Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

ASAP: Automated Style-Aware Similarity Measurement for Selection of Annotated Pre-training Datasets in 2D Biomedical Imaging

MIGUEL MOLINA-MORENO[1,2], MARCEL P. SCHILLING[3], MARKUS REISCHL[3] and RALF MIKUT[3]

[1]Department of Immunobiology, Yale University, New Haven CT 06511 USA (e-mail: miguel.molinamoreno@yale.edu)
[2]Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Madrid, Spain
[3]Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Baden-Württemberg 76344 Germany (e-mail: {marcel.schilling,markus.reischl,ralf.mikut}@kit.edu)

Corresponding author: Ralf Mikut (e-mail: ralf.mikut@kit.edu).

This work was supported by: the HoreKa Supercomputer through the Ministry of Science, Research, and the Arts Baden-Württemberg; the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition; the programs NACIP and MSE of the Helmholtz Association; the National Grant PID2020-118504GB-I00 of the Spanish Ministry of Science and Innovation (State Research Agency); and Madrid Regional Government Grant Y2020/NMT-6660 (synergistic R&D interdisciplinary project COMPANION-CM). Miguel Molina-Moreno was supported by the FPU Grant FPU18/02825 of the Spanish Ministry of Education, Culture and Sports. We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

ABSTRACT Medical imaging scenarios are characterized by varying image modalities, several organs/cell shapes, and little annotated data because of the expertise required for labeling. The successful use of state-of-the-art deep-learning approaches requires a large amount of annotated data or a pre-trained model. Despite the constant publication of new annotated datasets and pre-trained models, a vast subset of them remains untapped, owing to the challenges in effectively applying transfer learning or domain adaptation across varying scenarios. In this paper, we propose an automated style-aware framework for predicting the similarity value of a new biomedical dataset with respect to the state-of-the-art annotated datasets, selecting the most suitable annotated dataset for transfer learning or domain adaptation. Our pipeline, consisting of an autoencoder trained with self-supervised learning through a comprehensive loss function that considers the image reconstruction, style features, and dataset membership, does not need any kind of labels in training and test stages. The resulting 2D latent space represents a similarity measurement, which is demonstrated to correlate with the pre-training results in a task of binary semantic segmentation, and can provide the dataset that offers the optimal results for pre-labeling or pre-training a new biomedical task. Our results demonstrate the superior performance of this measurement with respect to manual selection and the state-of-the-art approaches. Therefore, ASAP can speed up the deployment processes of new biomedical applications. Our code is publicly available at https://github.com/miguel55/ASAP.

INDEX TERMS biomedical imaging, meta-learning, similarity, style transfer, transfer learning

I. INTRODUCTION

CONVOLUTIONAL Neural Networks (CNNs) have shown significant potential in complex tasks such as medical and biomedical imaging [1]–[6]. However, the parameters of traditional CNN pipelines are obtained through supervised learning, a process that requires labels obtained through human annotation. The annotation process for any dataset is tedious and time-consuming [7]. In the particular case of a biomedical dataset, this process must be done by individuals with significant expertise in the interpretation of biomedical imaging, a process that requires a notable amount of time and resources [8].

Nevertheless, the pool of annotated, available, and large-scale datasets for different biomedical imaging modalities is continuously growing, e.g., for dermatology [9], Magnetic Resonance Imaging (MRI) [10], and histopathology imaging [11]. On the one hand, foundation models [12] are able to make use of this large and diverse dataset pools, and useful in other applications, such as real image segmentation [13], but still restricted to specific image modalities (CT, MRI, endoscopy) in medical and biomedical image domains [14], [15]. On the other, the effort involved in the annotation process of a new biomedical dataset can be drastically reduced with transfer learning and domain adaptation [7].

However, in a biomedical context, these processes are particularly challenging due to several variations regarding the image modalities: i) object type (organs/tissues/cells,...), ii) color, iii) illumination, and iv) aspect ratio/spatial resolution. Hence, due to the difficulty of effectively performing transfer learning and/or domain adaptation, most of medical imaging datasets often remain unused [16].

In that regard, transfer learning or domain adaptation can benefit of a quantitative measurement of dataset similarity. In this way, by selecting the most similar dataset from the state of the art for a new biomedical image processing problem, the knowledge learned from the source domain can be optimally transferred to target domain. However, the literature on the similarity among image datasets is scarce despite this topic being a long-standing challenge in image processing. Current proposals about dataset similarity either use the associated text metadata as a Natural Language Processing (NLP) problem [17]; require labeled datasets to compute the similarity measurement, with the corresponding human effort devoted to annotation; or even rely on fixed pre-trained networks (e.g., trained on real-world datasets such as ImageNet [18]) for feature extraction, which do not consider the particularities of biomedical image data. In addition, state-of-the-art approaches focus on classification tasks [19], [20], whereas segmentation, which needs more human effort to generate the annotations, is not considered. In contrast to dataset similarity, image similarity has been thoroughly studied for Content-Based Image Retrieval (CBIR) tasks in real-world image scenarios and many metrics have been proposed [21]–[23]. Nevertheless, the proposed metrics have yet to be extensively proven to be useful in biomedical imaging.

In this paper, we propose an Automated Style-Aware similarity measurement for selection of Pre-training datasets (ASAP) in 2D biomedical imaging. To the best of our knowledge, this is the first self-supervised and image-based learning system predicting a similarity measurement among biomedical imaging datasets. ASAP is an extension of our previous work in [24], and their results have been demonstrated to improve dataset selection with respect to the state of the art in a semantic segmentation task.

The main contributions of our work are the following:

- We introduce a workflow to build a 2D latent-space embedding whose disposition is related to the similarity of biomedical imaging datasets, and correlated with the transfer learning performance in a supervised training task.
- We introduce a reliable, robust, and interpretable measurement for quantifying dataset similarity over our embedding, the Embedding Quality Criterion (EQC).
- Our proposed model, an autoencoder (AE) based on content and style features [25] trained with self-supervised contrastive learning [26], unveils that style features are a powerful source of similarity in the biomedical image field.
- The gathered knowledge enables identification of the closest-matching annotated biomedical dataset to a new one. This improves transfer learning and domain adaptation, even using a small subset of unlabeled images from the new dataset (our approach does not need labeled datasets).

The remainder of this paper is organized as follows: Section II briefly reviews the related work. In Section III, we provide a general description of our method for obtaining dataset similarity. Thereby, we offer a more detailed description of the style-aware contrastive learning system, delving into the formulation of our similarity measurement for the sake of clarity. Section IV describes and discusses the experimental results that support our method. Finally, we summarize our conclusions and outline future lines of research in Section V.

II. RELATED WORK

We present the state-of-the-art methods related to our approach in the next subsections. First, an overview of the similarity measurement approaches is provided, including image and dataset similarity. Second, a review of the existent approaches for self-supervised and contrastive learning is presented (our approach learns from the datasets' labels through these techniques). In addition, an analysis of the methods regarding domain adaptation and style transfer is performed (our approach uses the style of the images to measure the similarity among datasets).

IMAGE AND DATASET SIMILARITY

Image similarity measurement is an open problem in image processing, traditionally associated with image retrieval, e.g., in the case of search engines. Hence, the existing research mainly focuses on real-world images. There are different approaches to measure image similarity: global and local similarity measurements based on gray-level and color distributions [21], spectral- and residual-based measurements [22], perceptual measurements consistent with the ratings given by humans [23], or content-based similarity for CBIR tasks [27], together with deep learning approaches traditionally based on autoencoders [28] and Siamese networks [29]. A more recent field of study, associated with machine learning research, is dataset similarity (in this case, not necessarily restricted to image). In that case, the similitude among datasets is evaluated, for example, for speech emotion recognition [30] or text biomedical datasets [31].

Regarding image datasets, the literature is rather limited. Alvarez-Melis and Fusi propose OTDD, a distance measurement among datasets based on optimal transport, which is model-agnostic and does not involve training [19]. Developed for supervised learning, it relies on label occurrence for each dataset to their feature vectors in a latent space (obtained as a fixed-size representation through a Neural Network). The authors obtained promising results for both real-world and text classification datasets. Cheplygina et al. [32] quantified the similarity of medical imaging datasets by representing each one of them by performances of simple classifiers. This representation had an 89.3% accuracy in predicting the origin of each dataset, but they do not quantitatively analyze the similarity among different datasets. Godau and Maier-Hein introduce in [20] the concept of task fingerprinting, to capture domain- (image modality) and task- (medical instruments, pathology, etc.) similarities among biomedical image datasets, based on the work by Achille et al. [33]. Task fingerprinting uses supervised learning or a fixed feature extractor (e.g. a ResNet-50 [34] pre-trained in ImageNet) to convert a task, represented by image data and optionally their corresponding labels, to a fixed-length vector representation which can be compared with those of other tasks. Some metrics, e.g., based on the distributions of the features for each task, are used to quantify the task similarity.

To sum up, most state-of-the-art methods use labels or a fixed feature extractor to estimate the similarity among datasets, instead of proposing specific architectures that models the similarity among them. Moreover, the investigations are mainly done regarding classification tasks.

SELF-SUPERVISED LEARNING (SSL)

SSL is an emerging research area, which has gained attention in the most recent years [35]. In particular, Contrastive Learning (CL), such as SimCLR [36] and BarlowTwins [37], are widely used. The fundamental idea involves defining pretext tasks which can provide pseudo annotations. For example, augmentating the same sample via rotation or cropping, and training a network with the different versions of the image. This learning task must ensure that the original image and the modified one are encoded as closely as possible in the latent space given by an AE. However, the CL methods mentioned above often require more computational effort or the hyperparameters, e.g., to determine appropriate data augmentations, are difficult to set. In addition, the main drawback with traditional AEs is that they primarily concentrate on image reconstruction from the latent-space features, not leveraging all the potential of the low-dimensional representation.

