

Enhancing Interactive Annotation in Digital Pathology with Foundation Models

Thibaud Vanmechelen
University of Liège
Liège, Belgium

Noé Gille
University of Rouen
Normandie, France

Ba Thien Le
University of Liège
Liège, Belgium

Pierre Geurts
University of Liège
Liège, Belgium

Raphaël Marée
University of Liège
Liège, Belgium

Abstract—Manual annotation of large-scale biomedical images is a major bottleneck in digital pathology, since it requires significant expert time and effort. In this paper, we explore the use of foundation models for segmentation to facilitate and accelerate this annotation process. More particularly, we investigate the Segment Anything Model (SAM) and its successor SAM2, evaluating their performance across three in-house datasets and one public dataset which cover diverse segmentation tasks. In addition, we design and evaluate new architectures that build on these models and also incorporate domain-specific pathology encoders. Our experiments show that domain adaptation substantially improves segmentation results across various structures, with fine-tuning yielding the strongest gains and hybrid designs offering a complementary benefit, paving the way for more effective interactive annotation. Furthermore, we integrate the best model we obtained into an open-source, user-friendly, web platform.

Index Terms—Deep learning, interactive segmentation, SAM, digital pathology

I. INTRODUCTION

Deep learning is a promising approach for automating annotation tasks in digital pathology. However, it still faces a significant challenge: the scarcity of data. Segmentation datasets are typically limited in scale, while training modern models such as transformers [1] generally requires large amounts of data. Collecting these data is both costly and time-consuming as it often requires the intervention of healthcare professionals who must delineate precisely the target objects. Therefore, leveraging current deep learning methods to assist with this process is particularly important.

In this paper, we investigate the performance of two recent foundation models for segmentation, SAM [2] and SAM2 [3], applied to large digital pathology images, commonly referred to as Whole Slide Images (WSIs). Our study examines the performance of these models in different settings, including their combination with fine-tuning strategies, and also explores different combinations of SAM with pathology-specific encoder models. Our experiments are aimed at spatial prompt-based segmentation of arbitrary 2D target objects, without specifying the particular nature of the structures being segmented. We evaluate our approach on four datasets covering multiple tissue types. All datasets consist of patches extracted from WSIs for model training and evaluation. Moreover, we integrate the best model into the open-source Cytomine platform [4] to demonstrate the practical usability of our approach.

The remainder of this paper is organized as follows. **Section 2** reviews related work on the Segment Anything models and their applications in the biomedical domain. **Section 3** describes our proposed methodology, including fine-tuning strategies and hybrid architectures combining SAM with pathology-specific encoders. **Section 4** presents the datasets used and the results of our experiments. Finally, **Section 5** summarizes our findings and concludes the paper.

II. RELATED WORKS

The Segment Anything Model (SAM) [2] introduced the task of promptable segmentation, in which a model returns segmentation masks from user prompts such as points, bounding boxes, or masks. This model is a transformer-based architecture that learns a general mapping between image and prompt embeddings, allowing it to generate accurate segmentation masks for virtually any object. Its successor, SAM2 [3], extended the original framework to video and improved both its speed and accuracy, while SAM3 [5] introduces prompts combining short noun phrases and image exemplars.

While they were originally trained on natural images, SAM models have been subsequently evaluated in the biomedical domain [6]–[9]. These works show that bounding box prompts usually outperform text and point prompts, and that performance can vary depending on the target object size and the image type. The models typically perform best on well-circumscribed objects but struggle more with dense or fine-grained structures (e.g. poorly capturing nuclear structures [9]), a common challenge in pathology.

