International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01956-x

Learning Dynamic Prototypes for Visual Pattern Debiasing

Kongming Liang[1] · **Zijin Yin**[1] · **Min Min**[2] · **Yan Liu**[2] · **Zhanyu Ma**[1] · **Jun Guo**[1]

Received: 16 December 2022 / Accepted: 30 October 2023

Abstract
Deep learning has achieved great success in academic benchmarks but fails to work effectively in the real world due to the potential dataset bias. The current learning methods are prone to inheriting or even amplifying the bias present in a training dataset and under-represent specific demographic groups. More recently, some dataset debiasing methods have been developed to address the above challenges based on the awareness of protected or sensitive attribute labels. However, the number of protected or sensitive attributes may be considerably large, making it laborious and costly to acquire sufficient manual annotation. To this end, we propose a prototype-based network to dynamically balance the learning of different subgroups for a given dataset. First, an object pattern embedding mechanism is presented to make the network focus on the foreground region. Then we design a prototype learning method to discover and extract the visual patterns from the training data in an unsupervised way. The number of prototypes is dynamic depending on the pattern structure of the feature space. We evaluate the proposed prototype-based network on three widely used polyp segmentation datasets with abundant qualitative and quantitative experiments. Experimental results show that our proposed method outperforms the CNN-based and transformer-based state-of-the-art methods in terms of both effectiveness and fairness metrics. Moreover, extensive ablation studies are conducted to show the effectiveness of each proposed component and various parameter values. Lastly, we analyze how the number of prototypes grows during the training process and visualize the associated subgroups for each learned prototype. The code and data will be released at https://github.com/zijinY/dynamic-prototype-debiasing.

Keywords Dataset bias · Algorithmic fairness · Medical image analysis · Diversity · Polyp segmentation

1 Introduction

Deep learning-based models have achieved great success in computer vision and have become an indispensable part of modern systems like face analysis, medical imaging, and autonomous driving. However, a practical challenge in all applications is the model could often be biased, since it is trained overly dependent on the training dataset and tends to inherit the imbalance of data (Buolamwini & Gebru, 2018; Yoneyama et al., 2017; Tartaglione et al., 2021). In general, the bias issue is usually defined as one or a collection of extraneous protected or sensitive attributes that distort the relationship between the input and output and hence lead to erroneous conclusions (Pourhoseingholi et al., 2012).

Ranging from face recognition (Buolamwini & Gebru, 2018) to medical imaging analysis (Yoneyama et al., 2017; Seyyed-Kalantari et al., 2021a, b), data bias can be easily influenced by skew distributions with respect to different types of attributes (e.g. race, sex, and age) and distracts the model from learning the actual discriminative cues (Adeli

Communicated by Oliver Zendel.

Kongming Liang and Zijin Yin have contributed equally to this work.

✉ Kongming Liang
liangkongming@bupt.edu.cn

✉ Min Min
minmin823@sina.com

Zijin Yin
yinzijin2017@bupt.edu.cn

Yan Liu
13911798288@163.com

Zhanyu Ma
mazhanyu@bupt.edu.cn

Jun Guo
guojun@bupt.edu.cn

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China

[2] Department of Gastroenterology, Fifth Medical Center of Chinese PLA General Hospital, Beijing 100071, China

Published online: 08 December 2023

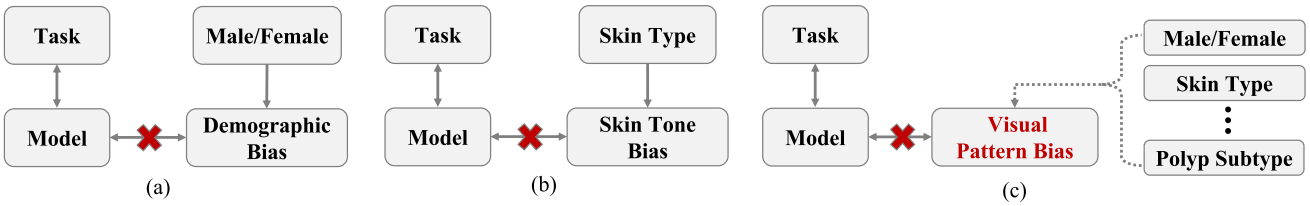

Fig. 1 Categorization of bias problems in segmentation. **a** Demographic Bias (Puyol-Antón et al., 2021; Ioannou et al., 2022) **b** Skin Tone Bias (Xu et al., 2022) **c** Visual pattern bias (Ours). We argue that the visual pattern diversity, which could be induced by tremendous factors (e.g. demographics, skin tones), is an immediate cause of bias. While conventional debiasing methods highly depend on annotations of bias labels, we aim to address pattern bias in an unsupervised way

et al., 2021). Specifically, Buolamwini and Gebru (2018) found a video-based gender classification model performing differently for various racial groups because of the under-representation of black populations in the training set. For applications in healthcare, examples of this are the studies carried out by Seyyed-Kalantari et al. (2021a, b) in which bias is examined in chest X-ray pathology classification. These works demonstrate underdiagnosis disparities between protected groups defined by sex, race, insurance, and age for three publicly available chest X-ray datasets. The Multi-Ethnic Study of Atherosclerosis (MESA) (Yoneyama et al., 2017) found that there are profound racial disparities among people with cardiovascular disease. To some extent, this could reflect how the algorithms are over-optimized by current evaluation metrics but ignore the properties such as fairness and diversity.

To fight against bias in deep learning-based models, a lot of studies have put efforts to evaluate and mitigate dataset bias. Generally, these methods can be divided into three categories: pre-processing, in-processing, and post-processing. Pre-processing techniques (Calmon et al., 2017) solve the issue of data itself. For example, the distributions of specific protected variables are discriminatory and imbalanced. These methods tend to transform the data before model training so that the underlying discrimination is eliminated or mitigated. In-processing techniques (Hong & Yang, 2021; Tartaglione et al., 2021) solve the issue during the training procedure. They tend to modify the learning algorithms, such as balancing multiple optimization objectives of both accuracy and fairness. Post-processing techniques (Chiappa, 2019) often mitigate the bias of the output of algorithms after the training procedure. They tend to perform a transformation to model prediction to mitigate the discrimination towards specific attributes.

While extensive research has shown great potential for handling classification bias, very few works (Puyol-Antón et al., 2021; Ioannou et al., 2022; Xu et al., 2022) concentrated on segmentation bias. In cardiac MR segmentation, Puyol-Antón et al. (2021) found significant racial bias in segmentation accuracy, caused by a racial imbalance in the training data. In brain MR segmentation, Ioannou et al. (2022) found that there are significant sex and race bias effects in model performances and biases have a strong spatial component with some brain regions exhibiting much stronger bias than others. In skin segmentation, Xu et al. (2022) proposed to learn color invariant features by color space augmentation since the training dataset is significantly biased toward lighter skin tones.

Previous segmentation debiasing works mainly focus on the demographic diversity of a dataset (e.g. gender, age, skin tone), as shown in Fig. 1a, b. In contrast to them, we argue that the immediate cause of segmentation bias is visual pattern diversity, as shown in Fig. 1c. Since deep learning models are prone to capture the major visual pattern and dismiss the minority (Dong et al., 2018) during the training process, they may produce biased segmentation results for different visual patterns and further lose robustness and generalization. Thus, the imbalance of visual patterns induces the inequality of model representation capability for different samples. One possible solution is to annotate all the types of visual patterns and balance the learning process by putting more emphasis on minor ones. However, the definition of the visual pattern is multifarious and far more than demographic diversity. For example, previous work (Moayeri et al., 2022) has collected 18 informative visual attributes (e.g. colored-eyes, hairy, patterned) to analyze the robustness and interpretability of deep networks. Du et al. (2022) and Xu et al. (2022) utilize six specialized skin subtypes to measure the bias of neural networks in dermatology. As the number of protected or sensitive attributes increases, it's laborious and costly to acquire manual annotation for all the visual patterns, especially in medical image analysis where only doctors can accomplish this job.

In this work, we attempt to answer the question: can a model automatically balance the learning of different visual patterns without the awareness of corresponding protected attributes? To this end, we propose a novel prototype-based framework that can adaptively discover and extract diverse visual patterns contained in the whole training dataset to enhance the representation ability of both majority and

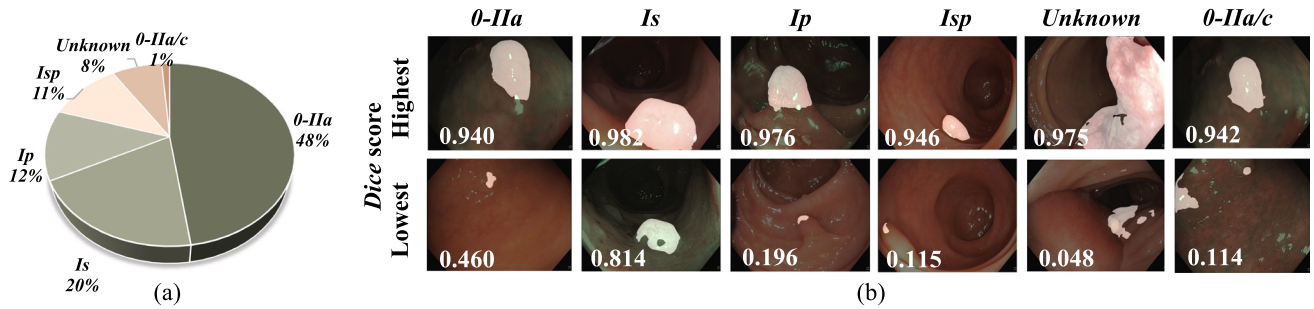

Fig. 2 **a** Distribution of subgroups in the PICCOLO dataset (Sánchez-Peralta et al., 2020). *0-IIa* to *0-IIa/c* refer to the six polyp subtypes described by *Paris Classification* (Vleugels et al., 2017). **b** Examples of highest and lowest segmentation accuracy in each subtype. The deep learning-based model generates more unstable predictions on minority subtypes

minority samples. Specifically, prototypes are constructed to store different visual pattern knowledge in a dynamic way. Given an input sample, we first calculate its similarities with all the prototypes. If the maximum of similarities is above a pre-defined threshold, the input sample will be categorized into the corresponding knowledge and further utilized to update its nearest prototype. Otherwise, the representation of the input sample will be considered an unknown pattern and allocated to construct a new prototype. In this way, the learned prototypes can aggregate the representation of similar visual patterns and maintain a large variety of visual patterns. For model inference, we enhance the representation of the input sample with the learned prototypes via a well-designed attention mechanism to increase the discriminative abilities of all the visual patterns. To evaluate the proposed framework, we conduct extensive experiments on polyp segmentation in colonoscopy images. All the polyp images are divided into different subgroups following the *Paris Classification* scale (Vleugels et al., 2017). The segmentation disparity on subgroups is utilized as the fairness evaluation metric. Figure 2 shows the imbalance distribution of subgroups in the PICCOLO dataset.

The main contributions of this paper can be summarized as follows:

1. We investigate the dataset bias of image segmentation in an unsupervised way. In contrast to the previous works, the labels of the protected or sensitive attributes are not needed during model training but are only necessary for model evaluation.
2. A novel prototype-based framework is proposed to discover and extract visual patterns correlated with the protected or sensitive attributes. The learned prototypes can dynamically balance the learning of both majority and minority subgroups.
3. With comprehensive experiments in terms of both effectiveness and fairness, we demonstrate the superiority of our model over other debiasing methods. In addition, extensive ablation studies are conducted to show the effectiveness of each proposed component and various parameter values.
4. We first propose to use a fairness metric for polyp segmentation evaluation. The fairness metric is able to measure the model performance on each polyp subtype and give more explanation on the algorithmic underdiagnosis. The experimental results indicate that the proposed framework can mitigate the dataset bias more effectively than the state-of-the-art models.

2 Related Work

2.1 Fairness and Debiasing

In the context of decision-making, fairness is defined as the absence of prejudice or favoritism towards an individual or group based on their inherent or acquired characteristics (Mehrabi et al., 2021; Saxena et al., 2019). In Verma and Rubin (2018), authors studied the taxonomy of fairness in algorithmic binary classification problems. There are many fairness definitions, and they are incompatible with each other (Barocas et al., 2017; Chouldechova, 2017). The three most common fairness are (1) equalized odds, (2) equal opportunity, and (3) statistical parity. Equalized Odds, provided by Hardt et al. (2016), states that prediction $\hat{Y}$ satisfies: $P(\hat{Y} = 1 \mid S = 1, Y = y) = P(\hat{Y} = 1 \mid S = 0, Y = y), y \in \{0, 1\}$, where S is protected attribute and Y is label. This implies that different protected groups should have an equal probability of true positives and false positives. Equal opportunity is formed in a relaxed notion: $P(\hat{Y} = 1 \mid S = 1, Y = 1) = P(\hat{Y} = 1 \mid S = 0, Y = 1)$, which only requires true positives are equal towards different protected groups. Statistical parity, also known as demographic parity, is defined to assure independence between predicted labels and protected attributes, formally: $P(\hat{Y} = 1 \mid S = 1) = P(\hat{Y} = 1 \mid S = 0)$.

There have been a wide collection of approaches developed to debias machine learning models and achieve group fairness. Generally, they can be divided into three categories as follows.

