A Dataset Generation Framework for Evaluating Megapixel Image Classifiers and Their Explanations

Gautam Machiraju(✉), Sylvia Plevritis, and Parag Mallick

Stanford University, Stanford, USA
gmachi@stanford.edu

Abstract. Deep learning-based megapixel image classifiers have exceptional prediction performance in a number of domains, including clinical pathology. However, extracting reliable, human-interpretable model explanations has remained challenging. Because real-world megapixel images often contain latent image features highly correlated with image labels, it is difficult to distinguish correct explanations from incorrect ones. Furthering this issue are the flawed assumptions and designs of today's classifiers. To investigate classification and explanation performance, we introduce a framework to (a) generate synthetic control images that reflect common properties of megapixel images and (b) evaluate average test-set correctness. By benchmarking two commonplace Convolutional Neural Networks (CNNs), we demonstrate how this interpretability evaluation framework can inform architecture selection beyond classification performance—in particular, we show that a simple Attention-based architecture identifies salient objects in all seven scenarios, while a standard CNN fails to do so in six scenarios. This work carries widespread applicability to any megapixel imaging domain.

Keywords: eXplainable AI · Interpretable ML · Salient object detection · Model selection · Synthetic data · Coarse supervision · Megapixel imagery

1 Introduction

Megapixel image datasets are increasingly common in multiple scientific and human-centered application domains (*e.g.*, histopathology [39,64], autonomous systems [117], remote sensing and atmospheric sciences [24], and cosmology [79]), but pose unique analytical challenges that are not present in standard image datasets (*i.e.*, those containing $<10^6$ pixels per image). Firstly, such datasets contain images that are often described as either coarsely labeled, weakly labeled,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19775-8_25.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13672, pp. 422–442, 2022.
https://doi.org/10.1007/978-3-031-19775-8_25

or *weakly annotated data* (WAD) [85], meaning that each image is only paired with an image-level label and lacks sub-image annotations. The WAD characterization also implies sufficiently high resolution to contain semantically distinct objects, or visual concepts [44], at multiple scales. However, such images are not typically annotated due to costs and required domain expertise. Secondly, these datasets typically have smaller sample sizes (of labeled images) than standard image datasets. Thirdly, megapixel image datasets now include multiplexed or multispectral images (*i.e.*, with high channel-wise dimensionality) generated in scientific domains including histopathology [10,11,22,36,37,42,72,86,106,111]. Finally, due to the scarcity of these datasets and the diversity of their channel spaces, the creation of pre-trained and foundation models [12] is uncertain. These four unique challenges, coupled with memory constraints of today's GPUs, have necessitated new machine learning approaches for classifying and understanding spatial systems imaged at high-resolution, megapixel scale.

While megapixel image classification has seen significant success with recent deep learning approaches, model explanations are largely still unreliable and uninterpretable. Recent studies have shifted classifiers toward end-to-end deep learning from traditional machine learning of hand-crafted (i.e. hypothesis-driven or keypoint-derived) featurization [39]. In particular, *Patch-based Convolutional Neural Networks* (*PatchCNNs*) [43,54] have reached state-of-the-art performances in domains such as cancer diagnostics and prognostics via histopathology [16,17,20,26,30,40,43,52,60,72,77,78,107] and remote sensing of geoeconomic indicators via multispectral satellite imagery [49,113,115,121]. Despite modeling success on deployment datasets—as commonly defined by classification performance statistics, *e.g.*, area under the Receiver Operating Characteristic (AUROC) and area under the Precision-Recall Curve (AUPRC)—studies may report qualitative and anecdotal assessment of model explanations or report a lack of interpretability [17,20,52,78]. To carry out this cursory assessment, Explanation Maps are commonly used to identify input-specific salient objects and regions-of-interest (ROIs) in test sets. However, Explanation Maps are rarely quantitatively assessed for correctness due to a lack of ground truth pixel-level annotations in WAD settings. Thus, challenging questions remain for megapixel image classifiers: Are models learning and explaining truly salient, human-interpretable objects? Or are those objects latent features or spurious correlations [120]? Are current modeling choices [88] impeding human-interpretable explanations? Which choices lead to enhanced interpretability? How should we assess interpretability? Conversely, can model explanations reveal learning mechanisms and behaviors [68,81], and if so, what mechanisms are desirable for megapixel imagery?