Several papers have attempted to improve the Segment Anything Models through fine-tuning or through some architectural adaptations. Frameworks such as MedSAM [10] and SAM^{Med} [11] adapt the model using a broad set of medical images, while other approaches such as SAM-Path [12] explored the combination of SAM with pathology-specific encoders. In our work, we reuse some pathology encoders from these previous studies (UNI [13]) as well as additional ones to assess the generalization capability of different encoder backbones. More recently, SAM2 has also been adapted in MedSAM-2 [14] and Path-SAM2 [15], enabling to further improve the previously obtained results. Both SAM-Path and Path-SAM2 focus on the segmentation of specific classes in pathology images through the use of automatic class prompts and are not designed for interactive use. In contrast, this study addresses

the interactive annotation setting in 2D histopathology without making assumptions about some predefined target classes. Concurrent with our work, SAM3 was likewise adapted to the medical domain (Medical SAM3 [16]), confirming that domain adaptation remains necessary for this newest generation too.

Overall, existing works confirm the potential of foundation segmentation models in the medical field but also highlight several limitations when applied to pathology, notably challenges related to the handling of high-resolution whole slide images, the segmentation of small or densely packed objects, and the lack of pathology knowledge in the base models.

III. METHODS

In this section, we present our approach for fine-tuning the Segment Anything Models along with the hybrid architectures we consider, and the post-processing method we propose. In all cases where learning is involved, our training set is composed of image patches, each patch is centred on a specific annotation (i.e. a region of interest). All patches are extracted at the native resolution of the whole slide images and then resized to 1024×1024 pixels to match SAM/SAM2's input resolution. We denote the segmentation dataset by $\mathcal{D} = (X, Y) \subset \mathcal{X} \times \mathcal{Y}$, where X and Y represent the set of input image patches and corresponding segmentation masks. For a given image patch, we write Ω for its set of pixels. We denote by N the batch size used during training. For sample j , the prediction at pixel $k \in \Omega$ is $\hat{y}_j^{(k)}$ and its ground-truth label $y_j^{(k)}$.

A. Fine-tuning of SAM and SAM2

While SAM and SAM2 were originally trained on natural images, as a baseline, we fine-tune both models on training patches extracted from the WSIs. We use the same supervised training scheme as SAM/SAM2 but we replace the original complex loss functions with simpler alternatives, since we found these to yield better results in our setting. Our loss is:

$$\frac{1}{N|\Omega|} \sum_{j=1}^N \left[- \sum_{k \in \Omega} (y_j^{(k)} \log \hat{y}_j^{(k)} + (1 - y_j^{(k)}) \log(1 - \hat{y}_j^{(k)})) \right. \\ \left. + \lambda \cdot |\hat{s}_{\text{IoU}}^{(j)} - s_{\text{IoU}}^{(j)}| \right]$$

The first term is the standard binary cross-entropy, which has already been employed in prior works fine-tuning SAM [10], [17]. The second term penalizes the difference between the predicted and the true intersection-over-union (IoU) scores of the predicted mask. Here, the predicted IoU $\hat{s}_{\text{IoU}}^{(j)}$ refers to the score produced by the model itself that estimates how well its segmentation matches the ground truth, while the true IoU $s_{\text{IoU}}^{(j)}$ is the actual overlap. λ controls the weight of the IoU term in the total loss. In our experiment, this additional calibration term was only included in SAM2, as it did not bring any noticeable benefit for SAM.

During the fine-tuning, we update only the prompt encoder and the mask decoder of SAM/SAM2, while keeping the image encoder frozen. This choice avoids the computational burden and overfitting risks that are associated with fully retraining the encoder. To ensure robustness across different

user prompting scenarios, we train the models with a variety of prompt types, including bounding boxes, positive and negative points, and masks. These prompts were generated automatically from the ground-truth masks, all with some degree of perturbations to simulate imperfect human inputs. In practice, we considered each prompt type separately as well as several of their combinations. For every training patch, we generated x augmented samples, where x corresponds to the total number of prompt settings that we considered.