Pre-processing techniques solve the issue of data itself, for example, the distribution of specific sensitive or protected variables is biased, discriminatory, and imbalanced. They tend to transform the data before training so that the underlying discrimination is eliminated or mitigated. For example, Feng et al. (2019) employ adversarial learning to capture the data distribution and generate fair latent representations to ensure that the distributions across different protected groups are equivalent. Calmon et al. (2017) propose to learn a data transformation with three optimization goals: controlling discrimination, limiting distortion in individuals, and preserving utility.

In-processing techniques solve the issue during the training procedure. They tend to modify the learning algorithms, such as incorporating a balance between multiple optimization objectives of both accuracy and fairness. For example, Hong and Yang (2021) designed a distance-weighted contrastive loss to pull a pair with the same target class but with different bias features. Tartaglione et al. (2021) propose a regularization term, whose aim is to regularize the deep features to prevent deep models from learning unwanted biases. The above two works both attempted to disentangle the correlation between biases and targets.

Post-processing techniques mitigate the output bias after the training procedure. They tend to perform a transformation to model prediction to mitigate the discrimination towards specific sensitive attributes. They can be attached to the end of any model and only need access to the predictions and sensitive attributes which makes them flexible and applicable to black-box applications. For example, Chiappa (2019) proposes to correct observations adversely affected by the sensitive attribute to form a new prediction.

However, all the prior works (Tartaglione et al., 2021; Chiappa, 2019; Thomas & Kovashka, 2021; Georgopoulos et al., 2021) focused on the group fairness definition which is specified as conditional independence statements in the binary classification setting. Contrary to them, our work seeks to study fairness in segmentation tasks.

2.2 Fairness in Medical Imaging

As deep learning models become increasingly integrated into medical imaging (Nie & Shen, 2020; Sitenko et al., 2021; Zhang & Ma, 2021), one primary concern is whether such algorithms are being employed in an ethical and fair way (Ahmad et al., 2020; Gichoya et al., 2021; Chen et al., 2021). Most of the previous works focus on the fairness issues in medical imaging classification (Zhang et al., 2022a; Seyyed-Kalantari et al., 2021a, b; Petersen et al., 2022). Seyyed-Kalantari et al. (2021b) demonstrate the disparities of algorithmic underdiagnosis between protected groups defined by sex, race, insurance, and age for three publicly available chest X-ray pathology classification datasets. Seyyed-Kalantari et al. (2021a) state that classifiers are found to consistently and selectively amplify the existing biases towards patients under-represented in the training set. These effects are worse on intersectional subpopulations, e.g. Black females, and persist across three large and multi-source chest X-ray datasets. Zhang et al. (2022a) benchmarked the performance of several debiasing models on the task of chest X-ray image classification, focusing on group fairness and minimax fairness. Petersen et al. (2022) assessed the robustness of the trained models in the face of varying dataset splits, sex composition, and stage of disease in MRI-based Alzheimer's disease classification. They found performances of deep networks for male and female test subjects are strongly dependent on the sex proportion in the training set, while the conventional linear regression method is robust to this variation. Du et al. (2022) studied the bias issue on sensitive attributes unrelated to demographic factors. They mitigate the performance disparity on different skin types using contrastive learning in dermatology classification.

Recently, some researchers studied group fairness in the field of medical imaging segmentation. These works (Puyol-Antón et al., 2021; Lee et al., 2022; Ioannou et al., 2022; Yuan et al., 2022) studied the bias towards different racial and sex populations. Puyol-Antón et al. (2021) performed extensive experiments to assess the segmentation performances of racial and gender groups with the cardiac MR image dataset. They are the first to show racial bias exists in deep learning-based segmentation models. Similarly, Lee et al. (2022) studied the effects of data imbalance on racial and sex bias in cardiac MR segmentation. Ioannou et al. (2022) trained multiple trials using different levels of sex imbalance in white subjects in brain MR segmentation task. They found that there are significant sex and race biases in the segmentation model and these biases have a strong spatial component with some brain regions exhibiting much stronger bias than others. Moreover, Puyol-Antón et al. (2021) proposed three debiasing baselines, which were inspired by works from the literature on fairness in classification.

However, previous works were devoted to measuring group fairness in terms of demographic factors (e.g. sex and race) but neglected other properties (e.g. pathology and morphology (Vleugels et al., 2017)) related to the object of interest. In our work, we focus on group fairness in terms of the morphological attributes of lesions.

2.3 Polyp Segmentation

Early solutions for automated polyp segmentation were mainly based on low-level features, for example, texture

(Mamonov et al., 2014), geometric features (Mamonov et al., 2014), and superpixels (Maghsoudi, 2017; Garcia-Peña et al., 2022). But they are far from satisfactory due to the poor representation ability of these conventional hand-crafted features.

In recent years, the development of polyp segmentation has been greatly promoted by deep learning techniques (Simonyan & Zisserman, 2014; Long et al., 2015; He et al., 2016; Li et al., 2019). Akbari et al. employed a fully convolutional neural network to solve the polyp segmentation, and their results are significantly better than traditional works. The encoder-decoder architectures, such as U-Net (Ronneberger et al., 2015), U-Net++ (Zhou et al., 2018) and ResUNet++ (Jha et al., 2019) showed their excellent performance in this field. Several works (Murugesan et al., 2019; Fang et al., 2019) attempted to employ auxiliary tasks or constraints to facilitate representation learning. Murugesan et al. (2019) proposed a multi-task learning framework that leverages distance estimation and boundary prediction tasks to assist polyp segmentation mask prediction. Similarly, Fang et al. (2019) used the area and boundary as constraints to guide learning better feature representations. ACSNet (Zhang et al., 2020) combines the global context and local details in the decoder to deal with the shape and size variance of polyps. PraNet(Fan et al., 2020) aggregates the multi-scale features and extract silhouette according to the local features. SCRNet(Wu et al., 2021) design the semantic calibration and refinement modules to bridge the semantic gap between different feature maps. Cheng et al. (2021) concentrated on the contour accuracy of predictions because of the blurred boundary between polyps and surroundings, and they refine the boundary by calculating eight oriented derivatives at each pixel. In Zhao et al. (2021), authors proposed a multi-scale subtraction network to eliminate redundancy and complementary information between the multi-scale features in conventional encoder-decoder architecture (Ronneberger et al., 2015). Zhang et al. (2022b) designed a dynamic kernel mechanism to use global context features to generate the segmentation head and iteratively update it by the extracted lesion features.

The recent vision transformer techniques (Dosovitskiy et al., 2020; Liu et al., 2021) significantly boosted the development of polyp segmentation tasks. Wang et al. (2022) used a pyramid Transformer encoder to improve the generalization ability. Dong et al. (2021) took into account the differences in contribution between different-level features, and designed an effective mechanism to fuse them in transformer architecture. TransFuse (Zhang et al., 2021) aggregated convolutional network with a transformer to obtain more discriminative feature representations.

However, the above previous works are mostly concerned with higher performance, such as the Dice similarity score. There is a growing interest in going beyond mere performance by measuring and addressing the robustness, fairness, interpretability, and generalization aspects of deep learning-based methods. To the best of our knowledge, our work is the first to explore more valuable metrics in colonoscopic polyp segmentation.

2.4 Prototype Learning

Prototype, also known as proxy (Movshovitz-Attias et al., 2017), or center (Wen et al., 2016), is the one representative of a class among training examples (Kim et al., 2021). Contrary the softmax weights in decision making, prototypes (Yang et al., 2018; Wang et al., 2019; Zhou et al., 2022), aim to learn a latent feature space where the prediction is made by calculating the distance between the test anchor and prototypes of each class. Prototype learning has been proved more robust on data scarcity paradigms, such as few-shot learning (Xu et al., 2022; Li et al., 2021), open-set recognition (Shu et al., 2020), incremental learning (Zhu et al., 2021) and object category discovery (Rambhatla et al., 2021).

Arik and Pfister (2019) provided an interpretable model that bases decisions on relevant prototypes. Zhou et al. (2022) proposed a non-parametric alternative based on non-learnable prototypes in semantic segmentation. The model represents each class as a collection of prototypes, relying on the mean features of several training pixels within that class. Different from building instance-based prototypes, in Chen et al. (2019), authors dissected an image into parts and designed prototypes for parts of each object category, then classified by combining evidence from part prototypes. Kim et al. (2021) present an attention mechanism in person re-identification by exploiting the prototype as guidance. Kwon et al. (2021) proposed a prototype-based framework using contrastive learning to learn discriminative representation such that features within the same class are close to each other while features from different classes are far away. In the domain adaptation field, Yue et al. (2021) leveraged prototypes to perform cross-domain instance-to-prototype matching to transfer knowledge from source to target domain. Moreover, Rambhatla et al. (2021) proposed a unified framework to iteratively memorize the past samples by prototypes and use prototypes to discover novel object discovery.

In this work, we demonstrate the debiasing ability of prototype learning. The proposed framework can adaptively update the known prototypes and discover the unknown or abnormal prototypes without corresponding sensitive attribute labels. Our update paradigm is similar to Kim et al. (2021) since both methods consider the hard negatives of prototypes to increase the discriminative ability of prototypes. The key distinction is that their method works in a fully-supervised manner with initialized fixed number of prototypes, while we design to update prototypes in an unsu-

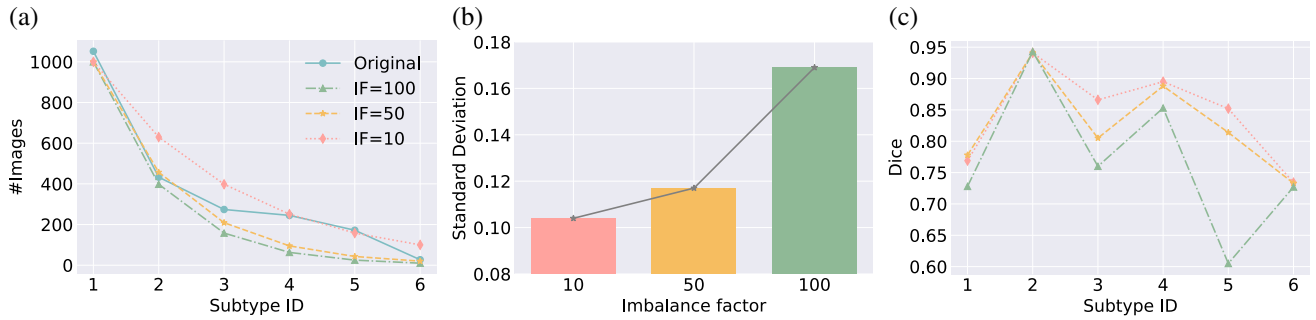

Fig. 3 **a** We resample images of each subtype by varying imbalance factors (IF) from 10 to 100. **b** The standard deviation of segmentation accuracy on six types varies along with imbalance factors. (c) The detailed segmentation accuracy on all subtypes with different imbalance factors

pervised manner and the number of prototypes is dynamic depending on encounter samples.

3 Pattern Bias

In this section, we first introduce the statistical bias previously studied in MR cardiac and brain segmentation tasks, then we delineate the pattern bias issues in polyp segmentation. Finally, we conduct an empirical study to further investigate the cause of the bias problem in polyp segmentation task.

3.1 Preliminary

Deep learning techniques have increasingly achieved remarkable performance in medical image segmentation applications. However, preliminary works (Ioannou et al., 2022; Lee et al., 2022; Puyol-Antón et al., 2021) demonstrate that, for brain and cardiac MR segmentation tasks, deep learning models have been shown to exhibit discrepant results towards demographic groups when trained in imbalanced datasets.

The structure and anatomy of the brain is widely known to vary between different demographic groups such as gender (Cosgrove et al., 2007) and race (Isamah et al., 2010). This situation is the same with cardiac structure (Kishi et al., 2015). Based on the above factors, prior findings (Ioannou et al., 2022; Puyol-Antón et al., 2021) suggest that such visual variations of structures of organs and lesions combined with imbalance distribution lead to statistical bias in segmentation performance. Moreover, Du et al. (2022) has demonstrated that other protected attributes besides demographic factors (e.g. visual skin type) can induce the biased prediction.

3.2 Polyp Pattern Bias

Based on the conclusions mentioned above, we argue that statistical bias exists not only among demographics (e.g. gender and race) but also among the properties that can directly affect the visual pattern of the objects. Therefore, we speculate that algorithms can exhibit disparity of performances towards specific visual patterns as they are usually distributed unevenly in a dataset.

However, the definitions of visual patterns are multifarious and changeable in various imaging fields and tasks (Moayeri et al., 2022). In polyp segmentation, *Paris Classification* (Axon et al., 2005; Vleugels et al., 2017) is a gold standard for the endoscopic classification of gastrointestinal superficial neoplastic lesions. It divides polyps into different subtypes by describing the morphology of superficial neoplastic polyps in the esophagus, stomach, and colon. Here, we utilize *Paris Classification* as the division criterion to conduct all subsequent experimental studies.

3.3 Empirical Study

We conduct an empirical study to further prove the above speculation about the cause of bias on the recently published dataset PICCOLO (Sánchez-Peralta et al., 2020) (More details in Sect. 5). Based on the *Paris Classification* criterion, we partition its training data into six subtypes. The data distribution is severely unbalanced as shown in Fig. 2.

Following dataset settings in long-tail (Alshammari et al., 2022) and imbalance learning (Cao et al., 2019), we reconstruct an imbalanced training set and a balanced validation set, by varying the Imbalance Factor (IF, the ratio of the number of samples in the largest subtype and that of the smallest). The number of the largest subtype is preserved and that of the other subtypes is re-sampled exponentially according to the value of IF. The distributions of all subtypes are shown in Fig. 3. Then we evaluate the effect of the imbalance factor on the performance of the vanilla U-Net model (Ronneberger et al., 2015) on all subtypes.