A growing focus on eXplainable Artificial Intelligence (XAI) and Interpretable Machine Learning (IML) [28,29,67,68,75,82] in human-centered and scientific application domains [4,52,105] has reframed interpretability as a priority for model development and selection. *Post-hoc model interpretability*—or a model's ability to make human-interpretable, input-specific explanations—is desired for decision-making but is often untenable due to discordance between

optimized model objectives (*e.g.*, predictive performance) and the end user's real-world objectives (*e.g.*, identifying salient ROIs) [68]. In order to quantify interpretability as a form of explanation correctness, a classification-adjacent task has emerged: *weakly supervised Salient Object Detection* (*wsSOD*). The wsSOD task can be conceptualized as a form of image segmentation targeting salient objects used for classification, but without annotated regions-of-interest (ROIs) as training inputs [13,23,109]. While Salient Object Detection (SOD) is the goal task evaluated via *post-hoc* Explanation Mapping, classification is the workhorse task used to define the learning objective and evaluation scheme. Using neural networks' *in situ* explanations, wsSOD presents an opportunity to both identify and validate salient objects held as ground truth, as well as discover objects *de novo* [30,33,44,109]. This opportunity is especially of interest in megapixel imagery to better understand large-scale spatial systems. This interest is compounded and even necessitated in multiplexed imagery (*e.g.*, spatial proteomics [61]) since Explanation Maps scale in channel-space. Alas, the lack of ground truth salient objects in such settings poses a roadblock to quantify interpretability and motivates wsSOD-based architectural evaluation prior to deployment.

As wsSOD gains popularity in megapixel imagery, the computer vision community needs benchmarking datasets and a quantitative evaluation framework for assessing classifiers and their explanations. To address this need, we present **MISO**, a novel dataset generator that creates **M**egapixel **I**mages with **S**alient **O**bjects for the wsSOD task. MISO creates synthetic control images with differentially expressed, class-specific properties and predefined ground truths to assess classification and wsSOD. We developed MISO and its derivative datasets to help understand the impact of modeling assumptions and design configurations (architectures, hyperparameters, etc.) [88] on these tasks, perform architecture selection, and coax out learning mechanisms. Finally, we demonstrate architecture selection through head-to-head comparisons between commonplace architectures and explanation methods. In summary, our contributions are as follows:

- A unifying and generalized framework for designing PatchCNNs, a popular approach to megapixel image classification
- MISO, a dataset generator and framework for creating synthetic megapixel images with multiple training and testing scenarios that simulate common properties of such datasets
- MISO-1, a benchmark dataset of synthetic controls generated by MISO
- MISO-2, a benchmark dataset derived from real-world histopathology data
- Interpretability Report Cards, a quantitative framework for measuring the average input-specific correctness of predictions and explanations

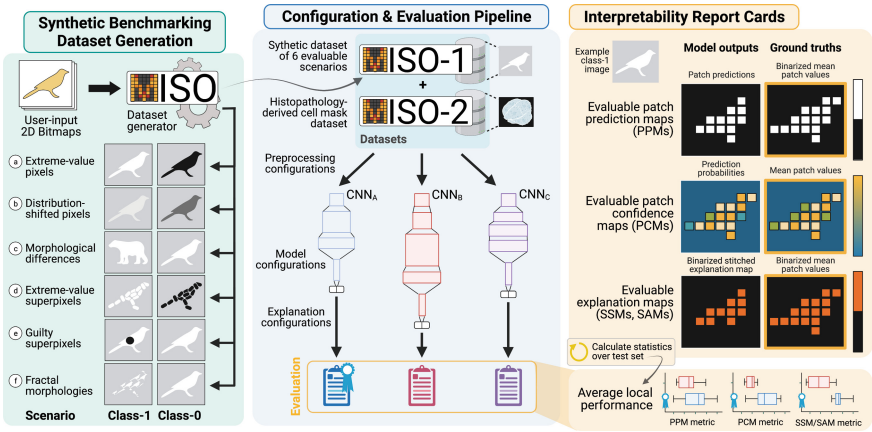


Fig. 1. Overview of the MISO generator and benchmarking datasets, MISO-1 and MISO-2. MISO-1 and MISO-2 alike evaluate models' abilities to detect salient objects and ROIs. Scenarios (a)–(f) of MISO-1 test models' abilities to detect differentially expressed properties and are trained and tested on separately. MISO-2 combines tested properties of MISO-1 to offer a more challenging wsSOD task.

