nature biomedical engineering

Article

https://doi.org/10.1038/s41551-022-00952-9

A deep-learning model for transforming the style of tissue images from cryosectioned to formalin-fixed and paraffin-embedded

Received: 1 July 2021

Accepted: 9 September 2022

Published online: 23 December 2022

Check for updates

Kutsev Bengisu Ozyoruk[1,2,3], Sermet Can[1,3], Berkan Darbaz[1,3,4], Kayhan Başak[5], Derya Demir[6], Guliz Irem Gokceler[1,3], Gurdeniz Serin[6], Uguray Payam Hacisalihoglu[7], Emirhan Kurtuluş[8], Ming Y. Lu[9], Tiffany Y. Chen[2], Drew F. K. Williamson[2], Funda Yılmaz[6], Faisal Mahmood[2,10,11,12] ✉ & Mehmet Turan[1,3,12] ✉

Histological artefacts in cryosectioned tissue can hinder rapid diagnostic assessments during surgery. Formalin-fixed and paraffin-embedded (FFPE) tissue provides higher quality slides, but the process for obtaining them is laborious (typically lasting 12–48 h) and hence unsuitable for intra-operative use. Here we report the development and performance of a deep-learning model that improves the quality of cryosectioned whole-slide images by transforming them into the style of whole-slide FFPE tissue within minutes. The model consists of a generative adversarial network incorporating an attention mechanism that rectifies cryosection artefacts and a self-regularization constraint between the cryosectioned and FFPE images for the preservation of clinically relevant features. Transformed FFPE-style images of gliomas and of non-small-cell lung cancers from a dataset independent from that used to train the model improved the rates of accurate tumour subtyping by pathologists.

Histologic examination of tissues by a pathologist is the gold standard for the diagnosis of many diseases. Although this examination is most often performed on formalin-fixed and paraffin-embedded (FFPE) tissues for final diagnosis, a faster alternative called cryosectioning (CS) is a crucial tool for intra-operative decision-making, usually for assessment of tumour margins, differentiation of malignant versus benign lesions and intra-operative staging. The process of formalin fixation and paraffin embedding can take 12–48 h, far exceeding the time limits of routine intra-operative decision-making. Instead, pathologists use CS, the immediate freezing and cutting of tissue, to accelerate the process of preparing slides from hours to minutes (Fig. 1). The trade-off of this increase in speed, however, is the introduction of numerous artefacts from freezing and cutting specimens, including distortion of cellular details, loss of tissue entirely due to ice crystal formation, folding and tearing of the delicate sections and large variances in staining due to changes in section thickness[1,2]. We show examples of artefacts common in cryosection whole-slide images (CS-WSIs) in Supplementary Figs. 1 and 2. These mostly irreversible artefacts can severely distort the

[1]Department of Computer Engineering, Bogazici University, Istanbul, Turkey. [2]Department of Pathology, Brigham and Women's Hospital and Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [3]Institute of Biomedical Engineering, Bogazici University, Istanbul, Turkey. [4]Virasoft Corporation, New York, NY, USA. [5]Department of Pathology, Sağlık Bilimleri University, Kartal Dr.Lütfi Kırdar City Hospital, Istanbul, Turkey. [6]Faculty of Medicine, Department of Pathology, Ege University, Izmir, Turkey. [7]Pathology Department, Istanbul Yeni Yuzyil University Medical Faculty, Gaziosmanpasa Hospital, Izmir, Turkey. [8]Stanford University, Stanford, CA, USA. [9]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [10]Cancer Data Science Program, Dana–Farber Cancer Institute, Boston, MA, USA. [11]Harvard Data Science Initiative, Harvard University, Cambridge, MA, USA. [12]These authors contributed equally: Faisal Mahmood, Mehmet Turan. ✉e-mail: faisalmahmood@bwh.harvard.edu; mehmet.turan@boun.edu.tr

https://doi.org/10.1038/s41551-022-00952-9

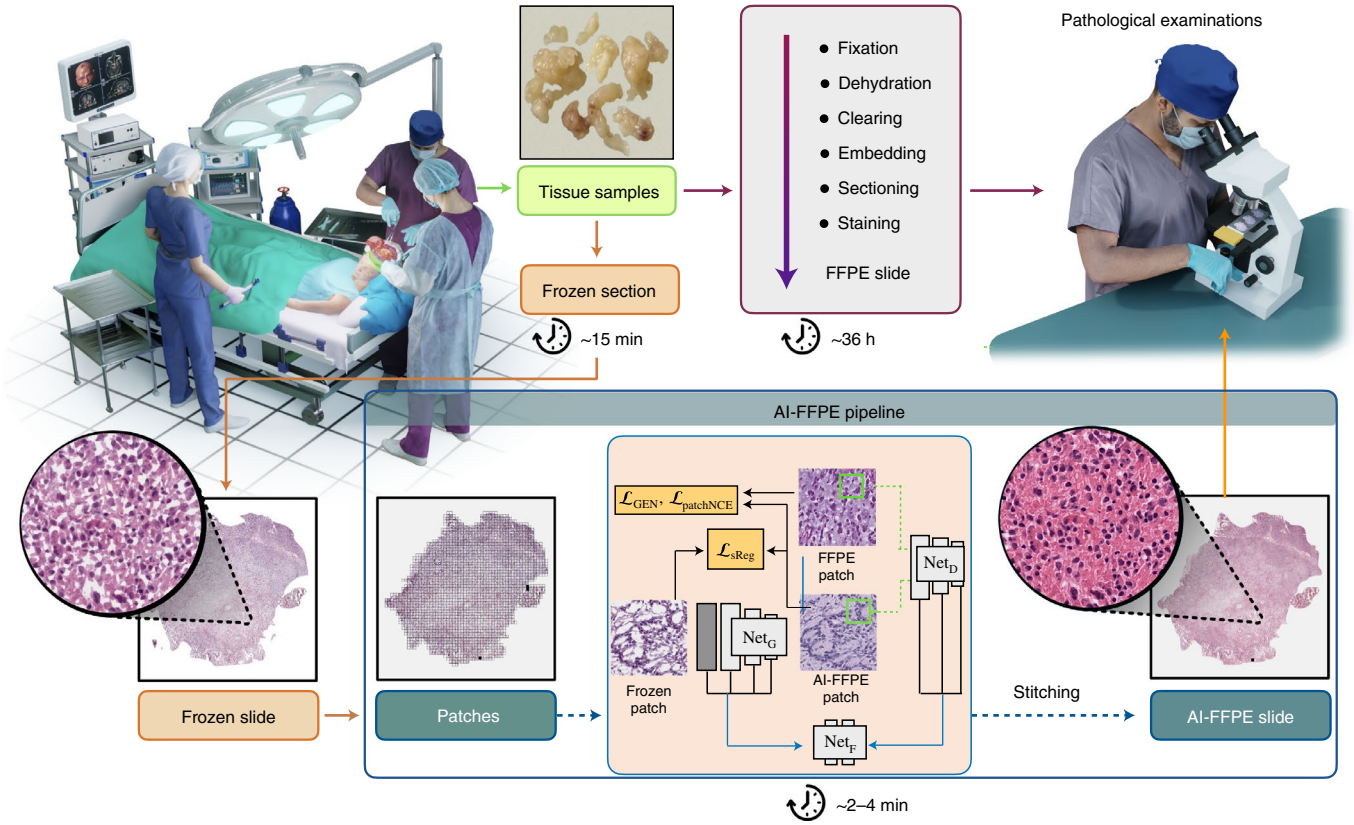


Fig. 1 | Cryosection to FFPE translation workflow. Diagram summarizing how AI-FFPE method fits into the routine preparation of surgically excised specimens for histopathological evaluation. The tissue samples are processed with either FFPE or CS. While the former takes around 36 h and provides higher quality images and definitive diagnosis, the latter takes only around 15 min, which is crucial for the intra-operative decision-making, but results in significantly compromised image quality and diagnostic accuracy. Our rapid 2–4 min AI-FFPE tool is integrated into the CS process as a last step, converting CS images into FFPE-style images to overcome CS-specific obstacles. The AI-FFPE pipeline receives digitized high-resolution CS-WSIs as an input. For each CS-WSI, $N \times N$ mini-patches are created, and the generator neural network, Net_G, works on each patch separately to fix the chemical and mechanical artefacts with the help of discriminator network, Net_D, and projection network Net_F under the guidance of GAN loss, $\mathcal{L}_{GAN}$, the patch-wise noise contrastive estimation loss, $\mathcal{L}_{patchNCE}$, and self-regularization loss, $\mathcal{L}_{sReg}$. At the last step, patches are stitched back for examination by pathologists providing improved-quality FFPE-style images at a speed that is compatible with the fast pace of intra-operative histopathological examinations.

appearance of tissue as compared with FFPE[3–5], disguising malignant cells and making benign cells look atypical. As such, cryosectioned tissue represents a significant diagnostic challenge for experts as documented by the relatively high discordance rate between diagnostic results from CS and FFPE tumour samples[6–16].

Deep learning has been used to address a variety of different tasks in diagnostic pathology[17–20], and deep-learning-based generative adversarial networks (GANs)[21,22] have been used in many different areas of medicine including image segmentation[23], resolution enhancement[24,25], domain adaptation[26], virtual staining of histology slides[27] and stain transfer between different stains[28]. GANs that are developed to translate between two relevant domains can be easily adapted to transferring between image modalities such as computed tomography and magnetic resonance imaging[29] or even between two tissue preparation techniques[30]. In traditional unpaired image-to-image translation methods, cycle consistency dictates the similarity between the images from the target domain and the reconstructed images generated by inverse mapping. In the typical state-of-the-art applications, adversarial loss[21] pushes the change in target appearance while the content is preserved by the cycle consistency loss. At a higher order, the model examines patches from source and target domains and penalizes the discrepancy on the diagnostically relevant regions between the frozen and synthesized patches to improve image quality by preserving relevant content. This paves the way for optimal stain transfer between domains, as it can be easily isolated from content-related features. However, the bijectivity assumption behind cycle consistency[31] restricts the model especially in cases where a large number of uncommon features exist across domains. one-sided unsupervised domain mapping (DistanceGAN)[32], transformation vector learning GAN (TraVeLGAN)[33] and geometry-consistency GAN (GcGAN)[34] propose one-way translation to overcome the circularity-based constraints. As an alternative approach, unsupervised image-to-image translation (UNIT)[35] and multimodel unsupervised image-to-image translation (MUNIT)[36] propose to learn a common content latent space by decomposing images into domain-invariant content representation and the domain-specific identities with respect to the style code. However, defining an objective function based only on a high-level signal that works for pixel-wise reconstruction leads to high computational complexity as well as blurry output images. Although Park et al.[37] have proposed a patch-based approach to avoid these burdens, it is a major bottleneck for applications in clinical pathology in cases where a balance between rectification of artefacts without changing the cellular formation is needed.

In this Article, we propose a GAN-based framework to translate between the rapid yet artefact-heavy cryosectioned tissue to the more commonly used FFPE tissue to address the key challenges with CS outlined above. Our pipeline first automatically segments the tissue region of each slide and divides it into many smaller patches (for example, 512 × 512 pixels) so they can serve as direct inputs to a generator to repair mechanical and chemical artefacts. The synthetically generated

images are then served to the discriminator after being concatenated with images from the target domain (FFPE), with an adversarial loss to promote the output images' resemblance to the target domain. Our proposed approach, artificial intelligence based formalin-fixed and paraffin-embedded (AI-FFPE) uses spatial attention-based learning to automatically identify artefact-enriched sub-regions to increase the diagnostic value of the whole slide while utilizing a self-regularization (SR) mechanism established between the cryosection input image and the synthesized FFPE-style image to preserve clinically relevant features. Moreover, the method is forced to retain the content of the input frozen images via contrastive learning without requiring auxiliary networks or memory banks[38]. As the contrastive visual representations are formulated by selecting non-corresponding patches within the image itself, the cross-domain content similarity exhibits a stronger signal, and the method becomes applicable even for a single image. Finally, patches are stitched back together for examination by pathologists, providing FFPE-style images of an improved quality at a speed that is compatible with the fast pace of intra-operative histopathological examinations. AI-FFPE is publicly available as an easy-to-use repository on GitHub (https://github.com/DeepMIALab/AI-FFPE).

In the following sections, we demonstrate the adaptability and diagnostic contribution of the method on two different computational pathology problems: (1) glioma subtyping and (2) non-small-cell lung cancer subtyping using both publicly available datasets and independent test cohorts. We also demonstrate the realistic appearance of the generated FFPE-like WSIs through a wide range of visual Turing tests (VTTs).

Results

Model for the translation of tissue-image styles from cryosectioned to FFPE

We constructed an unpaired neural-style transfer framework that is trained under the supervision of both patch-wise and pixel-wise signals where the content domain is CS and the style domain is FFPE images. We adopted an architecture (Fig. 2) typically found in GANs and that consists of a ResNet-based generator[39] and a PatchGAN[40] discriminator with the least square GAN loss-of-contrastive learning[41]. To design the final form of the algorithm, we assessed how integration of SR constraint and spatial attention block (SAB) and the combination thereof affects the performance of the algorithm (Fig. 3). Integration of SR-loss function enhanced the nuclear borders and staining quality and also prevented erroneous introduction of red blood cells into the image, a modification that could give a false impression of bleeding into the tissues (Fig. 3a,b). SR loss alone enabled the AI-FFPE to fill out the blank regions; however, this was executed in a relatively indiscriminate manner (Fig. 3c,d). The inadequate selectivity of the algorithm for extracellular matrix (ECM) repair has subsequently been resolved with the integration of an SAB modality. The combination of SR loss with SAB resulted not only in cumulative improvement but also in synergistic effect of incorporating textural details to the ECM. Sometimes, the combination of these modalities reduces the contrast between the nuclei and the cytoplasm. However, copious benefits of combining the two modalities significantly exceed its minimal disadvantages; therefore, we integrated SR loss and SAB into the final version of the model. A detailed illustration of the final AI-FPPE network architecture can be found in Fig. 2.