The experimental results of standard deviations and detailed segmentation accuracy towards different imbalance factors are presented in Fig. 3. From the results, we can obtain several vital observations. First, as the imbalance factor increases, the standard deviation of subtypes grows which

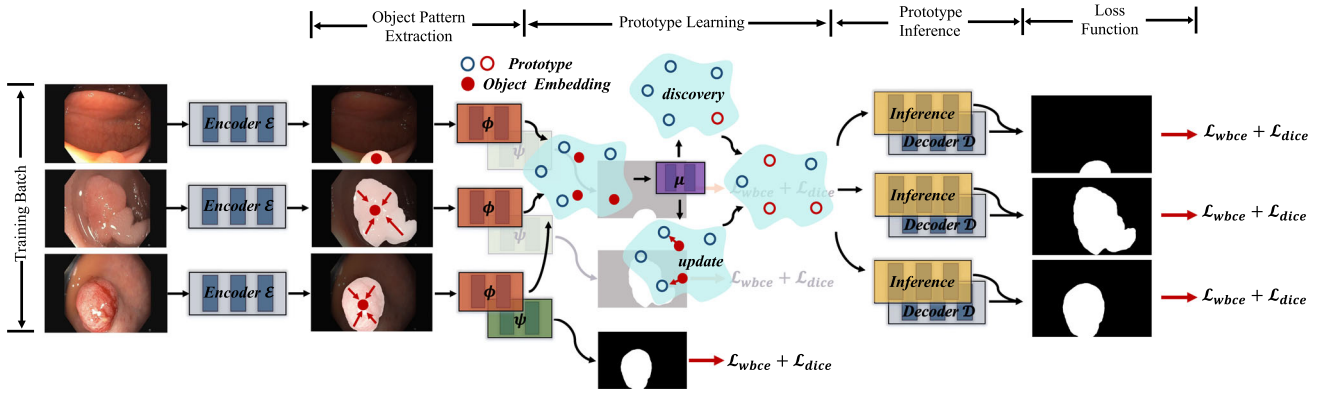

Fig. 4 The overall illustration of our proposed prototype-based framework. Images sampled from set $\mathcal{X}$ are passed through the encoder $\mathcal{E}$. Then the object pattern embeddings are computed by function ϕ. In prototype learning, we distinguish whether the semantics of pattern embeddings differ from existing prototypes based on computed similarity. **Discovery:** If the object is unknown (dissimilar), we utilize its content to construct a new prototype. **Update:** If the object is known (similar), we update its most relevant prototype. In prototype inference, we enhance the representation ability of features by incorporating similar prototypes. Finally, the predictions from decoder $\mathcal{D}$ and ψ are supervised by ground truths

indicates the increasing severity of segmentation disparity. Second, the segmentation accuracy on some minority subtypes (ID = 3 and 5) drastically decreases when reducing their samples while the performance of the major subtypes remains relatively stable. Third, there is no strictly positive correlation between the segmentation performance on subtypes and their sample size. For instance, the performance on subtype ID = 2 is much greater than that of subtype ID = 1, even though the number of samples of subtype ID=2 is much less. The result of a balanced setting, which achieves considerable bias towards subtype ID = 2, also supports this conclusion. This phenomenon is probably induced by the complexity of the visual pattern of the subtype, e.g. its intra-subtype variation is much smaller than that of the others.

In summary, we can see the performance is significantly biased under the imbalanced distribution of different subtypes. Constructing a completely balanced dataset seems to be a promising way to pursue the model's fairness. However, it is laborious as some minority groups are naturally scarce in normal scenarios, especially in clinical applications. In addition, to mitigate the statistical disparity issue in imbalanced datasets, collecting corresponding protected attributes of objects is time-consuming for experts. Therefore, it's important to build automated models that can make fair predictions without relying on additional annotations.

4 Method

4.1 Overview

Since the visual patterns are different across each subgroup, current algorithms tend to perform well in subjects with more common visual patterns and poorly in those with rare patterns. Therefore, we aim to utilize prototypes to discover and extract different visual patterns of objects from both majority and minority subgroups.

The proposed prototype-based framework gradually learns prototypes by aggregating similar object patterns and separating different object patterns. The stored prototypes are further used to enhance the feature representative ability. In particular, the network takes the input images and extracts visual patterns from object feature representations. And we adaptively model the similarities between different visual patterns and perform two operations simultaneously: (1) discover unknown patterns and (2) update known patterns. Finally, we employ prototype inference by taking the current state of the prototypes and features as input to enhance their semantic representation and discrimination via a well-designed attention mechanism.

The remainder of this section is structured as follows. In Sect. 4.2, we describe the base segmentation architecture. In Sect. 4.3, we present how to extract the visual patterns of objects from entire images. The prototype learning process is presented in Sect. 4.4, and the prototype inference is presented in Sect. 4.5. Finally, we describe the loss function in Sect. 4.6. The overview of the proposed framework is illustrated in Fig. 4.

4.2 Base Architecture

For a set of input images $\mathcal{X}$ and the corresponding segmentation masks $\mathcal{Y}$, our objective is to assign labels to the pixels belonging to foreground regions. We adopt conventional U-Net (Ronneberger et al., 2015) as the network backbone, which consists of an encoder $\mathcal{E} = \{e_1(\cdot), e_2(\cdot), \ldots, e_5(\cdot)\}$

and an decoder $\mathcal{D} = \{d_1(\cdot), d_2(\cdot), \ldots, d_5(\cdot)\}$. Given an input image $I \in \mathbb{R}^{3\times H\times W}$ sampled from $\mathcal{X}$, the extracted feature representation from the k^{th} encoder stage can be denoted as $\mathbf{f}_k^{'} = e_k(\mathbf{f}_{k-1}^{'}) \in \mathbb{R}^{C_k\times H_k\times W_k}$, where C_k is the number of the channels, $H_k = H/2^k$ and $W_k = W/2^k$ are the height and width of the feature respectively. We discard the notation k for clarity in the following sections. We cascade the proposed prototype-based framework after the last encoding block to discover the global semantics of objects since the last encoding block contains more high-level semantic information and fewer spatial details.

4.3 Object Pattern Extraction

The deep feature representation depicts the difference between the object and background at pixel-level granularity (Fu et al., 2019; Huang et al., 2019). Nevertheless, pixel-level features with spatial details can not model the visual pattern of the entire object. We argue that the feature representation could characterize the discrepancy between different subgroups of objects at condensed object-level granularity. Thus, we leverage to learn the object pattern embedding with the supervision of ground truth to facilitate the subsequent procedures.

Let the Object Pattern Extraction is denoted as $\phi(\cdot)$. For a given encoded feature $\mathbf{f}^{'} \in \mathbb{R}^{C\times H\times W}$, we first design a transformation function $\psi(\cdot)$, which is implemented by conv(3x3) $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ dropout $\rightarrow$ conv(1x1) $\rightarrow$ Sigmoid, to calculate a coarse segmentation map $\mathbf{s} = \psi\left(\mathbf{f}^{'}\right) \in \mathbb{R}^{H\times W}$ which indicates the probability of each pixel belonging to the object region. Then the object pattern embedding $\mathbf{f}$ is computed as below:

$$\mathbf{f} = \phi(\mathbf{f}^{'}) = GAP\left(\mathbf{s}\odot\mathbf{f}^{'}\right) \in \mathbb{R}^{C} \tag{1}$$

where $\odot$ is the element-wise multiply and GAP denotes the Global Average Pooling. With such an operation, we extract the global pattern information of the object and improve the semantic consistency inside the object region by suppressing possible background noise. Then we present details of each component of our proposed prototype-based framework.

4.4 Prototype Learning

We propose a novel prototype learning method to adaptively aggregate similar visual patterns and separate different visual patterns in an unsupervised way. Similar visual patterns will be allocated to the same prototype. And the rare patterns of minority subgroups will be stored in a unique prototype. Therefore, we can improve the representational capability of both the majority and minority, and hence mitigate the bias issue.

The prototypes are denoted as a collection of object pattern embeddings $\mathcal{M}_t = \{\mathbf{m}_t^i \in \mathbb{R}^C \mid i = 1, \ldots, N\}$ at time $t \in \{1, \ldots, T\}$. Our method performs two operations simultaneously: 1) *Unknown Pattern Discovery*: discovering the unknown (which can also be viewed as "dissimilar") visual patterns of objects based on existing prototypes, and 2) *Known Pattern Update*: updating the known patterns stored in existing prototypes using similar objects.

4.4.1 Unknown Pattern Discovery

Concretely, given the obtained object pattern embedding $\mathbf{f}$, we first establish the association with current prototypes by computing their correlation. The soft weight w_i is calculated using the cosine similarity as follows:

$$w_i = \frac{\mathbf{f}^\top \mathbf{m}_t^i}{\|\mathbf{f}\|_2 \left\|\mathbf{m}_t^i\right\|_2}, \quad i \in 1, ..., N \tag{2}$$

where $\mathbf{m}_t^i$ is the i-th prototype of $\mathcal{M}_t$ at the time step t. The weight $w_i, i \in 1, ...N$ denotes the correlation between the object pattern embedding and existing prototypes.

Threshold μ is then utilized to decide whether the object has a similar or dissimilar visual pattern with respect to prototypes. If $w_i < \mu, \forall i \in \{1, \ldots, N\}$, where $\mu \in [0, 1]$ is conventionally set to 0.5, the object $\mathbf{f}$ is considered outlier from the existing prototypes, which implies that the visual pattern is unknown for current knowledge stored in existing prototypes. This condition prompts us to initialize a new prototype using the content of the object pattern embedding. Thus, the new state of prototypes at the time step $t + 1$ is denoted as:

$$\mathcal{M}_{t+1} = \{\mathbf{m}_{t+1}^i \in \mathbb{R}^C \mid i = 1, \ldots, N + 1\} \tag{3}$$

where $\mathbf{m}_{t+1}^{N+1} = \mathbf{f}$ and $\mathbf{m}_{t+1}^i = \mathbf{m}_t^i$ for $i \in \{1, \ldots, N\}$.

4.4.2 Known Pattern Update

On one hand, we discover a new visual concept of objects from the coming training data; on the other hand, we also successively update and enrich the known visual knowledge stored in the prototypes by incorporating similar object pattern embeddings.

Concretely, if $w_i > \mu, \exists i \in \{1, \ldots, N\}$, the feature embedding $\mathbf{f}$ matches to existing prototypes, which implies that the visual appearance of the object is known for current knowledge. In this situation, we can update the matched prototype by aggregating its previous state with the current object pattern embedding. Then we detail the conventional naive prototype updating mechanism and our proposed Adaptive Momentum Update.

Preliminary. We first need to identify which prototype is the most relevant to the current object pattern embedding $\mathbf{f}$:

$$p = \underset{i \in \{1 \ldots N\}}{\operatorname{argmax}} w_i \tag{4}$$

The straightforward way to update the content of prototypes is using the exponential moving average as follows:

$$\mathbf{m}_{t+1}^{p} \leftarrow \eta \mathbf{m}_{t}^{p} + (1 - \eta)\mathbf{f} \tag{5}$$

where $\eta \in [0, 1]$ is an update momentum characterizing the amplitude of the adjusting distance of the prototype in latent space. It is normally set to a relatively large value, e.g. 0.9, the weight of the prototype $\mathbf{m}_t^p$ in Eq. (5) is extremely larger than that of object pattern embedding $\mathbf{f}$, which indicates that the adjusting distance of $\mathbf{m}_t^p$ is much less than $\mathbf{f}$. Thus the distribution of the prototypes in embedding space varies successively by updating the prototypes online. This operation allows contents to be retained in the corresponding prototype while progressively erasing older or irrelevant information, and hence can stabilize the remembrance and updating of long-term knowledge. The separability between prototypes is correlated to the discrimination of different object pattern embeddings.

However, in hard negative scenarios, we argue that the conventional updating mechanism obstructs the learning of discrimination between prototypes. As shown in Fig. 5a, the embedding $\mathbf{f}$ is used to update its nearest prototype $\mathbf{m}_t^p$, the small value of adjusting the distance of the prototype in the update procedure hampers the differentiation between prototypes and their close neighbors. Therefore, the traditional method encounters a dilemma: a large momentum value can stabilize the prototype updating but probably decrease separability between prototypes, while a small momentum value can increase separability between prototypes but instabilize the updating procedure. This situation necessitates an advanced update mechanism to explicitly increase separability among prototypes and ensure learning stability.