2 Related Work

2.1 Modern Megapixel Image Classifiers

Megapixel image classification has shifted toward deep learning-based approaches, driven by a well-posed data assumption. Implicit in the WAD characterization of megapixel images is the assumption of differentially expressed image features, often formulated as the *Multiple Instance* (*MI*) *assumption* [5,18,53,63,122]: that is, there exists at least one class-specific *instance* (*i.e.*, $\geq$1 pixels) within a class-1 image that corresponds with the class-1 image-level label. Inversely, a class-0 image reflects the absence of such instances. These "guilty" instances (*i.e.*, salient objects) are not known *a priori*, but the MI assumption can be encoded into a classifier's design [17,46,72,113]. This intuitively shifts the traditional classification task toward object detection and segmentation, albeit without ground truth ROI annotations (*e.g.*, bounding boxes). This assumption arises frequently in clinical settings: pathologists search for instances of cancer cells in healthy tissue (to determine cancer diagnosis and stage) and for anaplastic cancer cell morphology in tumor tissue (to determine tumor grade) [52,76].

In accordance with the MI assumption and a push to scalably process megapixel images as inputs to neural networks, classifiers have shifted the unit of analysis toward small sub-image croppings of source images. These croppings, often referred to as *patches* (as depicted in Fig. 2), sample source images and offer a granular, frequentist view of their heterogeneity with predictive utility in various settings [6,50,60]. With patches as inputs, popular models of today often take a disjoint two-stage, Transfer Learning approach to image classification: (1) patch-level predictions through the use of a Convolutional Neural

Network (CNN) followed by (2) patch aggregation and image-level prediction. The family of models in Stage-1 is sometimes referred to as a *Patch-based CNN* or *PatchCNN* [43,54] and can include Attention modules [48,58,94,103] as seen in recent architectures [46,51,72]. The implementation of Stage-2 has spanned a variety of models, ranging from simple decision rules to more sophisticated functions trained on any combinations of hidden vector representations, independent patch-level predictions, and the spatial arrangement of patch predictions (*i.e.*, a *Patch Prediction Map*, or *PPM*) and their probabilities [17,40] (*i.e.*, a *Patch Confidence Map*, or *PCM*). In summary, this two-stage strategy (Fig. 2) and its input patches are used for interconnected reasons: to operate under the memory constraints of GPUs, to significantly amplify the labeled dataset size while preserving image richness, and to perform fine-grained analyses.

However, PatchCNNs in their simplest forms make multiple inherently flawed modeling assumptions. Firstly, (I) patches are assumed and treated as statistically independent and identically distributed (IID), making image-level classification a combination of independent patch predictions. This is inherently false due to spatial autocorrelation between neighboring patches. Based on patch size, features captured may only represent local contextual information [13]. Secondly, because of the aforementioned WAD constraint and MI assumption, (II) patch labeling is often conducted via image-level label inheritance (ILI). ILI is a labeling strategy that labels all constituent patches of an image with their associated image-level label. Despite ILI being a noisy extension of the MI assumption through its "guilt by association" clause, this coarse supervision strategy has shown surprising classification success in the histopathology application domain [17,43]. Thirdly, as discussed, (III) PatchCNNs usually appear in decoupled, two-stage modeling pipelines. Thus, image-level predictions are not part of the PatchCNN learning objective when training on patches, meaning image-level features are not learned when constructing patch-level representations. This design implies (and enforces) that any learned salient objects are self-contained within IID patches. Despite operating on at least one of these limiting modeling assumptions, PatchCNN applications in patient prognostics via histopathology have achieved state-of-the-art classification performances with at AUROCs approaching 1.0 [17,40,43,77,107]. However, while qualitative analyses have been conducted for salient objects *in situ* [89], few examples quantitatively evaluate them [98] or objectively assess architectural or model interpretability.

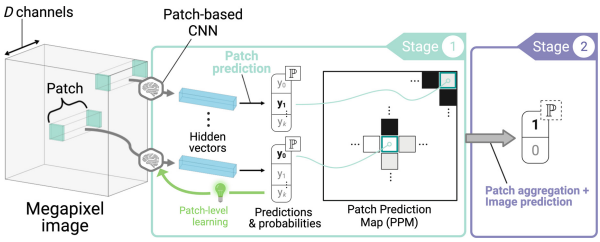


Fig. 2. Generalized schematic of the PatchCNN pipeline. Stage-1 operates on patches independently while Stage-2 operates on all patches per image.