Evaluation of the performance of the model

Frozen sectioning, when compared with FFPE, introduces additional and unique misleading artificial structures (histological artefacts) to the tissue (Supplementary Figs. 1 and 2). We first examined whether our model reverses these artefacts. Our results demonstrate that AI-FFPE efficiently corrects various frozen section artefacts in brain (Fig. 4) and lung (Fig. 5) sections, such as freezing, cutting, drying and staining artefacts. These artefacts often exist concurrently and exhibit shared elements. AI-FFPE rectifies each type of artefact by resolving various image quality issues together. For instance, by increasing the prominence of nuclear borders and generating more pronounced ECM texture, our method reversed blurring artefacts in lung and brain tissue slides (Figs. 4a and 5a). Similarly, in WSIs where regional cell densities have altered due to chattering (Fig. 4d), folding (Fig. 5d) and thickness variation (Figs. 4e and 5e) artefacts, AI-FFPE, by increasing the contrast between the nucleus and the cytoplasm, allows enhanced visualization of individual cells in the thicker regions of the slides where cells appear in high density because of being stacked on top of each other. Moreover, AI-FFPE reduces the differences in tissue thickness by filling the artificially occurring blank areas due to the freezing process.

One of the other more notable image corrections that AI-FFPE implements on these microscopic tissue images is the recovery of staining quality by improving the colour intensity, contrast and spectrum (Figs. 4c and 5c). Also, frozen sectioning tends to produce drying artefacts, which hinders the pathologists' ability to discern cellular structures, such as nuclear architecture and cytoplasmic borders (Fig. 6), all of which are reversed by our method by restoring structural contrasts and improving the colouring quality. AI-FFPE also corrects artefacts that are exclusive to frozen sectioning such as the presence of blank areas due to the formation of ice crystals (Figs. 4f and 5f). Notably, while reversing such diverse types of artefact individually or in combination, our method does not seem to introduce any misleading structures (Figs. 4 and 5).

We also compared the artefact-correcting performance of our method (final version, without SR loss, without SAB) with generic image translation models of fast contrastive unpaired loss translational learning (FastCUT) and cycle-consistent GAN (CycleGAN) on the brain (Extended Data Fig. 1) and the lung (Extended Data Fig. 2) WSIs. The results show that our model, custom-designed for CS to FFPE translation, outperforms these generic image-translation models, and the modalities that we have added to our method substantially contributed to its artefact-correcting performance.

To quantify the AI-FFPE model's efficiency in the transformation of CS images to FFPE-style images, we employed Frechet inception distance (FID), a well-established metric of similarity to assess AI-generated images' proximity to the target domain images. We compared AI-FFPE with generic image translation models of CycleGAN, contrastive unpaired loss translational learning (CUT) and FastCUT using the dataset that is detailed in the 'Description of the datasets' sub-section. We found that AI-FFPE's FID values were the lowest; hence, images generated by the AI-FFPE model were the closest to the real FFPE images, both for brain (AI-FFPE, 29.81 versus CUT, 32.28; FastCUT, 34.42; CycleGAN, 69.43) and lung (AI-FFPE, 28.15 versus CUT, 35.49; FastCUT, 35.71; CycleGAN, 39.19) images (Supplementary Table 2a). To determine the statistical significance of these results, we first examined the distribution of the results with Kolmogorov–Smirnov test in Supplementary Table 2b. As the data were found to have normal distribution ($P > 0.05$ for all groups), we analysed the data using two-sample two-tailed t-test which established that the differences are statistically significant (see Supplementary Table 2c demonstrating that our model's performance surpasses these generic models in translation of the images from CS to FFPE domain).

VTTs

To further corroborate AI-FFPE's image translation efficiency, we designed a series of VTTs performed by 27 board-certified pathologists. In the VTTs, participating pathologists were asked to classify the images as real or synthetic. To identify any bias because of being asked to classify the images into two categories, we first performed two pseudo-VTTs that consist entirely of real images, one for the brain and the other for the lung. The pseudo-VTTs showed that, on average, the pathologists assigned around half of the real FFPE images to real FPPE category (54.42% for brain, 61.36% for lung) and classified the rest as synthetic (Supplementary Table 3a).

https://doi.org/10.1038/s41551-022-00952-9

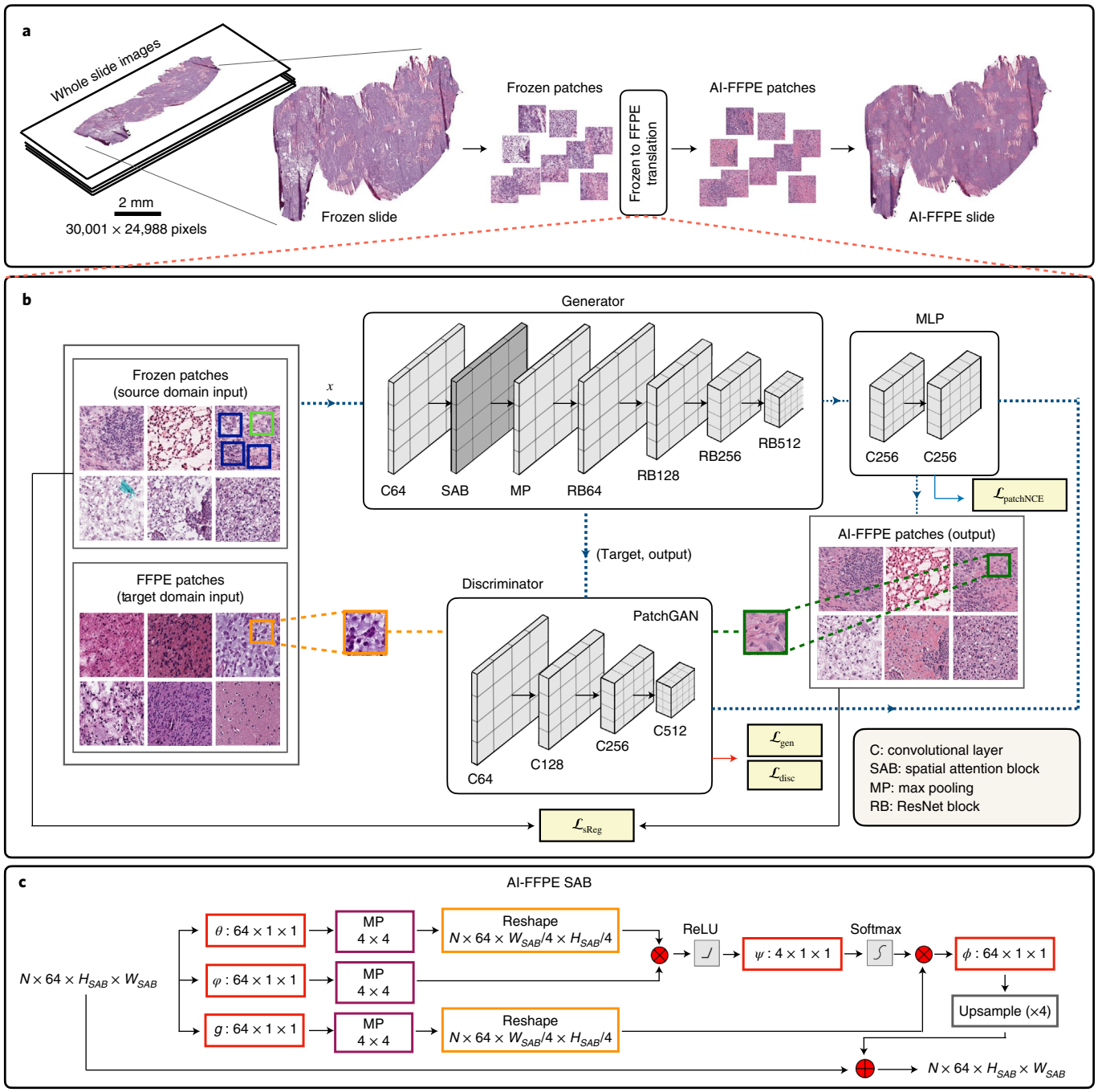


Fig. 2 | AI-FFPE method architecture. a, Process overview. CS-WSIs are first cropped into 512 × 512 pixel square patches for AI-FFPE networks to convert them into FFPE-like images which were, subsequently, stitched together to construct final WSIs to be examined by pathologists. **b**, A more detailed diagram of AI-FFPE networks. The input images from the source domain (CS) are fed into the SAB embedded in the ResNet-9 generator, consisting of ResNet Blocks (RB) with different filter sizes, where the filter sizes are indicated at the end of each RB, whose output is then served to the discriminator, consisting of convolutional layers (C) whose filter sizes are given as numbers at the end of each C (that is, C128 represents the convolutional block with the filter size 128) after being concatenated with the image from the target domain (FFPE). To promote the output images' resemblance to the target domain, we use the adversarial loss, which is a weighted summation of $\mathcal{L}_{gen}$ and $\mathcal{L}_{disc}$. To encourage content preservation by by making the networks concentrate on commonalities between two domains, SR and patch-wise noise contrastive estimation loss are integrated into the overall objective function. A two-layer multi-layer perceptron (MLP) network is added alongside the generator to improve the association between the input and the generated data. **c**, The flow diagram of SAB embedded in the ResNet-9 generator for the synthesis of the in silico alternative of FFPE sections.

We also conducted eight other VTTs with the same participants to compare the performance of AI-FFPE and generic image translation models in the literature (CycleGAN, FastCUT, CUT). A higher percentage of AI-FFPE-generated brain images are classified as real by the pathologists than those generated by the other image translation models (AI-FFPE, 54.84% versus CycleGAN, 44.84%; FastCUT, 47.34%; CUT, 50.56%) (Supplementary Table 3a). Slightly lower rates were found in the comparative VTTs with lung images but still with a more favourable rate for AI-FFPE (AI-FFPE, 52.28% versus CycleGAN, 42.5%; FastCUT, 45.7%; CUT, 46.96%) indicating that AI-FFPE produces more realistic FFPE images compared with the generic image translation models. We found slight agreement between pathologists on which images

https://doi.org/10.1038/s41551-022-00952-9

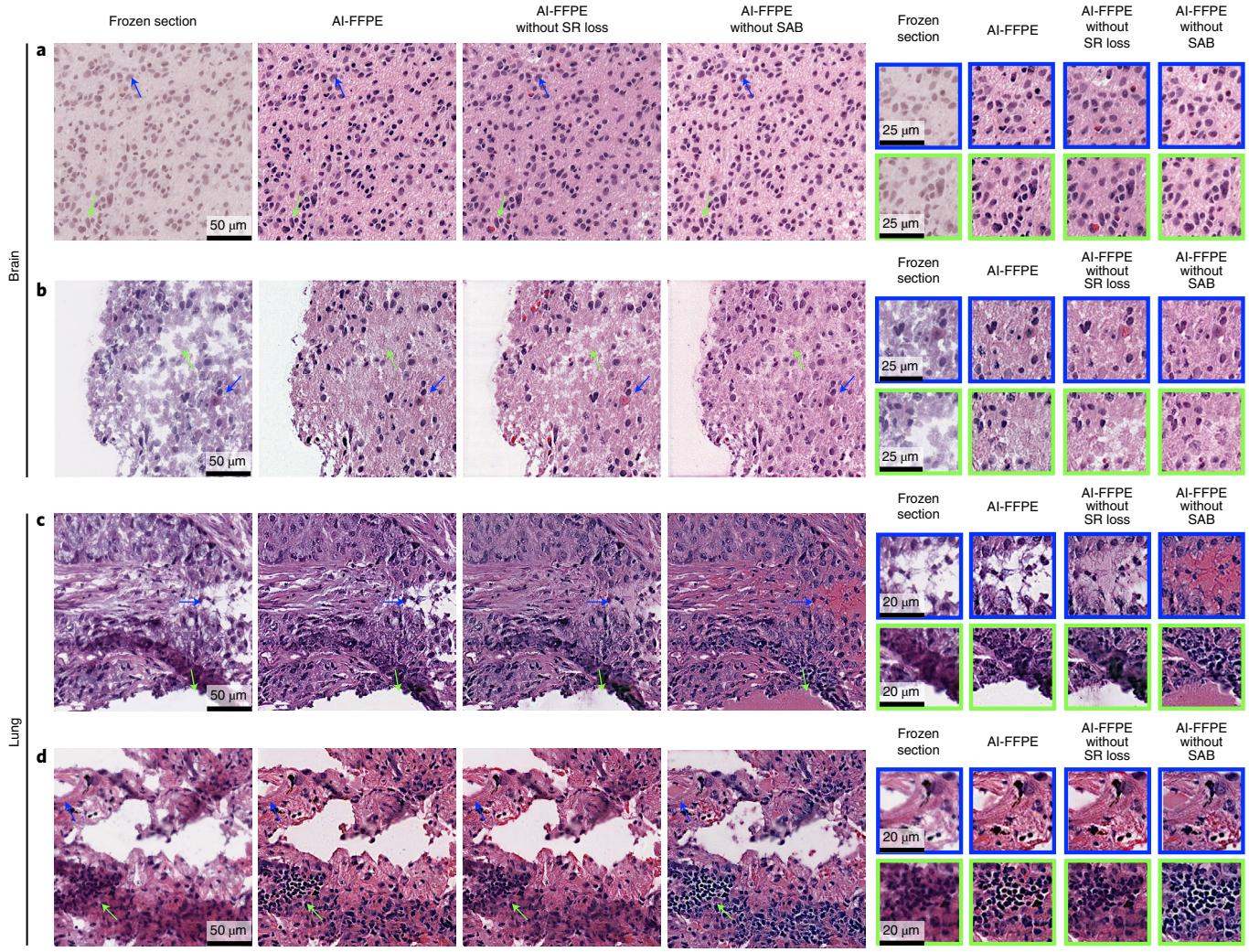


Fig. 3 | SAB and SR loss function effect on the model performance. a–d, AI-FFPE without SAB integration but with SR loss (AI-FFPE without SAB), and AI-FFPE without SR loss integration but with SAB (AI-FFPE without SR loss). The implementation of SR loss to the objective function prevents the network from adding clinically misguiding components such as erythrocytes (**a**,**b**), which might give a false impression of haemorrhage, enhances the appearance of nuclear borders (**d**) and improves overall staining quality (**a**). SR-loss implementation also repairs blank areas but without including textural details such as the presence of fibrils and lacks sufficient level of selectivity. With the integration of the SAB module, the model gains the ability to detect the regions (**b**,**c**) where ECM modifications should be avoided. When SR loss and SAB are combined, the images show not only cumulative improvement but also a synergistic effect such as adding fibrils and other textural components to the ECM and filling the areas that are lost by frozen sectioning (**c**). Sometimes, SR loss and SAB antagonize each other in image improvement; for example, SR alone is sometimes more efficient in enhancing nuclear appearance compared with SR loss–SAB combination (**c**, inside the picture frame). AI-FFPE also makes lymphocytes more prominent and hence easier to distinguish from the tumour cells (**d**). The SAB module also prevents the model from filling the vessels' lumen with ECM (**d**).