Adaptive Momentum Update. Therefore, we propose a novel Adaptive Momentum Update mechanism by taking into account the structural information of prototypes and input data. Concretely, for the prototype $\mathbf{m}_t^p$ involving the update process, we first identify its hardest negative by cosine similarity:

$$q = \underset{i \in \{1 \ldots N\} \setminus p}{\operatorname{argmax}} \frac{\mathbf{m}_t^{p\top} \mathbf{m}_t^i}{\left\|\mathbf{m}_t^p\right\|_2 \left\|\mathbf{m}_t^i\right\|_2} \tag{6}$$

When the prototype $\mathbf{m}_t^p$ is more similar to its hard negative $\mathbf{m}_t^q$ than to the object pattern embedding $\mathbf{f}$, the weight of $\mathbf{f}$ on update process should be greater compared to that of $\mathbf{m}_t^p$. So we define our Adaptive Momentum Update process as follows:

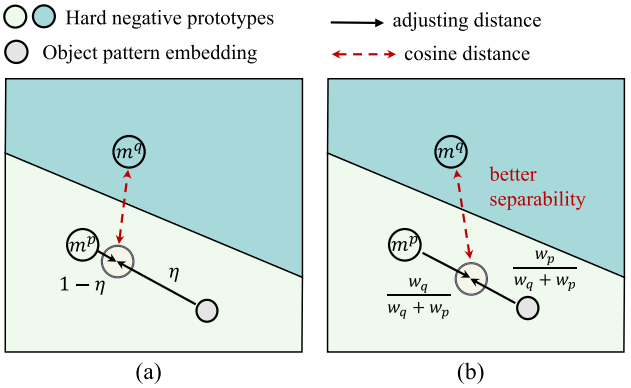

Fig. 5 Illustration of different prototype updating mechanisms in a hard negative scenario. **a** Naive mechanism, **b** Our adaptive mechanism. Our proposed adaptive mechanism can explicitly increase the separability between the prototype and its close neighbor

$$\mathbf{m}_{t+1}^{p} \leftarrow \frac{w_p}{w_q + w_p}\mathbf{m}_{t}^{p} + \frac{w_q}{w_q + w_p}\mathbf{f} \tag{7}$$

where w_p and w_q represent the similarity, computed from Eq. (2), between the object pattern embedding $\mathbf{f}$ with the prototype $\mathbf{m}_t^p$ and $\mathbf{m}_t^q$, respectively.

As shown in Fig. 5b, in a hard negative scenario, as $w_q \approx w_p$, the adaptive momentum can considerably pull away the prototype from its close neighbor, and hence increase separability between prototypes. The inter-difference of the learned prototypes can facilitate distinguishing discriminative visual patterns of both majority and minority groups of objects. Eventually, the stored knowledge especially rarely seen in training improves prediction performance on minority, and hence improves the fairness of the algorithm. Furthermore, as $w_q \ll w_p$, which means prototypes and their neighbors already have enough discrimination, the small adjusting distance maintains the stability of the prototype learning procedure.

4.5 Prototype Inference

It is critical to retrieve appropriate and relevant knowledge from the prototypes and assimilate it to enhance the representation of target features, especially for minority cases. The conventional attention mechanism (Fu et al., 2019; Wang et al., 2020; Yuan et al., 2020) achieves the adaptive spatial highlighting of the features at the pixel-wise granularity. But prototypes are supposed to own conceptual knowledge of visual patterns of integral objects without spatial details. Therefore, we propose incorporating global prototypes into the spatial details of feature representations.

4.5.1 Context Encoding

Concretely, we first pre-process the prototypes using the spatial content of the target feature $\mathbf{f}^{'}$:

$$\mathbf{a}_j = \frac{\sigma_1(\mathbf{W}_u(\mathcal{M})) \times \sigma_2(\mathbf{W}_v(\mathbf{f}^{'}))}{\sum_{k=0}^{HW} \sigma_1(\mathbf{W}_u(\mathcal{M})) \times \sigma_2(\mathbf{W}_u(\mathbf{f}^{'}))} \tag{8}$$

where $\mathbf{A} = \{\mathbf{a}_1, ..., \mathbf{a}_{HW}\} \in \mathbb{R}^{N \times HW}$, and $\times$ is matrix multiplication, $\mathbf{W}_u$ and $\mathbf{W}_v$ are 1x1 two convolution layers respectively, $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ denote tensor reshape operators. $\mathbf{A}$ depicts the encoded association of prototypes on the feature of each pixel. Then, we need to calculate the similarity coefficient between $\mathbf{f}^{'}$ and all the prototypes:

$$\mathbf{E} = \sigma_1(\mathbf{A}) \times \sigma_2(\mathbf{f}^{'}) \in \mathbb{R}^{C \times N} \tag{9}$$

Therefore, $\mathbf{E}$ represents the context information between knowledge contained by learned prototypes and the target feature.

4.5.2 Feature Enhancement

Then we need to select the useful context semantics to enhance the representation ability of $\mathbf{f}^{'}$. We leverage the global max-pooling-layer $\epsilon(\cdot)$ to identify the most context from $\mathbf{E}$:

$$\mathbf{e} = \text{Sigmoid}(\mathbf{W}_z(\epsilon(\mathbf{E}))) \tag{10}$$

Hence $\mathbf{e} \in \mathbb{R}^C$ reflects the semantic correlation of the target feature and the prototype most relevant to it. We enhance the feature representation by composition:

$$\mathbf{f}^{''} = \mathbf{f}^{'} \odot \mathbf{e} \tag{11}$$

where $\odot$ indicates the element-wise multiplication. This operation allows the most similar knowledge from the external prototypes to be incorporated into the features. For the minority subtypes with rare visual patterns, the recalled similar conceptual knowledge increase the representative ability of features and hence mitigates the bias issue.

4.6 Loss Functions

We utilize a deep supervision strategy for three intermediate maps of the decoder branch to jointly optimize the model parameters. In addition, the coarse segmentation map $\mathbf{s}$ is guided by the supervision which is acquired by downsampling the ground-truth segmentation mask.

Similar to the previous study (Fan et al., 2020), we employ the combination of a Weighted Binary Cross Entropy loss $\mathcal{L}_{wbce}$ and a *Dice* loss $\mathcal{L}_{dice}$ as the total loss function:

$$\mathcal{L} = \mathcal{L}_{wbce} + \mathcal{L}_{dice}. \tag{12}$$

In the weighted binary cross-entropy loss, each pixel (i, j) will be assigned a weight according to the difference between the center pixel and its surroundings:

$$\alpha_{ij} = \| \frac{\sum_{m,n \in U_{ij}} y_{mn}}{\sum_{m,n \in y_{ij}} 1} - y_{ij} \| \tag{13}$$

where U_{ij} is the area surrounding the target pixel and y_{ij} is the ground-truth label. Thus, hard pixels such as boundaries correspond to a larger weight and hence get more attention during training. In contrast, simple pixels like the inner area will be assigned a smaller weight. So the weighted binary cross entropy loss is as shown in:

$$\mathcal{L}_{wbce} = \frac{\sum_{i,j} \left(1 + \gamma \alpha_{ij}\right) \mathcal{L}_{bce}(i, j)}{\sum_{i,j} \gamma \alpha_{ij}} \tag{14}$$

where γ is a hyper-parameter, h_{ij} is the prediction of the pixel at location (i, j), and $\mathcal{L}_{bce}(i, j)$ is binary cross entropy function in pixel (i, j):

$$\mathcal{L}_{bce}(i, j) = -y_{ij} \log(h_{ij}) - (1 - y_{ij}) \log(1 - h_{ij})) \tag{15}$$

The *Dice* loss is calculated as follows:

$$\mathcal{L}_{dice} = 1 - \frac{2 \sum_{i,j} h_{ij} y_{ij}}{\sum_{i,j} h_{ij} + \sum_{i,j} y_{ij}} \tag{16}$$

$\mathcal{L}_{dice}$ can better compare the structural similarities between the prediction and ground truth. Therefore, our strategy could consider both pixel-level and region-level measurements.

5 Experiments

In this section, we perform extensive experiments to evaluate our methods from two perspectives of segmentation accuracy and fairness. We conduct ablation studies to verify the effectiveness of each component and explore the effect of various parameter values. In addition, we discuss further revealing the intrinsic properties of prototypes and verify our design intuition.

5.1 Datasets

To extensively verify the effectiveness of our method in segmentation fairness, we conduct experiments on various medical imaging domain: three popular benchmark datasets in colonoscopy for polyp segmentation, and one widely-used

benchmark dataset in dermatoscopy for skin lesion segmentation.

Kvasir-SEG (Jha et al., 2020) consists of images and annotations verified by experienced endoscopists, including several classes showing anatomical landmarks, pathological findings, or endoscopic procedures in the GI tract. The anatomical landmarks include Z-line, pylorus, cecum, etc., while the pathological finding includes polyps, colitis, etc. The data is collected using endoscopic equipment at Vestre Viken Health Trust in Norway. We choose the pixel-wise annotations of polyps, and we randomly split the 1000 samples into 600 training images, 200 validation images, and 200 test images.

EndoScene (Vázquez et al., 2017) is a widely-used benchmark for binary polyp segmentation in colonoscopy. It consists of 912 white-light images and the corresponding pixel-wise segmentation annotations. We follow the standard protocol in Vázquez et al. (2017) with the constraint that the lesions of the same patient should not appear in different sets, and the default split setting that 547 training images and 182 testing images, and 183 validating images. In order to investigate the bias issue of models' segmentation performance, we adopt the Paris Classification (Vleugels et al., 2017) which is perceived as the gold standard in colonoscopy, to categorize lesions into different subtypes according to polyp morphology for superficial neoplastic lesions in the esophagus, stomach, and colon. The corresponding categorical annotations on all images are extended by us and will be released to facilitate the following study.

PICCOLO (Sánchez-Peralta et al., 2020) is a recently published dataset collected in colonoscopy for polyp segmentation. It comprises 3433 manually annotated images (2131 white-light images and 1302 narrow-band images), originating from 76 lesions from 40 patients, which are distributed into training (2203 images), validation (897 images) and test (333 images) sets assuring patient independence between sets. In addition, clinical metadata involving the morphological and pathological attributes including Paris Classification is also provided for each lesion by default. It contains the largest number of samples, and the polyps appearing on it are much more complex and diverse than those on the other two datasets. Since the imbalance distribution of all subtypes inherently appears in itself, we decide to evaluate the fairness of all approaches in original datasets without re-manufacturing. Besides, the subtype named *0-IIb* exists in the test set but not in the training set, it hence can be regarded as an out-of-distribution subgroup.

ISIC-2017 (Codella et al., 2018) is a widely-used benchmark for binary skin lesion segmentation in dermatoscopy. It consists of 2750 high-resolution dermatoscopic images, along with corresponding binary segmentation masks, which are distributed into training (2000 images), validation (150 images), and test (600 images) sets assuring patient independence between sets. In addition, clinical metadata involving pathological diagnosis of skin lesions (three subtypes: melanoma, nevus, and keratosis) as well as patient demographics (age and gender) is also provided for each lesion by default. Similar to findings in polyp segmentation datasets, the data distribution on lesion diagnostic attributes is also severely unbalanced.

5.2 Experimental Settings

5.2.1 Comparison Methods

Following the experimental settings of the previous study (Fan et al., 2020), we adopt three medical image segmentation methods *i.e.,* UNet (Ronneberger et al., 2015), UNet++ (Zhou et al., 2018) and ResUNet++ (Jha et al., 2019) as the baseline comparisons. In the polyp segmentation task, several state-of-the-art frameworks, *i.e.,* PraNet (Fan et al., 2020), ACSNet (Zhang et al., 2020), SCRNet (Wu et al., 2021), SANet (Wei et al., 2021), CCBANet (Nguyen et al., 2021) and MSNet (Zhao et al., 2021) are adopted as the strong comparison approaches. Recently vision transformers demonstrated promising performance in medical image segmentation (Zhang et al., 2021; Dong et al., 2021; Valanarasu et al., 2021), thus we further include two transformer-based frameworks *i.e.,* Swin Transformer (Liu et al., 2021) and TransFuse (Zhang et al., 2021) for more exhaustive comparison. In the skin lesion segmentation task, several competitive methods, i.e., MedT (Valanarasu et al., 2021), UNext (Valanarasu & Patel, 2022) and FATNet (Wu et al., 2022) are adopted as the strong comparison approaches.

Following the experimental design of the preliminary work (Puyol-Antón et al., 2021), we adopt four bias mitigation algorithms to examine the fairness of segmentation accuracy on different polyp subtypes:

Group-Balanced Weighting (GBW): Being aware of the protected attribute, this strategy aims to manipulate the weight of the loss of individual samples to guarantee the balance of different groups in loss computation. Specifically, based on distributional frequencies of groups in training data, majority groups are assigned to smaller weights while minority groups are assigned to larger weights.

Group-Balanced Sampling (GBS): Being aware of the protected attributes, this strategy aims to manipulate the sampling strategy to balance the group distribution. For each mini-batch, the data are resampled by the protected attributes, *i.e.,* Paris Classification, and individuals are selected to ensure each protected group is equally represented.