DOMAIN ADAPTATION AND STYLE TRANSFER

Domain adaptation and style transfer are optimization techniques that aim to transform the data from a source domain in order to adopt the features (e.g., visual or acoustic) of a target domain. When the samples are represented in a latent space, this process manifests itself in a position shift: the target samples move from their original location to the source domain area, as demonstrated in [38]. Style transfer and domain adaptation have been widely employed in natural-scene scenarios in the literature: day-to-light image translation, segmentation of vegetation, or text style transfer [39], with promising results. However, in the case of medical imaging, the lack of labeled data and their heterogeneity (different scanners, scanning parameters, subject cohorts, etc.) reduce domain adaptation performance [40]. In addition, the existent domain adaptation methods predominantly concentrate on modifying the samples within the new dataset to enhance performance for a specific task, rather than attaining an interpretable representation in the latent space. The approach by Hou et al. [41], which better aligns with our work, performs a domain adaptation between a source domain that solely provides a pre-trained model without any source of data, and a target domain that contributes unlabeled data. In particular, they achieve domain adaptation by transforming the images from the target domain to align with the batch-wise feature statistics stored in the batch normalization layers of the pre-trained model.

This paper proposes an automated pipeline for predicting the similarity values of new datasets with respect to known annotated datasets, inspired by some of these approaches from the state of the art. Our scenario is an example of heterogeneous transfer learning, in the sense that the target and source domains are potentially different (in terms of image modality, color, resolution, shape of objects, etc.). We propose to use the style features developed by Gatys et al. [25], based on the Gram matrix of content features, as a source of similarity among medical imaging datasets. The resulting 2D latent similarity measurement is interpretable and correlated with the transfer learning results in a supervised task (in our case, binary semantic segmentation, but extensible to other applications).

III. MATERIALS AND METHODS

In this section, we first provide a description of the datasets used for this research. After that, we provide general description of our system, and, subsequently, a detailed explanation of its main module (the style-aware contrastive learning one) and the performance measurements developed for evaluating the embedding quality follow.

A. DATASETS

For this research, we consider a vast set of heterogeneous open-source datasets trying to represent the most-used 2D image modalities in a context of binary semantic segmentation. Table 1 summarizes the training and test dataset pools. We have selected 51 images per each one of the 9 known datasets to avoid unbalance during training, and a test set with 12 additional datasets to evaluate our approach. It is essential to note that, in this paper, we have used annotated datasets in test for the evaluation of our method; nonetheless, our approach does not need labels at the test stage.

Figure 2 shows some illustrative samples for specific training and test datasets. Differences among them are notable regarding image modality and style/aspect.

B. OVERVIEW OF ASAP

Figure 1 shows an overview of our pipeline. In our scenario, an annotated dataset $\mathcal{D}_i^{\mathrm{a}} = (\mathcal{X}_i, \mathcal{Y}_i)$ consists the image set $\mathcal{X}_i = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^{N_i}\}$ and the label set $\mathcal{Y}_i = \{y^1, y^2, ..., y^{N_i}\}$,with N_i images and N_i labels, respectively. The annotated dataset pool is composed of $I \in \mathbb{N}$ datasets $\mathcal{D}_1^{\mathrm{a}}, \ldots, \mathcal{D}_I^{\mathrm{a}}$ utilized to train a similarity estimator in an self-supervised manner. It is worth noting that each one of the datasets should have a similar number of samples to prevent class imbalance, as a requirement for the proper training of

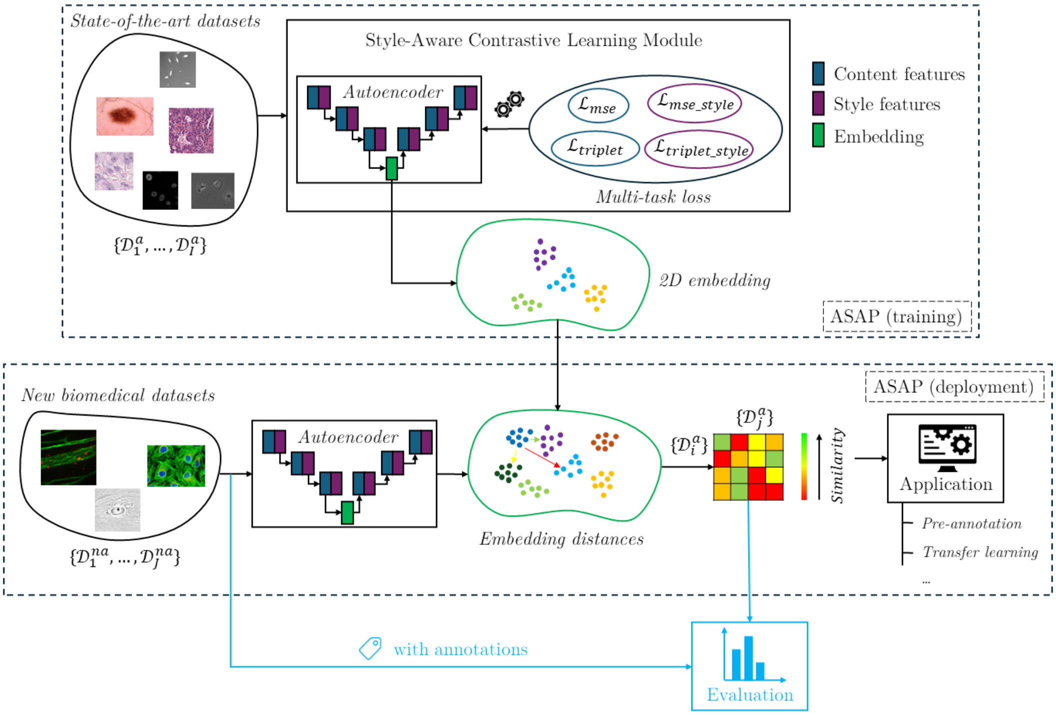

FIGURE 1: Diagram of the proposed method. In the training stage, the annotated dataset pool $\mathcal{D}_1^{\mathrm{a}}, \ldots, \mathcal{D}_I^{\mathrm{a}}$ is used to train the similarity estimator (Style Aware Contrastive Learning Module) in a self-supervised manner, with a multi-task loss which combines reconstruction, through the Mean Squared Error (MSE), and contrastive learning through a triplet loss, from both content and style features. In the deployment stage, this component, when trained, embeds the new datasets $\mathcal{D}_1^{\mathrm{na}}, \ldots, \mathcal{D}_J^{\mathrm{na}}$ in the 2D latent space and provides their similarity w.r.t. the state-of-the-art datasets. Hence, ASAP can select the most similar dataset for either pre-training or pre-labeling an unseen dataset. Only for evaluation purposes (shown in blue), in this article, we assess the quality of the similarity estimator by comparing its ability to select the optimal dataset from the state of the art with a supervised task (e.g., binary semantic segmentation) [24].

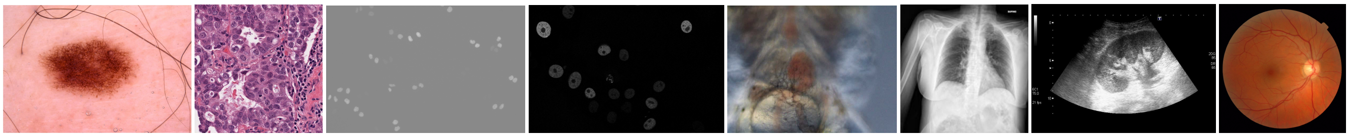

FIGURE 2: Illustrative examples of different datasets (training and test) used in our study. From left to right: ISIC2017 (dermatology), MonuSeg (histology), Fluo-N2DL-HeLa and Fluo-N2DH-GOWT1 (microscopy), heartSeg (RGB images), LUNG (X-ray), URI-CADS (ultrasound) and RIGA (fundus images). As observed, there is a significant variability in the aspect of the images and sometimes the imaging modality is not related to the aspect of the images (microscopy).

our system. In any case, if this requirement cannot be fulfilled with a sufficient number of samples, data augmentation techniques can be used to increase the number of samples of the underrepresented datasets.

Furthermore, a not annotated dataset, denoted by $\mathcal{D}_j^{\mathrm{na}} = (\mathcal{X}_j)$, comprises solely a set of images (without labels). During the testing phase, some new and unlabeled datasets $\mathcal{D}_1^{\mathrm{na}}, \ldots, \mathcal{D}_J^{\mathrm{na}}$, where $J \in \mathbb{N}$ represents denotes the total number of new datasets, are encoded into the latent space using the already trained similarity estimator. The most similar (closest) known dataset to a given new dataset $\mathcal{D}_j^{\mathrm{na}}$ is then determined by evaluating the embedding distances (with the embedding describing the similarity among the datasets in an abstract feature space). At this point, the pre-trained model (in the closest known dataset) can be leveraged: 1) without adaptations, such as for pre-labeling the new dataset [7]; 2) as the initial model to perform transfer learning.

Besides this task, and only for evaluation purposes, the known annotated datasets $\mathcal{D}_1^{\mathrm{a}}, \ldots, \mathcal{D}_I^{\mathrm{a}}$ are used to train an actual image processing task (e.g., segmentation) via supervised learning. In our particular case, the segmentation CNN is trained on the state-of-the-art datasets (annotated) but used for interpreting unlabeled datasets in test. Hence, annotations are required to evaluate the CNN results, but not to determine the most similar training dataset. The achieved results in the supervised task can be compared to the similarity values of

TABLE 1: Train and test dataset pools including modality.*HAM10000 and ISIC2017 do not present overlap images [42].

Set	Identifier	Name	Modality
Train	$\mathcal{D}^{\text{a}}_1$	HAM10000* [43]	Dermatology, skin lesions
	$\mathcal{D}^{\text{a}}_2$	MoNuSeg [44]	Histology
	$\mathcal{D}^{\text{a}}_3$	PanNuKe [45]	
	$\mathcal{D}^{\text{a}}_4$	Fluo-N2DL-HeLa [46]	Fluorescence microscopy
	$\mathcal{D}^{\text{a}}_5$	Cellpose [47]	Assorted microscopic images
	$\mathcal{D}^{\text{a}}_6$	PhC-C2DH-U373 [46]	Phase-Contrast microscopy
	$\mathcal{D}^{\text{a}}_7$	URI-CADS [48]	Ultrasound
	$\mathcal{D}^{\text{a}}_8$	heartSeg [49]	RGB cardiac images
	$\mathcal{D}^{\text{a}}_9$	DENTAL [50]	X-Ray (dental)
Test	$\mathcal{D}^{\text{na}}_1$	ISIC2017* [9]	Dermatology, skin lesions
	$\mathcal{D}^{\text{na}}_2$	CryoNuSeg [51]	Histology
	$\mathcal{D}^{\text{na}}_3$	Fluo-N2DH-GOWT1 [46]	Fluorescence microscopy
	$\mathcal{D}^{\text{na}}_4$	Fluo-C2DL-Huh7 [46]	
	$\mathcal{D}^{\text{na}}_5$	PhC-C2DL-PSC [46]	Phase-Contrast microscopy
	$\mathcal{D}^{\text{na}}_6$	LIVECell [52]	
	$\mathcal{D}^{\text{na}}_7$	DIC-C2DH-HeLa [46]	Difference-Interference Contrast microscopy
	$\mathcal{D}^{\text{na}}_8$	LUMINOUS [53]	Ultrasound
	$\mathcal{D}^{\text{na}}_9$	USNerve [54]	
	$\mathcal{D}^{\text{na}}_{10}$	RIGA [55]	Fundus images
	$\mathcal{D}^{\text{na}}_{11}$	CVC-ClinicDB [56]	RGB colonoscopy images
	$\mathcal{D}^{\text{na}}_{12}$	LUNG [57]	X-Ray (lung)

our system (the self-supervised task), as a metric to evaluate the performance of our method.

C. STYLE-AWARE CONTRASTIVE LEARNING MODULE

Figure 3 shows a detailed diagram of our developed architecture for the 2D representation of the datasets. The objective is to obtain a latent representation $\mathbf{z}$, for each given arbitrary image $\mathbf{x}$, that represents the aspect/style of the image $\mathbf{x}$ and whose distances to the other latent representations relate to their mutual similarity. To do so, we adopt an encoder-decoder CNN ($P \in \mathbb{N}$) built on a generic backbone. The backbone is composed of several Convolutional Blocks (CB), i.e., sub-networks of the architecture, each one devoted to provide deep features at a different depth and resolution. We denote the content feature maps provided by each CB as $\mathbf{C}_p \in \mathbb{R}^{H_p \times W_p \times K_p}, p = 1, \dots, P$, being P the total number of CBs (or equivalently, considered feature maps), for the downsampling path of the network (encoder). Each feature map has a spatial resolution $H_p = H/2^{(p+1)}; W_p = W/2^{(p+1)}$ defined by the accumulated spatial stride of the sub-network until its corresponding layer (e.g., stride = 4, 8, 16 and 32 for $P = 4$). The number of channels is denoted by K_p (e.g., 64, 128, 256 and 512 for $P = 4$). Analogously, we denote $\hat{\mathbf{C}}_p \in \mathbb{R}^{H_p \times W_p \times K_p}$ as the content features for the upsampling path (decoder).

The corresponding vectorized content feature maps are denoted by: $\mathbf{F}_p = \mathbf{C}_p \rightarrow \mathbb{R}_{H_p \cdot W_p \times K_p}$ and $\hat{\mathbf{F}}_p = \hat{\mathbf{C}}_p \rightarrow \mathbb{R}_{H_p \cdot W_p \times K_p}$, and are then used to obtain the encoder style features $\mathbf{S}_p \in \mathbb{R}^{K_p \times K_p}$ and the decoder style features $\hat{\mathbf{S}}_p \in \mathbb{R}^{K_p \times K_p}$, through the Gram matrix computed over every level of the architecture. Each element of the matrices $S_p^{kk'}$ and $\hat{S}_p^{kk'}$, with $k, k' \in 1, \dots, K_p$, is the inner product between the channels k and k' of the vectorized content feature maps $\mathbf{F}_p$ and $\hat{\mathbf{F}}_p$ in the form

$$S_p^{kk'} = \sum_{l}^{H_p \cdot W_p} F_p^{lk} F_p^{lk'}, \tag{1}$$

$$\hat{S}_p^{kk'} = \sum_{l}^{H_p \cdot W_p} \hat{F}_p^{lk} \hat{F}_p^{lk'}. \tag{2}$$

The encoder's output is then transformed into the latent embedding $\mathbf{z} \in \mathbb{R}^{N_{\text{emb}}}$, where $N_{\text{emb}} \in \mathbb{N}$ refers to the dimension of the embedding (cf. Figure 3). The reconstruction $\hat{\mathbf{x}}(\mathbf{x})$ follows after the decoder path.

The designed style-aware contrastive learning system combines a set of objective loss functions that receive as inputs: the input image and its features, computed style features, dataset association of the image (as self-supervision, using the dataset membership labels y), and the reconstructed image.

First, the reconstruction loss

$$\mathcal{L}_{\text{mse}} = (\hat{\mathbf{x}}(\mathbf{x}) - \mathbf{x})^2 \tag{3}$$

trains the AE to extract the pivotal features of the image $\mathbf{x}$ that better summarize it and hold significance for the reconstruction of the image. In addition, this loss has an analogous version for the style features:

$$\mathcal{L}_{\text{mse-style}} = \sum_{p=1}^{P} w_p (\mathbf{S}_p - \hat{\mathbf{S}}_p)^2, \tag{4}$$

with the hyperparameters $\mathbf{w} = \{w_p\}$, $p \in \{1, \dots, P\}$, $w_p \in \mathbb{R}^+$ (subject to $\sum_p w_p = 1$) are factors that weigh the contribution of each style layer to the loss. Defining the style enhances the AE's ability to bring similar datasets closer within the latent space while simultaneously pulling apart those that are dissimilar.

In the same way, according to the standard formulation of [26], the triplet loss follows:

$$\mathcal{L}_{\text{triplet}} = \max\left(||\mathbf{z} - \mathbf{z}^+||^2 - ||\mathbf{z} - \mathbf{z}^-||^2 + \gamma, 0\right), \tag{5}$$

where $\mathbf{z}^+$ is a positive anchor (a sample coming from the same dataset as the embedding $\mathbf{z}$, $y = y^+$), $\mathbf{z}^-$ a negative anchor (a sample belonging to a different dataset as the embedding $\mathbf{z}$, $y \neq y^+$); and $\gamma \in \mathbb{R}^+$ is the margin between positive and negative pairs (for this research, we use $\gamma = 1$). By combining the triplet loss with this self-supervised learning strategy, the AE achieves to separate the samples of different datasets within the latent space. It does so by maximizing the difference of the Euclidean distances between: 1) the positive anchor and $\mathbf{z}$, and 2) the negative anchor and $\mathbf{z}$. Therefore, in the latent space, it forces the input-image latent representation to be close to the positive and far from the negative anchor.

The style triplet loss is calculated in a similar way:

$$\mathcal{L}_{\text{triplet-style}} = \sum_{p=1}^{P} w_p \max\left(||\mathbf{S}_p - \mathbf{S}_p^+||^2 - ||\mathbf{S}_p - \mathbf{S}_p^-||^2 + \gamma, 0\right). \tag{6}$$

Here, $\mathbf{S}_p^+(\mathbf{x})$ and $\mathbf{S}_p^-(\mathbf{x})$ represent the encoder style features corresponding to the positive and negative anchor in

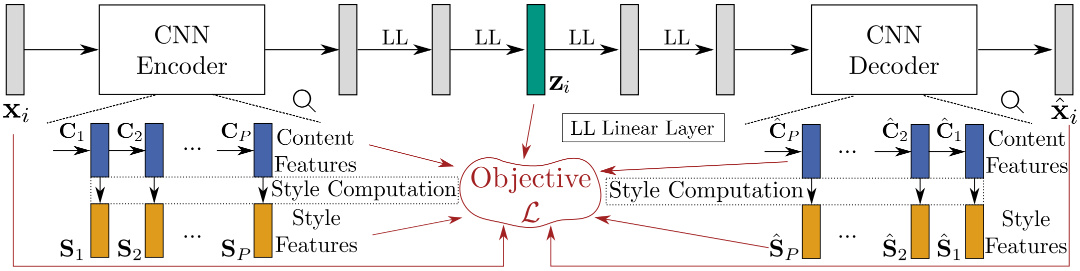

FIGURE 3: Detailed diagram of our Style-aware Contrastive Learning Module for similarity estimation. We propose to use an AE, built over a generic backbone with P convolutional blocks and adapted to provide both content ($\mathbf{C}_p/\hat{\mathbf{C}}_p$) and style ($\mathbf{S}_p/\hat{\mathbf{S}}_p$) features. The AE receives and processes the input image and then, two linear layers (LL) generate the latent embedding $\mathbf{z}_i$. The same, but inverted structure, provides the reconstruction. The objective loss function considers image reconstruction, and content/style features.

the p-th layer and $\gamma = 1$. In this way, we also enforce our latent representation to capture the dissimilar styles/aspects of the databases and translate these differences into datasets' distances.