are in one category or the other (Supplementary Table 3b). As average values do not adequately reflect the difference in evaluations of the pathologists because the distribution and randomness in the factors stemming from the pathologist are not taken into account, we made further analysis using a generalized linear mixed-effects model. We compared the inclination of pathologists classifying AI-FFPE-generated images as real FFPE image versus the same decision probability for images generated by CycleGAN, FastCUT and CUT. The results are given in Supplementary Table 4. In brain VTTs, the percentage difference in odds is significantly in favour of AI-FFPE and reaching as high as 33.1%, 16.7% when compared with its closest competitor, CUT. The difference becomes more noticeable in lung tissue tests, ranging between 21.18% and 43.22% in favour of AI-FFPE, demonstrating the higher efficiency of AI-FFPE in translating the images to the FFPE domain in both tissues.

We also employed whole-slide-level VTTs. Fifty AI-FFPE WSIs and 50 FFPE WSIs from lung and brain are shown together to pathologists. Because the whole slides are structured as a stitch of thousands of patches even if a single cue exists in one patch, one can deduce whether the WSI is processed by AI. Therefore, the probability of identifying a WSI consisting of 1,000 patches as real is equal to the probability of identifying all 1,000 patches real at once. For the evaluation of the result, we have calculated the ratio of evaluating synthetic WSI as real and real WSI as real WSI by dividing the former with the latter for ten pathologists. The mean of the ratios was found to be 0.81 (with a 95% confidence interval of 0.70–0.90). This indicates that identifying a synthetic image as real is only 19% lower on average compared with identifying a real image as real. This result is much more promising than expected.

We later examined whether AI-based CS-to-FFPE image transformation leads to improvement of diagnostically valuable visual patterns, particularly of those that deteriorate due to CS. Figure 6 compiles examples of improved visual patterns of importance that are specific to each of the four cancer types that we examined in this

https://doi.org/10.1038/s41551-022-00952-9

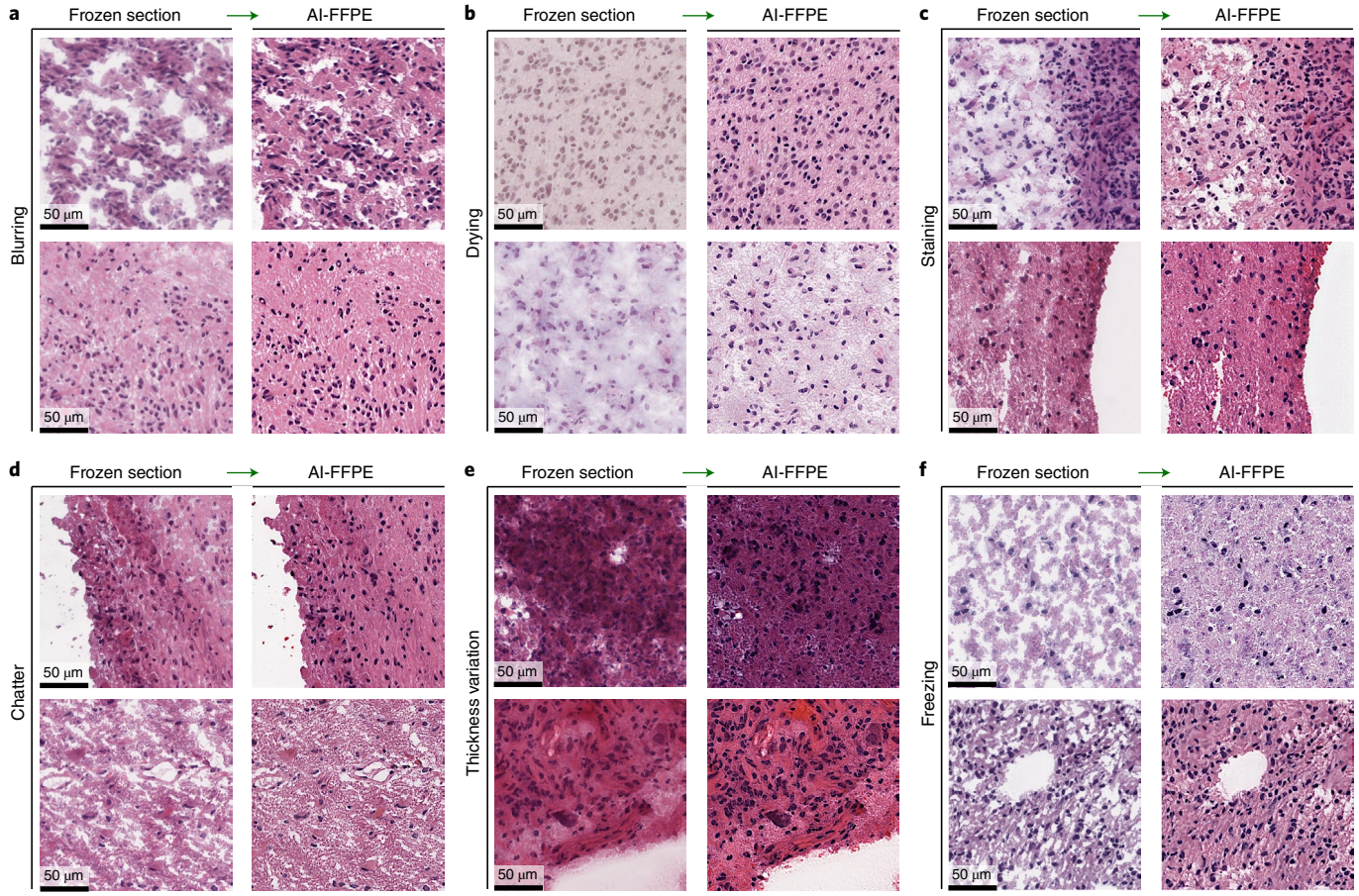


Fig. 4 | Improvement of artefacts in the brain frozen sections. Examples of AI-FFPE correction of different frozen sectioning artefacts. The patches were selected to represent one class of artefact; however, these artefacts do not exist in isolation, and different artefacts show overlapping features. AI-FFPE corrects these artefacts simultaneously. **a**, Reversal of blurring artefact with rectification of the losses in the ECM. **b**, Distorted cellular details due to drying artefacts are corrected in the AI-FFPE column. Improvements in the staining quality and ECM texture are also present in the same column. **c**, Correction of staining artefacts by increasing colour intensity, range and contrast, accompanied by minor improvements in the ECM. **d**, Correction of chatter artefacts by ECM restoration in the thinner areas of the tissue and the improvement of cellular details in all regions. **e**, Thickness variation artefacts are improved by enhancing nuclear borders in the thicker parts and the repair of ECM gaps in the thinner regions. **f**, Blank areas appear because large ice crystal formations are restored. AI-FFPE also sharpens nuclear borders and improves staining quality.

study. Increased nuclear colour intensity (hyperchromasia), diverse nuclear morphology (pleomorphism) and substantially increased mitotic rate manifested with the presence of mitotic figures are the differentiating histological features of glioblastoma multiforme (GBM), the highest grade of gliomas, from lower-grade gliomas. A lower-grade glioma, however, also has to be differentiated from gliosis, a benign condition. Nevertheless, in many cases, differential diagnosis based on haematoxylin-and-eosin staining alone could be difficult and requires further molecular characterization, which is a time-consuming process to inform intra-operative decision-making. In other cases of low-grade glioma (LGG), the presence of certain neoplastic patterns, such as nuclear atypia, mitotic figures and neovascularization can inform diagnosis using haematoxylin-and-eosin staining alone. AI-FFPE makes such patterns in CS images more noticeable by improving the appearance of the smallest vessels (capillaries) and areas surrounding these vessels (pericapillary area), the ECM and structures in the nuclei (chromatin and nucleoli). In the lung, AI-FFPE also improves the clarity of diagnostic patterns for adenocarcinomas and squamous cell carcinomas. For adenocarcinomas, rectification of freezing artefacts in tumour stroma, enhancing nuclear details of the tumour cells and overall improvement in distinction of stroma and tumour cells make features of diagnostic significance more visible, whereas for squamous carcinomas, AI-FFPE's improvement of epithelial features, such as squamous appearance, non-keratinized pattern and the presence of intercellular bridges, makes squamous cancer-specific features more prominent and easily recognizable. The diversity in types of improvement shows AI-FFPE's ability to repair and enhance cancer-type specific diagnostic patterns.

Evaluation of FFPE-trained models on AI-FFPE

Finally, we assessed whether the results we presented above would render an increased diagnostic performance by comparing the classification performance of our recently published clustering-constrained attention multiple instance learning (CLAM) algorithm[20] with the inputs of AI-FFPE pre-processed and regular CS WSIs. For classification of non-small-cell lung cancers into adenocarcinoma or squamous cell carcinoma subtypes, CLAM achieved a significantly higher test area under the curve (AUC) of 0.952 on AI-FFPE WSIs, compared with an AUC of 0.9061 on CS WSIs. For the subtyping of gliomas as second-grade (LGG) and fourth-grade (GBM) tumours, the classification model has benefited from AI-FFPE's image improvements, achieving an AUC score of 0.9837 for AI-FFPE WSIs, significantly higher than the AUC score of 0.9122 for CS WSIs.

Reader study on an independent test cohort

WSIs can vary greatly in image appearance due to different standards and protocols for tissue processing, slide preparation and digitization

https://doi.org/10.1038/s41551-022-00952-9

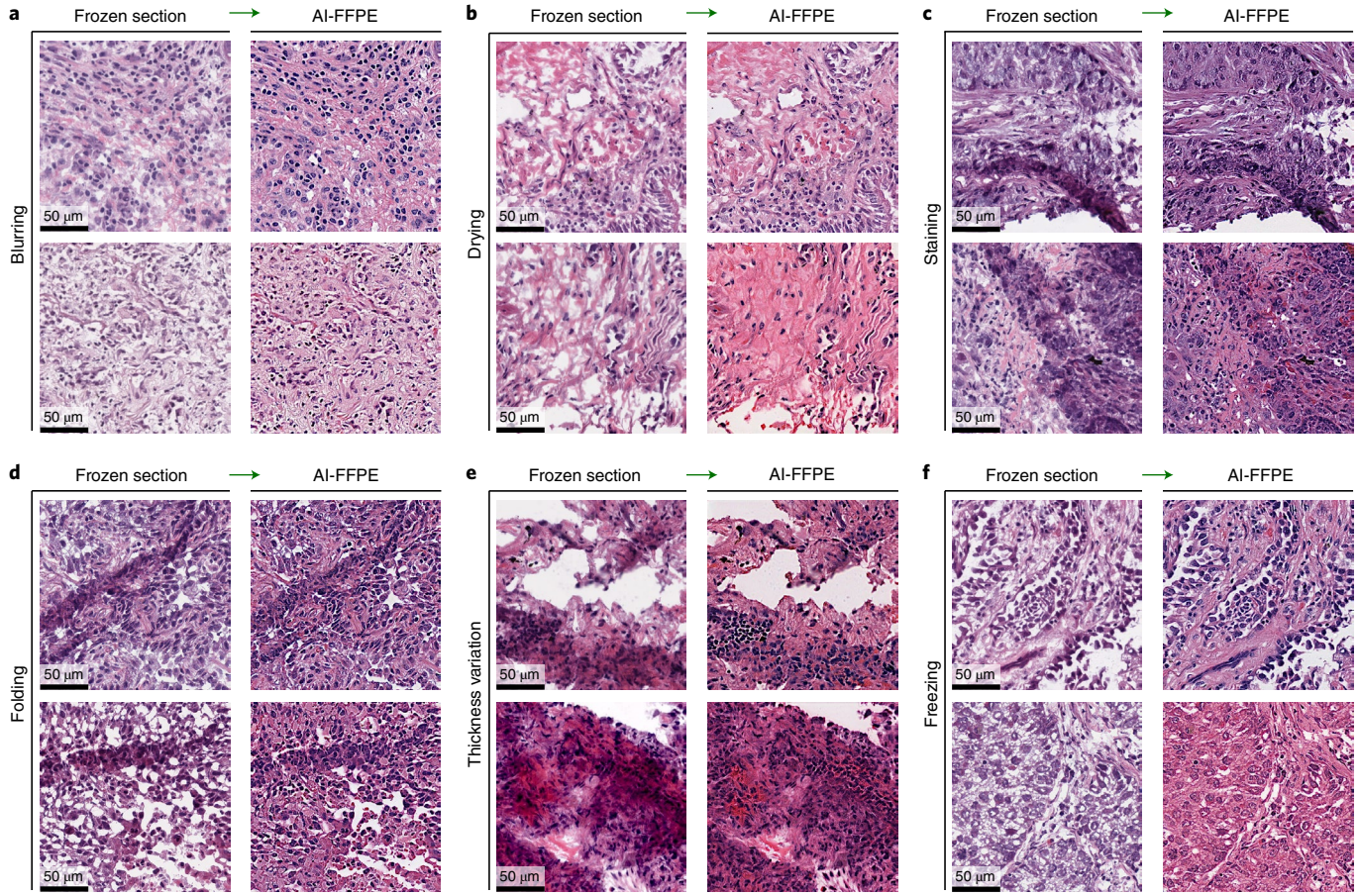


Fig. 5 | Improvement of artefacts in the lung frozen sections. Similar to the brain images, artefacts highlighted in each patch coexist with other types of artefact. **a**, AI-FFPE enhanced the resolution of blurred patches. Other improvements, such as ECM repair and improved staining quality, are also present. **b**, Compromised cellular details due to drying artefacts and distorted ECM are restored together. **c**, Improved colouring of the images with folding and thickness variation artefacts. AI-FFPE also significantly rectified image quality issues that occur due to the latter two. **d**, Correction of folding artefacts improved the appearance of the glandular pattern. The areas where ECM is lost are repaired. **e**, Augmentation of the appearance of cells in the thicker areas. Upper row patches also show enhanced visualisation of tissue infiltrating lymphocytes. **f**, Empty spaces due to ice crystal formation in the ECM (both patches) and inside the cells (lower patch) are corrected, and the staining quality is improved.