2.2 Explanation Maps and Salient Object Detection

Current explanation methods for neural network classifiers are predominantly *post-hoc* and input-specific (*i.e.*, local) [2,28,82], lending themselves to anecdotal and qualitative interpretations. These explanations often take the form of Explanation Maps (*e.g.*, Saliency Maps [62,95,100], Class Activation Maps [91], and Visual Attention Maps [51]). However, neural networks often fail to provide reliable and human-interpretable explanations through Explanation Mapping [57]. This discordance has been quantitatively studied in standard image domains (*e.g.*, applications in clinical decision support [8]), but is largely missing for megapixel image domains to our knowledge. Studies often omit large-scale assessment of model explanations and instead typically show meaningful ROIs for cherry-picked examples. Ultimately, quantifying explanation *correctness* is required to quantify a model's *post-hoc* interpretability—and is currently unobtainable without human-derived ground truth salient objects. Due to expensive annotation costs beyond image-level labels, wsSOD is an important task due to its wide applicability in assessing model explanations [108,109]. Conventional SOD, much like conventional classification approaches, uses hypothesis-driven, hand-crafted featurizers or keypoint detectors [1,9] to extract domain-specific image features (*e.g.*, intensity, color, texture) and identify salient objects. In contrast, deep learning-based wsSOD is often driven by an end-to-end classification task followed by Explanation Mapping to identify salient objects or ROIs [38,108].

2.3 Benchmarks for Salient Object Detection

While datasets for SOD (and wsSOD) exist for standard image domains, they do not exist for megapixel image domains to our knowledge. Current SOD datasets span a wide range of scenarios—simpler scenarios can contain a few objects overlaid on a solid background, while more complicated scenarios contain multiple objects with varying backgrounds [109]. While these datasets all have ground truths for salient regions (*e.g.*, pixel-wise annotated masks, bounding boxes, human eye fixation locations, etc.), they usually lack class structure and thereby any differentially expressed, class-specific objects for (weakly) supervised learning. Additionally, to our knowledge, no SOD datasets contain megapixel images and thus do not contain any data-relevant properties encountered in real-world datasets (Sect. 3.1). On the other hand, there exist very few real-world megapixel image datasets that have ground truth annotations to compare against any detected salient objects. Typically used for tasks such as ROI segmentation and classification (*e.g.*, in histopathology [69]), such datasets only test models' abilities to identify true positive expert annotations as defined by limited *a priori* domain knowledge (albeit often with low inter-rater reliability [87,97]), but fail to test detected false negatives beyond its scope. Realistic data synthesized from generative models can fall short for similar reasons—defining ground truth salient objects is difficult to condition on, define, and verify, and (ii) even if successful, such salient objects will still only ever reflect domain knowledge. In

reality, many real-world scientific settings will likely never have truly exhaustive annotations for ground truth salience. Finally, to our knowledge, there exists no dataset and evaluation framework to systematically evaluate a patch-based model's textural (*i.e.*, pixel-value heterogeneity), morphological (*i.e.*, edge shape between foreground and background), and contextual awareness (*i.e.*, patch co-localization) at patch and image scales.

3 Proposed Benchmarking Datasets

3.1 Data-Relevant Properties for Megapixel Imagery

Megapixel images often contain differentially expressed properties that classifiers should be attuned to. One property is differential (A) pixel values, where pixel intensities can be indicative of class membership. An analogy from histopathology includes channel expression indicative of a phenotype (*e.g.*, benign versus malignant tumor tissue [74]). Additionally, (B) local abnormalities in intensity, texture, and morphology can exist, or relatively small objects found within larger ones (*i.e.*, MI assumption). An illustrative example from histopathology includes cancer cells in healthy tissue [59]. Another property we may see is differential (C) large-scale morphologies, *i.e.*, patterning and different edge shapes. In histopathology, this can be exemplified by invasive cancer cells altering the stromal patterning in tissue [55]. Finally, the global texture or degree of (D) object clustering (*i.e.*, the size, number, and density of objects) can differ between classes. This is demonstrated by the tightly packed organization of cells in healthy tissue versus cellular anaplasticity in high-grade tumors (*e.g.*, Gleason grade [99]).