For SAM, we pre-compute image embeddings to accelerate the training and use a batch size of 64 with a learning rate of 1×10^{-5} for 30 epochs. For SAM2, the batch size was reduced to 16 while keeping the same learning rate, and the optimization was performed with Adam and weight decay 10^{-4} . In both cases, we use small learning rates to avoid destabilizing the pretrained weights, while we choose the batch sizes to maximize the GPU capacity.

B. Hybrid Architectures

To further adapt SAM to histopathology, we investigate hybrid architectures combining it with pathology-specific encoders. We consider three domain-specific, state-of-the-art, encoders: **H-Optimus-0** (1.1B parameters, trained on 500k slides) [18], **UNI** (303M parameters, trained on 100k WSIs) [13], and **UNI-2H** (681M parameters, trained on a broader dataset) [13]. Since all these encoders are based on standard Vision Transformers (ViTs [19]) and do not provide multi-scale features, they are only integrated with SAM, whose image encoder also follows a standard ViT architecture.

We explore two main strategies for designing our hybrid architectures that integrate pathology-specific encoders with SAM. The first strategy (Section III-B1) simply incorporates the domain-specific encoder alongside SAM's original one through a fusion mechanism, leaving the rest of the processing unchanged. The second strategy (Section III-B2) leverages the domain-specific encoder in a post-processing role: the fine-tuned SAM model first produces low-resolution segmentation masks that are then refined using the additional features extracted by the pathology encoder. The first approach thus injects pathology knowledge before the segmentation prediction, the second after it.

1) *Fusion Architecture*: In the fusion architecture, the features from the pathology-specific encoder are combined with SAM's features through a learnable fusion operator ϕ : $z = \phi(f_{\text{SAM}}(x), f_{\text{path}}(x))$, $z \in \mathbb{R}^{d'' \times h \times w}$, with $f_{\text{SAM}} : \mathcal{X} \rightarrow \mathbb{R}^{h \times w \times d}$ the frozen SAM image encoder, and $f_{\text{path}} : \mathcal{X} \rightarrow \mathbb{R}^{h' \times w' \times d'}$ the domain-specific encoder, also frozen. We use standard feature concatenation for fusion, which maximizes validation Dice/IoU while keeping the architecture light.

In this setting, SAM's image encoder produces an embedding of shape $h \times w \times d$, with $h = w = 64$ for an input resolution of 1024×1024 (ViT/16 architecture), and $d = 768$ in the ViT-B variant we used here. On the other hand, the pathology-specific encoders (UNI, UNI-2H, H-Optimus-0) work on 224×224 inputs and follow either ViT/14 (UNI-2H,

H-Optimus-0) or ViT/16 (UNI) backbone architectures, which yields feature maps of size $n \times n \times d'$ with $n \in \{14, 16\}$.

To make these representations compatible, we first upsample the pathology feature map to $64 \times 64 \times d'$ using a deconvolution operator $\Theta(\cdot) : \mathbb{R}^{n \times n \times d'} \rightarrow \mathbb{R}^{64 \times 64 \times d'}$, where we adapt the kernel and stride depending on the value of n . The SAM and pathology embeddings are then concatenated along the channel dimension, which results in a tensor of shape $64 \times 64 \times (d + d')$. Finally, the neck module (SAM's channel-projection block) projects this to the standard $256 \times 64 \times 64$ mask-decoder input.

For training, we keep both image encoders frozen, and fine-tune the upsampling module, the prompt encoder, and the mask decoder. This setup is illustrated in Figure 1a, with frozen components in blue and trainable ones in red.

2) *Refinement Architecture*: In the refinement architecture (Figure 1b), the pathology-specific features are combined with the predicted low-resolution masks, within a new module: the refinement module. This module is based on attention mechanisms [20], and can be formally described by an operator: $\nu : \mathbb{R}^{3 \times 256 \times 256} \times \mathbb{R}^{n \times n \times d'} \rightarrow \mathbb{R}^{3 \times 256 \times 256}$, which takes the low-resolution masks together with the pathology encoder embeddings, and outputs refined masks at the same resolution.