between different laboratories. Therefore, it is important to validate that the AI-FFPE models trained on The Cancer Genome Atlas (TCGA) (USA) are robust to data-source-specific variables and can be generalized to real-world clinical data from the sources that are not encountered during the training. Therefore, we collected and scanned a total of 132 brain samples (GBM, 90; LGG, 42) from Sağlik Bilimleri University, Kartal Dr. Lütfi Kırdar City Hospital (Turkey) and 166 lung samples (lung adenocarcinoma (LUAD), 93; lung squamous cell carcinoma (LUSC), 73) from Ege University Faculty of Medicine (Turkey) as independent test cohorts to evaluate the generalization performance of our TCGA-trained AI-FFPE models. Demographic statistics of all datasets are provided in Supplementary Tables 5 and 6. We conducted a reader study using WSIs from the independent cohort to evaluate whether AI-FFPE improves the diagnostic accuracy. Participating pathologists were asked to diagnose CS-WSIs, AI-FFPE WSIs generated from the same CS-WSIs and FFPE-WSIs from the same tumour. We found that AI-FFPE improves the odds of accurate tumour subtyping of LGG and GBM by 19% and of LUSC and LUAD by 16%.

Discussion

Computational histopathology offers tools to improve and evaluate microscopic tissue images, and there have been a growing number of studies focusing on histopathological applications of AI. These studies range from developing new tools to assist pathologists by automatized counting elements of interest on tissue slides to computational algorithms that guide the pathologist to diagnostically relevant regions, and attempts to replace chemical staining of tissues with virtual staining to tumour sub-typing by AI. However, improvement of frozen section image quality, which is inherently susceptible to deterioration because of the sample preparation techniques, has remained as an unexplored territory with the exception of a previous attempt to translate CS renal cell carcinoma images to FFPE with a CycleGAN model trained and tested on very limited data (20 cases for each)[42]. In our study, we developed an AI algorithm called AI-FFPE that efficiently resolves many frozen-sectioning-related image quality issues in brain and lung tissues. Frozen sectioning introduces artefacts with diverse characteristics. Some artefacts obscure cellular and ECM details, and a few others (folding, chattering and thickness variation) produce uneven cellular density; while freezing artefacts introduce blank spaces into the tissue, and staining artefacts compromise colouring quality. Our method's efficiency in correcting the aforementioned artefacts of different nature demonstrates its versatility, which is also reflected in its ability to highlight patterns of diagnostic importance in different types of tumour and tissue. For example, AI-FFPE's transformation enhanced such diagnostic patterns for lung adenocarcinomas and squamous carcinomas; both are epithelial in origin but exhibit distinct patterns of diagnostic importance. Despite the distinct architecture and embryonic origins of brain gliomas from the lung carcinomas,

https://doi.org/10.1038/s41551-022-00952-9

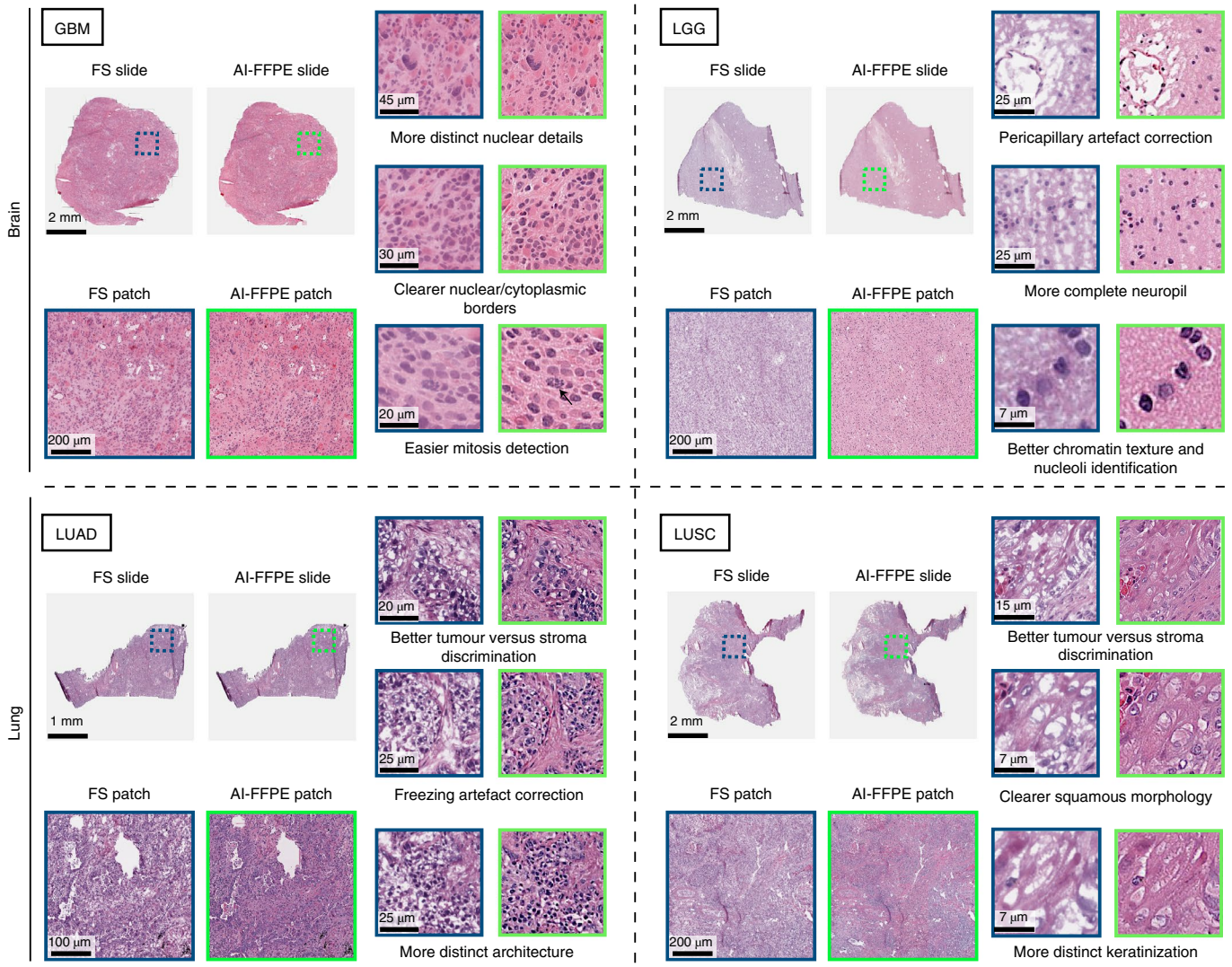


Fig. 6 | AI-FFPE enhancement of tumour-specific diagnostic patterns. Comparison of frozen section (FS) versus AI-FFPE images demonstrating how AI-FFPE reveals and augments diagnostic characteristics in GBM, LGG, LUAD and LUSC slide images. AI-FFPE's CS artefact correction methods such as image sharpening and rectification of empty spaces help clearer identification and evaluation of nuclear borders, diffuse growth pattern, neovascularization, necrosis and mitotic figures, allowing more precise grading and easier diagnosis of atypia for brain tumours. Similarly, AI-FFPE transformation of lung patches makes squamous-carcinoma- and adenocarcinoma-specific morphology more visible by correcting stromal freezing artefacts, producing more prominent nuclear details and stroma and tumour distinctions for the former and highlighting epithelial architecture and cellular characteristics for the latter.

AI-FPPE also improves the appearance of diagnostically significant patterns in brain glioma WSIs. These results suggest promising prospects for the application of AI-FFPE to WSIs of other tumours and organs and WSIs with more rare artefactual alterations that we have not assessed here. Our cross-organ adaptability experiments showed that AI-FFPE trained on one organ (brain or lung) can improve the images from the other (Extended Data Fig. 3) but not to an extent of success that is achieved by models that are trained on the same organ. Fine tuning of the models with images from the other organ(s) could resolve the sub-optimal cross-organ adaptability. Further studies involving other tumours of brain and lung, and other organs and artefact types, are required to show malleability of our method to the aforementioned extended applications.

In this work, we also performed comparative studies of AI-FFPE with generic image translation models, such as CUT, FastCUT and CycleGAN. Our model showed better performance as demonstrated by the comparative FID scores and output image quality of slide patches that harbour various types of artefact. Furthermore, we also conducted extensive VTTs to determine whether the FID results translate into meaningful performance difference in generating FFPE-style images when evaluated by expert eyes. Analysis of the VTT results shows that AI-FFPE-generated images are significantly more likely to be classified as real FFPE images than the images generated by the generic image translation models. As documented in the corresponding figures (Fig. 3 and Extended Data Figs. 1 and 2), the integration of SR loss and SAB into the final AI-FFPE version seems to greatly contribute to the better performance observed in these evaluations.

Crucially, our results not only establish clear improvements in the images but also demonstrate that the improvements culminate in increased diagnostic accuracy by showing that AI-FFPE-transformed images increase CLAM's cancer sub-typing performance. More notably, AI-FFPE significantly improves the rates of accurate tumour subtyping (LGG versus GBM in the brain, LUSC versus LUAD in the lung) of the pathologist in the reader study. These results can be further improved in future studies by addressing the potential failure cases that we presented in Extended Data Fig. 4. In future, prospective clinical studies can validate AI-FFPE's contribution to intra-operative diagnostic accuracy of CS samples and surgical decision-making in real hospital settings.

Methods

AI-FFPE method

Network architecture. We translate the images with the dimension (H, W, 3) from input CS domain, $\mathbf{X}$, $\mathbf{X} \subset \mathbb{R}^{H\times W\times 3}$ to appear like an image from the output domain, $\mathbf{Y}$, $\mathbf{Y} \subset \mathbb{R}^{H\times W\times 3}$. We are given a dataset of unpaired instances $\mathbf{x} \in \mathbf{X}$, $\mathbf{y} \in \mathbf{Y}$. Our method can operate even when $\mathbf{X}$ and $\mathbf{Y}$ only contain a single image each. Our method only requires learning the mapping in one direction and avoids using inverse auxiliary generators and discriminators. This can largely simplify the training procedure and reduce the training time. Our generator function, G, consists of encoder part, G_{enc}, and the decoder part, G_{dec}, which are applied sequentially to produce output image $G(\mathbf{x}) = G_{\text{dec}}(G_{\text{enc}}(\mathbf{x}))$. The generator is followed by a two-layer multi-layer perceptron (MLP). At the end, G_{disc} discriminates between the CS input image and the generated FFPE-like image.