3.2 MISO Dataset Generator and MISO-1 Benchmark

We present **MISO** to offer a principled approach to systematically evaluate megapixel image classifiers and their explanations. MISO is a dataset generator and generalizable framework for creating synthetic grayscale control images. Generated using standard techniques in image processing (Appendix A.3), images are partitioned into six datasets, i.e. scenarios, that each simulate one or more differentially expressed data properties (A)-(D) between classes. While trivial to classify at the image-level, a scenario's images may pose difficulty in determining class membership at the patch-level. Each scenario is intended to be trained and tested on individually. We also present **MISO-1**, a benchmarking dataset generated by MISO for the wsSOD task (Fig. 1, Fig. 3). Each of MISO-1's six scenarios contain $n \approx 100$ weakly annotated megapixel images for each of the training and test sets (Fig. 3). Dataset specifications are outlined in Appendix A.3. Because we define ground truth salient objects in the creation of MISO-1, we can test models' capabilities to identify *all* differentially expressed salient objects—a necessity for selecting interpretable models when ground truth annotations are unavailable or non-exhaustive. Regardless of scale, Scenarios (a), (b), and (d) test awareness of pixel value, while (c), (e), and (f) test awareness of morphology:

(a) **Extreme-value pixels (EVP):** This scenario tests data property (A). Images in this scenario contain 0- or 1-valued objects of varying large-scale morphologies for their respective 0-class and 1-class images. Since image textures are relatively smooth (with "salt or peppered" fuzziness in some cases), this scenario tests a patch-based model's ability to map patches to classes via their pixel values alone, regardless of what parts of the image's objects are contained within the patches.

(b) **Distribution-shifted pixels (DSP):** This scenario also tests data property (A), albeit with increased difficulty than the EVP scenario. Images in this scenario contain objects with pixel values that are sampled from non-overlapping class distributions. More concretely, there exist varying large-scale morphologies with pixels ranging from [0.6,1] for 1-class images and [0,0.4] for 0-class images. Similar to the EVP scenario, this scenario tests a patch-based model's ability to map patches to classes by pixel values alone.

(c) **Morphological differences (MD):** This scenario tests data property (C). Images in class-0 have a shared large-scale morphology (Fig. 3). Images in class-1 take on all other large-scale morphologies derived from the other 23 input bitmaps. Because data property (A) is tested by both EVP and DSP scenarios, MD has 1-valued objects for both classes. This scenario tests a patch-based model's contextual and morphological awareness.

(d) **Extreme-value superpixels (EVSP):** This scenario tests data property (A) like the EVP scenario, but with added difficulty—this scenario simulates settings with varying large-scale morphologies comprised of many small objects. To achieve this, we took coordinates from in-house cell segmentation data (masks provided with MISO-1) and overlaid 0- and 1-valued circles (of radius 10 pixels) within our large-scale morphologies for 0- and 1-class images, respectively. This scenario also tests a patch-based model's ability to map patch pixel values to classes.

(e) **Guilty superpixels (GSP):** This scenario tests data property (B). Class-0 images contain 1-valued objects with varying large-scale morphologies and class-1 images contain those objects but with randomly placed 0-valued circles of (randomly selected) radii between [100] pixels within them. This tests a patch-based model's textural and contextual awareness.

(f) **Fractal morphologies (FM):** This scenario tests data property (D). Class-0 images contain 1-valued objects of varying large-scale morphologies, while class-1 images contain mosaics of those same objects but shrunken, repeatedly tiled, and mapped within their original morphology boundaries. This approach creates sufficient class balance between class-0 and class-1 patches. This tests a patch-based model's textural, morphological, and contextual awareness.

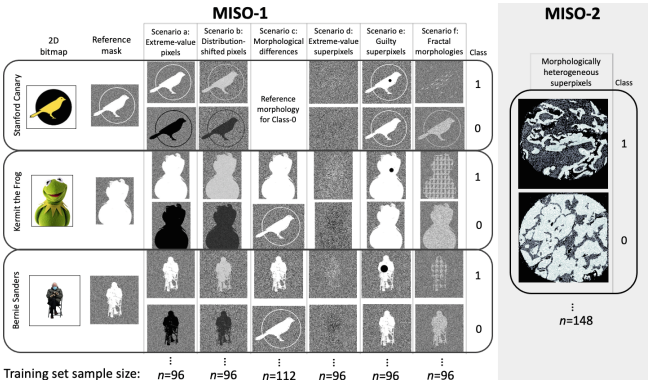


Fig. 3. MISO-1 example images for three reference masks. MISO-2 example images.