The refinement module (Figure 2) consists of a sequence of convolutional, linear, and multi-head attention layers. The input masks are first projected to a 256-channel representation using a convolutional layer. In parallel, the pathology embeddings are linearly projected to the same channel dimension¹. These two inputs are then passed through a multi-head attention layer, where the low-resolution masks correspond to the queries and the pathology features to the keys. The attention output is processed by two convolutional layers with ReLU activations, which produces a tensor of size $3 \times 256 \times 256$. Finally, this result is added to the original mask through a skip connection, ensuring that the module only refines the masks rather than overwriting them completely.

During training, only the refinement module is updated, while all the other components remain frozen. We reuse the SAM weights obtained after the fine-tuning step described in Section III-A, in order to see if the refinement module is able to further improve the results.

Hybrid Architectures Training Details. For both the fusion and refinement architectures, we use the binary cross-entropy loss, and perform training with a batch size of 32, a learning rate of 1×10^{-5} , and the Adam optimizer for 30 epochs.

C. Post-processing

While fine-tuning and hybrid architectures improve the segmentation accuracy, the resulting masks are not necessarily user-friendly, often appearing noisy. To address this, we evaluated several post-processing strategies aimed at producing cleaner and more regular masks, and retained the simplest and most effective one which is based on OpenCV [21].

The procedure consists of three steps. First, a morphological opening (erosion followed by dilation) is applied to remove

¹The embeddings are flattened to $(n \times n) \times 256$ for processing, rather than $n \times n \times 256$, for notation convenience.

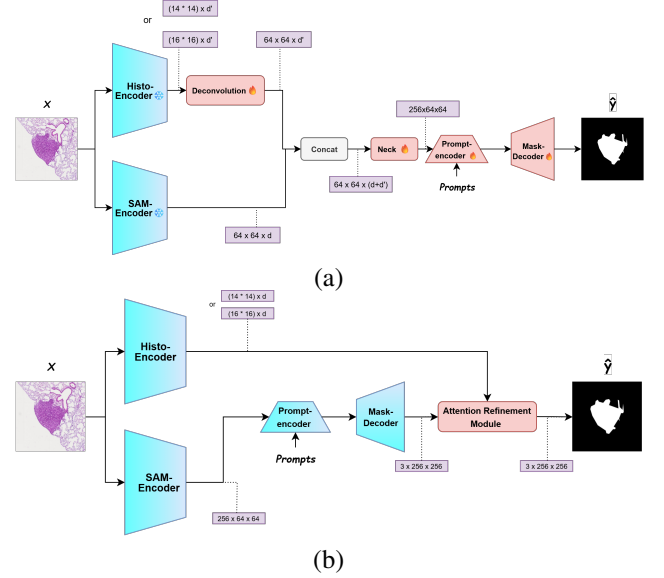


Fig. 1. **Hybrid Architectures** combining SAM with a pathology-specific encoder. **(a) Fusion**: SAM and histopathology encoder features are combined at the patch level (deconvolution, concatenation, neck module) before the prompt encoder and mask decoder. **(b) Refinement**: the SAM encoder generates initial masks, then refined using histopathology encoder features, without modifying SAM itself.

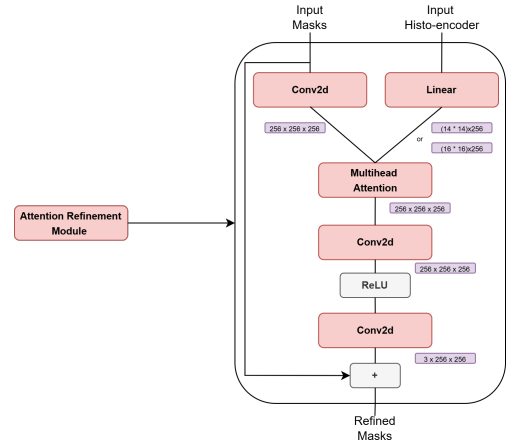


Fig. 2. **Refinement Module** used in the architecture of Fig 1. The module takes as input the low-resolution masks predicted by the SAM decoder and the features extracted by the histopathology encoder. These are fused through a multi-head attention block followed by several convolutional layers, which refine the mask boundaries.