Hyperparameters and training details. After randomly sampling slides, we have trained a GAN under the supervision of adversarial, contrastive and SR loss using a mini-batch size of one patch. To avoid adding clinically irrelevant information to the images, we enforced the network to create synthetic FFPE-like images with a closer content to frozen samples using a SR loss with the aim of preserving the spatial orientation of the nuclei and other diagnostically relevant features. However, the intensity and ratio of the integrated SR functionality are kept adjustable by a weight hyperparameter. Unlike the typical GAN network architecture that consists of ResNet-based generator and PatchGAN discriminator with the least square GAN loss of CUT models[37], our architecture employs a generator with artefact-aware attention block and patch-based self-aware contrastive loss. As our method simplifies the training procedure by operating just in one direction, that is, from source domain to target, the training time drastically decreases compared with traditional cycle consistency-based unpaired image-to-image translation methods in the literature. Standard adversarial loss, $\mathcal{L}_{\text{GAN}}$, is the first component of our hybrid loss function:

$$\mathcal{L}_{\text{GAN}}(G, G_{disc}, \mathbf{X}, \mathbf{Y}) = \mathbb{E}_{\mathbf{y}\sim\mathbf{Y}} \log G_{disc}(\mathbf{y}) + \mathbb{E}_{\mathbf{x}\sim\mathbf{X}} \log(1 - G_{disc}(G(\mathbf{x}))), \quad (1)$$

where the discriminator G_{disc} attempts to recognize whether the patch is a synthetically generated AI-FFPE sample by $G(\mathbf{x})$ or a real FFPE sample, $\mathbf{y}$. $\mathbb{E}_{\mathbf{y}\sim\mathbf{Y}}$ is the expected value over all real data instances and $\mathbb{E}_{\mathbf{x}\sim\mathbf{X}}$ is the expected value over all random inputs to the generator. The adversarial loss function dictates to learn and eliminate the style differences, which triggers the use of a noise contrastive estimation function that guarantees the content preservation in patch level[43]. Therein, we employ a patch-wise noise contrastive loss to find the resemblance between input CS and output image AI-FFPE patches, by taking an output patch from a generated image in the FFPE target domain and matching it with a corresponding frozen sample image patch at the same location with the expectation that they will form a positive pair, whereas other patches that are dissimilar will form negative pairs. One positive with corresponding input is being mapped to L-dimensional real vector space, $\mathbb{R}^L$, $v, v^+ \in \mathbb{R}^L$. M negative non-corresponding inputs are mapped to $M \times L$ dimensional real vector space, $\mathbb{R}^{M\times L}$, $v^- \in \mathbb{R}^{M\times L}$, where $v_k^- \in \mathbb{R}^L$ signifies the kth negatives in M samples. We normalize the vectors onto unit spheres to stop the space from collapsing or even growing. The problem is set as $(M+1)$-way classification, where scaling the distance between the output and the examples by a temperature $\nu = 0.07$ also passes as logit[44]. The probability of positive examples selected over the negative ones is formulated as a cross-entropy loss and is calculated as:

$$\mathcal{L}(v, v^+, v^-) = -\log\left[\frac{\exp(v \cdot v^+/\nu)}{\exp(v \cdot v^+/\nu) + \sum_{n=1}^{M} \exp(v \cdot v_n^-/\nu)}\right]. \quad (2)$$

where ν is the temperature scaling the distance between output samples. As per defined probability, $\mathcal{L}(v, v^+, v^-)$, the patches from the frozen section tissue boundary are expected to be more closely associated with the synthesized FFPE section's boundary than the patches from other regions of the slide. Once L layers of interest are selected from the generator, feature maps coming from the generator are given as input to the MLP, H_l, layers as introduced in SimCLR: a simple framework for contrastive learning of visual representations[45]. Similarly, synthesized images are encoded with these two networks as features, z, $\{\hat{z}\}_L = \{H_l(G^l_{\text{enc}}(G(x)))\}_L$ where the layers of encoder indexed by $l \in \{1, ..., L\}$. The patch-wise noise contrastive estimation loss, $\mathcal{L}_{\text{patchNCE}}$, is defined based on the final features, $\{z_l\}'s$:

$$\mathcal{L}_{\text{patchNCE}}(G, H, \mathbf{X}) = \mathbb{E}_{\mathbf{x}\sim\mathbf{X}} \sum_{l=1}^{L} \sum_{s=1}^{S_l} \mathcal{L}(\hat{z}_l^s, z_l^s, z_l^{S\setminus s}). \quad (3)$$

where $s \in S = \{1, ..., S_l\}$, and S_l is the number of spatial locations in each l layer. To prevent the network from adding any clinically misleading information, we penalize the deviations from real input images by a SR pixel-wise L_1 loss function, $\mathcal{L}_{\text{sReg}}$:

$$\mathcal{L}_{\text{sReg}}(G, \mathbf{X}) = ||\mathbf{X} - G(\mathbf{X})||_1 \quad (4)$$

For each patch, the final objective, $\mathcal{L}_{\text{AI-FFPE}}$, is a weighted sum of these loss functions with constant:

$$\begin{aligned}\mathcal{L}_{\text{AI-FFPE}} &= \mathcal{L}_{\text{GAN}}(G, D, \mathbf{X}, \mathbf{Y}) + \lambda_{\text{sReg}}\mathcal{L}_{\text{sReg}}(G, \mathbf{X}) \\ &\quad + \lambda_{\mathbf{X}}\mathcal{L}_{\text{patchNCE}}(G, H, \mathbf{X}) + \lambda_{\mathbf{Y}}\mathcal{L}_{\text{patchNCE}}(G, H, \mathbf{Y}).\end{aligned} \quad (5)$$

where $\mathbf{X}$ and $\mathbf{Y}$ stand for input CS and output FFPE domain, respectively. $\mathcal{L}_{\text{sReg}}$, $\lambda_{\mathbf{X}}$, and $\lambda_{\mathbf{Y}}$ are constant weight multipliers for corresponding loss functions. That way, we aim to achieve diagnostically more informative and interpretable images after translation by the network under the $\mathcal{L}_{\text{AI-FFPE}}$ supervision. Our model was trained using the Adam optimizer[46] with an initial learning rate of 0.0002 as well as the momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for five epochs, and the outputs throughout the iterations are given in Extended Data Fig. 5. We use a batch size of 1, instance normalization[47] and Xavier weight initialization[48]. ResNet with nine residual blocks[39] is chosen as a generator, PatchGAN as discriminator[40] and least square GAN loss[41]. $\mathcal{L}_{\text{sReg}}$ is set as 0.03, $\lambda_{\mathbf{X}}$ and $\lambda_{\mathbf{Y}}$ as 1 for identity loss with temperature (τ) 0.08 and enqueued 512 patches for each image for each iteration. For the sake of fairness, all compared methods share the same parameter set with the AI-FFPE in the training and testing phases.

SAB. The SAB mechanism can be seen as a non-local convolution operation extracting the relative weights of all positions on the feature maps for any given SAB input, $\mathbf{X}_{SAB}$, from the real vector space with dimensions (batch size, 64, the height of the $\mathbf{X}_{SAB}$, the width of the $\mathbf{X}_{SAB}$), $\mathbb{R}^{N\times 64\times H_{SAB}\times W_{SAB}}$, $\mathbf{X}_{SAB} \in \mathbb{R}^{N\times 64\times H_{SAB}\times W_{SAB}}$.

After applying convolutional operations θ and φ (followed by max-pooling), we employ the dot product operation on the output tensors, which is activated by the rectified linear unit (ReLU) function, σ_{relu}:

$$\mathbf{P} = \psi(\sigma_{\text{relu}}(\theta(\mathbf{X}_{SAB})\varphi(\mathbf{X}_{SAB})^{\top})). \quad (6)$$

The product, $\theta(\mathbf{X}_{SAB})\varphi(\mathbf{X}_{SAB})^{\top}$, provides input covariance measurement, which can be interpreted as a degree of inclination between two feature maps at different channels. After convolving with ψ, the output P is normalized via the softmax function, σ_{softmax}, to produce a spatial attention map that is multiplied with the output of the convolution operator g. Then, the result of multiplication is convolved by ϕ then upsampled to produce $\mathbf{S}$. Finally, an element-wise sum operation between $\mathbf{S}$ and the input $\mathbf{X}_{SAB}$ generates the output $\mathbf{F} \in \mathbb{R}^{N\times 64\times H_{SAB}\times W_{SAB}}$:

$$\mathbf{S} = \phi(\sigma_{\text{softmax}}(\mathbf{P})g(\mathbf{X}_{SAB})), \quad (7)$$

$$\mathbf{F} = \mathbf{S} + \mathbf{X}_{SAB}. \quad (8)$$

Short connection between the input $\mathbf{X}_{SAB}$ and $\mathbf{S}$ finalizes the block output $\mathbf{F}$ by strengthening the residual signals. The detailed flow diagram of block operations of the SAB module is given in Fig. 2c.

Description of the datasets

The TCGA open-source database was used to train and test the AI-FFPE algorithm. For the brain, TCGA-GBM (highest-grade brain tumours) and TCGA-LGG (low-grade brain tumours) datasets, consisting of tumours that belong to the most common cancer of brain—gliomas—and representing two distinct histopathological, biological and clinical patterns, were used. For the lung, TCGA-LUSC and TCGA-LUAD projects, which are composed of WSIs from two most common but histologically distinct lung cancer types, were utilized. We used a subset of the dataset from these projects. Our subset consisted of 97,271 (46,058 GBM, 51,213 LGG) frozen and 110,087 (52,435 GBM, 57,652 LGG) FFPE patches from 590 (294 GBM, 296 LGG) patients for the brain, and 135,785 (68,789 LUSC, 66,996 LUAD) frozen and 71,311 (34,599 LUSC, 36,712 LUAD) FFPE patches from 650 (323 LUSC, 327 LUAD) patients for the lung.

Description of the datasets of the independent test cohorts. A total of 132 brain samples (GBM, 90; LGG, 42) from Sağlik Bilimleri University, Kartal Dr. Lütfi Kırdar City Hospital (Turkey) and 166 lung samples (LUAD, 93; LUSC, 73) from Ege University Faculty of Medicine (Turkey) are collected as independent test cohorts for evaluating the generalization performance of our trained models. The brain cohort includes 66 cryosections and the lung cohort includes 83 cryosections, and each patient might have several sections. For the brain dataset, the demographic distribution is as follows: sex, 18 females, 23 males; cancer type, 29 GBM, 12 LGG; biopsy years, 2 in 2017, 4 in 2018, 20 in 2019, 3 in 2020, 12 in 2021. For the lung dataset, the demographic distribution is as follows: sex, 12 females, 71 males; cancer type, 47 cases of LUAD, 36 cases of LUSC; biopsy years, 2 in 2015, 10 in 2016, 5 in 2017, 30 in 2018, 13 in 2019, 23 in 2020. A demographic summary of the in-house dataset is included in Supplementary Table 5. Apart from descriptive statistics, the observed artefacts in the datasets are also given in Supplementary Table 1 for the comparative statistics of artefact frequency in CS and FPPE slides in our in-house and TCGA datasets.

Tissue preparation for the independent test cohorts. Each sample collected in the independent test cohort is divided into two samples, one for the preparation of frozen section slides and the other for the FFPE slides. CS and FFPE sections that are closely matched to each other are selected for further processing. All in-house slides are later digitized at Ege University using an Aperio AT2 scanner at ×20 magnification.

Segmentation and patching of WSI. The biopsy tissue in each WSI were segmented using the CLAM WSI analysis toolbox[20]. A binary tissue mask denoting the tissue and non-tissue regions were computed for each downsampled input image in hue, saturation, value (HSV) colour space by thresholding the saturation channel median blurring. The estimated contours of the denoted tissue and the cavities of tissue were then filtered depending on their area to generate the final segmentation mask. The model was trained on ×20 magnifications which was segmented into 512 × 512 pixel patches without overlap using the segmentation contour.

After processing frozen patches individually, AI-FFPE patches are stitched back to reconstruct the WSIs. The model is trained at the patch level, which allows the training dataset size to be scaled up and makes the process memory efficient.

Patch-level evaluations. For the quantitative assessment of the GAN performance at patch level, the FID[49,50] is utilized as a metric that is most compatible with human perception. The multidimensional Gaussian distribution of the real FFPE images and the generated AI-FFPE images in a deep network space as well as the difference of the mean and standard deviations of these two distributions are computed. As the generated images start becoming more realistic throughout the iterations, their statistics resemble the real FFPE images from the target domain, and the FID score decreases gradually. The FID, $\mathcal{I}_{\mathrm{FID}}$, can be formulated as:

$$\mathcal{I}_{\mathrm{FID}} = \| \mu_1 - \mu_2 \|_2^2 + \mathrm{Tr}(C_1 + C_2 - 2 \times \sqrt{(C_1 \times C_2)}) \tag{9}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ refer to the feature-wise mean of the real and generated images, C_1 and C_2 are co-variance matrices for the real and generated feature vector and Tr stands for trace function, for example, the sum of element along the main diagonal of the square matrix.

VTTs

Comparative VTTs. Eight VTTs are designed to test the comparative performance of AI-FFPE, CUT, FastCUT and CycleGAN on brain and lung patches. Each VTT consists of 100 patches from a single tissue (brain or lung) and includes equal numbers (50) of LUAD and LUSC patches for lung VTT, and LGG and GBM patches for brain VTT. Half of the 50 patches for each cancer type consist of real images, and the others comprise output images from the corresponding model. The same 27 board-certified pathologists who participated in the pseudo-VTTs participated in the comparative VTTs.

Pseudo-VTTs. Two separate pseudo-VTTs are developed for each tissue. Each test comprises 100 real patches: 50 LUAD and 50 LUSC for the lung baseline VTT, and 50 LGG and 50 GBM for the brain pseudo-VTT. Twenty-seven board-certified pathologists participated in the VTTs.

We performed Fleiss Kappa statistics to evaluate the inter-observer agreement between the participating pathologists. Before the VTTs, to determine the required number of pathologists and samples, we employed power analysis as explained in detail in the Fleiss' Kappa power analysis subsection.