3.3 MISO-2 Benchmark

To evaluate PatchCNNs in a more realistic setting, we also present MISO-2, a binary image dataset derived from histopathology data. This dataset includes morphologically heterogeneous superpixels in the form of segmented cell masks from $n = 184$ images. The masks span two histologic types of lung cancer, defining our class structure for image-level labels: 92 adenocarcinoma samples as class-0 and 92 squamous cell carcinoma samples as class-1. Histologic type is used as a label for its ability to describe superpixel patterns represented throughout MISO-2's images. Cell morphology heterogeneity arises from both intra-sample heterogeneity (via multiple cell types, *i.e.*, stromal and cancer cells) and inter-sample heterogeneity (via histologic type). We split the data using an 80–20 split, resulting in 148 training images and 36 test images. This benchmark simultaneously tests data properties (B)–(D).

4 Experiments and Results

4.1 Baseline Methods and Evaluable Outputs

To establish baseline classifiers for MISO-1 and MISO-2, we chose two commonplace PatchCNN architectures to perform the wsSOD task. We used *VGG-19* [96] and *VGG with Attention* (*VGG-Att*) with pre-pooled modules [51] with shared patch sizes (96×96 pixels for MISO-1 and 224×224 pixels for MISO-2) and background filtration. Both architectures were trained (over 10 epochs for MISO-1 and 20 epochs for MISO-2) and tested on all scenarios separately, resulting in fourteen evaluable models. Both architectures output PPMs and PCMs, which spatially contextualize patch predictions and prediction probabilities per image. For explanations, we also chose standard baseline methods: Saliency and Attention Maps. VGG-19 can compute Saliency Maps per patch, while VGG-Att can compute both Saliency and Attention Maps per patch. To

speed up run-time and smooth out ROIs, we averaged absolute saliency or attention scores [3] per patch. We then concatenated scores to construct an array per image that we refer to as *Stitched Saliency Maps* (*SSMs*) and *Stitched Attention Maps* (*SAMs*). Binarization of the SSMs and SAMs was performed with the popular adaptive threshold of double the mean saliency or attention score [13]. All other design configurations (*e.g.*, hyperparameters) for preprocessing, modeling, and explanations are shared between architectures. Refer to Appendix A.4 for a summary of configurations that define benchmarks for architecture selection.

4.2 Evaluation Framework: Interpretability Report Cards

Because ground truths are defined by the control images in MISO-1 and MISO-2, quantitative evaluation includes the correctness of spatially resolved predictions and explanations. To evaluate both model predictions and explanations, we use performance statistics from classification, image segmentation, and saliency analysis [13,38,65,109,116] to score outputs against ground truths (Fig. 1). Three types of analyses are conducted in this section: (i) independent patch predictions (of their constituent image's class) over the whole patch dataset and the resulting image-level predictions based on decision rules; (ii) PPMs and PCMs to assess the correctness of predictions and prediction probabilities in space; and (iii) SSMs and SAMs to assess wsSOD capabilities, *i.e.*, correctness of explanations. We define correctness using the notion of *explanation plausibility*, *i.e.*, the quality of alignment between model explanations and human interpretations [47], and use it as our proxy for interpretability. Thus, we compute similarity scores between an explanation and its corresponding ground truth salient objects. We do not assess *explanation faithfulness*, *i.e.*, how accurately an explanation reflects a model's true reasoning process [47], due to its relative difficulty to evaluate quantitatively. To summarize evaluation, we provide *Interpretability Report Cards* per model and per scenario, which consist of average statistics and confidence intervals over test set images and visualizations of example predictions and explanations (Fig. 4, Appendix A.6). Specifically for Explanation Mapping, this strategy moves toward a global measure of *post-hoc* model interpretability through the aggregate evaluation of local, or input-specific, explanations. It should be noted that the following statistics are all inflated by background patch filtration, a commonly performed preprocessing step in several application domains.