Attribute-aware Meta Learning (AML): This strategy is originally proposed by Dwork et al. (2012) which includes separate networks or branches trained for classifying different sensitive attributes. It is employed in facial expression recognition for bias mitigation (Xu et al., 2020). In our experiments, we use a shared encoder and separate decoders to

Table 1 Comparison results on three benchmarks of polyp segmentation

Method	EndoScene				Kvasir-SEG				PICCOLO			
	$MAE\downarrow$	$Dice\uparrow$	$IoU\uparrow$	$\mathcal{F}\uparrow$	$MAE\downarrow$	$Dice\uparrow$	$IoU\uparrow$	$\mathcal{F}\uparrow$	$MAE\downarrow$	$Dice\uparrow$	$IoU\uparrow$	$\mathcal{F}\uparrow$
U-Net (15') (Ronneberger et al., 2015)	0.032	0.839	0.772	0.688	0.520	0.842	0.760	0.679	0.061	0.605	0.519	0.422
U-Net++ (18') (Zhou et al., 2018)	0.045	0.729	0.646	0.637	0.052	0.842	0.760	0.703	0.054	0.682	0.615	0.581
ResUNet++ (19') (Jha et al., 2019)	0.063	0.524	0.443	0.436	0.056	0.811	0.727	0.648	0.058	0.602	0.537	0.470
PraNet (20') (Fan et al., 2020)	0.035	0.817	0.744	0.758	0.031	0.892	0.836	0.779	0.030	0.717	0.698	0.659
ACSNet (20') (Zhang et al., 2020)	0.030	0.852	0.787	0.816	0.032	0.893	0.838	0.790	0.035	0.788	0.733	0.696
SCRNet (21') (Wu et al., 2021)	0.029	0.853	0.788	0.821	0.036	0.886	0.825	0.804	0.049	0.666	0.594	0.492
CCBANet (21') (Nguyen et al., 2021)	0.034	0.839	0.765	0.781	0.031	0.894	0.834	0.783	0.028	0.763	0.717	0.686
SANet (21') (Wei et al., 2021)	0.031	0.842	0.772	0.789	0.029	0.902	0.845	0.789	0.035	0.792	0.730	0.696
MSNet (21') (Zhao et al., 2021)	0.035	0.808	0.745	0.760	0.034	0.890	0.831	0.787	0.032	0.813	0.753	0.722
LDNet (22') (Zhang et al., 2022b)	0.031	0.844	0.778	0.792	0.031	0.893	0.835	0.775	0.028	0.829	0.772	0.737
Swin (21') (Liu et al., 2021)	0.032	0.839	0.723	0.782	0.028	0.902	0.838	0.712	0.028	0.823	0.764	0.735
TransFuse (21') (Zhang et al., 2021)	0.033	0.824	0.764	0.735	0.032	0.898	0.834	0.759	0.029	0.777	0.709	0.645
Ours	**0.028**	**0.858**	**0.795**	**0.826**	**0.027**	**0.912**	**0.859**	**0.812**	**0.018**	**0.860**	**0.825**	**0.786**

The best and second best results are **highlighted** and underlined

perform object segmentation and protected attribute classification.

Stratified Group Model (SGM): Contrary to the above two approaches, this strategy assumes that the protected attributes are accessible at inference as well as training time. It applies an independent segmentation model for each group. We initially train the vanilla U-Net (Ronneberger et al., 2015) using the unbalanced full training data and then fine-tune separated models using samples of protected groups.

Compared to the above debiasing approaches, our method does not need any supervision from the protected attributes during training. For a fair comparison, we implement the above methods using the same U-Net architecture.

5.2.2 Evaluation Metrics

The performance of all methods is evaluated from three aspects: effectiveness which identifies the overall segmentation accuracy, fairness which measures the segmentation disparity on different subgroups and trade-off which combines the effectiveness and fairness.

Effectiveness: Following the previous work (Fan et al., 2020), we evaluate segmentation performances from pixel precision, region similarity, and contour accuracy. We employ mean absolute error (*MAE*) to calculate pixel-level errors. To measure the region-based segmentation similarity, we utilize the Dice similarity coefficient (*Dice*), and the intersection-of-union coefficient (*IoU*). For contour accuracy, we apply the boundary Dice measurement ($\mathcal{F}$). Concretely, let the contours of the prediction mask h and ground truth y are denoted as c_h and c_y, respectively. The precision P_c and recall R_c between c_h and c_y can be calculated by a bipartite graph matching (Martin et al., 2004). Thus, the boundary Dice coefficient is defined as:

$$\mathcal{F} = \frac{2P_c R_c}{P_c + R_c} \tag{17}$$

Among these evaluation metrics, a higher value of *Dice*, *IoU*, $\mathcal{F}$ and a lower value of *MAE* indicate better segmentation results.

Fairness: Following the previous study (Puyol-Antón et al., 2021), we employ the same metrics to evaluate fairness: the standard deviation (*SD*) and the skewed error ratio (*SER*) of the *Dice* value. The standard deviation measures the amount of disparity of the *Dice* values between different subtypes.

Table 2 Quantitative results of fairness on PICCOLO dataset

Method	Effectiveness								Fairness		Trade-off
	0-IIa	*Is*	*Ip*	*Isp*	*Unknown*	*0-IIa/c*	*0-IIb*	*Overall*↑	*SD*↓	*SER*↑	*CAI*↑
U-Net (15') (Ronneberger et al., 2015)	0.679	0.932	0.658	0.610	0.661	0.350	0.326	0.605	0.193	0.342	0
Group-Balanced Weighting	0.683	0.953	0.843	0.799	0.646	0.722	0.402	0.667	0.162	0.415	0.047
Group-Balanced Sampling	0.728	0.943	0.866	0.895	0.605	0.734	0.376	0.659	0.182	0.392	0.033
Attribute-aware Meta Learning	0.631	0.947	0.626	0.856	0.687	0.740	0.352	0.650	0.177	0.365	0.031
Stratified Group Model	0.693	0.954	0.853	0.756	0.612	0.699	0.399	0.652	0.164	0.412	0.038
PraNet (20') (Fan et al., 2020)	0.788	0.951	0.914	0.896	0.703	0.821	0.374	0.717	0.183	0.386	0.061
ACSNet (20') (Zhang et al., 2020)	0.855	**0.962**	0.916	**0.950**	0.813	0.777	0.483	0.788	0.153	0.496	0.112
SCRNet (21') (Wu et al., 2021)	0.759	0.934	0.874	0.664	0.709	0.795	0.190	0.666	0.227	0.195	0.014
CCBANet (21') (Nguyen et al., 2021)	**0.876**	0.953	0.925	0.879	0.786	0.859	0.365	0.763	0.187	0.377	0.082
SANet (21') (Wei et al., 2021)	0.856	0.952	0.875	0.932	0.788	0.825	0.569	0.792	0.118	0.594	0.131
MSNet (21') (Zhao et al., 2021)	0.870	0.959	0.917	0.949	0.783	0.858	0.647	0.813	0.102	0.671	0.150
LDNet (22') (Zhang et al., 2022b)	0.835	0.952	**0.931**	0.918	0.815	0.835	0.718	0.829	0.076	0.751	0.171
Swin (21') (Liu et al., 2021)	0.862	0.951	0.924	0.912	0.849	**0.880**	0.569	0.823	0.119	0.594	0.146
TransFuse (21') (Zhang et al., 2021)	0.820	0.954	0.911	0.866	0.841	0.842	0.387	0.777	0.175	0.399	0.095
Ours	0.858	0.953	0.900	0.862	**0.914**	0.864	**0.746**	**0.860**	**0.060**	**0.780**	**0.194**

Best and second best results are **highlighted** and underlined. All methods are reproduced using the officially released codes

The skewed error ratio is calculated by the ratio of the highest error rate to the lowest error rate among different subtypes, which can be denoted as

$$SER = \frac{\min_g \left(1 - Dice_g\right)}{\max_g \left(1 - Dice_g\right)} \tag{18}$$

where g represents different subgroups. Among these two metrics, a higher value of *SER* and a lower value of *SD* indicate better fair results.

Trade-off: We modify the Conjunctive Accuracy Improvement (*CAI*) (Paul et al., 2022) which is proposed to measure both the effectiveness and fairness of algorithms. The new *CAI* is defined as the weighted linear combination of two terms including the (signed) standard deviation among subtypes decrement and the (signed) overall accuracy improvement, whose computed with respect to a baseline and the candidate debiased algorithm:

$$CAI = \frac{(SD^b - SD^d) + (Dice^d - Dice^b)}{2} \tag{19}$$

where SD^b and SD^d denote the standard deviation of the baseline and the debiased model. Similarly, $Dice^b$ and $Dice^d$ are the $Dice$ scores of the baseline and the debiasing model, respectively. A higher value of *CAI* represents greater superiority of the debiasing method.

5.2.3 Implementation Details

Our model is implemented in Pytorch and trained on a single NVIDIA RTX 3090. We adopt the pretrained ResNet-34 (He et al., 2016) as the encoder backbone of U-Net architecture. Wang et al. (2020) reveals the feature *"slow drift"* phenomena, which speculates features change drastically at the early phase of training but become relatively stable within a certain number of training iterations. Based on this observation, we leverage the warm-up strategy to apply prototype learning and inference after 30 epochs, allowing the model to reach a certain local optimal field where feature embeddings and the learned prototypes become more stable.

Table 3 Quantitative results of fairness on EndoScene dataset

Method	Effectiveness				Fairness		Trade-off
	Is	*Isp*	*Ip*	*Overall*↑	*SD*↓	*SER*↑	*CAI*↑
U-Net (15') (Ronneberger et al., 2015)	0.873	**0.916**	0.512	0.839	0.181	0.554	0
Group-Balanced Weighting	0.909	0.905	0.542	0.848	0.172	0.591	0.009
Group-Balanced Sampling	0.906	0.908	0.543	0.847	0.172	0.593	0.009
Attribute-aware Meta Learning	0.886	0.914	0.522	0.844	0.178	0.567	0.003
Stratified Group Model	0.887	0.852	0.453	0.799	0.197	0.505	−0.028
PraNet (20') (Fan et al., 2020)	0.899	0.908	0.472	0.837	0.204	0.515	−0.013
ACSNet (20') (Zhang et al., 2020)	0.894	0.902	0.552	0.844	0.163	0.608	0.011
SCRNet (21') (Wu et al., 2021)	0.826	0.790	0.306	0.724	0.237	0.363	−0.086
CCBANet (21') (Nguyen et al., 2021)	0.901	0.876	0.594	0.839	0.139	0.656	0.021
SANet (21') (Wei et al., 2021)	0.892	0.912	0.510	0.842	0.185	0.555	−0.001
MSNet (21') (Zhao et al., 2021)	0.891	0.861	0.428	0.808	0.212	0.474	−0.031
LDNet (22') (Zhang et al., 2022b)	0.886	0.914	0.523	0.844	0.178	0.567	0.004
Swin (21') (Liu et al., 2021)	**0.929**	0.898	0.504	0.848	0.193	0.538	−0.004
TransFuse (21') (Zhang et al., 2021)	0.872	0.885	0.583	0.833	0.139	0.655	0.018
Ours	0.928	0.888	**0.631**	**0.858**	**0.132**	**0.676**	**0.034**

The best and second best results are **highlighted** and underlined. All methods are reproduced using the officially released codes

Table 4 Quantitative results of fairness on ISIC-2017 dataset

Method	Effectiveness				Fairness		Trade-off
	Nevus	*Melanoma*	*Keratosis*	*Overall*↑	*SD*↓	*SER*↑	*CAI*↑
U-Net (15') (Ronneberger et al., 2015)	0.829	0.686	0.590	0.766	0.099	0.415	0
Group-Balanced Weighting	0.858	0.749	0.737	0.818	0.055	0.532	0.048
Group-Balanced Sampling	0.846	0.723	0.698	0.800	0.065	0.509	0.034
Attribute-aware Meta Learning	0.844	0.753	0.716	0.807	0.054	0.551	0.043
Stratified Group Model	0.852	0.709	0.655	0.795	0.083	0.429	0.023
UNet++ (18') (Zhou et al., 2018)	0.865	0.786	0.760	0.835	0.044	0.559	0.061
ResUNet++ (19') (Jha et al., 2019)	0.799	0.661	0.512	0.729	0.117	0.411	−0.027
MedT(21') (Valanarasu et al., 2021)	0.741	0.618	0.554	0.689	0.078	0.579	−0.028
UNext (22') (Valanarasu & Patel, 2022)	0.852	0.792	0.755	0.826	0.040	0.603	0.059
FATNet (22') (Wu et al., 2022)	0.857	0.781	0.774	0.830	0.038	**0.629**	0.063
Ours	**0.874**	**0.795**	**0.805**	**0.848**	**0.035**	0.617	**0.073**

The best and second best results are **highlighted** and underlined. All methods are reproduced using the officially released codes

To enlarge the data diversity, we utilize data augmentation strategies such as random horizontal and vertical flips, zoom, shift, and rotation. All augmented images are then resized to 352×352 for training. We deploy the Adam optimizer with an initial learning rate of 1e-4, a batch size of 32, and a maximum epoch number of 150.

5.3 Results

5.3.1 Performance on Effectiveness

We validate the effectiveness of the proposed approach on three widely-used benchmarks, by comparing it with previous state-of-the-art polyp segmentation methods including both convolutional networks and transformers.