Fleiss' Kappa power analysis. When the numbers of ratings per subject are equal, Fleiss, Nee and Landis[51] derived and confirmed the following formulae for the approximate standard errors of Fleiss' Kappa, $\hat{\kappa}$, appropriate for testing the hypothesis that the underlying value is zero:

$$\widehat{\mathrm{s.e.}_0}(\hat{\kappa}) = \frac{\sqrt{2}}{\sum_{j=1}^{k} \bar{p}_j \bar{q}_j \sqrt{nm(m-1)}} \times \sqrt{\left(\sum_{j=1}^{k} \bar{p}_j \bar{q}_j\right)^2 - \sum_{j=1}^{k} \bar{p}_j \bar{q}_j (\bar{q}_j - \bar{p}_j)} \tag{10}$$

where k is the number of categories, m is the number of raters, n is the number of subjects (patches) and $\bar{p}_j$ ($\bar{q}_j = 1 - \bar{p}_j$) is the overall proportion of ratings in category j. The hypothesis may be tested by referring to the critical ratio:

$$Z = \frac{\hat{\kappa} - \kappa}{\widehat{(\mathrm{s.e.}_0)}(\hat{\kappa})} \sim \mathcal{N}(0,1) \tag{11}$$

$$1 - \textbf{power} = \mathbb{P}(Z < z_\beta) \tag{12}$$

$$1 - \alpha = \mathbb{P}(Z < z_\alpha) \tag{13}$$

where α is the probability of rejecting the null hypothesis even though it is accurate and z_α is the critical value based on Type-I error and z_β is the critical value based on Type-II error. If the following inequality is satisfied, then one can deduce the sufficiency of the sample size of the test for the given number of pathologists:

$$z_\beta > \frac{z_\alpha \widehat{\mathrm{s.e.}_0}(\hat{\kappa}) - \kappa_1}{\widehat{\mathrm{s.e.}_0}(\hat{\kappa})}. \tag{14}$$

Therefore, to reach the 95% test power that is the probability of detecting agreement between pathologists when there is an agreement, we need to assess the Kappa values on 19 images from each group for 27 pathologists (for further analysis, see Supplementary Table 7).

Generalized linear mixed-effects model to compare the tendency of identifying synthetic images as real based on VTTs. We build a generalized mixed-effects model to compare how pathologists who classify synthetic WSI as real differ from each other based on the generation methods: FastCUT, CUT and CycleGAN:

$$\begin{aligned}\log\left(\frac{\mathbb{P}(Z=1)}{\mathbb{P}(Z=0)}\right) = \beta_0 &+ \beta_1 I(\text{Slide} = \text{Synth}_{\text{CycleGAN}}) \\ &+ \beta_2 I(\text{Slide} = \text{Synth}_{\text{FastCUT}}) \\ &+ \beta_3 I(\text{Slide} = \text{Synth}_{\text{CUT}}) \\ &+ W\end{aligned} \tag{15}$$

where Z is a Bernoulli random variable that stands for the truth value of pathologists who answer for the synthetic images in the VTTs, Slide stands for an image in the test and $W \sim \mathcal{N}(0, \sigma^2)$ represents the randomness effects that stem from the pathologists. The synthetic images generated by CycleGAN, FastCUT, and CUT are represented by $\text{Synth}_{\text{CycleGAN}}$, $\text{Synth}_{\text{FastCUT}}$ and $\text{Synth}_{\text{CUT}}$, respectively. The indicator functions, I, for the three compared groups are defined as follows:

$$I(\text{Slide} = \text{Synth}_{\text{FastCUT}}) = \begin{cases} 1 & \text{if the patch is synthetic in FastCUT test}, \\ 0 & \text{otherwise}. \end{cases} \tag{16}$$

$$I(\text{Slide} = \text{Synth}_{\text{CycleGAN}}) = \begin{cases} 1 & \text{if the patch is synthetic in CycleGAN test}, \\ 0 & \text{otherwise}. \end{cases} \tag{17}$$

$$I(\text{Slide} = \text{Synth}_{\text{CUT}}) = \begin{cases} 1 & \text{if the patch is synthetic in CUT test}, \\ 0 & \text{otherwise}. \end{cases} \tag{18}$$

$$\text{Differences} = \begin{cases} 100(e^{(-\text{Estimate})} - 1) & \text{if Estimate} < 0, \\ 100(e^{(\text{Estimate})} - 1) & \text{otherwise}. \end{cases} \tag{19}$$

We considered two-sided hypothesis tests to determine whether there is any statistically significant difference between our method and CycleGAN (H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$), CUT (H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$) and FastCUT (H_0: $\beta_3 = 0$ versus H_1: $\beta_3 \neq 0$). Null hypotheses were rejected if the associated P values are less than the significance level of 0.05.

WSI VTT. Fifty AI-FFPE WSIs and 50 FFPE WSIs from lung and brain are shown together to pathologists. As the AI-FFPE WSIs shrink throughout the AI-FFPE pipeline as described above in the WSI processing section, we also employed the same steps for the FFPE slides to equate the conditions. First, they were divided into patches by down-scaling and then were stitched back as whole slides without any modification by the AI-FFPE model.

Reader study

Study design. The reader study is based on the images collected in the independent cohort described above. Pathologists were asked to diagnose frozen-section WSIs, AI-FFPE-improved versions of the same CS images and the FFPE images from the same tumours. FFPE and CS images were obtained using the method detailed above in the 'Tissue preparation for the independent test cohorts' section to produce CS to FFPE sections closely matching each other. Twenty-five patients having one CS, one FFPE and one AI-FFPE WSIs for each cancer subtype (LUAD, LUSC, LGG and GBM) were evaluated by each pathologist, totalling 300(= 25 × 3 × 4) WSIs.

Generalized linear mixed-effects model. We have built a logistic regression to model the probability of giving the true diagnosis on CS and AI-FFPE WSIs.

$$\log\left(\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}\right) = \beta_0 + \beta_1 I(\text{Slide} = \text{CS}) + U + V \tag{20}$$

where Y is a Bernoulli random variable that stands for the truth value of pathologists' diagnosis.

$$I(\text{Slide} = \text{CS}) = \begin{cases} 1 & \text{if the slide is CS}, \\ 0 & \text{if the slide is AI-FFPE}. \end{cases} \tag{21}$$

$U \sim \mathcal{N}(0, \sigma_1^2)$ random variable stands for the random effects of pathologists and $V \sim \mathcal{N}(0, \sigma_2^2)$ for the random effects of the patients. U and V are included in the model to account for the dependencies between the repeated answers from the same pathologists and the repeated slides from the same patients, respectively. To prove the power of our method on the diagnostic accuracy improvement, one has to reject the H_0 hypothesis claiming that β_1 is equal to 0. If one fails to reject the null hypothesis, then β_1 is equal to 0 implies that there is no statistically significant difference for diagnostic accuracy between frozen and AI-FFPE slides.

Details of the CLAM evaluations. The vanilla CLAM[20] pipeline was used for training the classification models. During training, slides are randomly sampled and provided to the model using a batch size of 1. Weights and bias parameters of the attention module are initialized randomly and trained end to end with the rest of the model using the slide-level labels as no ground truth attention is available. Adam optimizer was used with a learning rate of 2×10^{-4} and ℓ_2 weight decay of 1×10^{-5}. Default values for computing the running averages of the first and second moments of the gradient were used (that is, $\beta_1 = 0.9$ and $\beta_2 = 0.999$), and ϵ term (for numerical stability) was set to 1×10^{-8}. All models in the study were trained for at least 50 epochs and up to a maximum of 200 epochs if the early stopping criterion is not met.

Computational hardware and software. WSIs were processed on Intel Xeon multi-core central processing units and two NVIDIA 2080 Ti graphics processing units using the publicly available CLAM[20] whole-slide processing pipeline implemented in Python (version 3.7.5). Deep-learning models were trained on NVIDIA GeForce RTX 3090 graphics processing units using Pytorch (version 1.7.0).

Ethics oversight

The study was approved by the Ege University Ethics Committee(s) (reference E-99166796-050.06.04-425014 and ID number 21-11.1T/45).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The TCGA diagnostic whole-slide data (GBM, LGG, LUAD and LUSC) and the corresponding labels are available from the NIH genomic data commons (https://portal.gdc.cancer.gov). Restrictions apply to the availability of in-house data, which were used with institutional permission for the purposes of this project. All requests for access to in-house data may be addressed to the corresponding authors and will be processed in accordance with institutional guidelines. Data can only be shared for academic research purposes and will require a material-transfer or data-transfer agreement with the receiving institution.

Code availability

All codes were implemented in Python using PyTorch as the primary deep-learning library. The complete pipeline for processing whole-slide

images and for training and evaluating the deep-learning models is available at the AI-FFPE repository at https://github.com/DeepMIALab and can be used to reproduce the experiments reported in this paper.

References

1. Brown, R. W. *Histologic Preparations: Common Problems and Their Solutions* (College of American Pathologists, 2009).
2. Jaafar, H. Intra-operative frozen section consultation: concepts, applications and limitations. *Malays. J. Med. Sci.* **13**, 4–12 (2006).
3. Oh, E. et al. Comparison of accuracy of whole-exome sequencing with formalin-fixed paraffin-embedded and fresh frozen tissue samples. *PLoS ONE* **10**, e0144162 (2015).
4. Pichat, J., Iglesias, J. E., Yousry, T., Ourselin, S. & Modat, M. A survey of methods for 3D histology reconstruction. *Med. Image Anal.* **46**, 73–105 (2018).
5. Renne, S., Redaelli, S. & Paolini, B. Cryoembedder, automatic processor/stainer, liquid nitrogen freezing, and manual staining for frozen section examination: a comparative study. *Acta Histochem.* **121**, 761–764 (2019).
6. Patil, P., Shukla, S., Bhake, A. & Hiwale, K. Accuracy of frozen section analysis in correlation with surgical pathology diagnosis. *Int. J. Res. Med. Sci.* **3**, 399 (2015).
7. Bittar, H., Incharoen, P., Althouse, A. & Dacic, S. Accuracy of the IASLC/ATS/ERS histological subtyping of stage I lung adenocarcinoma on intraoperative frozen sections. *Mod. Pathol.* **28**, 1058–1063 (2015).
8. Rogers, C., Klatt, E. C. & Chandrasoma, P. Accuracy of frozen-section diagnosis in a teaching hospital. *Arch. Pathol. Lab. Med.* **111**, 514–517 (1987).
9. Cho, H. J., Lim, S., Choi, G. & Min, H. Neural stain-style transfer learning using GAN for histopathological images. *JMLR: Workshop and Conference Proceedings* **80**, 1–10 (2017).
10. Tofte, K., Berger, C., Torp, S. & Solheim, O. The diagnostic properties of frozen sections in suspected intracranial tumors: a study of 578 consecutive cases. *Surg. Neurol. Int.* **5**, 170 (2014).
11. Adesina, A. M. Frozen section diagnosis of pediatric brain tumors. *Surg. Pathol. Clin.* **3**, 769–796 (2010) .
12. Predina, J., Keating, J., Patel, N., Nims, S. & Singhal, S. Clinical implications of positive margins following non-small cell lung cancer surgery. *J. Surg. Oncol.* **113**, 264–269 (2015).
13. Marchevsky, A. M. et al. Frozen section diagnoses of small pulmonary nodules: accuracy and clinical implications. *Ann. Thorac. Surg.* **78**, 1755–1759 (2004).
14. Zin, A. A. M. & Zulkarnain, S. Diagnostic accuracy of cytology smear and frozen section in glioma. *Asian. Pac. J. Cancer Prev.* **20**, 321–325 (2019).
15. Obeidat, F. et al. Accuracy of frozen-section diagnosis of brain tumors: an 11-year experience from a tertiary care center. *Turk. Neurosurg.* **29**, 242–246 (2018).
16. Xiang, Z. et al. An effective inflation treatment for frozen section diagnosis of small-sized lesions of the lung. *J. Thorac. Dis.* **12**, 1488–1495 (2020).
17. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
18. Chen, P.-H. C. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453–1457 (2019).
19. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1 (2019).
20. Lu, M. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
21. Goodfellow, I. et al. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **3**, 2672–2680 (2014).
22. Creswell, A. et al. Generative adversarial networks: an overview. *IEEE Signal Process. Mag.* **35**, 53–65 (2017).
23. Bentaieb, A. & Hamarneh, G. Adversarial stain transfer for histopathology image analysis. *IEEE Trans. Med. Imaging* **37**, 792–802 (2018).
24. Bobrow, T. L., Mahmood, F., Inserni, M. & Durr, N. J. Deeplsr: a deep learning approach for laser speckle reduction. *Biomed. Opt. Express* **10**, 2869–2882 (2019).
25. Almalioglu, Y. et al. Endol2h: deep super-resolution for capsule endoscopy. *IEEE Trans. Med. Imaging* **39**, 4297–4309 (2020).
26. Mahmood, F., Chen, R. J. & Durr, N. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE Trans. Med. Imaging* **37**, 2572–2581 (2018).
27. Rivenson, Y. et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nat. Biomed. Eng.* **3**, 466–477 (2019).
28. de Haan, K. et al. Deep learning-based transformation of H&E stained tissues into special stains. *Nat. Commun.* **12**, (2021).
29. Sorin, V., Barash, Y., Konen, E. & Klang, E. Creating artificial images for radiology applications using generative adversarial networks (GANS)—a systematic review. *Acad. Radiol.* **27**, 1175–1185 (2020).
30. Siller, M. et al. On the acceptance of "fake" histopathology: a study on frozen sections optimized with deep learning. *J. Pathol. Inform.* **13**, 6 (2022).
31. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)* 2242–2251 (2017).
32. Benaim, S. & Wolf, L. One-sided unsupervised domain mapping. In *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, 2017). https://proceedings.neurips.cc/paper/2017/file/59b90e1005a220e2ebc542eb9d950b1e-Paper.pdf
33. Amodio, M. & Krishnaswamy, S. Travelgan: image-to-image translation by transformation vector learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8975–8984 (2019).
34. Fu, H. et al. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2242–2251 (IEEE, 2019).
35. Liu, M.-Y., Breuel, T. & Kautz, J. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, 2017). https://proceedings.neurips.cc/paper/2017/file/dc6a6489640ca02b0d42dabeb8e46bb7-Paper.pdf
36. Huang, X., Liu, M.-Y., Belongie, S. & Kautz, J. Multimodal unsupervised image-to-image translation. In *Computer Vision – ECCV 2018* (eds. Ferrari, V., Hebert, M., Sminchisescu, C. & Weiss, Y.) 179–196 (Springer, 2018).
37. Park, T., Efros, A., Zhang, R. & Zhu, J.-Y. Contrastive learning for unpaired image-to-image translation. In *ECCV* 319–345 (2020).
38. Wu, Z., Xiong, Y., Yu, S. X. & Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3733–3742, (IEEE, 2018).
39. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016).
40. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5967–5976 (2017).