Classification Performance Statistics: For independent patch predictions, ground truth annotations are derived from ILI labeling. Patch-level AUROC, AUPRC, and Average Precision (AP) are computed over all IID-assumed patches, thus reflecting class membership prediction of individual patches. Image-level AUROC, AUPRC, and AP are also computed over all image-level labels using patch aggregation functions (*i.e.*, image-level decision rules described in Appendix A.4). Image-level prediction probabilities are generated with each decision rule's pooling strategy. Results are found in Appendix A.5 Table 3.

Patch Prediction and Confidence Maps: Ground truth annotations for PCMs are derived from patch means, while ground truth annotations for PPMs are derived from patch means and an applied manual binarization threshold specific to image label and scenario (Fig. 1). To assess PPMs on an IID patch-level, we calculate set-theoretic statistics including F_β-measure (with $\beta^2 = 0.3$ [13]), Dice (*i.e.*, F_1-measure), Jaccard, and Overlap coefficients, as well as Sensitivity, Specificity, and Mean Absolute Error (MAE). To assess PPMs structurally, we use the E-measure [32] and also introduce a new metric called the *Scagnostics Distance* (*ScagDist*). ScagDist simply featurizes binary masks by their topological properties using techniques from computational geometry and graph theory [112] and takes the cosine distance between them. It should be noted that we tallied default ScagDist values of 1.0 for blank SSM or SAM outputs. We assess PCMs on a patch-level via the MAE and structurally via the Structural Similarity Index Measure (SSIM) [110]. Results are shown in Table 2 (and Appendix A.5 Table 4). It should be noted that PPMs and PCMs were not evaluated for the GSP scenario due to the difficulty in defining a single notion of correctness for ground truths, given the ILI labeling scheme used for training (Appendix A.6 Fig. 10).

Stitched Saliency and Attention Maps: Due to the differentially expressed nature of the image properties in MISO-1 and MISO-2, ground truth annotations for SSMs and SAMs (salient ROIs) are conveniently the same as those constructed for PPMs (patch labels)—they are derived from patch means and an applied manual binarization threshold specific to image labels and scenarios (Fig. 1, Appendix A.3 Fig. 6). For ease, we only assess class-1 test images. To assess binarized SSMs and SAMs on an IID patch-level, we again calculate

Table 1. Average statistics over test-set PPMs. A **boldface** result indicates a superior score between architectures for a given scenario (directionality denoted by $\uparrow, \downarrow$). †Generated mean value required image sampling to avoid runtime or memory issues. $^\clubsuit$Denotes binary map evaluation. $^\diamondsuit$Denotes structural evaluation. $^\heartsuit$Denotes IID patch-level evaluation.

Scenario	Sensitivity$^{\clubsuit\heartsuit\uparrow}$	Specificity$^{\clubsuit\heartsuit\uparrow}$	ScagDist$^{\clubsuit\diamondsuit\downarrow}$	F_β-measure$^{\clubsuit\heartsuit\uparrow}$	E-measure$^{\clubsuit\diamondsuit\uparrow}$	MAE$^{\clubsuit\heartsuit\downarrow}$
VGG-19						
EVP	0.999 ± 0.000	0.926 ± 0.016	0.011 ± 0.006	0.931 ± 0.016	0.952 ± 0.005	0.038 ± 0.008
DSP	0.500 ± 0.100	0.929 ± 0.016	0.138 ± 0.035	0.499 ± 0.100	0.666 ± 0.036	0.246 ± 0.056
MD	0.500 ± 0.093	0.938 ± 0.011	**0.004** ± 0.001	0.499 ± 0.092	0.837 ± 0.015	0.137 ± 0.025
EVSP	0.498 ± 0.100	0.923 ± 0.019	0.535 ± 0.094	0.321 ± 0.068	0.514 ± 0.028	0.160 ± 0.024
FM	0.500 ± 0.100	**0.923** ± 0.018	0.508 ± 0.099	0.429 ± 0.086	0.575 ± 0.033	0.205 ± 0.039
MISO-2	**0.557**† ± 0.167	**0.890**† ± 0.040	**0.171**† ± 0.094	0.572† ± 0.171	**0.628**† ± 0.072	**0.323**† ± 0.121
VGG-Att						
EVP	0.999 ± 0.000	0.926 ± 0.016	0.011 ± 0.006	0.931 ± 0.016	0.952 ± 0.005	0.038 ± 0.008
DSP	**0.999** ± 0.000	**0.928** ± 0.016	**0.013**† ± 0.006	**0.927** ± 0.018	**0.951** ± 0.005	**0.038** ± 0.008
MD	**0.721** ± 0.048	**0.938** ± 0.011	0.008† ± 0.001	**0.719** ± 0.051	**0.908** ± 0.008	**0.099** ± 0.017
EVSP	**0.998** ± 0.001	0.923 ± 0.019	**0.035** ± 0.012	**0.819** ± 0.041	**0.888** ± 0.013	**0.057** ± 0.013
FM	**0.986** ± 0.009	0.922 ± 0.017	**0.009** ± 0.004	**0.924** ± 0.016	**0.944** ± 0.006	**0.045** ± 0.009
MISO-2	0.400† ± 0.086	0.865† ± 0.040	0.317† ± 0.071	**0.601**† ± 0.095	0.461† ± 0.040	0.437† ± 0.068