The quantitative results are exhibited in Table 1. On EndoScene, our proposed method outperforms all methods with a relatively marginal increment of all metrics. On Kvasir-SEG, in particular, our model further raises the previous best results from 0.902/0.845/0.804 to 0.912/0.859/0.812 in terms of $Dice/IoU/\mathcal{F}$, respectively. On PICCOLO, our method significantly advances the state-of-the-art results from 0.829 to 0.860 in $Dice$. The improvement in IoU, MAE, and $\mathcal{F}$ are also substantial, achieving 0.825, 0.018, and 0.786, respectively. It is noteworthy that the performance of the vision transformer on PICCOLO is obviously

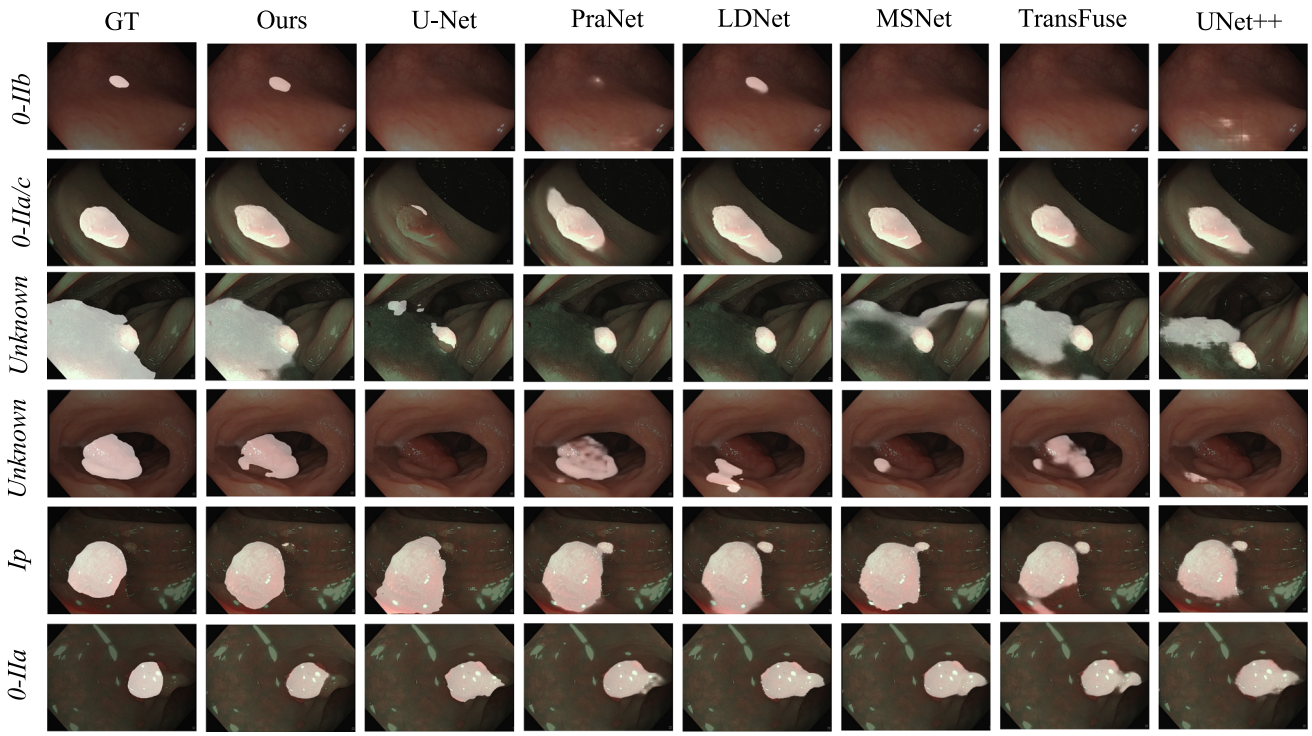

Fig. 6 Qualitative results of colonoscopy polyp segmentation. Especially on minority subtypes, e.g. *0-IIb*, *0-IIa/c* and *Unknown*, our method produces more accurate segmentation results compared with other State-of-the-art models

better than that on EndoScene. Since PICCOLO has more diverse lesions and complex scenes than those on the other two datasets, it indicates the vision transformer owns the capability to better handle massive samples and complex circumstances which is consistent with the recent finding (Bhojanapalli et al., 2021). These observations suggest that our prototype-based representation learning method is more applicable in more diverse lesions in the practical clinical scenario.

5.3.2 Performance on Fairness

Experimental results are summarized in Tables 2, 3 and 4. The *Dice* accuracy on different subtypes is reversely sorted according to their amount in the training set.

State-of-the-Art Methods. First, the quantitative results exhibit that our approach is fairer than state-of-the-art methods, especially in minority subtypes. For example, in Table 2, CCBANet, SCRNet, PraNet, and our method achieve 0.187, 0.227, 0.183, and 0.060 on *SD*, respectively. Second, we also notice that overall segmentation performance is approximately related to the fairness of the model. For example, in Table 2, from the SCRNet (Wu et al., 2021) to the LDNet (Zhang et al., 2022b), the overall *Dice* grade increases from 0.666 to 0.829, meanwhile the *SD* score decreases from 0.227 to 0.076. However, there are also exceptions. For example, in Table 3, as the overall *Dice* of ACSNet is higher than that of CCBANet, the *SD* score of ACSNet is still higher than that of CCBANet. These observations indicate that the introduction of advanced techniques, e.g. context information, attention mechanisms, and strong pretrained backbone, enhances the learning ability on rarely seen hard cases, achieves more improvement on minority subtypes, and eventually contributes to the progress of fairness. But not all advanced mechanisms benefit from fairness, e.g. in Table 3 PraNet owns equivalent overall *Dice* but heavily worse *SD* compared with U-Net. As illustrated in Fig. 6, our method can produce accurate segmentation masks on various polyp subtypes.

Bias Mitigation Algorithms. The first five rows show the comparison between the baseline and four approaches for bias mitigation. We can notice that all bias mitigation strategies gain considerable decrement on *SD* and increment on *SER*. And they have great improvement on minority subtypes while marginal raising on majority subtypes. For example, in Table 2, the Group-Balanced Weighting significantly promotes *Dice* score on *0-IIa/c* from 0.350 to 0.722, but marginally increases *Dice* score on *0-IIa* from 0.679 to 0.683. In Table 4, the Attribute-aware Meta Learning improves *Dice* score on *Keratosis* from 0.590 to 0.716, but marginally increases *Dice* score on *Nevus* from 0.829 to 0.844. Among all debiasing methods, Group-Balanced Weighting achieves the best segmentation parity on all subgroups. Besides, we can observe that Attribute-aware Meta Learning, which is

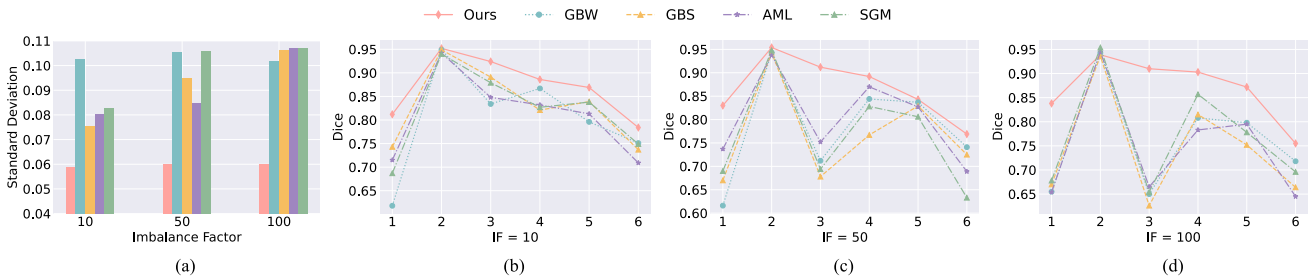

Fig. 7 **a** The standard deviations of all bias mitigation algorithms under different values of Imbalance Factor (IF) varying from 10 to 100. **b**, **c**, **d** The detailed results on all subtypes. The segmentation disparity of our method even obtains a minor decline with the increase of the imbalance factor. And our method consistently outperforms other models by a significant margin

a widely-used debiasing method in the classification task, can't mitigate the bias issue in segmentation well. This is because of the intrinsic difference between bias issue in classification and our setting: the bias in classification is mainly caused by the entanglement of deep features on sensitive attributes and target attributes (Creager et al., 2019; Hong & Yang, 2021; Tartaglione et al., 2021), while bias in segmentation is induced by the imbalance distribution among sensitive attributes. In summary, all debiasing methods can reduce the bias issue at a certain intensity.

The experimental results in the last row of Table 2 and Table 3 indicate that our proposed approach significantly outperforms all debiasing comparisons in terms of all fairness metrics. It is noteworthy that, compared to other methods, our approach obtains more significant improvement in minority subtypes. For example, our method achieves the best or second best performances on minority subtypes of *unknown*, *0-IIa/c*, and *0-IIb*. This demonstrates that our learned prototypes mainly represent minority subtypes, and further verifies that our framework is able to explore and store rare visual patterns from massive and complex data.

Effect of Imbalance Factor. Consistent with the settings of Sect. 3.3, we vary the Imbalance Factor (IF) of the training set of PICCOLO to further evaluate the debiasing ability of our method under different circumstances. From the results illustrated in Fig. 7, we can obtain several observations. First, our method consistently outperforms other debiasing algorithms by a significant margin, achieving the best performance on all subtypes under different imbalance distributions. Second, our method owns a greater ability against the increasing data imbalance. For instance, as IF increases, the standard deviation of our method only obtains a minor increment, while those of the three debiasing algorithms(GBS, AML, and SGM) gain a considerable increment. It is noteworthy that the segmentation disparity of Group-Balanced Weighting (GBW) approximately also remains unchanged. These conclusions suggest the superiority of our approach against the imbalance distribution.

Table 5 Ablation analysis for coarse segmentation map **s** in Object Pattern Extraction function on three benchmarks

Method	EndoScene		Kvasir-SEG		PICCOLO	
	Dice	*IoU*	*Dice*	*IoU*	*Dice*	*IoU*
Ours	**0.858**	**0.795**	**0.912**	**0.859**	**0.860**	**0.825**
w/o **s**	0.815	0.743	0.881	0.824	0.786	0.727

The best results are highlighted in bold

5.4 Ablation Study

In this section, we conduct extensive studies to explore the effects of each proposed component and various parameter values.

5.4.1 Object Pattern Extraction

In object pattern extraction, a coarse segmentation map **s** is designed to force the network to concentrate on the visual pattern of the interesting object and eliminate the interference of irrelevant information, e.g. background. We evaluate the effectiveness of the above operation on segmentation performance on three benchmarks. As shown in Table 5, as the coarse segmentation map **s** is discarded, the accuracy drastically decreases from 0.858/0.912/0.860 to 0.815/0.881/0.786 in terms of *Dice*, on EndoScene/Kvasir-SEG/PICCOLO, respectively. The ablation results clearly demonstrate that concentrating on the foreground pattern can facilitate the subsequent prototype learning and improve the segmentation performance.

5.4.2 Unknown Pattern Discovery

The Unknown Pattern Discovery mechanism identifies whether coming samples are dissimilar from existing prototypes. If a feature is considered dissimilar from the current prototypes, it will be utilized to construct a new prototype. Therefore, our method can dynamically determine the num-

Table 6 Ablation results of different prototype-based strategies on PICCOLO

Method	Effectiveness								Fairness		Trade-off
	0-IIa	*Is*	*Ip*	*Isp*	*Unknown*	*0-IIa/c*	*0-IIb*	*Overall*↑	*SD*↓	*SER*↑	*CAI*↑
Baseline	0.679	0.932	0.658	0.610	0.661	0.350	0.326	0.605	0.193	0.342	0
# proto=6	+ 0.178	**+ 0.022**	**+ 0.251**	+ 0.174	+ 0.159	+ 0.472	+ 0.264	+ 0.212	0.114	0.615	0.146
# proto=12	**+ 0.180**	+ 0.018	+ 0.243	**+ 0.327**	+ 0.185	+ 0.474	+ 0.274	+ 0.216	0.110	0.624	0.150
# proto=18	+ 0.173	+ 0.017	+ 0.223	+ 0.313	+ 0.225	+ 0.494	+ 0.301	+ 0.234	0.098	0.658	0.165
# proto=24	+ 0.156	+ 0.011	+ 0.234	+ 0.321	+ 0.228	+ 0.480	+ 0.233	+ 0.221	0.121	0.588	0.146
Ours (# proto=6)	+ 0.178	+ 0.021	+ 0.242	+ 0.252	**+ 0.253**	**+ 0.514**	**+ 0.420**	**+ 0.255**	**0.060**	**0.780**	**0.194**

We select three prototype amounts: # proto=6, 10, 20. Best and second best results are **highlighted** and underlined. Our dynamic discovery mechanism gets a relatively small increment in the majority whilst a tremendous raising in minority subtypes

Table 7 Ablation analysis of different clustering algorithms on ISIC-2017

Methods	Clustering quality		Effectiveness				Fairness		Trade-off
	SC↑	*DBI*↓	*Nevus*	*Melanoma*	*Keratosis*	*Overall*↑	*SD*↓	*SER*↑	*CAI*↑
Kmeans (#clusters=5)	0.41	1.31	0.866	0.772	0.771	0.833	0.045	0.584	0.061
GMM (#clusters=5)	0.45	1.39	0.826	0.746	0.725	0.796	0.044	**0.630**	0.042
Mean Shift (#clusters=4)	0.34	1.65	0.827	0.735	0.714	0.792	0.049	0.605	0.038
DBSCAN (#clusters=3)	0.04	1.50	0.857	0.750	0.734	0.817	0.065	0.539	0.048
FINCH (#clusters=5) (Sarfraz et al., 2019)	0.34	1.41	0.866	0.770	0.752	0.830	0.050	0.539	0.056
Ours (#proto=5)	**0.47**	**1.18**	**0.874**	**0.795**	**0.805**	**0.848**	**0.035**	0.617	**0.073**

We report the results in terms of both clustering quality and segmentation effectiveness and fairness. The best and second best results are **highlighted** and underlined

ber of prototypes in the training procedure. We hence conduct an ablation analysis to compare our discovery mechanism with the approaches utilizing a fixed number of prototypes and explore the impact of prototype amounts on fairness. All comparison methods use the same prototype updating and inference schemes as us, and own initialized prototypes following a uniform distribution. We manually select three amounts: # proto=6 (same prototype quantity as us), and 12, 18, and 24.