41. Mao, X. et al. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)* 2813–2821 (2017).
42. Falahkheirkhah, K. et al. A generative adversarial approach to facilitate archival-quality histopathologic diagnoses from frozen tissue sections. *Lab. Invest.* **102**, 554–559 (2021).
43. Gutmann, U. & Hyvärinen, A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Eds. Teh, Y.W. and Titterington, M.). PMLR **9**, 297–304, (Proceedings of Machine Learning Research, 2010).
44. Wu, Z., Xiong, Y., Yu, S. X. & Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3733–3742 (2018).
45. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. E. A simple framework for contrastive learning of visual representations. *ICML'20: Proceedings of the 37th International Conference on Machine Learning*, **149**, 1597–1607, 2020.
46. Kingma, D. & Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations* (ICLR, 2014).
47. Ulyanov, D., Vedaldi, A. & Lempitsky, V. S. Instance normalization: the missing ingredient for fast stylization. Preprint at https://arxiv.org/abs/1607.08022 (2016).
48. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res. Proc. Track* **9**, 249–256 (2010).
49. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing System* **30**, 6629–6640 (2017).
50. Hodoglugil, U. & Mahley, R. Turkish population structure and genetic ancestry reveal relatedness among eurasian populations. *Ann. Hum. Genet.* **76**, 128–41 (2012).
51. Fleiss, J. L., Levin, B. and Paik, M. C. *Statistical Methods for Rates and Proportions* (Wiley, 2003).

Acknowledgements

M.T., K.B.O. and G.I.G. are grateful to the Scientific and Technological Research Council of Turkey (TUBITAK) for a 2232 International Outstanding Researcher Fellowship and to TUBITAK Ulakbim for the Turkish National e-Science e-Infrastructure (TRUBA)-cluster and data-storage services. We also thank Ö. Asar and H. Okut for their guidance and assistance in evaluation of the results.

Author contributions

M.T., F.M. and K.B.O. conceived the study and designed the experiments. K.B.O. performed the experimental analysis. B.D., G.I.G., M.Y.L., E.K., D.D., K.B. and T.Y.C. curated the training and test datasets. K.B.O., S.C., K.B., D.D., G.S., M.T. and F.M. analysed the results. K.B.O., S.C., U.P.H., F.Y., D.F.K.W., M.T. and F.M. prepared the manuscript. M.T. and F.M. supervised the research.

Competing interests

The authors declare no competing interests.

Additional information

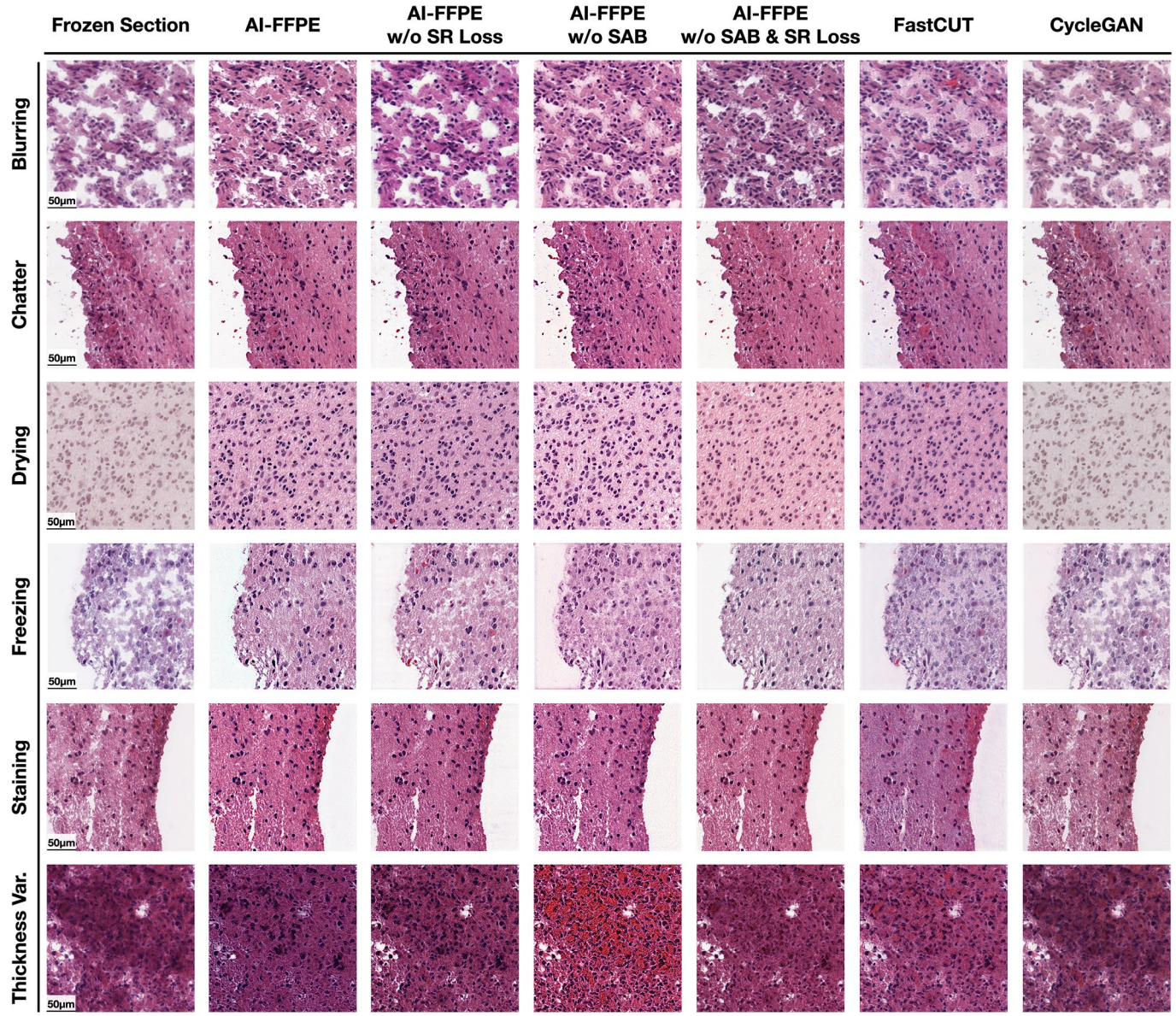
Extended data is available for this paper at https://doi.org/10.1038/s41551-022-00952-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-022-00952-9.

Correspondence and requests for materials should be addressed to Faisal Mahmood or Mehmet Turan.

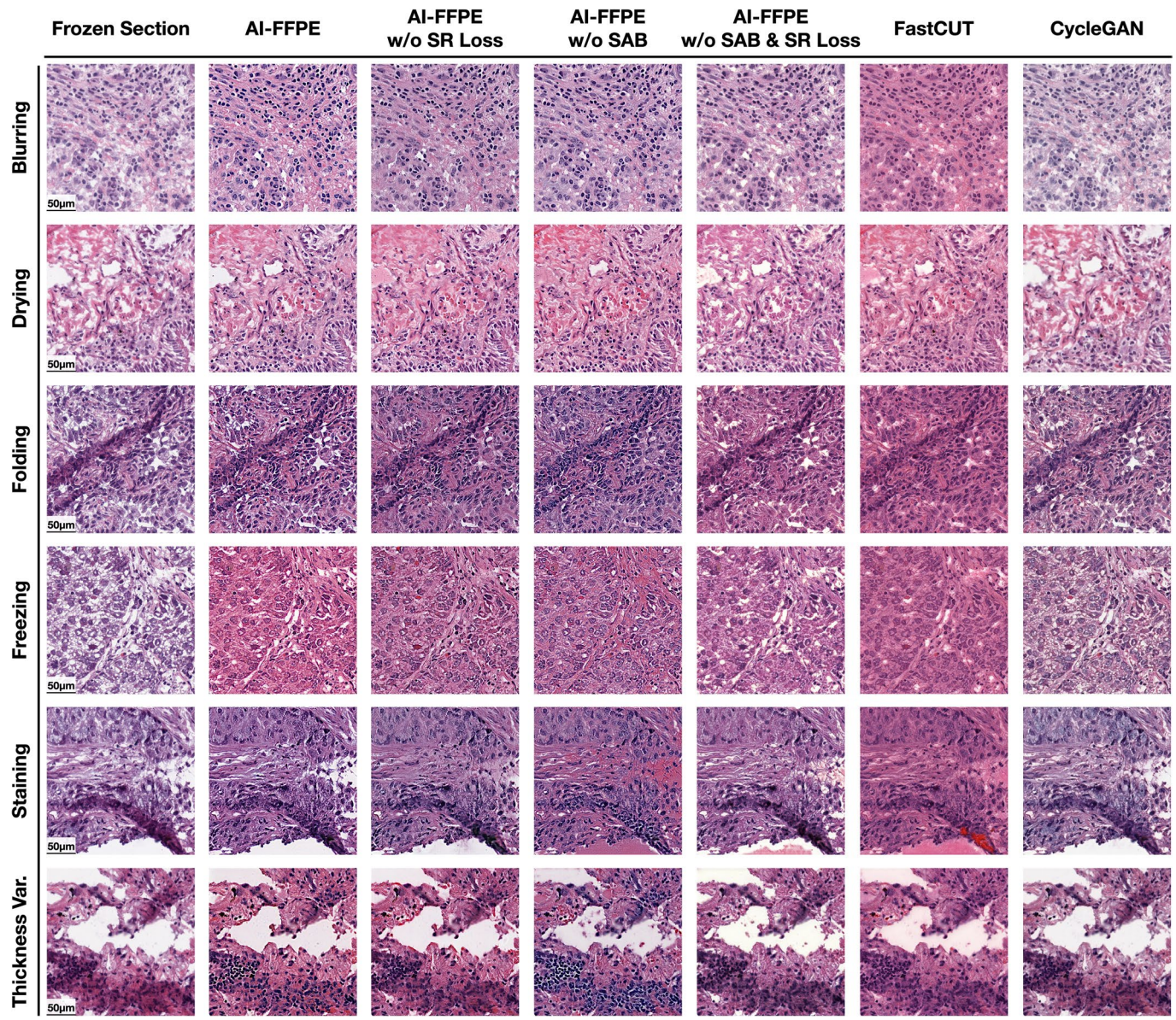
Peer review information *Nature Biomedical Engineering* thanks Geert Litjens and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41551-022-00952-9



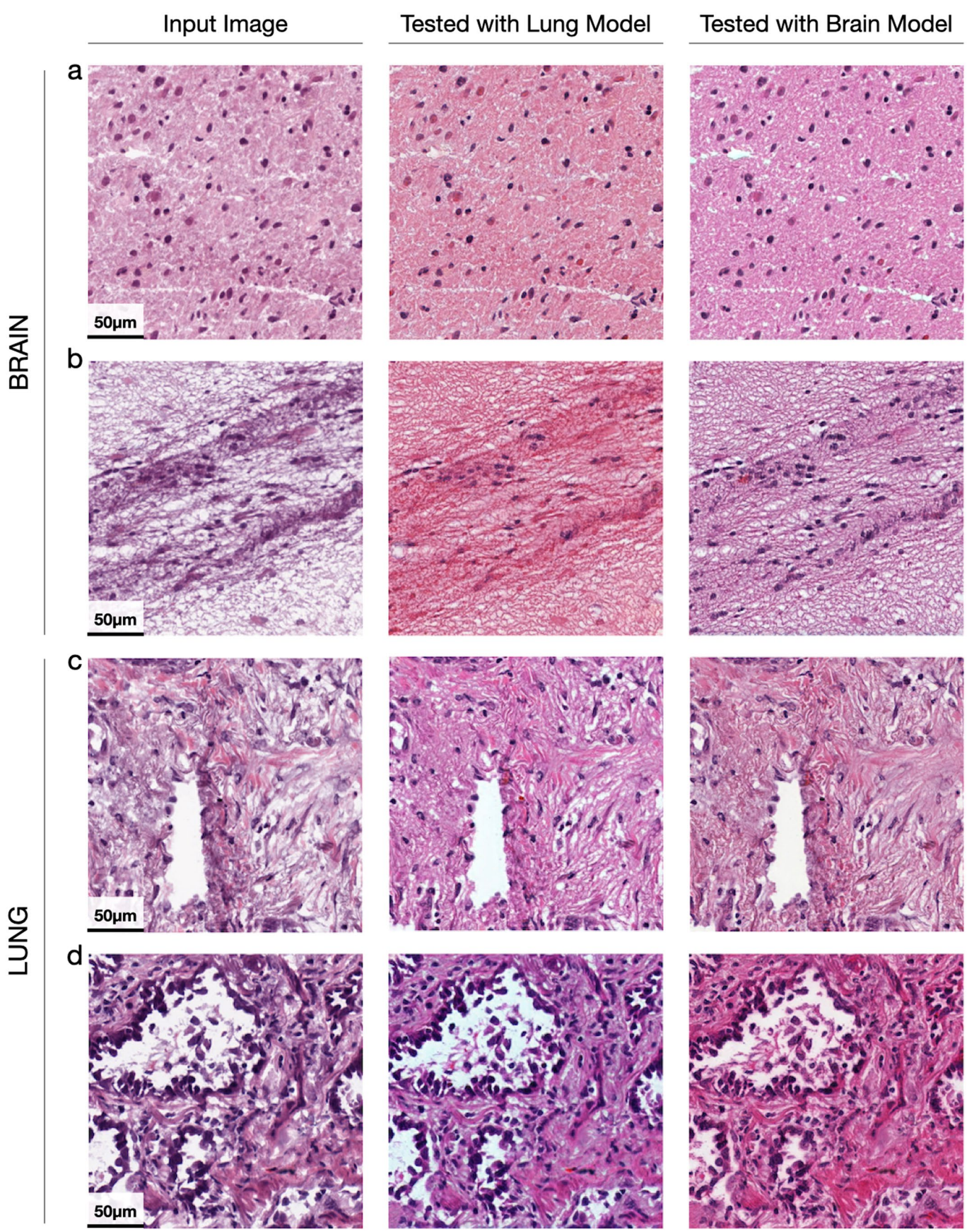
Extended Data Fig. 1 | Comparison of all bench-marked methods' improvement of various artefacts in brain tissue sections. Comparison of all bench-marked methods' improvement of various artefacts in brain tissue sections} AI-FFPE are compared to several unsupervised image-to-image translation methods such as CyleGAN, FastCUT, AI-FFPE without spatial attention block integration but with SR-Loss (AI-FFPE w/o SAB), AI-FFPE without SR loss integration but with SAB (AI-FFPE w/o SR loss), AI-FFPE without both SR loss and SAB (AI-FFPE w/o SAB and SR loss).