The marginal objective loss functions form a combined loss function:

$$\begin{aligned}\mathcal{L} =& \alpha\left(\beta\mathcal{L}_{\text{mse}} + (1-\beta)\mathcal{L}_{\text{mse-style}}\right) \\ &+ (1-\alpha)\left(\beta\mathcal{L}_{\text{triplet}} + (1-\beta)\mathcal{L}_{\text{triplet-style}}\right),\end{aligned} \tag{7}$$

with hyperparameters $\alpha \in [0,1]$ and $\beta \in [0,1]$. The first one, α, balances the effects of the reconstruction and triplet losses; the second one, β, weighs the influence of each stage of the AE in the style loss.

Finally, it is noteworthy that our similarity estimator performs in an end-to-end basis and is trained with a trio of images in each iteration: the one under study, a positive anchor (belonging to the same dataset), and a negative anchor (belonging to a different dataset). In addition, Gram matrices, despite being computationally less efficient than more recent approaches for style computation based on instance normalization [58], provide explicit correlations between feature maps that better capture global style information and tend to preserve overall style characteristics throughout the resulting image [59]. Hence, they are especially appropriate for the purpose of summarizing the global style of images in a few components. We will demonstrate in Section IV-B that their computational cost is negligible for our particular case.

D. PERFORMANCE MEASUREMENTS

In this section, we provide some measurements to evaluate the quality of the proposed embedding and the performance of selecting the best available annotated dataset.

1) Quality of the Embedding

The quality of an embedding is related to some properties that will be desirable when obtaining a latent space representation for a specific task: e.g., a good quality embedding for classification must discriminate among samples from different classes, but a good quality embedding for dataset selection must depict the datasets in different and compact clusters where similar datasets for a human's perspective are grouped close to each other. Our particular case falls into the second category. In this case, a meaningful embedding should have the following properties:

1) *Coherence* - grouping the samples of the same dataset in compact areas of the latent space.
2) *Semantic similarity* - placing the samples of similar datasets, i.e., same image modalities, in proximity within the latent space, while simultaneously pulling apart the samples from highly dissimilar datasets.

We propose a single measurement, drawn upon the Ward criterion [60] and denoted as the Embedding Quality Criterion (EQC), to assess the quality of embeddings (see Eq. 8). This metric evaluates the latent space representations of the state-of-the-art annotated datasets (denoted by $\mathbf{z}_{a,i}$ for the i-th dataset) by averaging the ratios between: 1) the sum of distances among samples within each database of a pair (in the numerator of Eq. 8), and 2) the distances among samples of the pair of databases (in the denominator of Eq. 8), for all the pairs. Hence, it follows

$$\text{EQC} = \frac{2}{(I^2-I)} \sum_{i=1}^{I} \sum_{k=i+1}^{I} \underbrace{\frac{\xi+\chi}{\psi}}_{\text{EQC}_{ik}=\text{EQC}_{ki}}, \tag{8}$$

where the dataset assignment is denoted by y. The terms ξ and χ denote the intra-database distances and the term ψ the inter-database distance, being:

$$\xi = \frac{2}{N_i(N_i-1)} \sum_{l=1,m=l+1}^{N_i} \|\mathbf{z}_{a,i}^{l} - \mathbf{z}_{a,i}^{m}\|, \tag{9}$$

$$\chi = \frac{2}{N_k(N_k-1)} \sum_{n=1,o=n+1}^{N_k} \|\mathbf{z}_{a,k}^{n} - \mathbf{z}_{a,k}^{o}\|, \tag{10}$$

$$\psi = \frac{1}{N_i N_k} \sum_{l=1}^{N_i} \sum_{n=1}^{N_k} \|\mathbf{z}_{a,i}^{l} - \mathbf{z}_{a,k}^{n}\|. \tag{11}$$

We use the EQC as a measurement to select the top-performing configuration for our model (optimal α, β and $\mathbf{w}$

hyperparameters), robust across different experiments. Conversely, the raw embedding distances can vary across experiments due to our non-normalized embedding vectors, so they are not robust for hyperparameter selection. The metric EQC spans from 0 to $\inf$, exhibiting commutativity between two datasets *i*/*k* ($\mathrm{EQC}_{ik} = \mathrm{EQC}_{ki}$), and excluding the elements along the matrix diagonal. A lower EQC_{ik} (preferred) indicates that datasets i and k form compact clusters in the latent space (with a greater distance of samples between the datasets compared to the distance of samples within the single datasets).

Note that the EQC is similar to the Davies-Bouldin criterion [61] for cluster separation, but it does not consider the worst case for every cluster (the closest cluster). In our scenario, there can coexist similar datasets (from the same image modality and style/aspect), and their placement in the same area of the latent space should not largely penalize the evaluation measurement.

Finally, our EQC measurement considers each dataset as a collection of individual images, so it can be adapted to image-to-dataset similarity if we consider single test images instead of complete datasets. This property is particularly useful in the case of heterogeneous test datasets.

2) Performance on Dataset Selection

Additionally, we can evaluate the degree of adjustment between the dataset distances in our embedding and the pre-trained model results in a supervised task. For this purpose, we propose the correlation between the results of the test datasets $\mathcal{D}_j^{\mathrm{na}}$ with the pre-trained models in each one of the training datasets $\mathcal{D}_i^{\mathrm{a}}$, in terms of the Dice-Sørensen Coefficient (DSC) for image segmentation; and the embedding distances calculated per combination $\mathcal{D}_i^{\mathrm{na}}/\mathcal{D}_j^{\mathrm{a}}$ as the first performance measurement. We denote the pre-training results and embedding distances between a train dataset i and a test dataset j as DSC_j^i and ED_j^i, respectively.

If the correlation is positive, the embedding disposition can be interpreted as a dataset similarity measurement. In this case, the embedding distances among objective/known datasets are related to the effectiveness of transfer learning or pre-annotation for test in the objective datasets after pre-training in the known datasets. Besides, we can employ two useful measurements for retrieval tasks, such as Rank Correlation Index (RCI) [62] and top-K accuracy, given that our objective is to retrieve the best-performing training dataset for a test dataset. RCI evaluates the monotony in the relationship between the rankings of DSC_j^i and $-\mathrm{ED}_j^i$ (note that we consider the negative distances, i.e., the proximity), i.e., if the best pre-training datasets for a given test dataset j are the closest ones in the embedding, and the worst datasets are the furthest ones. Top-K accuracy considers the k closest datasets in the latent space (smaller ED_j^i) and analyzes if the best pre-training dataset is inside this set.

In our scenario, if similar results are obtained in the test stage from certain training datasets, it can be penalized by top-K accuracy measurement if the selected dataset is not the best. However, from our application's perspective, this situation is not harmful. We propose to include the loss in segmentation performance between the selected dataset and the optimal dataset (denoted by $*$) as a performance measurement, in terms of the DSC:

$$\Delta_{\mathrm{DSC}} = \sum_{j=1}^{J} \mathrm{DSC}_j^* - \mathrm{DSC}_j^{\mathrm{selected}} \qquad (12)$$

for all J test datasets. Note, if our system selects the best-performing dataset, it yields $\mathrm{DSC}_j^* - \mathrm{DSC}_j^{\mathrm{selected}} = 0$. In any other case, its value will be positive, and smaller values indicate better segmentation performance (lower loss with respect to pre-train with the optimal dataset).

IV. EXPERIMENTS AND DISCUSSION

A. EXPERIMENTAL SETUP

To implement our approach, we use: 1) an AE based on a lightweight ResNet18 backbone [34] with $P = 4$ feature maps for the embedding calculation (replaceable for other backbones), and 2) an U-Net [63] with Dice objective function for the supervised segmentation task, both developed and deployed using PyTorch. The raw data is min-max normalized, scaling it to the range [0,1]. For the training process, we adopt a random split ratio of 80% for training and 20% for validation. An exhaustive data augmentation policy (flipping, random crop & resize, brightness/contrast adjustment, rotate/shift/scale, Gaussian noise/blur) is performed to capture all the variability for the training datasets. We determine the hyperparameters through random search. Our CNN training involves the use of Adam optimizer with early stopping and learning rate scheduling, and is performed on multiple cluster nodes equipped with Intel Xeon Platinum 8368 CPU and four NVIDIA A100 Tensor Core GPU.

B. EXPERIMENTAL RESULTS

We compare different ablated versions of our proposed approach with some of the most relevant approaches from the state of the art, and additionally we demonstrate its superior performance with respect to a group of human experts performing the same task. The description of these approaches is given below:

1) **CC** and **NMI** - Two classical approaches for measuring image similarity, namely cross-correlation (CC) and normalized mutual information (NMI), between the images from the training and test datasets [21].
2) **FRC** - A spectral-based measurement, the Fourier Ring Correlation (FRC) [22].
3) **OTDD** - The best-performing supervised approach from the state of the art [19]. This approach obtains a latent-space representation of the datasets through a pre-trained ResNet-18 CNN. Then, it defines an Optimal Transport problem to compute the cost of matching the representations of each pair of datasets (distributions), which is solved through the Wasserstein distance between these representations. Although in the original

TABLE 2: EQC for different sizes of the latent space.

N_{emb}	1	2	3	4	5
EQC	0.591	**0.384**	0.406	0.412	0.388

version of the paper they use classification datasets and consider a multi-modal distribution formulation (image classes), we assume an uni-modal distribution (our scenario does not provide object classes).

4) **KLD-TF** - The best-performing non-supervised approach from the state of the art [20], consisting of obtaining the fingerprint of the different datasets with a pre-trained ResNet-34 CNN and a similarity measurement based on Kullback-Leibler divergence.
5) **SL+PCA+UMAP** - A direct approach for measuring task similarity, composed of a dimensionality reduction technique applied over the final feature layer of a ResNet-18 network pre-trained in ImageNet. In this case, PCA [64] is used in conjunction with UMAP [65], due to the high dimensionality reduction of the output layer of ResNet-18 (512 channels). Hence, PCA first selects the number of features that explains the 95% of the input variance, and then UMAP is applied to the reduced set of features to obtain the 2D embedding.
6) **Group of experts** - 7 experts from different disciplines in close relation with biomedical imaging (biomedical engineers, biological and biomedical sciences' researchers and medical doctors) have been selected to perform the most similar dataset selection.
7) **Best expert** - the best expert' value for each performance measurement is also shown for comparison. Please note that these values may (and indeed) come from different experts.
8) **Ours-AE** - Our autoencoder trained only to reconstruct the samples ($\mathcal{L}_{\text{mse}}$).
9) **Ours-AE-T** - Our autoencoder trained in a self-supervised manner with dataset labels ($\mathcal{L}_{\text{mse}}$ + $\mathcal{L}_{\text{triplet}}$).
10) **Ours-AE-S** - Our autoencoder trained in the style-aware manner proposed in this paper ($\mathcal{L}_{\text{mse}}$ + $\mathcal{L}_{\text{mse-style}}$).
11) **ASAP** - Our complete proposal, trained in a style-aware and self-supervised manner ($\mathcal{L}_{\text{mse}}$ + $\mathcal{L}_{\text{mse-style}}$ + $\mathcal{L}_{\text{triplet}}$ + $\mathcal{L}_{\text{triplet-style}}$).

1) Hyperparameter Validation

We validate the parameters of our method N_{emb}, α, β and w_p, $p \in \{1, \ldots, 4\}$, regarding our non-supervised measurement, the EQC. First, we fix the embedding dimension to $N_{\text{emb}} = 2$ because the EQC does not vary significantly with this parameter when considering a latent-space dimension above 1D, and it allows for visualization purposes without a subsequent dimensionality reduction algorithm (see Table 2). The reconstructions are not accurate with such reduced latent-space dimension, but as we will demonstrate after, this does not affect the performance of our embedding when representing new samples.

Second, we validate the remaining parameters of our method in terms of the EQC. The results are shown in Figure 4. The optimum value for the α parameter is $\alpha = 0.5$. The best configuration of our approach equally balances the triplet and reconstruction parts of our loss. The optimal value $\beta = 0.6$ also indicates that the contribution of the style and content layers to our approach is similar (slightly higher for the content layers). Finally, the weight values $\mathbf{w} = [0.2, 0.2, 0.3, 0.3]$ indicate the need of both the shallowest style layers (capturing low-level features such as texture) and the deepest layers (encoding illumination, color, and high-level features such as shape/number of objects, more in consonance with the segmentation task) for creating a meaningful latent representation of datasets. However, there is a slight preference for the deepest layers. Additionally, the curves demonstrate that around the optimal hyperparameter values remains relatively stable. This fact endorses the robustness of our method: small variations in the hyperparameters yield embeddings with consistent properties.

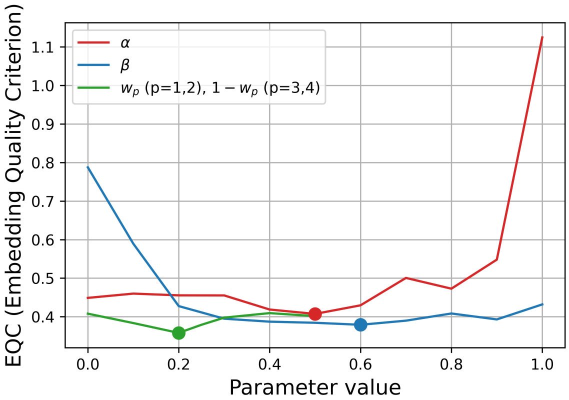

FIGURE 4: Hyperparameter validation based on EQC. Curves for the weighting parameters w.r.t. the loss function (α and β) and weights of style layers w_p are shown for $N_{\text{emb}} = 2$.

2) Ablation Study and Comparison with the State of the Art

Table 3 shows the performance measurements for both the different ablated versions of our approach with the optimal hyperparameters and the described state-of-the-art approaches. Our style-aware proposal, trained using the triplet loss, demonstrates superior performance compared to other methods, in terms of EQC and key information retrieval metrics like top-1 accuracy, top-2 accuracy, and Δ_{DSC}. Some conclusions can be extracted from the results:

- First, our approach outperforms manual selection (the group of experts and the best of them) for each measurement. By analyzing the experts' errors, we can conclude that they mainly perform an attribute matching over the datasets (selecting the one in the training set that has the same image modality or task as the test dataset). This choice based on experience is reasonable for some datasets with the same image modality, but our approach transcends this attribute matching and also considers the style and aspect of the images, even in more difficult

TABLE 3: Performance measurements (direction of the optimal value of the metric is denoted by ↓/↑) for the different versions of the proposed system with optimal hyperparameters ($\mathbf{w} = [0.2, 0.2, 0.3, 0.3]$, $\alpha = 0.5$, $\beta = 0.6$, and $N_{\text{emb}} = 2$) and the state-of-the-art. The best result is marked in bold.

Approach	**EQC** ↓	**Correlation** ↑	**RCI** ↑	**Top-1 acc. (%)** ↑	**Top-2 acc. (%)** ↑	**Top-3 acc. (%)** ↑	$\mathbf{\Delta_{DSC}}$ **(%)** ↓	**# of GFLOPS** ↓
CC [21]	–	**0.360**	**0.321**	33.33	41.67	50.00	21.29	$1.843 \cdot 10^{-3}$
NMI [21]	–	0.190	0.092	16.67	25.00	66.67	25.94	$3.744 \cdot 10^{-3}$
FRC [22]	–	0.094	0.110	00.00	41.67	50.00	27.46	$8.410 \cdot 10^{-2}$
OTDD [19]	–	0.314	0.244	16.67	41.67	0.5	20.05	1.129
KLD-TF [20]	–	0.048	0.128	33.33	66.67	**91.67**	18.43	3.661
SL+PCA+UMAP	0.365	0.131	0.178	41.67	50.00	50.00	–	2.510
Group of experts	–	–	–	36.91	61.91	79.76	15.26	–
Best expert	–	–	–	50.00	75.00	91.67	10.67	–
Ours-AE	1.125	−0.143	−0.048	33.33	50.00	75.00	21.87	0.148
Ours-AE-T	0.449	0.095	0.143	50.00	75.00	75.00	13.61	0.148
Ours-AE-S	0.788	−0.098	−0.017	33.33	58.33	66.67	26.26	0.156
ASAP	**0.358**	0.148	0.185	**75.00**	**83.33**	83.33	**3.89**	0.156

cases.

- Second, the correlation and RCI are positive for our approach, indicating the correspondence between the segmentation results and the distances in the embedding. In particular, our proposal successfully identifies the optimal pre-training dataset for 9 out of 12 datasets, with a performance loss compared to the oracle system of approximately 4% (in terms of DSC).
- Third, some other approaches, such as cross-correlation and OTDD, achieve reasonable correlation with the pre-training results (better than ours), but their accuracy in identifying the best dataset is low. The other approaches, especially the KLD-TF approach [20], obtain acceptable results for top-K accuracy. Still, their errors are less coherent than in the case of our approach, so the DSC loss is high (around 20%).
- Fourth, our approach improves the DSC coefficient of the pre-annotations by a margin of 15% in Dice. According to different approaches from the state of the art, the impact of these results is remarkable. They can reduce the annotation time by a margin of at least 50% [66], [67]; or notably improve the detection accuracy of subsequent algorithms, e.g., for cell detection [2], [68], [69], by a margin of 3-30%, depending on the pre-annotation Dice coefficient for each dataset.
- Lastly, it is shown that both the contrastive learning ablated version (Ours-AE-T) and style one (Ours-AE-S) improve the baseline autoencoder. The combination of both in our proposed system improves the consistency of the embedding, especially at the finer resolution, increasing the top1-accuracy.

In addition, Table 3 shows the computational complexity of all compared methods. Our method has higher computational costs than traditional approaches (CC, NMI and FRC, the ones offering the worst results), but improves computational efficiency with respect to the rest of deep-learning approaches (due to the use of a smaller image size and a lightweight neural network). It is remarkable that the style computation with the Gram matrix does not imply a significant increase in the computational burden of our method.

Finally, we have performed a study of robustness to noise of ASAP vs. the state-of-the-art approaches shown in Table 3. We have compared the performance measurements across datasets with different proportions of noisy labels (from 5% to 50%). This study is presented in a Supplementary Data file due to its extension, and demonstrates that our approach, despite being slightly more sensitive to noise than the approaches from state-of-the-art (because of the need for training with the noisy samples), still performs better than the compared methods for every noise level.

3) Analysis of our Approach

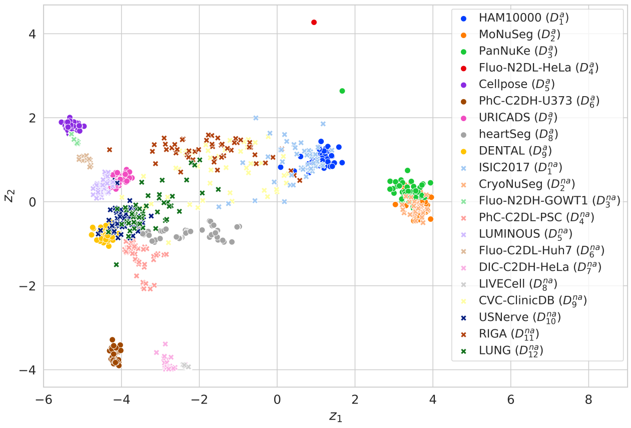

FIGURE 5: 2D train/test latent-space embedding for our approach. Training datasets are denoted with 'o' markers and test datasets with different pastel colors and 'x' markers to facilitate distance comparison visually. Note that most of the datasets (train/test) yield to compact portrayals (especially Fluo-N2DL-HeLa). Distances in the embedding are directly related with dataset similarity.

Figure 5 represents the latent space of the train/test datasets. Different datasets are placed in distinct areas of the latent space. Furthermore, similar datasets, i.e., PhC-C2DH-373, LIVECell, and DIC-C2DL-HeLa, with the same image modality and/or aspect, are close in the latent space (see the bottom left corner area). Furthermore, the test datasets occupy specific areas in the latent space, except for RIGA and CVC-ClinicDB (very dissimilar to the training datasets), whose samples are spread in the central area of the embedding.

When dealing with datasets from the same image modality (e.g., CryoNuSeg, a histology one), our method identifies the corresponding datasets in the training set (MoNuSeg, PanNuke, the histology ones) as the most suitable ones for pre-training. However, our proposal goes beyond a modal matching between the train and test datasets. For example, PhC-C2DL-PSC dataset (a phase-contrast microscopy one) is not assigned to PhC-C2DH-U373 (with the same image modality), but it is assigned to DENTAL, an X-ray image dataset. In some difficult cases, as the CVC-ClinicDB dataset, a modal matching approach could select HAM10000 as the dataset for pre-training (both have RGB images and roundish objects), but our approach selects the best one, heartSeg, a less evident choice.

In addition, Table 4 presents the embedding distances in combination with the DSC coefficients obtained for the test datasets (pre-labeling case). The results are accurate for most cases and the system selects the best pre-training dataset. Nevertheless, for $\mathcal{D}_2^{\text{na}}$ (CryoNuSeg), a histology dataset, the system identifies MoNuSeg as the best dataset for pre-training instead of PanNuKe. However, both pre-training datasets offer a similar performance. For $\mathcal{D}_{10}^{\text{na}}$ (RIGA), the system has difficulties with determining the best pre-training dataset, but all of them offer similar and discrete results. Finally, in the case of $\mathcal{D}_{12}^{\text{na}}$ (LUNG), the system opts for selecting a dataset with the same image modality (X-Ray), but the best results are provided by a dermatology dataset (HAM10000), which has a completely different aspect.

The efficiency in dataset selection can also be noticed in the typical pre-labeling examples shown in Figure 6. For most of our imaging datasets, the pre-annotations are a good starting point to refine them through a manual labeling process. However, annotations are not accurate enough for USNerve, CVC-ClinicDB, or LUNG datasets. In these cases, especially for USNerve and CVC-ClinicDB, the annotation process requires significant expert knowledge, due to the fact that the most of the objects (nerves and tumors, respectively) have a similar aspect to their background and are barely distinguishable for a non-expert annotator. Hence, a pre-labeling system is not supportive during image annotation in these cases.

On the other hand, we obtain the representations of some unseen training datasets images to evaluate the coherence of our system representations. We have included two RGB real-image datasets, VOC2012 [70] and COCO [71], as $\mathcal{D}_{13}^{*\text{na}}$ and $\mathcal{D}_{14}^{*\text{na}}$, to evaluate if there is some correspondence between real-world and biomedical image domains. Figure 7 shows the representations of these new samples over our embedding. First, we can see that our exhaustive data augmentation policy is enough to capture the variability of the problem (allowing us to avoid more complex architectures, such as variational strategies). Second, we can observe that the real-world images from VOC2012 and COCO do not locate in coherent areas of the embedding, even not in the RGB image area, so we conclude that biomedical and real-world image domains present inherent differences that hinder a direct application of our algorithm to real-world domains.

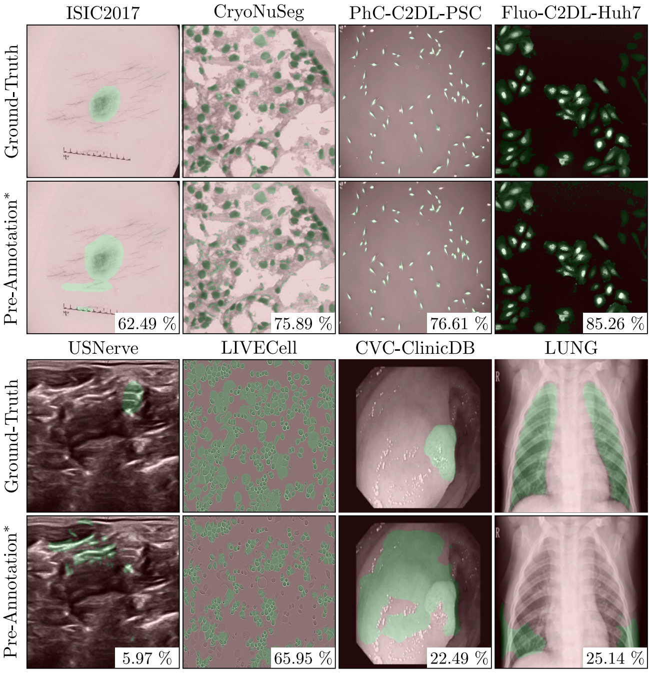

FIGURE 6: Pre-annotation results for different datasets with DSC coefficient. The results of the selected pre-training models, obtained using ASAP, are shown in comparison to the ground truth annotation for all test datasets. The DSC score and overlays (green: foreground, red: background) are represented.

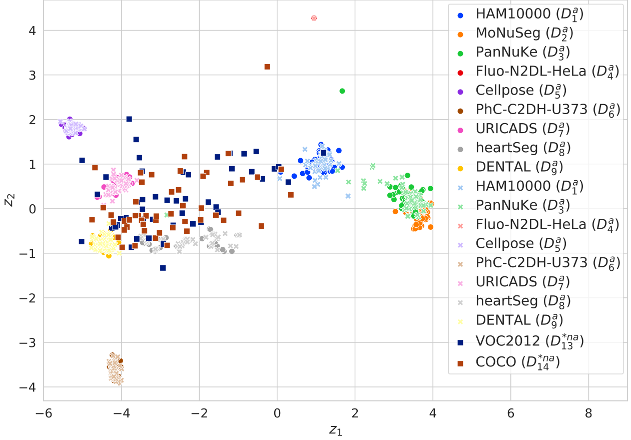

FIGURE 7: Coherence study. Representations of unseen images from the training datasets, with ’x’ markers, and some real-world image datasets, VOC2012 and COCO, with square marker and dark colors. Dataset disposition in the latent space does not change with the introduction of new samples from the training datasets. Real-image datasets offer very spread portrayals.

TABLE 4: Latent-space distances for our approach vs. DSC (in %) using pre-training (training datasets) for the test datasets. The selected dataset combination and most similar dataset combination is marked in bold. Except for $\mathcal{D}_{12}^{\text{na}}$ (LUNG), the rest of the errors are made in difficult cases ($\mathcal{D}_{10}^{\text{na}}$, RIGA), or do not suppose a high loss in accuracy ($\mathcal{D}_{2}^{\text{na}}$, CryoNuSeg). In addition, for $\mathcal{D}_{9}^{\text{na}}$ (USNerve), $\mathcal{D}_{10}^{\text{na}}$ (RIGA) and $\mathcal{D}_{11}^{\text{na}}$ (CVC-ClinicDB), with very discrete segmentation results, the pre-labeling process is not useful.

Train		$\boldsymbol{\mathcal{D}_1^{\text{a}}}$	$\boldsymbol{\mathcal{D}_2^{\text{a}}}$	$\boldsymbol{\mathcal{D}_3^{\text{a}}}$	$\boldsymbol{\mathcal{D}_4^{\text{a}}}$	$\boldsymbol{\mathcal{D}_5^{\text{a}}}$	$\boldsymbol{\mathcal{D}_6^{\text{a}}}$	$\boldsymbol{\mathcal{D}_7^{\text{a}}}$	$\boldsymbol{\mathcal{D}_8^{\text{a}}}$	$\boldsymbol{\mathcal{D}_9^{\text{a}}}$
Test	$\boldsymbol{\mathcal{D}_1^{\text{na}}}$	**0.80 / 62.77**	3.35 / 5.66	3.00 / 10.87	3.40 / 25.15	5.89 / 25.43	6.59 / 47.44	4.61 / 1.33	3.36 / 37.42	5.24 / 3.08
	$\boldsymbol{\mathcal{D}_2^{\text{na}}}$	2.75 / 53.94	**0.33 / 69.21**	**0.57 / 74.07**	5.16 / 47.13	9.08 / 19.95	8.49 / 51.61	7.66 / 52.48	5.92 / 33.48	8.01 / 0.41
	$\boldsymbol{\mathcal{D}_3^{\text{na}}}$	6.37 / 11.43	9.04 / 10.75	8.70 / 10.78	6.76 / 62.58	**0.31 / 67.37**	5.30 / 11.32	1.57 / 17.27	3.74 / 0.08	2.45 / 38.53
	$\boldsymbol{\mathcal{D}_4^{\text{na}}}$	6.03 / 39.30	8.65 / 35.82	8.32 / 36.94	6.71 / 70.19	**0.90 / 74.58**	4.71 / 38.97	1.01 / 54.91	3.17 / 4.09	1.85 / 37.67
	$\boldsymbol{\mathcal{D}_5^{\text{na}}}$	5.23 / 1.62	7.33 / 1.64	7.13 / 2.78	7.17 / 13.21	3.55 / 13.30	2.46 / 1.23	1.89 / 15.84	1.47 / 0.06	**1.03 / 63.55**
	$\boldsymbol{\mathcal{D}_6^{\text{na}}}$	6.03 / 0.11	7.08 / 13.02	7.07 / 56.93	8.82 / 52.40	6.47 / 52.25	**1.89 / 61.22**	4.78 / 27.35	3.27 / 0.02	3.80 / 28.05
	$\boldsymbol{\mathcal{D}_7^{\text{na}}}$	6.30 / 0.71	7.48 / 67.94	7.45 / 68.20	8.99 / 67.05	6.25 / 67.24	**1.42 / 83.79**	4.61 / 32.40	3.29 / 0.02	3.54 / 0.98
	$\boldsymbol{\mathcal{D}_8^{\text{na}}}$	5.64 / 12.96	7.08 / 13.01	7.86 / 12.74	6.70 / 23.40	1.68 / 21.37	4.02 / 13.08	**0.51 / 52.13**	2.47 / 8.00	1.14 / 0.41
	$\boldsymbol{\mathcal{D}_9^{\text{na}}}$	5.28 / 3.41	7.62 / 2.19	7.37 / 2.26	6.80 / 2.63	2.63 / 2.84	3.24 / 2.94	1.02 / 3.45	1.71 / 0.17	**0.65 / 4.97**
	$\boldsymbol{\mathcal{D}_{10}^{\text{na}}}$	2.79 / 0.89	5.47 / 0.00	5.12 / 0.00	4.10 / 1.30	3.70 / 1.26	5.56 / 0.91	2.55 / 0.04	**2.40 / 1.58**	**3.48 / 12.31**
	$\boldsymbol{\mathcal{D}_{11}^{\text{na}}}$	2.65 / 15.37	5.22 / 1.40	4.90 / 1.56	4.34 / 19.82	4.04 / 16.19	5.25 / 15.33	2.69 / 0.17	**2.11 / 22.96**	3.38 / 7.76
	$\boldsymbol{\mathcal{D}_{12}^{\text{na}}}$	**4.57 / 59.15**	6.96 / 47.41	6.69 / 45.73	6.15 / 37.27	2.91 / 34.89	3.62 / 2.79	**1.23 / 27.97**	1.42 / 0.03	1.33 / 0.55