Several observations can be made from the ablation results on the PICCOLO dataset in Table 6. First, as the number of prototypes increases from 6 to 18, the performance on fairness also increases. For example, the improvements on minority subtypes *0-IIa/c*/*0-IIb* significantly improve from 0.472/0.264 to 0.494/0.301, respectively. However, as the number of prototypes continuously increases to 24, the performance on fairness starts to deteriorate. Second, from the last row, we can see that the proposed adaptive discovery mechanism obtains a relatively small increment in the majority but a tremendous raise in minority subtypes, and achieves the best performance on fairness.

Moreover, since the essence of prototype learning entails aggregating knowledge of similar visual patterns within the deep feature space, our approach is similar with conventional unsupervised clustering algorithms. Therefore, we perform a ablation analysis comparing our proposed method with conventional clustering techniques. We select both parametric and non-parametric clustering techniques as comparative methods. Parametric clustering methods include K-Means and Gaussian Mixture Model (GMM), which require the number of clusters as input. We use the number of discovered prototypes from our approach for a fair comparison. Non-parametric clustering methods include Mean Shift, DBSCAN, and FINCH (Sarfraz et al., 2019). These methods can automatically discover groupings in the data based on different statistical criteria. We select the clustering that is closest to our discovered number of prototypes for a fair comparison. In our analysis, we additionally adopt internal evaluation schemes to assess the quality of clustering results without requiring ground truth cluster assignments. The Silhouette Score (SC) is calculated as the ratio of the average distance to other data points within the same cluster to the minimum distance to data points in other clusters. A higher value indicates better clustering quality. The Davies-Bouldin Index (DBI) is calculated as the average maximum ratio of the within-cluster distance and the between-cluster distance for each cluster. A lower value indicates better clustering quality. These two metrics can both measure the separation and compactness of clusters. Experimental results on the ISIC-2017 dataset are exhibited in Table 7. First, from the

Table 8 Quantitative results of various discovery thresholds on PICCOLO dataset

Threshold	Effectiveness		Fairness		# proto
	Dice	*IoU*	*SD*	*SER*	
0.3	0.723	0.696	0.256	0.204	3
0.4	0.768	0.725	0.215	0.282	4
0.5	**0.860**	**0.825**	**0.060**	**0.780**	6
0.6	0.822	0.770	0.096	0.682	23
0.7	0.828	0.772	0.121	0.555	30

The best results are highlighted in bold

last row, we can see that the proposed adaptive discovery mechanism obtains the best performance in both clustering quality and segmentation effectiveness and fairness. Second, the parametric clustering methods demonstrate superior performance compared to non-parametric ones. Since the non-parametric methods highly depend on the chosen of other hyper-parameters, e.g. distance functions, the maximum distance between two samples for one to be considered as in the neighborhood of the other, the minimum number of samples to be a cluster, they are sensitive in deep feature spaces where high dimensionality and data sparsity cause unstable distance measuring and numerous outliers.

In summary, the experimental results validate that our dynamic prototype learning can better explore scarce visual patterns than a fixed prototype scheme and conventional clustering algorithms, by measuring the similarities between the past known knowledge and current features.

5.4.3 Discovery Threshold

During pattern discovering, feature assignment to a prototype in $\mathcal{M}$ is based on a threshold μ using the cosine similarity of the object pattern embedding with prototypes. If the computed similarity is smaller than μ, the object pattern embedding is considered different from existing prototypes. The larger value of μ, the easier it is for our framework to discover different visual patterns. Therefore, the threshold μ indirectly controls the number of prototypes, thereby measuring the ability of visual pattern modeling.

In Table 8, we conduct experiments with various thresholds. First, we notice that the amount of discovered prototypes increases as the threshold increases, which is consistent with our intuition. Second, adopting threshold $\mu = 0.5$ reaches the greatest performance on both effectiveness and fairness. Third, we notice that the performance of $\mu = 0.6, 0.7$ is much greater than that of $\mu = 0.3, 0.4$. These observations indicate that the visual pattern knowledge captured by a relatively small threshold is deficient, while that captured by a relatively large threshold is adequate but tends to be redundant.

Table 9 Analysis of naive update with various coefficients adaptive momentum update on PICCOLO

Methods	η	Effectiveness		Fairness	
		Dice	*IoU*	*SD*	*SER*
w/ naive update	0.9	0.844	0.793	0.104	0.646
	0.7	0.830	0.773	0.112	0.603
	0.3	0.804	0.733	0.153	0.435
w/ adaptive update	–	**0.860**	**0.825**	**0.060**	**0.780**

The best results are highlighted in bold

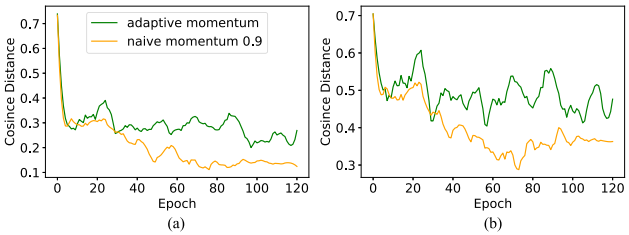

Fig. 8 Comparison of average cosine distance between **a** prototypes and their hardest negative, and **b** all prototypes per epoch on training procedure. The curves are properly smoothed for clarity

5.4.4 Known Pattern Update

The prototype update mechanism successively enriches the known prototypes by incorporating similar objects. Our proposed Adaptive Momentum Update can increase the separability among prototypes. In Table 9, we compare our proposed update strategy in Eq. (7) with the conventional updating method using fixed momentum update coefficient in Eq. (5). We notice that with the increase of the update momentum, the model has witnessed considerable performance promotion on both effectiveness and fairness. That is because a high value of η reduces the changing amplitude of prototypes during updating, and hence stabilizes the accumulation of long-term knowledge. In the last row, our proposed strategy improves the performance of *Dice* by 0.016 and *SD* by 0.044 and achieves the best performance on all metrics of effectiveness and fairness.

Moreover, we conduct an experiment to validate the capability of our Adaptive Momentum Update for increasing the separability among prototypes. Concretely, we record the average cosine distance between prototypes and their hardest negatives, as well as between all prototypes in the training procedure. In Fig. 8, obviously, prototypes learned by our adaptive momentum have larger separability. Besides, the learning procedure of the naive updating mechanism is more stable than ours, because the naive momentum is relatively larger than that of ours, which is consistent with the intuition that the large updating momentum can stabilize prototype learning. In summary, the ablation results demonstrate that explicit consideration of separability between prototypes improves the fairness of the approach.

Table 10 Comparison of different distance functions in prototype learning on three benchmarks

Distance function	EndoScene		Kvasir-SEG		PICCOLO	
	Dice	*IoU*	*Dice*	*IoU*	*Dice*	*IoU*
euclidean	84.42	77.36	90.27	82.69	85.22	78.95
cosine	**85.83**	**79.48**	**91.16**	**85.87**	**86.01**	**82.47**

The best results are highlighted in bold

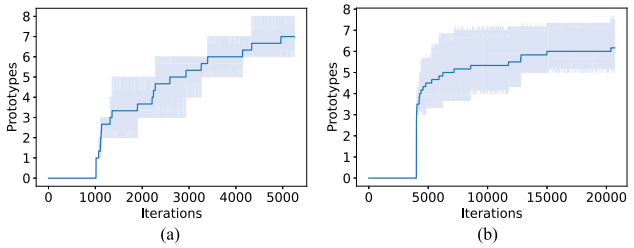

Fig. 9 The normalized statistics of images belong to different prototypes on subtypes. There are great distributional variations among different prototypes

5.4.5 Distance Function

In prototype learning, we utilize the cosine distance function to measure similarities between object pattern embeddings and prototypes. Moreover, we conduct experiments by replacing cosine distance with euclidean distance. As shown in Table 10, the performance of cosine distance is better than euclidean distance on all three benchmarks. That is because, euclidean distance, which pays more attention to the absolute numerical differences between embeddings, probably can not mine actual differences between pattern embeddings.

5.5 Discussion

In this section, we discuss and analyze the intrinsic properties of our prototype-based algorithm.

5.5.1 Prototype Sequential Analysis

To understand the behavior of prototype growth during the training procedure, we deliver the time sequential analysis of prototypes on EndoScene and PICCOLO datasets. The results are presented in Fig. 9. First, the number of prototypes dramatically increases at the start and then gradually stabilizes. This observation reflects that the stored knowledge is vacant at an early stage of training, and eventually saturated at the late stage, which is consistent with our intuition. Second, the error band indicates the differences in the behavior of the prototype growing between trials. When the training data is shuffled, there is a discrepancy between the sequence of training samples in each trial. We can observe that the error band of our prototype amount suppresses under $[-1,1]$, which implies that our dynamic prototype is robust to shuffle training data.

5.5.2 Prototype Visualization

In this part, we further interpret the prototypes learned from our framework. We calculate the number of images belonging to each prototype in terms of subtypes. The distributions are normalized to [0, 1]. From illustrations in Fig. 10, we observe that there are great distributional variations among different prototypes. For example, the last prototype characterizes the majority subtypes (e.g. *0-IIa* and *Is*), while the second and third prototypes mainly characterize the minority subtypes (e.g. *0-IIb* and *0-IIa/c*). These significant observations demonstrate that, without the supervision of corresponding protected attributes, our learned prototypes can still consciously correlate to different subtypes. And our method can explore inherent distinctions between embeddings with different visual patterns.

6 Limitations

Dataset bias (Puyol-Antón et al., 2021; Khosla et al., 2012) and task bias (Tartaglione et al., 2021) are the two most prevalent biases in recent research. Dataset bias is often induced due to the imbalance distribution of data with respect to protected attributes (Adeli et al., 2021). Task bias is induced by the intrinsic dependency between protected attributes and the target task. For instance, hair length has distorted associations with gender in face recognition (Adeli et al., 2021). Since the proposed framework aims to discover the visual patterns in an unsupervised way, it can only mitigate the dataset bias introduced by imbalance distribution, but can not deal with the scenarios where protected variables and target tasks are entangled in feature space.

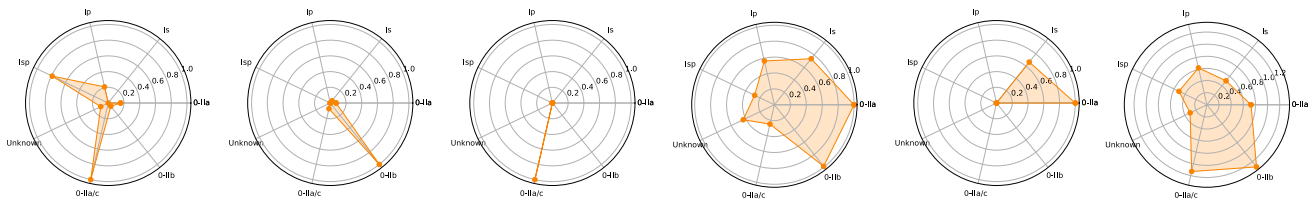

Fig. 10 The number of prototypes growing during the training process on **a** EndoScene and **b** PICCOLO

In our work, the proposed method aims to balance the learning of visual pattern diversity. In fact, the cause of diversity may come from many kinds of aspects such as gender, race, object subtypes, environmental conditions, and so on. Therefore, it is very interesting to evaluate the model fairness in several aspects simultaneously. However, since most of the segmentation datasets do not contain enough meta information, our evaluation of fairness is limited to polyp subtypes.

7 Conclusions

In this paper, we extend the model bias from demographic diversity to visual pattern diversity. We argue that demographic bias can also be attributed to visual pattern diversity in the segmentation task. To this end, we propose a prototype-based network that can balance the learning of different groups for a given dataset. We first propose an object pattern embedding mechanism to make the network focus on the foreground region. Then we design a prototype learning method to discover and memorize different visual patterns contained in data to dynamically balance the learning of both majority and minority groups. Moreover, our proposed network can build and update prototypes in an unsupervised manner and the number of prototypes is dynamic depending on encounter samples. Therefore, our method is more applicable to scenarios where the labels of visual patterns are various and costly to acquire.

We evaluate the proposed prototype-based network on three widely used polyp segmentation datasets with abundant qualitative and quantitative experiments. The performance is quantified on both effectiveness and fairness. For effectiveness, our proposed method outperforms both the convolutional network-based and the transformer-based state-of-the-art methods. Especially on the large-scale PICCOLO dataset, our method significantly surpasses the state-of-the-art results from 82.91% to 86.78% in $Dice$. Since PICCOLO has more diverse lesions and complex scenes, the improvement suggests that our prototype learning method is more applicable to more diverse lesions which are more commonly seen in the clinical scenario. For fairness, compared to other methods, our approach obtains significant improvement in minority subtypes (e.g. *unknown*, *0-IIa/c*, and *0-IIb*). This demonstrates that part of the learned prototypes can enhance the representation of minority subtypes, and further verifies that our framework is able to discover and extract rare visual patterns from massive and complex data.

Moreover, extensive ablation studies are conducted to show the effectiveness of each proposed component and various parameter values. Lastly, we analyze how the number of prototypes grows during the training process and visualize the nearest images for each learned prototype. This further verifies that different prototypes store various visual patterns and the diversity constraint of prototypes can better mitigate the bias.

In the further work, we aim to find more evidence to show the intermediate cause of bias in the imbalance distribution is different visual patterns. Furthermore, we plan to further evaluate the model segmentation fairness on several protected attributes simultaneously.

Data Availability The datasets generated during the current study are available in the Github repository, https://github.com/zijinY/dynamic-prototype-debiasing.