https://doi.org/10.1038/s41551-022-00952-9



Extended Data Fig. 2 | Comparison of all bench-marked methods' improvement of various artefacts in lung tissue sections. Under the constrain of cycle consistency loss, CycleGAN, in most of the cases, does not implement any changes on the input FS images. Also, FastCUT's contrastive learning is useful to maximize shared content-related features in between input and synthesized image patches, however, lacks an essential quality required for the improvement of lung WSIs and that is to determine the tissue edges, such as boundaries of vessels or airways, beyond which has to remain untouched.

https://doi.org/10.1038/s41551-022-00952-9

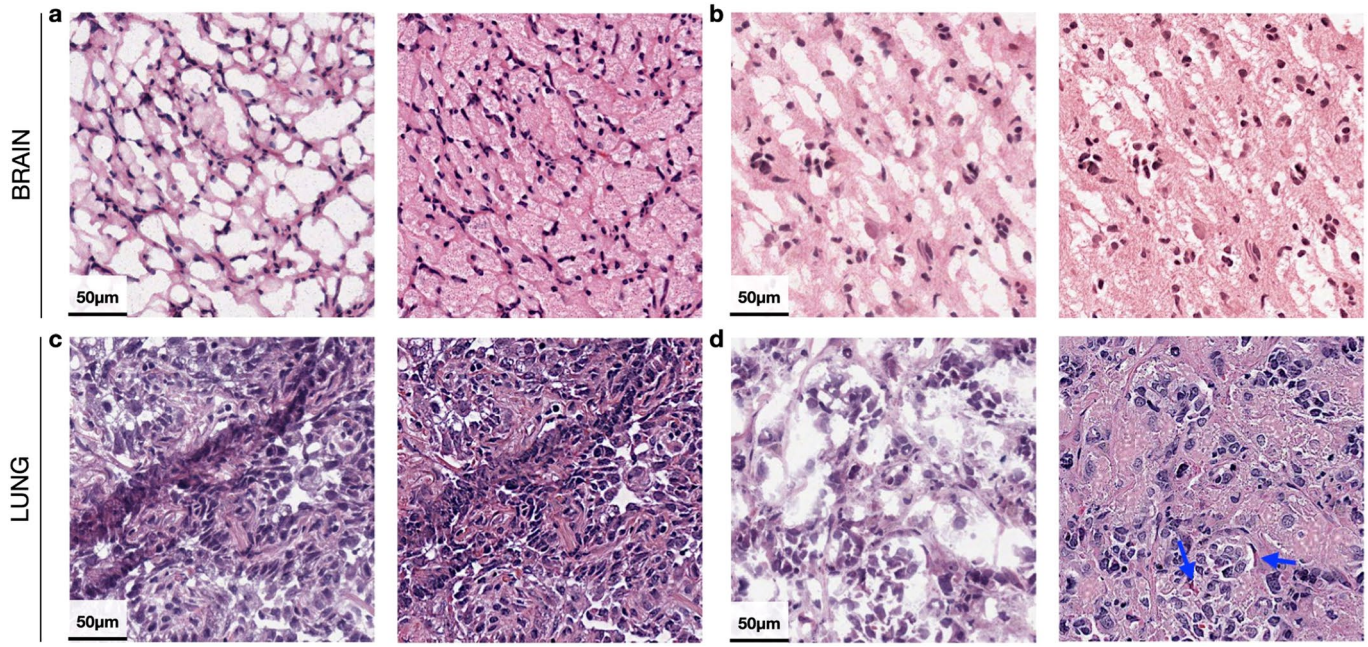


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41551-022-00952-9

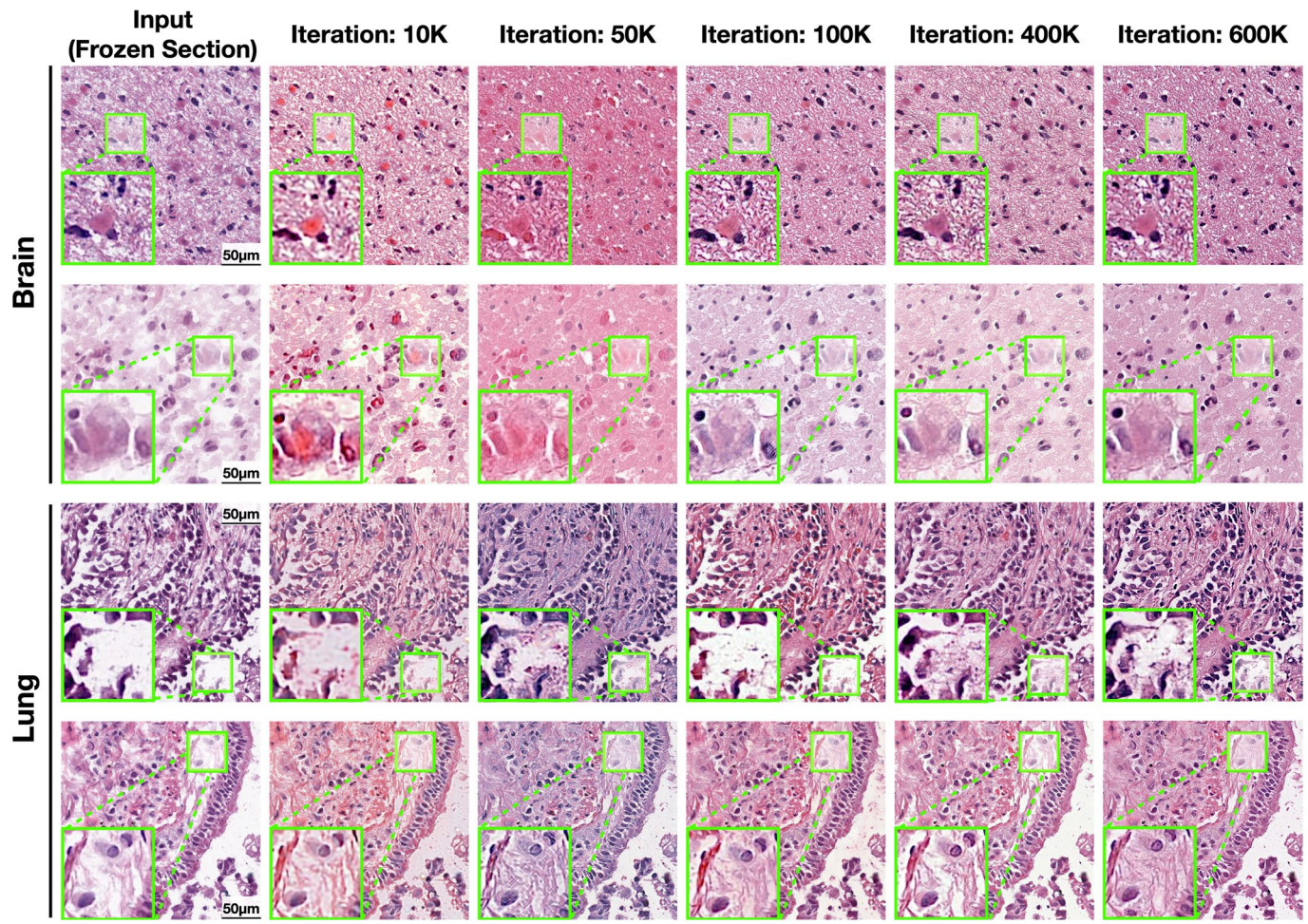
Extended Data Fig. 3 | Cross-organ adaptability of AI-FFPE. a When the brain patches were processed with the lung-trained AI-FFPE model, many improvements that are observed in the same images processed by the brain-trained model were replicated. However, the staining appeared more red/pink. **b** Similar but more severe colouring problem is present here. Some cells and nuclei became pink and lost their cellular characteristics becoming unrecognizable. The folding artefact was not rectified as efficiently as it is in the brain-model tested version. However, slightly more of empty spaces are filled with extracellular matrix(ECM) in the lung-model tested version, which did not add any significant diagnostic information compared its brain-model tested version. **c** When the lung patch was tested on some colouring difference and slight differences in ECM patterns in the parenchyma. Both models preserved the empty space in the middle **d** staining in more red side of the spectrum, the empty areas are more intensively filled probably due to more dense nature of brain tissue which does not harbour empty spaces filled with air.

https://doi.org/10.1038/s41551-022-00952-9



Extended Data Fig. 4 | Exemplars of failure cases. a In rare cases of severe freezing artefacts, the model cannot reverse the dislocation of the cells/nuclei resulting from expanding ice crystals pushing the tissue away from its original location. The empty clefts where the ice crystals are formed are filled with the ECM, creating a mesh-like appearance. **b** The models sometimes show sub-optimal performance in correcting severe chatter artefacts because the images in the FFPE target domain also exhibit a relatively high frequency of chatter artefacts. **c** Although, in the vast majority of cases, corrected images with folding artefacts show clear clues indicating that the original patches contain folding artefacts, rarely, it might take a bit longer for the examiner to recognise the increase in cell densities is actually due to the corrected folding artefacts. **d** The arrows show examples of unnecessary orange-pink colouring of some cells/areas. However, these colour aberrations are uncommon and do not seem to affect the diagnostically meaningful patterns present in the images.

https://doi.org/10.1038/s41551-022-00952-9



Extended Data Fig. 5 | Improvement of output image quality throughout the network training. Network output images for the brain and lung tissue section at different stages of the learning process, that is, after 10k, 50k, 100k, 200k, 400k and 600k In brain sections, the visibility of astrocytic glial neoplastic and stromal cell nuclei, as well as the fibrillar structures improve through iterations. Even though the visual enhancement in the lung tissue sample is highly challenging due to alveolar architecture, significant restoration of the connective tissue is observed as the training progress. At the beginning of training, the diagnostically misguiding regions such as artificial presence of bleeding and blurred nuclear boundary were frequently observed in AI-FFPE patches. However, these issues have been resolved at the end of five epoch.

nature portfolio

Corresponding author(s): Faisal Mahmood, Mehmet Turan

Last updated by author(s): Jul 22, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used all available slides from the TCGA-GBM, TCGA-LGG, TCGA-LUSC and TCGA-LUAD projects, which are openly available from the NIH genomic data commons: https://portal.gdc.cancer.gov /projects/TCGA-GBM, https://portal.gdc.cancer.gov /projects/TCGA-LGG, https://portal.gdc.cancer.gov /projects/TCGA-LUSC, https://portal.gdc.cancer.gov /projects/TCGA-LUAD All in-house slides were scanned by Aperio AT2 at Ege University Hospital.
Data analysis	Slide pre-processing is based on the opensource CLAM tool: https://github.com/mahmoodlab/CLAM. All code for developing and testing AI-FFPE models is available at https://github.com/DeepMIALab/AI-FFPE.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The TCGA diagnostic whole-slide data (GBM, LGG, LUAD and LUSC) and the corresponding labels are available from the NIH genomic data commons (https://portal.gdc.cancer.gov). Restrictions apply to the availability of in-house data, which were used with institutional permission for the purposes of this project. All requests for access to in-house data may be addressed to the corresponding authors, and will be processed in accordance with institutional guidelines. Data can only be shared for academic research purposes and will require a material-transfer or data-transfer agreement with the receiving institution.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Supplementary tables 5 and 6 provide the patient-sex distributions corresponding to the TCGA and in-house datasets used in the study.
Population characteristics	Population characteristics including sex, age, self reported race and ethnicity (for the TCGA data) are available in Supplementary tables 5 and 6.
Recruitment	No patient recruitment was necessary for the retrospective use of histology whole-slide images.
Ethics oversight	The Ege University Ethics Committee approved the study (reference E-99166796-050.06.04-425014 and ID number 21-11.1T/45).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used all available data in the TCGA for model development. Power analysis was used to predetermine the sample size for the reader study. We used all data from publicly available repositories for model development. The in-house dataset was collected the test phase.
Data exclusions	Exclusion criteria included slides with significant marking covering the tissue area, as well as damaged and missing tissue slides. Slides with markings that do not predominantly cover tissue regions were also excluded.
Replication	Replication was successful under all conditions for which results are reported.
Randomization	For the TCGA-GBM, TCGA-LGG, TCGA-LUSC and TCGA-LUAD data, patients were randomly divided into three groups: training, validation and test sets.
Blinding	During the reader study, the experts were blinded to the slide labels, and to the FFPE/AI-FFPE status of the slide.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging