images. In: 16th International Symposium on Medical Information Processing and Analysis. SPIE.
Dou, Q., et al., 2018. Unsupervised cross-modality domain adaptation of ConvNets for biomedical image segmentations with adversarial loss. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 691–697.
Tran, L., et al., 2019. Gotta Adapt'em all: joint pixel and feature-level domain adaptation for recognition in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
Tzeng, E., et al., 2017. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition.
Yu, M., et al., 2022. Domain-prior-induced structural MRI adaptation for clinical progression prediction of subjective cognitive decline. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer.
Chen, C., et al., 2019. Synergistic image and feature adaptation: towards cross-modality domain adaptation for medical image segmentation. In: Proceedings of the AAAI conference on artificial intelligence.
Kumari, S., Singh, P., 2024. Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. Comput. Biol. Med. 170, 107912.
Fang, Y.Q., et al., 2024. Source-free unsupervised domain adaptation: a survey. Neural Netw. 174, 106230.
Luo, X., et al., 2024. Crots: Cross-domain teacher–student learning for source-free domain adaptive semantic segmentation. Int. J. Comput. Vis. 132 (1), 20–39.
Cao, Y., et al., 2024. Towards source-free domain adaptive semantic segmentation Via importance-aware and prototype-contrast learning. IEEE Trans. Intell. Veh. 1–13.
Ritter, H., Botev, A., Barber, D., 2018. Online structured laplace approximations for overcoming catastrophic forgetting. In: Advances in Neural Information Processing Systems 31 (Nips 2018), 31.
Schwarz, J., et al., 2018. Progress & compress: a scalable framework for continual learning. In: International Conference on Machine Learning. PMLR.
Rebuffi, S.-A., et al., 2017. iCaRL: incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.
Dhar, P., et al., 2019. Learning without memorizing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
Lopez-Paz, D., Ranzato, M., 2017. Gradient episodic memory for continual learning. Advances in neural Information Processing Systems, p. 30.
Shin, H., et al., 2017. Continual learning with deep generative replay. Adv. Neural Inf. Process. Syst. 30 (Nips 2017), 30.
Wu, C.S., et al., 2018. Memory replay gans: Learning to generate new categories without forgetting. Adv. Neural Inf. Process. Syst. 31.
Chaudhry, A., et al., 2018. arXiv preprint.
Farajtabar, M., et al., 2020. Orthogonal gradient descent for continual learning. In: International Conference on Artificial Intelligence and Statistics. PMLR.
Cha, H., Lee, J., Shin, J., 2021. Co2l: contrastive continual learning. In: Proceedings of the IEEE/CVF International conference on computer vision.
Mehta, S.V., et al., 2023. An empirical investigation of the role of pre-training in lifelong learning. J. Mach. Learn. Res. 24 (214), 1–50.
Ramasesh, V.V., Lewkowycz, A., Dyer, E., 2021. Effect of scale on catastrophic forgetting in neural networks. In: International Conference on Learning Representations.
Rusu, A.A., et al., Progressive neural networks. arXiv preprint arXiv:1606.04671, 2016.
Mallya, A., Davis, D., Lazebnik, S., 2018. Piggyback: adapting a single network to multiple tasks by learning to mask weights. In: Proceedings of the European conference on computer vision (ECCV).
Ebrahimi, S., et al., 2020. Adversarial continual learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer.
Li, Z., Hoiem, D., 2018. Learning without forgetting. IEEe Trans. Pattern. Anal. Mach. Intell. 40 (12), 2935–2947.
Vázquez, D., et al., 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthc. Eng. 1–9. **2017**.
Silva, J., et al., 2014. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. 9 (2), 283–293.
Chen, J., et al., 2020. Generative adversarial networks for video-to-video domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence.
English, 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data 7 (1), 283.
Vente, C.d., et al., 2021. Deep learning regression for prostate cancer detection and grading in Bi-parametric MRI. IEEE Trans. Biomed. Eng. 68 (2), 374–383.
Khan, Z., et al., 2021. Recent automatic segmentation algorithms of MRI prostate regions: a review. IEEe Access. 9, 97878–97905.
Liu, Q., Dou, Q., Heng, P.-A., 2020. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. Springer.
Liu, Q.D., et al., 2020. MS-Net: Multi-Site network for improving prostate segmentation with heterogeneous MRI data. IEEe Trans. Med. ImAging 39 (9), 2713–2724.
Alzahrani, S.M., Al Doghaither, H.A., Al-Ghafari, A.B., 2021. General insight into cancer: an overview of colorectal cancer (Review). Mol. Clin. Oncol. 15 (6), 271.
Biffi, C., et al., 2022. A novel AI device for real-time optical characterization of colorectal polyps. NPJ. Digit. Med. 5 (1), 84.
Yildiz, V.M., et al., 2020. Plus disease in retinopathy of prematurity: convolutional neural network performance using a combined neural network and feature extraction approach. Transl. Vis. Sci. Technol. 9 (2), 10-10.
Blencowe, H., et al., 2013. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. Pediatr. Res. 74 (1), 35–49.
Gopal, L., et al., 1995. Retinopathy of prematurity: a study. Indian J. Ophthalmol. 43 (2), 59–61.
Roth, D.B., et al., 2001. Screening for retinopathy of prematurity employing the RetCam 120 - sensitivity and specificity. Arch. Ophthalmol. 119 (2), 268–272.