Finally, we present a 2D latent-space representation for a system trained with the complete collection of datasets (training and test) presented in this study, in Figure 8. In this way, we can qualitatively analyze whether their disposition in the 2D space is related to the datasets' aspect and modality. As observed, the embedding disposition is coherent and consistent. The RGB datasets are placed at the top and the gray-level ones are at the bottom. Likewise, we can identify some areas with datasets belonging to the same image modality (see the dermatology and histology clusters at the upper-right corner of Figure 8), and others that group datasets from different image modalities but with the same style/aspect (see the colonoscopy/fundus images area at the upper-left corner, or the bottom one, combining ultrasound and microscopy). Therefore, this embedding could be used in the deployment stage of the system to select the best pre-training dataset for a new one.

V. CONCLUSIONS AND FUTURE WORK

This paper presents ASAP, an automated style-aware system for selecting pre-training datasets in 2D biomedical imaging. The system aims to speed up the deployment process of a new biomedical application, especially in the early stages, by selecting the corresponding dataset that offers the best results for pre-labeling or pre-training.

In that sense, our approach, based on an autoencoder relying on content and style features and self-supervised contrastive learning, can: 1) obtain a 2D latent-space embedding of training datasets, whose disposition is related to the similarity of the datasets; 2) determine the best-performing embedding with the help of a performance measurement, the embedding quality criterion, that evaluates the relative disposition of the datasets in the latent space; and 3) select the best dataset to pre-train a model among the training datasets according to the distances in the embedding, demonstrating a positive correlation between this similarity measurement, as represented in the latent space, and the test results.

Our experiments demonstrate the interpretability of our 2D embedding and its correspondence with transfer learning performance across the training and test datasets. Moreover, we have unveiled style features as a crucial source of similarity, enhancing embedding quality, and aligning dataset disposition with the results of transfer learning. Consequently, our approach holds promise for potential applications, serving as a pre-annotation framework for new datasets or helping to select optimal models for transfer learning in medical/biomedical imaging. In addition, our method can be also interpreted as an image-to-dataset similarity measurement, which can identify samples in a known dataset that are not conducive to transfer learning or domain adaptation.

The main research directions for further work include, first, the deployment of a more comprehensive latent space with additional 2D datasets, than can be enlarged with the release of new state-of-the-art datasets. In addition, a study and improvement of the scalability w.r.t. the amount of data is part of the future work. ASAP needs to compute the latent space every time a new dataset is considered. We contemplate two different strategies to mitigate this shortcoming: 1) class-incremental learning [72], to incorporate the knowledge of the new datasets incrementally, reducing the computational cost without compromising the learned information from previous datasets; and 2) few-shot learning [73], to further reduce the number of training images per class to a few examples accelerating the deployment of updated ASAP models. We also plan to give answer to the questions about when the model can benefit from pre-training with multiple datasets. Moreover, we plan to extend the work to 3D image modalities (such as CT and MRI), either using 3D architectures or analyzing the 3D volumes as a z-stack (or with the central slice).

ACKNOWLEDGMENT

This paper is an extended and revised version of a preliminary conference article that was presented and awarded the Best Paper Award at ISBI 2023 [24]. We thank the anonymous reviewers from the conference and this journal for their meaningful insights about the improvement of this work. We also kindly thank the group of experts for their inestimable help in the experimental part of this article.

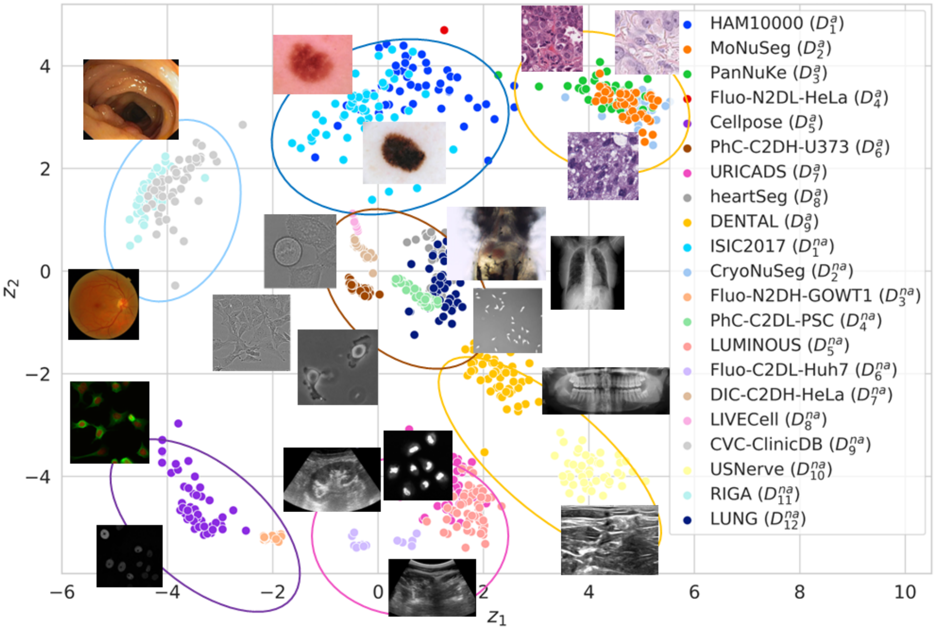

FIGURE 8: Deployment of ASAP: complete 2D latent-space embedding, trained with the complete collection of training and test datasets. In elliptical shapes of different colors, embedding areas grouping different datasets regarding its modality, aspect, color, etc. We include a representative sample from every dataset next to its representation for the purpose of similarity visualization.

REFERENCES

[1] A. Doerr, N. Vogt, and L. Tang, “Deep learning gets scope time,” *Nature Methods*, vol. 16, no. 12, p. 1195, 2019.

[2] J. C. Caicedo, A. Goodman, K. W. Karhohs, *et al.*, “Nucleus segmentation across imaging experiments: The 2018 Data Science Bowl,” *Nature Methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[3] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[4] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen, “Deep learning for cellular image analysis,” *Nature Methods*, vol. 16, pp. 1233–1246, Dec. 2019.

[5] F. Ahmed, S. Abbas, A. Athar, *et al.*, “Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence,” *Scientific Reports*, vol. 14, no. 6173, 2024.

[6] F. Ahmed, A. Fatima, M. Mamoon, and S. Khan in *2024 2nd International Conference on Cyber Resilience (ICCR)*, pp. 1–6, 2024.

[7] M. P. Schilling, S. Schmelzer, L. Klinger, and M. Reischl, “KaIDA: a modular tool for assisting image annotation in deep learning,” *Journal of Integrative Bioinformatics*, 2022.

[8] L. Zhou, N. Heller, and Y. Shi, *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*. Lecture Notes in Computer Science, Springer Nature, 2019.

[9] N. C. F. Codella, D. Gutman, M. E. Celebi, *et al.*, “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 *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, pp. 168–172, 2018.

[10] B. H. Menze, A. Jakab, S. Bauer, *et al.*, “The multimodal brain tumor image segmentation benchmark (BRATS),” *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[11] Z. Meng, Z. Zhao, B. Li, F. Su, and L. Guo, “A cervical histopathology dataset for computer aided diagnosis of precancerous lesions,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1531–1541, 2021.

[12] R. Bommasani, D. A. Hudson, E. Adeli, *et al.*, “On the opportunities and risks of foundation models.” http://arxiv.org/abs/2108.07258, 2021.

[13] A. Kirillov, E. Mintun, N. Ravi, *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

[14] M. Moor, O. Banerjee, Z. Abad, *et al.*, “Foundation models for generalist medical artificial intelligence,” *Nature*, vol. 616, pp. 259–265, 2023.

[15] M. J. Cardoso, W. Li, R. Brown, *et al.*, “MONAI: An open-source framework for deep learning in healthcare,” 2022.

[16] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big Data*, vol. 3, no. 9, 2016.

[17] D. Bernhauer, M. Nečasiký, P. Škoda, J. Klímek, and T. Skopal, “Open dataset discovery using context-enhanced similarity search,” *Knowledge and Information Systems*, vol. 64, 2022.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems (NIPS)* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), vol. 25, pp. 1097–1105, 2012.

[19] D. Alvarez-Melis and N. Fusi, “Geometric dataset distances via optimal transport,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS’20, 2020.

[20] P. Godau and L. Maier-Hein, “Task fingerprinting for meta learning in biomedical image analysis,” in *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 436–446, Springer International Publishing, 2021.

[21] H. B. Mitchell, *Image Similarity Measures*. Image Fusion: Theories, Techniques and Applications, Springer Nature, 2010.

[22] M. Van Heel, “Similarity measures between images,” *Ultramicroscopy*, vol. 21, no. 1, pp. 95–100, 1987.

[23] T. Frese, C. A. Bouman, and J. P. Allebach, “Methodology for designing image similarity metrics based on human visual system models,” in *Proceedings of the SPIE*, vol. 3016, pp. 472–483, 1997.

[24] M. Molina-Moreno, M. P. Schilling, M. Reischl, and R. Mikut, “Automated style-aware selection of annotated pre-training databases in biomedical imaging,” in *Proceedings of the IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2023.

[25] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

[26] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, “Large scale online learn-

ing of image similarity through ranking," *Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.

[27] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704, 2022.

[28] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChe Journal*, vol. 37, pp. 233–243, 1991.

[29] A. Malhotra, "Single-shot image recognition using Siamese Neural Networks," in *Proceedings of the 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 2550–2553, 2023.

[30] I. Siegert, R. Böck, and A. Wendemuth, "Using a PCA-based dataset similarity measure to improve cross-corpus emotion recognition," *Computer Speech & Language*, vol. 51, pp. 1–23, 2018.

[31] X. Wang, Z. Huang, and F. van Harmelen, "Evaluating similarity measures for dataset search," in *Proceedings of the Web Information Systems Engineering (WISE)*, pp. 38–51, 2020.

[32] V. Cheplygina, P. Moeskops, M. Veta, B. D. Bozorg, and J. Pluim, "Exploring the similarity of medical imaging classification problems," in *Proceedings of the CVII-STENT@MICCAI*, pp. 59–66, 2017.

[33] A. Achille, M. Lam, R. Tewari, A. Ravichandran, S. Maji, C. C. Fowlkes, S. Soatto, and P. Perona, "Task2Vec: Task embedding for meta-learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

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

[35] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, 2021.

[36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, pp. 1597–1607, 2020.

[37] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow Twins: Self-supervised learning via redundancy reduction," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320, 2021.

[38] M. Schutera, M. Hussein, J. Abhau, R. Mikut, and M. Reischl, "Night-to-day: Online image-to-image translation for object detection within autonomous driving by night," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 3, pp. 480–489, 2021.

[39] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," in *Advances in Data Science and Information Engineering*, pp. 877–894, 2021.

[40] S. Kumari and P. Singh, "Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives," *Computers in Biology and Medicine*, vol. 170, 2024.

[41] Y. Hou and L. Zheng, "Source free domain adaptation with image translation," *ArXiv*, vol. abs/2008.07514, 2020.

[42] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Medical Image Analysis*, vol. 75, 2022.

[43] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, 2018.

[44] N. Kumar, R. Verma, D. Anand, *et al.*, "A multi-organ nucleus segmentation challenge," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1380–1391, 2020.

[45] J. Gamper, N. A. Koohbanani, K. Benet, A. Khuram, and N. M. Rajpoot, "PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification," in *Proceedings of the European Congress on Digital Pathology*, pp. 11–19, 2019.

[46] V. Ulman, M. Maška, K. E. G. Magnusson, *et al.*, "An objective comparison of cell-tracking algorithms," *Nature Methods*, vol. 15, pp. 1141–1152, 2017.

[47] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, pp. 100–106, 2021.

[48] M. Molina-Moreno, I. González-Díaz, M. Rivera Gorrín, V. Burguera Vion, and F. Díaz-de María, "URI-CADS: A fully automated computer-aided diagnosis system for ultrasound renal imaging," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 4, pp. 1458–1474, 2024.

[49] M. Schutera, L. Rettenberger, C. Pylatiuk, and M. Reischl, "Methods for the frugal labeler: Multi-class semantic segmentation on heterogeneous labels," *PLOS ONE*, vol. 17, no. 2, pp. 1–14, 2022.

[50] A. H. Abdi, S. Kasaei, and M. Mehdizadeh, "Automatic segmentation of mandible in panoramic X-ray," *Journal of Medical Imaging*, vol. 2, no. 4, 2015.

[51] A. Mahbod, G. Schaefer, B. Bancher, C. Löw, G. Dorffner, R. Ecker, and I. Ellinger, "CryoNuSeg: A dataset for nuclei instance segmentation of cryosectioned H&E-stained histological images," *Computers in Biology and Medicine*, vol. 132, 2021.

[52] C. Edlund, T. R. Jackson, N. Khalid, N. Bevan, T. Dale, A. Dengel, S. Ahmed, J. Trygg, and R. Sjögren, "LIVECell-a large-scale dataset for label-free live cell segmentation," *Nature Methods*, vol. 18, pp. 1038–1045, 2021.

[53] C. J. Belasso, B. Behboodi, H. Benali, M. Boily, H. Rivaz, and M. Fortin, "LUMINOUS database: lumbar multifidus muscle segmentation from ultrasound images," *BMC Musculoskeletal Disorders*, vol. 21, pp. 703–727, 2020.

[54] H. Health, "Ultrasound BP nerve segmentation dataset." https://www.kaggle.com/competitions/ultrasound-nerve-segmentation/data, 2016.

[55] A. Almazroa, S. Alodhayb, E. Osman, *et al.*, "Retinal fundus images for glaucoma analysis: the RIGA dataset," in *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, vol. 10579 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 2018.

[56] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[57] D. Kermany, K. Zhang, and M. Goldbaum, "Large dataset of labeled optical coherence tomography (OCT) and chest X-Ray images." https://data.mendeley.com/datasets/rscbjbr9sj/3, 2018.

[58] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1510–1519, 2017.

[59] Y. Zhang, Y. Tian, and J. Hou, "CSAST: Content self-supervised and style contrastive learning for arbitrary style transfer," *Neural Networks*, vol. 164, pp. 146–155, 2023.

[60] H. Joe and J. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[61] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[62] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[63] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

[64] I. Jolliffe, *Principal Component Analysis*. Springer Verlag, 1986.

[65] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *The Journal of Open Source Software*, vol. 3, no. 29, pp. 861–862, 2018.

[66] M. Pachitariu and C. Stringer, "Cellpose 2.0: how to train your own model," *Nature Methods*, vol. 19, pp. 1634–1641, 2022.

[67] R. Hollandi, A. Diósdi, G. Hollandi, N. Moshkov, and P. Horváth, "AnnotatorJ: an ImageJ plugin to ease hand annotation of cellular compartments," *Molecular Biology of the Cell*, vol. 31, pp. 2179–2186, 2020.

[68] A. Zargari, G. Lodewijk, N. Mashhadi, *et al.*, "DeepSea: An efficient deep learning model for single-cell segmentation and tracking of time-lapse microscopy images," *Cell Reports Methods*, vol. 3, no. 6, 2023.

[69] P. Shrestha, N. Kuang, and J. Yu, "Efficient end-to-end learning for cell segmentation with machine generated weak annotations," *Communications Biology*, vol. 6, no. 232, 2023.

[70] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results." http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[71] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference in Computer Vision (ECCV)* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), pp. 740–755, 2014.

[72] G. M. Van de Ven, T. Tuytelaars, and A. S. Tolias, "Three types of incremental learning," *Nature Machine Intelligence*, vol. 4, pp. 1185–1197, 2022.

[73] J. Lu, P. Gong, J. Ye, J. Zhang, and C. Zhang, "A survey on machine learning from few samples," *Pattern Recognition*, vol. 139, p. 109480, 2023.

MIGUEL MOLINA-MORENO was born in Rus, Jaén, Spain, in 1993. He received the Ph. D. degree in Multimedia and Communications from Universidad Carlos III de Madrid, Madrid, Spain, in 2023, and since then he has worked as a Postdoctoral Associate in the Department of Immunobiology at Yale University. He received the Best Paper Award at the IEEE 20th International Symposium on Biomedical Imaging (ISBI 2023). His research interests focus on machine learning, deep learning and computer vision for biological/biomedical/medical tasks.

MARCEL P. SCHILLING received the bachelor's and master's degrees in mechanical engineering from the Karlsruhe Institute of Technology, Karlsruhe, Germany, 2018 and 2020, and his Ph.D. degree from the same institution in 2023. After his period in the Research Group Machine Learning for High-Throughput and Mechatronics, Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology; he is now a Machine Learning Scientist at ZEISS. His research interests include image processing, data-centric deep learning, and machine learning for high-throughput screening

MARKUS REISCHL received the Dipl.-Ing. and Ph.D. degrees in mechanical engineering from the University of Karlsruhe, Karlsruhe, Germany, in 2001 and 2006, respectively. Since 2020, he is an Adjunct Professor with the Faculty of Mechanical Engineering and is heading the research group Machine Learning for High-Throughput and Mechatronics with the Institute for Automation and Applied Computer Science, Karlsruhe Institute of Technology, Karlsruhe, Germany. His research interests include man–machine interfaces, image processing, machine learning, and data analytics.

RALF MIKUT received the Diploma degree in automatic control from the University of Technology, Dresden, Germany, in 1994, and the Ph.D. and Habilitation degrees in mechanical engineering from the University of Karlsruhe, Karlsruhe, Germany, in 1999 and 2007, respectively. Since 2011, he has been an Adjunct Professor at the Faculty of Mechanical Engineering and the Head of the Research Group Automated Image and Data Analysis, Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology (KIT), Germany. His current research interests include machine learning, image processing, life science applications, and smart grids.

• • •