Declarations

Conflict of interest The authors declared that they have no conflicts of interest to this work.

References

Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E. V., Fei-Fei, L., Niebles, J.C., & Pohl, K. M. (2021). Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2513–2523).

Ahmad, M. A., Patel, A., Eckert, C., Kumar, V., & Teredesai, A. (2020). Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3529–3530).

Alshammari, S., Wang, Y. X., Ramanan, D., & Kong, S. (2022). Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6897–6907).

Arik, S. O., & Pfister, T. (2019). Protoattend: Attention-based prototypical learning. arXiv preprint arXiv:1902.06292

Axon, A., Diebold, M., Fujino, M., Fujita, R., Genta, R., Gonvers, J.-J., Guelrud, M., Inoue, H., Jung, M., Kashida, H., et al. (2005). Update on the Paris classification of superficial neoplastic lesions in the digestive tract. *Endoscopy, 37*(6), 570–578.

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *Nips Tutorial, 1*, 2.

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., & Veit, A. (2021). Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10231–10241).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77–91). PMLR.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems,30*.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems,32*.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems,32*.

Cheng, M., Kong, Z., Song, G., Tian, Y., Liang, Y., & Chen, J. (2021). Learnable oriented-derivative network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 720–730). Springer.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science, 4*, 123–144.

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence,* (Vol. 33, pp. 7801–7808).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., & Kittler, H., & Halpern, A. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 168–172). IEEE.

Cosgrove, K. P., Mazure, C. M., & Staley, J. K. (2007). Evolving knowledge of sex differences in brain structure, function, and chemistry. *Biological Psychiatry, 62*(8), 847–855.

Creager, E., Madras, D., Jacobsen, J. H., Weis, M., Swersky, K., Pitassi, T., & Zemel, R. (2019). Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning* (pp. 1436–1445). PMLR.

Dong, B., Wang, W., Fan, D. P., Li, J., Fu, H., & Shao, L. (2021). Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932

Dong, Q., Gong, S., & Zhu, X. (2018). Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(6), 1367–1381.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S., & Uszkoreit, J. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Du, S., Hers, B., Bayasi, N., Hamarneh, G., & Garbi, R. (2022). FairDisCo: Fairer AI in dermatology via disentanglement contrastive learning. arXiv preprint arXiv:2208.10013

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226).

Fan, D. P., Ji, G. P., Zhou, T., Chen, G., Fu, H., Shen, J., & Shao, L. (2020). Pranet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 263–273). Springer.

Fang, Y., Chen, C., Yuan, Y., & Tong, K. Y. (2019). Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 302–310). Springer.

Feng, R., Yang, Y., Lyu, Y., Tan, C., Sun, Y., & Wang, C. (2019). Learning fair representations via an adversarial framework. arXiv preprint arXiv:1904.13341

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3146–3154).

Garcia-Peña, R. M., Teevno, M. A., Ochoa-Ruiz, G., & Ali, S. (2022). Supra: Superpixel guided loss for improved multi-modal segmentation in endoscopy. arXiv preprint arXiv:2211.04658

Georgopoulos, M., Oldfield, J., Nicolaou, M. A., Panagakis, Y., & Pantic, M. (2021). Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision, 129*(7), 2288–2307.

Gichoya, J. W., McCoy, L. G., Celi, L. A., & Ghassemi, M. (2021). Equity in essence: A call for operationalising fairness in machine learning for healthcare. *BMJ Health & Care Informatics, 28*(1), e100289.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems,29*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Hong, Y., & Yang, E. (2021). Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems, 34*, 26449–26461.

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., & Liu, W. (2019). Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 603–612)

Ioannou, S., Chockler, H., Hammers, A., & King, A. P. (2022). A study of demographic bias in cnn-based brain MR segmentation. In *International Workshop on Machine Learning in Clinical Neuroimaging* (pp. 13–22). Springer.

Isamah, N., Faison, W., Payne, M. E., MacFall, J., Steffens, D. C., Beyer, J. L., Krishnan, K. R., & Taylor, W. D. (2010). Variability in frontotemporal brain structure: The importance of recruitment of African Americans in neuroscience research. *PLoS One, 5*(10), 13642.

Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., Lange, T. D., Johansen, D., & Johansen, H. D. (2020). Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling* (pp. 451–462). Springer.

Jha, D., Smedsrud, P. H., Riegler, M. A., Johansen, D., De Lange, T., Halvorsen, P., & Johansen, H. D. (2019). Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)* (pp. 225–2255). IEEE.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., & Torralba, A. (2012). Undoing the damage of dataset bias. In *European Conference on Computer Vision* (pp. 158–171). Springer.

Kim, H., Joung, S., Kim, I. J., & Sohn, K. (2021). Prototype-guided saliency feature learning for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4865–4874).

Kishi, S., Reis, J. P., Venkatesh, B. A., Gidding, S. S., Armstrong, A. C., Jacobs, D. R., Jr., Sidney, S., Wu, C. O., Cook, N. L., Lewis, C. E., et al. (2015). Race-ethnic and sex differences in left ventricular structure and function: The coronary artery risk development in young adults (cardia) study. *Journal of the American Heart Association, 4*(3), 001264.

Kwon, H., Jeong, S., Kim, S., & Sohn, K. (2021). Dual prototypical contrastive learning for few-shot semantic segmentation. arXiv preprint arXiv:2111.04982

Lee, T., Puyol-Anton, E., Ruijsink, B., Shi, M., & King, A. P. (2022). A systematic study of race and sex bias in cnn-based cardiac MR segmentation. arXiv preprint arXiv:2209.01627

Li, G., Jampani, V., Sevilla-Lara, L., Sun, D., Kim, J., & Kim, J. (2021). Adaptive prototype learning and allocation for few-shot segmen-

tation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8334–8343).

Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 510–519).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022)

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440).

Maghsoudi, O. H. (2017). Superpixel based segmentation and classification of polyps in wireless capsule endoscopy. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1–4). IEEE.

Mamonov, A. V., Figueiredo, I. N., Figueiredo, P. N., & Tsai, Y.-H.R. (2014). Automated polyp detection in colon capsule endoscopy. *IEEE Transactions on Medical Imaging, 33*(7), 1488–1502.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(5), 530–549. https://doi.org/10.1109/TPAMI.2004.1273918

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1–35.

Moayeri, M., Pope, P., Balaji, Y., & Feizi, S. (2022). A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 19087–19097).

Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S., & Singh, S. (2017). No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 360–368).

Murugesan, B., Sarveswaran, K., Shankaranarayana, S. M., Ram, K., Joseph, J., & Sivaprakasam, M. (2019). Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 7223–7226). IEEE.

Nguyen, T. C., Nguyen, T. P., Diep, G. H., Tran-Dinh, A. H., Nguyen, T. V., & Tran, M. T. (2021). Ccbanet: Cascading context and balancing attention for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 633–643). Springer.

Nie, D., & Shen, D. (2020). Adversarial confidence learning for medical image segmentation and synthesis. *International Journal of Computer Vision, 128*(10), 2494–2513.

Paul, W., Hadzic, A., Joshi, N., Alajaji, F., & Burlina, P. (2022). Tara: Training and representation alteration for AI fairness and domain generalization. *Neural Computation, 34*(3), 716–753.

Petersen, E., Feragen, A., Zemsch, L. D. C., Henriksen, A., Christensen, O. E. W., & Ganz, M. (2022). Feature robustness and sex differences in medical imaging: A case study in MRI-based Alzheimer's disease detection. arXiv preprint arXiv:2204.01737

Pourhoseingholi, M. A., Baghestani, A. R., & Vahedi, M. (2012). How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology from Bed to Bench, 5*(2), 79.

Puyol-Antón, E., Ruijsink, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R., & King, A. P. (2021). Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 413–423). Springer.

Rambhatla, S. S., Chellappa, R., & Shrivastava, A. (2021). The pursuit of knowledge: Discovering and localizing novel categories using dual memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9153–9163).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention* (pp. 234–241). Springer.

Sánchez-Peralta, L. F., Pagador, J. B., Picón, A., Calderón, Á. J., Polo, F., Andraka, N., Bilbao, R., Glover, B., Saratxaga, C. L., & Sánchez-Margallo, F. M. (2020). Piccolo white-light and narrow-band imaging colonoscopic dataset: A performance comparative of models and datasets. *Applied Sciences, 10*(23), 8501.

Sarfraz, S., Sharma, V., & Stiefelhagen, R. (2019). Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8934–8943).

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., Liu, Y. (2019). How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 99–106).

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I., & Ghassemi, M. (2021a). Medical imaging algorithms exacerbate biases in underdiagnosis

Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I. Y., & Ghassemi, M. (2021b). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine, 27*(12), 2176–2182.

Shu, Y., Shi, Y., Wang, Y., Huang, T., & Tian, Y. (2020). P-odn: Prototype-based open deep network for open set recognition. *Scientific Reports, 10*(1), 1–13.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Sitenko, D., Boll, B., & Schnörr, C. (2021). Assignment flow for order-constrained oct segmentation. *International Journal of Computer Vision, 129*(11), 3088–3118.

Tartaglione, E., Barbano, C. A., & Grangetto, M. (2021). End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13508–13517).

Thomas, C., & Kovashka, A. (2021). Predicting visual political bias using webly supervised data and an auxiliary task. *International Journal of Computer Vision, 129*(11), 2978–3003.

Valanarasu, J. M. J., & Patel, V. M. (2022). Unext: Mlp-based rapid medical image segmentation network. arXiv preprint arXiv:2203.04967

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 36–46). Springer.

Vázquez, D., Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., López, A. M., Romero, A., Drozdzal, M., & Courville, A. (2017). A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering,2017*.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (fairware)* (pp. 1–7). IEEE.

Vleugels, J. L., Hazewinkel, Y., & Dekker, E. (2017). Morphological classifications of gastrointestinal lesions. *Best Practice & Research Clinical Gastroenterology, 31*(4), 359–367.

Vleugels, J. L., Hazewinkel, Y., & Dekker, E. (2017). Morphological classifications of gastrointestinal lesions. *Best Practice & Research Clinical Gastroenterology, 31*(4), 359–367.

Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., & Song, S. (2022). Stepwise feature fusion: Local guides global. arXiv preprint arXiv:2203.03635

Wang, K., Liew, J. H., Zou, Y., Zhou, D., & Feng, J. (2019). Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9197–9206).

Wang, X., Zhang, H., Huang, W., & Scott, M. R. (2020). Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6388–6397).

Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S. K., & Cui, S. (2021). Shallow attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 699–708). Springer.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision* (pp. 499–515). Springer.

Wu, H., Zhong, J., Wang, W., Wen, Z., & Qin, J. (2021). Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 2916–2924).

Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., & Wen, Z. (2022). FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis, 76*, 102327.

Xu, H., Sarkar, A., & Abbott, A. L. (2022). Color invariant skin segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2906–2915).

Xu, T., White, J., Kalkan, S., & Gunes, H. (2020). Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision* (pp. 506–523). Springer.

Xu, W., Xian, Y., Wang, J., Schiele, B., & Akata, Z. (2022). Attribute prototype network for any-shot learning. *International Journal of Computer Vision,* 1–19.

Yang, H. M., Zhang, X. Y., Yin, F., & Liu, C. L. (2018). Robust classification with convolutional prototype learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3474–3482).

Yoneyama, K., Venkatesh, B. A., Bluemke, D. A., McClelland, R. L., & Lima, J. A. (2017). Cardiovascular magnetic resonance in an adult human population: Serial observations from the multi-ethnic study of atherosclerosis. *Journal of Cardiovascular Magnetic Resonance, 19*(1), 1–11.

Yuan, Y., Chen, X., & Wang, J. (2020). Object-contextual representations for semantic segmentation. In *European Conference on Computer Vision* (pp. 173–190). Springer.

Yuan, H., Hadzic, A., Paul, W., de Flores, D. V., Mathew, P., Aucott, J., Cao, Y., & Burlina, P. (2022). Edgemixup: Improving fairness for skin disease classification and segmentation. arXiv preprint arXiv:2202.13883

Yue, X., Zheng, Z., Zhang, S., Gao, Y., Darrell, T., Keutzer, K., & Vincentelli, A.S. (2021). Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13834–13844).

Zhang, H., Dullerud, N., Roth, K., Oakden-Rayner, L., Pfohl, S., & Ghassemi, M. (2022a). Improving the fairness of chest x-ray classifiers. In *Conference on Health, Inference, and Learning* (pp. 204–233). PMLR.

Zhang, R., Lai, P., Wan, X., Fan, D. J., Gao, F., Wu, X. J., & Li, G. (2022b). Lesion-aware dynamic kernel for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 99–109). Springer.

Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., & Yu, Y. (2020). Adaptive context selection for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 253–262). Springer.

Zhang, Y., Liu, H., & Hu, Q. (2021). Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 14–24). Springer.

Zhang, H., & Ma, J. (2021). Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision, 129*(10), 2761–2785.

Zhao, X., Zhang, L., & Lu, H. (2021). Automatic polyp segmentation via multi-scale subtraction network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 120–130). Springer.

Zhao, X., Zhang, L., & Lu, H. (2021). Automatic polyp segmentation via multi-scale subtraction network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 120–130). Springer.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 3–11). Springer.

Zhou, T., Wang, W., Konukoglu, E., & Van Gool, L. (2022). Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2582–2593).

Zhu, K., Cao, Y., Zhai, W., Cheng, J., & Zha, Z. J. (2021). Self-promoted prototype refinement for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6801–6810).