Interactive Ground-Truth-Free Image Selection for FLIM Segmentation Encoders

Matheus A. Cerqueira*, Flávia Sprenger†, Bernardo C. A. Teixeira†,
Silvio Jamil F. Guimarães‡, and Alexandre X. Falcão*
*Institute of Computing, University of Campinas, Campinas, São Paulo, Brazil
†Hospital de Clínicas, Universidade Federal do Paraná, Curitiba, Paraná, Brazil
‡Computer Science Department, Pontifical Catholic University of Minas Gerais, Belo Horizonte, Brazil

Abstract*—Deep Learning has shown impressive results in computer vision tasks at the cost of becoming increasingly extensive and requiring a voluminous amount of labeled images. Feature Learning from Image Markers (FLIM) is a method to build convolutional encoders with minimal user effort in image annotation. FLIM estimates the encoder's filters directly from markers drawn by the user (scribbles, clicks) in significant regions of a few representative images (*e.g.*, 9). However, the selection of those few images significantly impacts the model's performance, and image selection is mainly addressed by either the user inspecting all images and subjectively selecting the most relevant ones or by using a method that requires pixel-wise annotation (ground-truth) for recommending images. The main goal of our work is to make this task less subjective by presenting a ground-truth-free method for recommending representative images to train FLIM encoders while guiding the user's choice with a visual explanation. Our method suggests new images based on the distance to the previously selected images, using a proposed descriptor that generates different patterns according to filter attention. The user selects the next image among the suggested ones, which repeats until the user's satisfaction. According to our experiments, the proposed method can outperform expert manual selection and an interactive method based on pixel-wise annotation. These results are demonstrated for brain tumor and gastrointestinal parasite segmentation.

I. Introduction

Deep learning has boosted the automation of various activities using images, such as object detection, instance segmentation, and image classification. However, deep models usually rely on large annotated datasets and have become more complex, either by increasing the number of layers or by using different models together, demanding more powerful hardware, intensive human effort in data annotation, and energy consumption [1], [2].

Feature Learning from Image Markers (FLIM) is a method that estimates the filters of convolutional encoders from user-drawn markers (scribbles, clicks) on a few representative images (1-9). The filters of all convolutional layers are learned directly by clustering from patches (*i.e.*, filter candidates) centralized at marker pixels [3]. FLIM has been used to train shallow encoders with remarkable performance in multiple tasks, but such performance is related to selecting images with distinct appearances for training [4]. In [4]–[6], this selection is done manually, with the user inspecting the training set for the most diverse images, a subjective and time-consuming method. An example of variability in object appearance for interactive image selection can be seen in Figure 1 for brain tumor and gastrointestinal parasite (Schisto) segmentation.

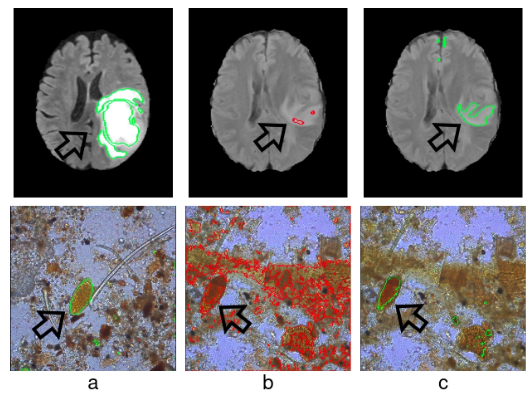


Fig. 1. Example of variability in object appearance for image selection and interactive selection for (top) brain tumor and (bottom) Schisto segmentation. In (a) the first selected images and their feature map, (b) new images whose objects are not adequately captured from the feature map of (a). Finally, (c) shows the new images and their feature maps after using FLIM.

Typically, there is a recommendation function, and users only annotate images as an oracle annotator without exploring their knowledge for the learning process [7], [8]. Other techniques embrace user knowledge in image selection rather than only annotating selected images. For example, in [9], users properly select and label samples based on visual projections. Also, in [10], the authors show that those interactions could be propagated to other samples according to the projection, improving their results. For FLIM networks, recent work has proposed an interactive method for selecting images based on already learned filters and image ground-truth (GT) [11]. However, the method only partially exploits the user's knowledge since the user also selects images solemnly on a metric computed by a *recommendation function* and then draws markers on them. Furthermore, they used a recommendation function that requires a complete annotation, even if it is not used for the encoder learning.

This work proposes a novel interactive image selection method for training FLIM encoders without needing ground-truth (GT) to recommend images. Our method relies on the

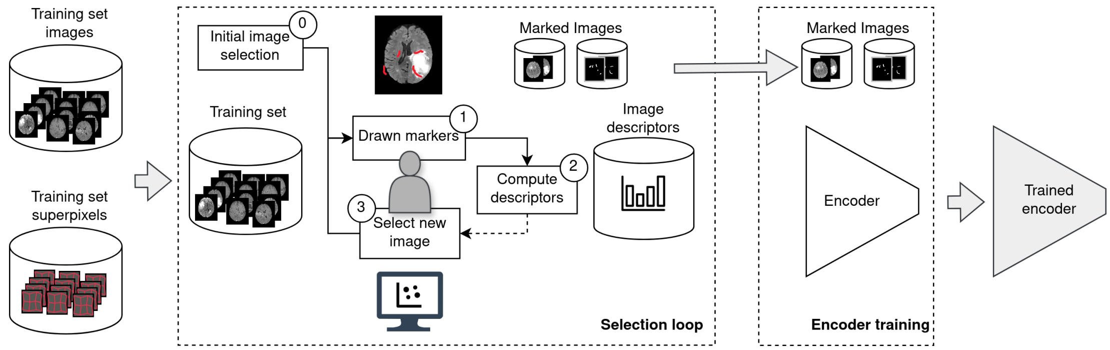


Fig. 2. Our interactive image selection method for encoder training.

proposed image descriptor and a visual interface that guides user decisions. The proposed image descriptor generates different patterns according to the attention of feature maps, so we recommend images using the distance of descriptors from already selected images. Moreover, the user selects a new image based on the pool of recommended ones, guided by a 2D projection and visual explanations. We compared our method with (i) subjective manual selection, (ii) a similar approach that relies on pixel-wise annotation for sampling, and (iii) using all training data with backpropagation. We have applied the proposed method in two biomedical image tasks: a 3D brain multi-label tumor segmentation and 2D gastrointestinal parasite binary segmentation. The proposed method with the new descriptor outperforms the compared state-of-the-art (based on ground-truth) and manual selection for these applications.

II. Related Works

A. Human-in-the-loop image selection

Recent works have provided more freedom for the user to select images for annotation from a pool of recommendations. In [12], the user selects images according to erroneous classification, taking into account the contextual information that led to such result – *e.g.*, whether there was extensive sky in the image or if there is a shadow on the object to be detected. In [13], the authors used a recommendation function based on attention regions of the network so the user can select images based on such visual explanations. In [14], the authors also used visual explanations to guide image selection, but the user actions will be penalized directly as a loss function regularization term.

Despite progress in providing more user control over the learning process, such methods are still strongly linked to a model-centric view, with recommendation functions typical of active learning methods that suffer from the initialization (cold start) problem, for instance. In [15], the authors attempt to solve these problems by adapting an extensive network to a new data set (domain adaptation) based on user clicks in different regions, thus providing more user control. However, the method focuses on the user's correcting erroneous annotations in small regions of an immense remote-sensing image.

B. Feature Learning from Image Markers (FLIM)

FLIM enables shallow convolutional encoders (1-3 layers) from fewer weakly annotated images (1-9) while maintaining its competitive performance compared to standard deep models [3], [5], [6]. FLIM starts with (i) selecting a few images, then (ii) the user draws markers on significant regions of a few selected images, patches are extracted from all marked pixels for normalization and clustering, and (iii) one filter candidate is derived from the center of each cluster. Finally, (iv) the process continues for the entire encoder from the marker onto the previous layer's output – without backpropagation.

However, the selection of a few representative images is challenging. Most works, like [4]–[6], use manual selection with the user inspecting the whole training pool and selecting the most diverse images – a time-consuming and subjective task. One work used raw image projection, and the user manually selected images from this projection to mark without considering what was learned [3]. A recent study, presented in [11], showed that image selection could be made by using a loop that samples images according to their performance according to the ground-truth (GT). The work presents promising results, but the pipeline is limited because it requires GTs, which requires the availability of a professional to annotate all images, and those GTs are used only for image selection, not using it for the rest of the encoder training.

III. Interactive Ground-Truth-Free Image Selection

A. Pipeline

As illustrated in Figure 2, our method for encoder training is separated into two main parts: the **selection loop** and **encoder training**. The **selection loop** starts by sampling a random image (0). In the following, the user draws markers on it and employs FLIM to learn convolutional filters from those markers (1). We compute image descriptors for all training images based on new convolutional filters (2), and the user selects a new image according to the distance from previously selected images using our descriptor and with guidance from a visualization system (3). The newly selected image is added to the marked image set, restarting another cycle.

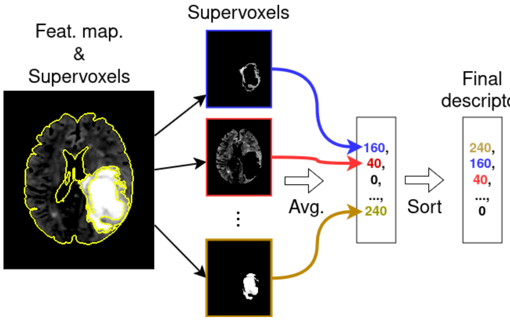


Fig. 3. Ilustration of our image descriptor computation.

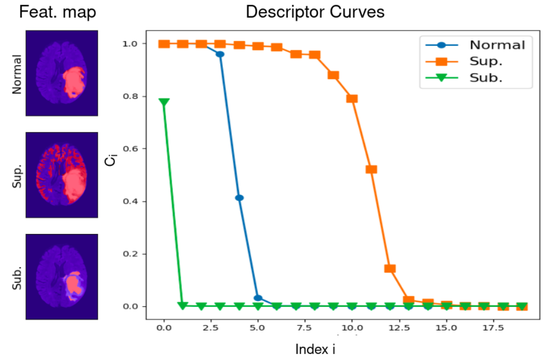


Fig. 4. Heatmaps of Feature maps that activate differently for the image/object and the corresponding descriptor curves c_i for each i.

Once we have a small set of relevant and diverse images (5 in this work), we start the **encoder training**, in which we train a proper multi-layer FLIM encoder with marked images.

B. Proposed Descriptor

Descriptors are computed based on filters that detect objects (tagged by the user) and previously computed superpixels (or supervoxels for 3D images) without pixel-wise annotation. Given an image, a previously computed superpixel, and the feature map, the *image descriptor* is defined as the average activation of each superpixel region in the feature map.

Let Image I be defined as a pair $\langle D, \mathbf{I} \rangle$, $\mathbf{I}(q) \in \mathbb{R}^m$ mapping each pixel (voxel for 3D) element $q \in D \subset \mathbb{Z}^n$ given $m \in \mathbb{N}^*$, $d \in \mathbb{N}^*, d \geq 2$, being m the image bands and d its dimensions. Also, $\mathcal{S} = \{S_1, S_2, ..S_{N_s}\}$ is a superpixel image with N_s superpixels, with each superpixel $S_i \subseteq D$ as a set of image pixels. Also, $F = \langle D, \mathbf{F} \rangle$ is a feature map that highlight the object, with $\mathbf{F}(q) \in \mathbb{R}$. Therefore the image descriptor for a given filter F is defined as the vector $\vec{\mathbf{c}} = [c_1, c_2, .., c_{N_s}]$, in which

$$c_i = \frac{1}{|S_i|} \sum_{q \in S_i} \mathbf{F}(q) \tag{1}$$

with $|S_i|$ being the total number of pixels in S_i for $i \in [1, N_s]$. Since a (precomputed) superpixel partitions the image domain into non-overlapping regions, the descriptor computation will iterate for each pixel of a feature map. Therefore, the computational cost of the descriptor is $O(n)$, with n being the total number of pixels.

There is no guarantee that a given superpixel will be in the exact location for both images; for example, S_1 could be located in the tumor for one image but on the background for others. Therefore, we sort each descriptor into non-increasing order so that each descriptor will be a characteristic curve of how it activates for many/few elements (ilustrated in Figure 3).

Figure 4 contains the primary motivation behind our *image descriptor*. It shows descriptor curves for three feature maps that capture the object (a tumor in this case) differently. In (sub), we have a scenario in which the feature map did not completely capture the tumor, (ok) the feature map better captures the tumor, and (sup.) the feature map overestimates the tumor. Note that there is a difference in the presentation of the descriptor for each scenario, and we explored this difference in the selection of the other images. It is worth mentioning that a descriptor similar to the case (sup.) could occur due to the 'leaking' of the superpixel algorithm [16].

Moreover the number of feature maps changes as we select more images, for two feature maps F^1 and F^2 with vectors $\vec{\mathbf{c}}^1$ and $\vec{\mathbf{c}}^2$, the new image descriptor is defined as a new vector that is the concatenation of both, $\vec{\mathbf{c}} = [\vec{\mathbf{c^1}}, \vec{\mathbf{c^2}}]$.

C. Image selection with visual guidance

We use a visual projection from image descriptor and heatmaps with feature map attention to guide the image selection as presented the image recommendation interface[1] of Figure 5. The user can then see where the marked image lies in the projection while noting the recommended **closest** and the most **distant** images from marked images (A). Furthermore, the interface shows all recommended images, ordered by distance to the marked image. Note that the closest images (C) had their tumors detected by the already learned filters, which did not happen with distant images (D).

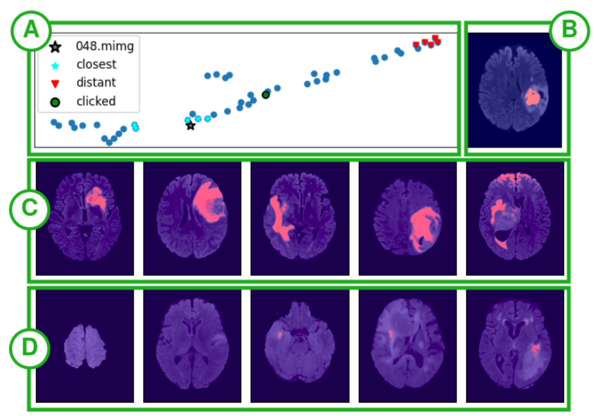


Fig. 5. Image recommendation interface: (A) The projection with one marked (star) and recommended images, separated into closest (C) and distant images (D). Also, (B) contains an image the user clicks for inspection colored according to filters attention.

The user can **click** on an image from the projection (A) and inspect its feature maps (B), allowing them to decide whether to mark it or not. The goal of the visualization is for the user to spatially identify and inspect images individually in (B). After a selection, the new image will be displayed as a

[1] Available in github.com/abrMatheus/GTFreeImgSelection

star as a new projection (A), and the interface will highlight new recommended images in (B) and (C). Also, the interface does not have a visualization of selected images, to inspect a selected image, we need to click on its stars in (A) and visualize each image in (B).

It is worth mentioning that we used t-SNE [17] as image descriptor projection in our work, but other techniques could be applied. Also, we adopted the L2 distance in this work, recommending the most distant and close images as a form of exploration/exploitation.

IV. Experimental Setup

Our experiments involved selecting a reduced set of images to train an encoder for a segmentation network. We compared our selection with other methods and varied the initial image (splits). In the following, we describe the datasets, architecture, and selection methods.

A. Datasets

The first used dataset is Brain Tumor Segmentation (BTS), a private dataset containing 80 3D images of Glioblastoma (GBM) with two MRI scans (FLAIR and T1Gd) per patient. As a second dataset, the Schistosoma Mansoni eggs dataset (Schisto) is also proprietary (soon to be public) with 625 labeled images, and specialists annotated both datasets. We randomly divided the datasets into 60% for training, 10% for validation, and 30% for testing. In the training set, we applied the image selection method, obtaining a very small number of images (5) for marking and then training the network encoder. For the GBM and Schisto dataset we define $N_s = \{40, 20\}$ and t-SNE perplexity of $\{5, 10\}$, respectively.

We chose the GBM dataset because of the tumors' different sizes and shapes and because capturing the tumors' internal subregions: Enhancing Tumor (ET), Tumor Core (TC) and Whole Tumor (WT). On the other hand, the Schisto dataset shows eggs of standard size and shape, but the images often have a cluttered background, and fecal impurities sometimes cover the parasites.

B. Architecture

For the architecture, we adopted a small 3D U-Net architecture, sU-Net, consisting of three convolutional layers encoder and a symmetrical decoder [6]. We also used FLIM to train the encoder using 5 marked images obtained with the proposed method. We applied backpropagation for the decoder training, using the rest of training images and using learning rate of $2.5e^{-3}$ with linear decay, loss (average of Cross-Entropy and Dice), and a total of 100 epochs. We also used ADAM optimizer and a batch size equal to one.

C. Image selection methods

For the image selection methods, we compared our interactive method against the user's *manual selection*, with the user manually inspecting the training set and selecting the most relevant to be marked, and also against the *labeled* criterion that uses the ground- truth of the images as a form of selection, starting from the same initial images [11]. Finally, we verified the model's performance trained with backpropagation, using all training images rather than only the marked images.

TABLE I

Mean and Standard deviation for GBM tumor (mean value of all three classes) and Schisto segmentation for the validation set.

	GBM Avg.			Schisto		
Split	Split1	Split2	Split3	Split1	Split2	Split3
Manual	0.72 ± 0.04	0.72 ± 0.05	$\mathbf{0.73 \pm 0.04}$	0.80 ± 0.16	0.81 ± 0.16	0.82 ± 0.16
Labeled	$\mathbf{0.73 \pm 0.04}$	$\mathbf{0.72 \pm 0.05}$	0.72 ± 0.06	0.80 ± 0.16	0.81 ± 0.15	0.81 ± 0.16
Ours	0.72 ± 0.04	0.72 ± 0.05	0.71 ± 0.05	$\mathbf{0.82 \pm 0.16}$	$\mathbf{0.82 \pm 0.16}$	$\mathbf{0.84 \pm 0.15}$

V. Quantitative results

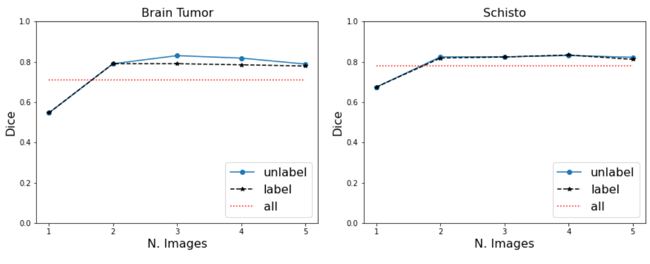
Initially, the goal is to check the variability of the selection methods for multiple data splits. Therefore, we check the sU-Net performance on the validation set for three different splits for each method. We verified the performance of manual, unlabeled selection and our proposed method, whose selection criterion does not use the ground-truth.

Table I contains the performance of the sU-Net segmentation model for validation sets of different splits and image selection methods. Note how the different data splits of each method had slight variations in the mean, demonstrating the constancy of the selections. There was also little difference in the selection methods, demonstrating the quality of our proposed method. The proposed method obtained the best metrics in the Schisto dataset and competitive values compared to other methods for the GBM dataset, showing the average labels. Also, the higher standard deviation in the Schisto dataset is related to the impurity of those images that look similar to parasites, justifying the motive for evaluation or method in this dataset.

Finally, we took the best model for each method/split (bold values in Table I) and verified its performance for the testing set of both datasets (Table II). We also compared the multiple image selection methods with the model training from scratch with *all* training images (using backpropagation).

Table II demonstrates the superiority of using FLIM to train the encoder. In all selection methods, the performance was notably better than that achieved through backpropagation. It's worth noting that in FLIM methods, we trained the encoder with FLIM and then froze it, using only 5 marked images. The rest of training images was used only to update the sU-Net decoder. Results illustrated in Table II also underscore the effectiveness of our proposed selection method, which produced results comparable to labeled and manual selection. It even outperformed them, achieving the best values for two classes of the GBM dataset and the best value for Schisto. Our method does not require the effort of looking at the entire test set as in manual selection, nor does it require ground-truth information to train the encoder.

Since both the labeled and the proposed methods are interactive, differing only in the selection criterion, we also check the performance of the final model (sU-Net) for each newly selected image starting from the same initial image. Figure 6 shows the performance of the U-Net model with each new



(a) (b)

Fig. 6. Dice vs interaction for labeled and unlabeled (ours) selection

choice of training images. Note that there is similar behavior for both methods, which also corroborates the strength of the proposed method. Moreover, in Figure 6, we can see small oscillations, with the most significant difference of 0.04 for the third brain image, which could be either because they are the first different image (images 1 and 2 are the same for labeled and unlabeled) or because of the training of the rest of the model (since we used backpropagation to train the decoder).

VI. Discussions

In general, the tables' results demonstrated our method's strength, both due to the closeness with the splits and the comparison with other image selection methods, in particular with [11] that uses a ground-truth (GT). That said, our proposed method is an excellent tool, since we do not use GT to train the encoder, only the encoder.

Furthermore, it would be possible to use the proposed descriptor for other methods besides image selection. For example, we could use such descriptor for inference, checking the distance from a test image to the marked images. If the value of this distance is high, we will have a model that does not generalize to this test image.

Our proposed method recommends images using two factors: (i) the projection originating from the descriptor and (ii) the distance to the annotated images. We checked if there is a relationship between (ii) the distance to an annotated image and (iii) the performance the performance of the filters using ground-truth as done in [11].

Figure 7 presents a 2D projection in which each point is an image from the training set, with the color representing the performance of such images (from blue (0) to bright yellow (1)). Note that for brain images, there are clusters of images with good (light) and bad (dark) performance. For Schisto images, these clusters are also present; however,

TABLE II
MEAN DICE AND STANDARD DEVIATION FOR TEST DATASET FOR ALL IMAGE SELECTION METHODS AND USING ALL IMAGES FOR TRAINING (BACKPROPAGATION)

	Brain			Schisto
Sel.	ET	TC	WT	
All	0.6668 ± 0.1362	0.7023 ± 0.1684	0.7640 ± 0.0081	0.7814 ± 0.1692
Manual	0.7093 ± 0.0495	0.8015 ± 0.0806	0.7824 ± 0.0983	0.8248 ± 0.1526
Labeled	0.6886 ± 0.0774	**0.8387 ± 0.0680**	0.8278 ± 0.0744	0.8127 ± 0.1490
Ours	**0.7595 ± 0.0714**	0.8346 ± 0.0907	**0.8347 ± 0.0784**	**0.8385 ± 0.1418**

outliers are present here, that is, images with low performance (dark) in groups of good images (light) and vice versa. This phenomenon occurred due to impurities in the images, making it challenging to generate codes either by superpixels (due to variation in impurities) or by impurities that look like parasites (also showing as a high standard deviation in Tables I and II). Figure 8 shows the schisto projection comparing the marked image with two outliers. One has high performance but is distant due to a leak in the superpixel, and the second (Near) has an image that detects the impurity rather than the parasite.

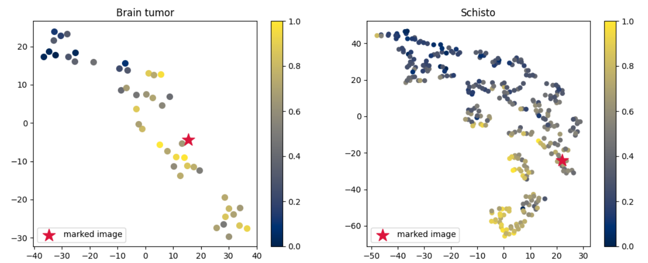


Fig. 7. 2D projection of the training images using our proposed descriptor and showing its relationship with segmentation performance for one marked image. Performance vary from 0 (blue) to 1 (yellow)

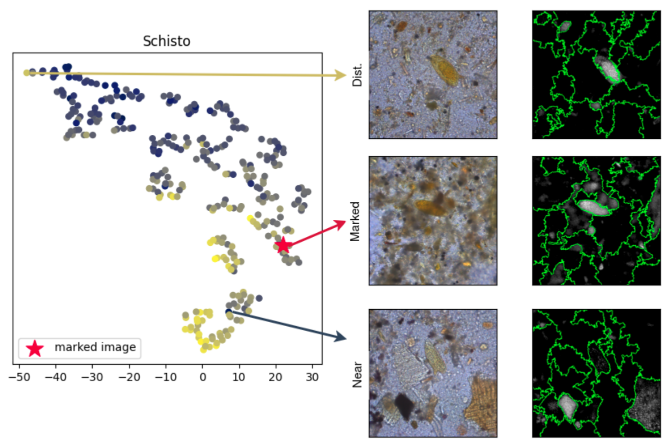


Fig. 8. Outliers for the relationship between the proposed descriptor and image performance using GT (Schisto dataset).

One way to inspect the relationship between the proposed descriptor is to compute the correlation between the inverse of the distance to the marked image and the performance for each image – it is expected that images closer to the marked image have a good performance. Figure 9 presents the result of such correlation, and for both datasets the correlation was positive, that is, more distant images have lower performance. Furthermore, we check which line best represents this correlation (in black on the graph).

Differently, Figure 10 compares the marked image with its recommended images to be annotated (which are not outliers), with two distant images (Dist.) and two near images (Near). Notice how the nearby images activate for the parasite while the distant images activate for impurities. Thus, once one of the distant images is selected, the model will learn filters that capture its particularities.

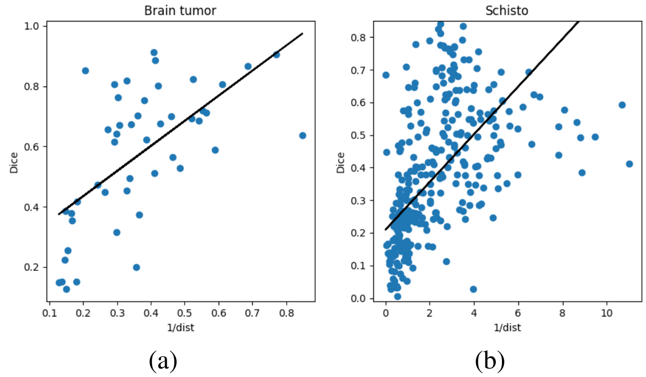


Fig. 9. Correlation between the images performance and the inverse of the distance using proposed descriptor.

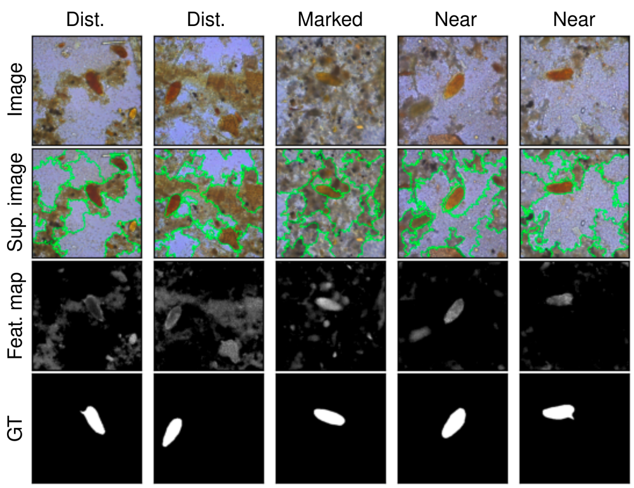


Fig. 10. Image, superpixels, feature map, and GT for recommended images of Schisto dataset (Near and Distant images).

VII. Conclusion

Training an encoder with a limited set of images is challenging, especially finding which images are most valuable for such an activity. The present work proposed an interactive method for image selection and training FLIM encoders based on a new descriptor that generates different patterns according to the attention of the filters. In each interaction, the user selects an image from a pool of recommended images and utilizes a 2D projection and explanations as visual guidance. We demonstrate its superiority over the subjective manual expert selection and a similar approach that relies on pixel-wise annotation (ground-truth) for sampling images. Our method, by contrast, is less subjective, more efficient, and does not require ground-truth, making it a game-changer in the field of image selection for training. For future work, we will address some limitations, such as exploring the method with other datasets, not only biomedical, and defining a stopping criterion for the number of selected images.

Acknowledgment

The authors thank CAPES (88887.820768/2023-00 and STIC-AMSUD 88887.878869/2023-00), FAPESP (2023/09210-0, 2023/14427-8, and 2013/07375-0), CNPq (304711/2023-3, 407242/2021-0, 306573/2022-9, 442950/2023-3), and FAPEMIG (APQ-01079-23, PCE-00417-24 and APQ-05058-23) for financial support.

References

[1] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, pp. 13 693–13 696, 2020.

[2] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: a survey of case studies," *ACM computing surveys*, vol. 55, no. 6, pp. 1–29, 2022.

[3] I. E. De Souza and A. X. Falcão, "Learning cnn filters from user-drawn image markers for coconut-tree image classification," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[4] A. M. Sousa, F. Reis, R. Zerbini, J. L. Comba, and A. X. Falcão, "Cnn filter learning from drawn markers for the detection of suggestive signs of covid-19 in ct images," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3169–3172.

[5] I. E. de Souza, C. L. Cazarin, M. R. Veronez, L. Gonzaga, and A. X. Falcão, "User-guided data expansion modeling to train deep neural networks with little supervision," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[6] M. A. Cerqueira, F. Sprenger, B. C. Teixeira, and A. X. Falcão, "Building brain tumor segmentation networks with user-assisted filter estimation and selection," in *18th International Symposium on Medical Information Processing and Analysis*, vol. 12567. SPIE, 2023, pp. 202–211.

[7] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, 2022.

[8] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: A state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, 2023.

[9] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, and M. Sedlmair, "Comparing visual-interactive labeling with active learning: An experimental study," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 298–308, 2017.

[10] B. C. Benato, J. F. Gomes, A. C. Telea, and A. X. Falcão, "Semi-automatic data annotation guided by feature space projection," *Pattern Recognition*, vol. 109, p. 107612, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320304155

[11] M. A. Cerqueira, F. Sprenger, B. C. Teixeira, and A. X. Falcão, "Interactive image selection and training for brain tumor segmentation network," *arXiv preprint arXiv:2406.03225*, 2024.

[12] Z. Zhao, P. Xu, C. Scheidegger, and L. Ren, "Human-in-the-loop extraction of interpretable concepts in deep learning models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 780–790, 2021.

[13] K. Uehara, H. Nosato, M. Murakawa, and H. Sakanashi, "Object detection in satellite images based on active learning utilizing visual explanation," in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2019, pp. 27–31.

[14] P. Serrano e Silva, R. Cruz, A. Shihavuddin, and T. Gonçalves, "Interpretability-guided human feedback during neural network training," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2023, pp. 276–287.

[15] G. Lenczner, A. Chan-Hon-Tong, B. Le Saux, N. Luminari, and G. Le Besnerais, "Dial: Deep interactive and active learning for semantic segmentation in remote sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 3376–3389, 2022.

[16] F. C. Belém, S. J. F. Guimaraes, and A. X. Falcão, "Superpixel segmentation using dynamic and iterative spanning forest," *IEEE Signal Processing Letters*, vol. 27, pp. 1440–1444, 2020.

[17] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.