small noisy regions. This is then followed by a morphological closing (dilation followed by erosion), which fills small holes and reconnects adjacent regions that should belong to the same object. Finally, any remaining small connected components are removed if their area is below 10% of the largest component, ensuring that only the main structure is preserved.

As our experiments will demonstrate, this method improves the regularity and usability of masks while having a negligible impact on quantitative performance metrics on our datasets.

IV. EMPIRICAL EVALUATION

A. Datasets

We train and evaluate our models on four datasets: three in-house collections (LBTD-AGDC10, LBTD-NEO04, and CHU-ANAPATH-NST-DL) and one public dataset, Camelyon16 [22]. Among these, we used CHU-ANAPATH-NST-DL exclusively as an external test set.

LBTD-AGDC10 from the Laboratory of tumor and development biology contains 1199 annotations of tissue areas in whole-slide images of mice lung categorized as either adenocarcinoma or non-adenocarcinoma regions. LBTD-NEO04 contains 1538 diverse annotations including bronchi, cartilage, and red blood cells. Camelyon16 provides whole-slide images of sentinel lymph nodes with 2597 pixel-level annotations of metastatic regions. CHU-ANAPATH-NST-DL is a breast cancer dataset collected at the pathology department of CHU of Liège, containing 13290 annotations corresponding to various tissue areas such as calcifications, healthy glands, in situ carcinoma, and infiltrated regions.

The annotations of the other three datasets combined were split into a 70% training, 15% validation, and 15% test sets, with the split performed at the whole-slide (patient) level.

B. Models Performance Analysis

We evaluated all approaches on the CHU-ANAPATH-NST-DL dataset, which was excluded from training, and therefore enables to evaluate the generalization performance, as well as on the test sets of all other three datasets. To ensure a controlled and fair comparison, all evaluated variants are trained and assessed using the same datasets and prompting protocol; therefore, we do not include medical SAM adaptations (such as MedSAM and SAM^{Med}) that rely on additional external medical corpora or other method-specific fine-tuning process, as this would mix up the impact of architectural choices with differences in the training data. The results are summarized in Table I. For CHU-ANAPATH-NST-DL, we report four standard segmentation metrics: Dice score, IoU, precision, and recall. For the other three datasets, we only report the dice score. Our evaluation is carried out using the same prompt combinations that we used during training, including bounding boxes, positive points, negative points, and masks. Table I reports the scores for the original SAM/SAM2 models, their fine-tuned versions, and the proposed hybrid architectures. For these architectures, we only report the results obtained with the pathology-specific encoder that achieved the best performance.

From Table I, we observe that the baseline performance of the original SAM and SAM2 models is very poor, with SAM2 slightly outperforming SAM. This is in line with observations reported in other studies [3], [9], [14]. Fine-tuning clearly improves the performance for both models, with SAM showing an improvement of around 50%. This proves that even without updating the image encoder, introducing some pathology-specific knowledge through fine-tuning already significantly enhances the segmentation accuracy. The hybrid architectures also provide notable performance gains, with both fusion and

TABLE I
PERFORMANCE OF ALL MODELS ON CHU-ANAPATH-NST-DL (ALL METRICS, IN %) AND ON THE TEST SETS OF LBTD-AGDC10 (D1), LBTD-NEO04 (D2) AND CAMELYON16 (C16) (DICE SCORE ONLY, COLUMNS 6-8). FT: FINE-TUNED. "FUSION" AND "REF" DENOTE RESP. THE FUSION AND THE REFINEMENT ARCHITECTURES.

Model	Dice	IoU	Prec	Rec	D1	D2	C16
SAM	56.1	44.3	52.1	78.2	53.4	63.1	60.1
SAM2	65.6	52.5	62.9	82.0	63.5	70.1	65.7
SAM-FT	87.5	78.5	85.1	91.3	86.4	92.4	89.8
SAM2-FT	88.6	80.3	86.1	92.5	87.5	93.1	90.5
Fusion-H-opt-0	83.5	72.9	79.9	90.1	84.4	89.9	85.3
Ref-UNI-2H	87.5	78.5	85.5	91.0	86.5	92.4	89.9

refinement approaches improving the results by roughly 50%. This confirms that adding pathology-specific encoders can also improve the performance of SAM. However, fine-tuning alone improves more the performance than the fusion hybrid approach. This is somewhat surprising, as neither method involves updating the encoder weights. A possible explanation for this is that the pathology encoders that we used operate on a much lower input resolution (224×224) compared to SAM's encoder (1024×1024). Therefore, they may capture less fine details, and the upsampling required to align the two feature maps could also act as a bottleneck, limiting the performance. Finally, the refinement architecture provides only marginal gains over the SAM-FT model.

In any case, this result should be interpreted as a key finding: with our current design, hybrid integration of low-resolution pathology features does not consistently outperform end-to-end fine-tuning, highlighting that the main bottleneck is likely due to resolution misalignment. Future work might explore pathology encoders trained on higher-resolution and/or multi-scale patches and more principled cross-scale fusion mechanisms to reduce the information loss when matching SAM's 1024×1024 feature resolution.

While annotations vary considerably across datasets, we observe very similar level of performance across all test sets, with a slightly better dice score on 2 of the 3 datasets that were used for training the models. Therefore, one can expect the best models to perform well on new annotation tasks.

In Figure 3, we present an example produced by the fine-tuned SAM2 model. As expected from the quantitative metrics in Table I, the predicted masks align well with the ground truth in all cases, confirming the good overall quality. However, some minor noise can still be observed in the predictions such as random holes in the masks, and stray pixels.

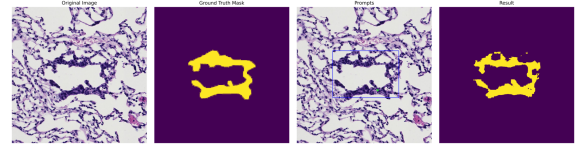


Fig. 3. Segmentation example produced by the fine-tuned SAM2 model. The example comes from LBTD-AGDC10. The prompt is a single positive point (in green) and a box (in blue).

TABLE II

PERFORMANCE EVALUATION OF THE POST-PROCESSING METHOD. THIS TABLE REPORTS THE RESULTS OF THE APPLICATION OF THE POST-PROCESSING METHOD TO BOTH FINE-TUNED SAM AND SAM2, EVALUATED ON THE LBTD-AGDC10, LBTD-NEO04, AND CAMELYON16 TEST SETS, DESCRIBED IN SECTION IV. THE LABEL "POST-PRO" INDICATES THAT POST-PROCESSING WAS APPLIED.

Metric	SAM-FT	SAM-FT (Post-pro)	SAM2-FT	SAM2-FT (Post-pro)
Dice	0.897	0.896	0.906	0.904
IoU	0.824	0.823	0.837	0.835
Precision	0.900	0.900	0.904	0.904
Recall	0.903	0.901	0.915	0.913
Compactness	0.362	0.548	0.323	0.545
Solidity	0.564	0.891	0.496	0.887
Perimeter Smoothness	0.854	1.188	0.795	1.187
Connected Component	5.399	1.017	7.196	1.018
Size Retention	/	0.996	/	0.995

C. Post-processing Performance Analysis

Our post-processing method is designed to improve the usability of these masks by addressing the remaining artifacts, such as stray pixels and small holes, while preserving the strong quantitative accuracy observed above. To evaluate this, we report the same performance metrics as before, and additionally measure compactness, solidity, perimeter smoothness, the number of connected components, and the size retention ratio. Together, these metrics provide a better idea of mask regularity. Although they do not directly measure usability in terms of user effort (e.g., the number of prompts or the time needed to obtain a satisfactory mask), they can be considered reasonable proxies: more regular and compact masks are generally easier to edit and require fewer corrections.

Table II presents our results. We observe that our proposed post-processing method is highly effective. The quantitative metrics (Dice, IoU, precision, and recall) remain nearly unchanged, with differences of only a few thousandths. While on the other hand, all qualitative metrics show substantial improvement. For instance, the compactness and solidity increase by nearly 50%, while the perimeter smoothness tends to 1 (which indicates more regular and polygon-like boundaries). The number of connected components is highly reduced, by a factor of 5 to 7, making the masks far less fragmented. And most importantly, the size retention remains close to 1, meaning that only a small fraction of the pixels is removed while significant gains in mask quality and usability are obtained. Overall, this demonstrates that our post-processing approach successfully preserves the quantitative performance while producing cleaner, more regular, and user-friendly masks.

Figure 4 shows the effect of the post-processing on a predicted mask. The first mask corresponds to the ground truth, the second to the raw prediction, and the third to the post-processed output. As shown, the refined mask is more regular: the stray pixels are removed, some small holes are filled, and the overall structure is cleaner. This visual improvement is consistent with the trends that we observed in Table II.

V. CONCLUSION

We evaluated SAM and SAM2 on histopathology data, fine-tuned them, and proposed hybrid architectures incorporating pathology-specific encoders for interactive segmentation.

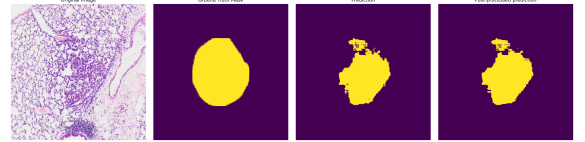


Fig. 4. **Post-Processing Example.** Post-processing method applied to a SAM2 prediction from LBTD-NEO04. From left to right: original image, ground truth, raw prediction, and post-processed prediction.

Experiments on four datasets covering diverse pathological structures show that domain adaptation is essential for segmentation in digital pathology: fine-tuning consistently yielded the largest performance gains, whereas hybrid architectures provided more modest gains. We also introduced a lightweight post-processing method that generates cleaner and more user-friendly masks while preserving high accuracy. Although studies involving pathologists would better quantify how much our approaches accelerate annotation, we proposed several measures as proxies for these gains. A systematic study of robustness to staining protocols, scanners, institutions, and unseen tissue types is left for future work. These results offer practical guidance for deploying foundation-model-based interactive segmentation in digital pathology, and suggest that future work should prioritize effective domain adaptation over increasingly complex architectural modifications.

The best model was integrated into the open-source Cytomine platform for practical deployment in annotation workflows. A dedicated tool was added to the web interface and runs interactively on a single GPU: it automatically refines bounding-box annotations through image extraction, SAM2 inference, post-processing, and automatic annotation updates. To encourage further research, validation, and real-world adoption, the complete implementation together with the trained SAM2 models will be released under the Apache 2.0 license.

VI. ACKNOWLEDGMENTS

Datasets LBTD* were acquired by the team of Prof. Didier Cataldo, and dataset CHU* by Michel Reginster in the team of Prof. Philippe Delvenne. This work was supported by the IMI 2 Joint Undertaking (Grant No. 945358, BIGPICTURE) and the Service Public de Wallonie Recherche (Grant No. 2010235, ARIAC by DIGITALWALLONIA4.AI).

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.
- [2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.
- [3] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," in *The Thirteenth International Conference on Learning Representations*, 2025.
- [4] R. Marée, L. Rollus, B. Stévens, R. Hoyoux, G. Louppe, R. Vandaele, J.-M. Begon, P. Kainz, P. Geurts, and L. Wehenkel, "Collaborative analysis of multi-gigapixel imaging data using cytamine," *Bioinformatics*, pp. 1395–1401, January 2016.
- [5] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. S. Coll-Vinent, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. HAZRA, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollar, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer, "SAM 3: Segment anything with concepts," in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: <https://openreview.net/forum?id=r35clVtGzw>
- [6] R. Deng, C. Cui, Q. Liu, T. Yao, L. W. Remedios, S. Bao, B. A. Landman, L. E. Wheless, L. A. Coburn, K. T. Wilson, Y. Wang, S. Zhao, A. B. Fogo, H. Yang, Y. Tang, and Y. Huo, "Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging," in *IS&T International Symposium on Electronic Imaging*, 2025, pp. COIMG–132.
- [7] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Medical Image Analysis*, p. 102918, Oct. 2023.
- [8] Y. Zhang, Z. Shen, and R. Jiao, "Segment anything model for medical image segmentation: Current applications and future directions," *Computers in Biology and Medicine*, p. 108238, 2024.
- [9] Q. Kong, S. Sharifi, Y. Wang, M. Cristani, and Z. Ruan, "Is sam3 ready for pathology segmentation?" 2026.
- [10] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, Jan. 2024.
- [11] C. Wang, D. Li, S. Wang, C. Zhang, Y. Wang, Y. Liu, and G. Yang, "SAM^{Med}: A medical image annotation framework based on large vision model," *arXiv preprint arXiv:2307.05617*, 2023.
- [12] J. Zhang, K. Ma, S. Kapse, J. Saltz, M. Vakalopoulou, P. Prasanna, and D. Samaras, "Sam-path: A segment anything model for semantic segmentation in digital pathology," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops*, M. E. Celebi, M. S. Salekin, H. Kim, S. Albarqouni, C. Barata, A. Halpern, P. Tschandl, M. Combalia, Y. Liu, G. Zamzmi, J. Levy, H. Rangwala, A. Reinke, D. Wynn, B. Landman, W.-K. Jeong, Y. Shen, Z. Deng, S. Bakas, X. Li, C. Qin, N. Rieke, H. Roth, and D. Xu, Eds. Cham: Springer Nature Switzerland, 2023, pp. 161–170.
- [13] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, B. Chen, A. Zhang, D. Shao, A. H. Song, M. Shaban *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, 2024.
- [14] J. Zhu, A. Hamdi, Y. Qi, Y. Jin, and J. Wu, "Medical sam 2: Segment medical images as video via segment anything model 2," *arXiv preprint arXiv:2408.00874*, 2024.
- [15] M. Zhang, L. Wang, Z. Chen, Y. Ge, and X. Tao, "Path-sam2: Transfer sam2 for digital pathology semantic segmentation," *arXiv preprint arXiv:2408.03651*, 2024.
- [16] C. Jiang, T. Ding, C. Song, J. Tu, Z. Yan, Y. Shao, Z. Wang, Y. Shang, T. Han, and Y. Tian, "Medical sam3: A foundation model for universal prompt-driven medical image segmentation," 2026.
- [17] Y. Song, Q. Zhou, X. Lu, Z. Shao, and L. Ma, "Su-sam: A simple unified framework for adapting segment anything model in underperformed scenes," *arXiv preprint arXiv:2401.17803*, 2024.
- [18] C. Saillard, R. Jenatton, F. Llinares-López, Z. Mariet, D. Cahané, E. Durand, and J.-P. Vert, "H-optimus-0," 2024. [Online]. Available: <https://github.com/biioptimus/releases/tree/main/models/h-optimus/v0>
- [19] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.
- [20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.
- [21] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
- [22] B. Ehteshami Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. van der Laak, and C. Consortium, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, pp. 2199–2210, 2017.