set similarity coefficients as described in the previous subsection. To assess binarized maps structurally, we again use the E-measure and ScagDist. We use MAE to assess non-binarized maps on a patch-level. Finally, to assess non-binarized maps structurally, we use the S-measure [31] and SSIM. Results are shown in Table 2 (and Appendix A.5 Table 5). Testing every patch explanation against its patch-level ground truth helps measure image-level awareness of differential expression.

4.3 Analysis of Results

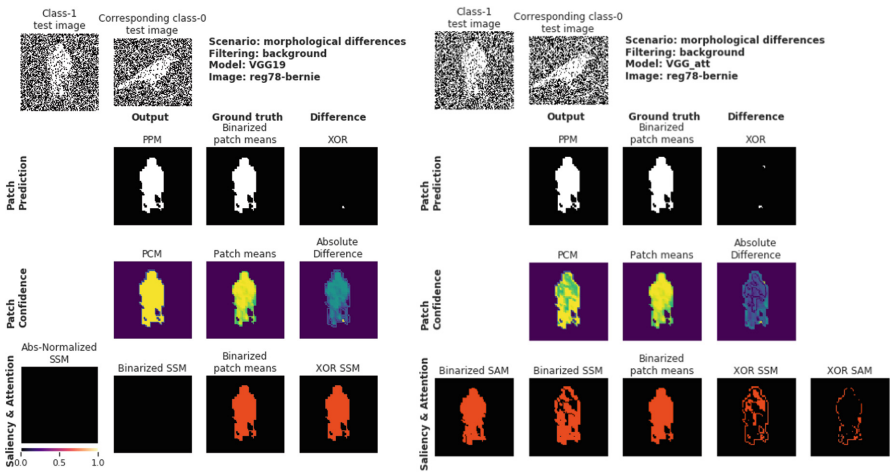
Are models learning and explaining truly salient, human-interpretable objects? Yes, our wsSOD experiments suggest that a subset of models are able to do so despite the aforementioned flawed modeling assumptions. However, our experiments highlight divergent patterns of classification and explanation performance over a number of evaluation statistics. *Are current modeling choices impeding human-interpretable explanations? Which choices lead to enhanced interpretability?* For MISO-1, while both architectures perform fairly well with patch- and image-level class predictions (Table 1, Appendix A.5 Table 3 and Table 4), VGG-Att still generally outperforms VGG-19 in both generating accurate class predictions and human-interpretable explanations. VGG-Att's relatively high performance is increasingly clear as the scenarios become intuitively

Table 2. Average statistics over test set Explanation Maps. Stitched Saliency Maps (SSMs) are constructed for both models, while Stitched Attention Maps (SAMs) are also constructed for VGG-Att. A **boldface** result indicates a superior score between architectures for a given scenario (directionality denoted by ↑, ↓). *Indicates that SAMs yielded a top score for VGG-Att. †Generated mean value required image sampling to avoid runtime or memory issues. ♣Denotes binary map evaluation. ♠Denotes non-binary map evaluation. ♢Denotes structural evaluation. ♡Denotes IID patch-level evaluation.

Scenario	MAE$^{\spadesuit\heartsuit\downarrow}$	F_β-measure $^{\clubsuit\heartsuit\uparrow}$	ScagDist$^{\clubsuit\diamond\downarrow}$	S-measure$^{\spadesuit\diamond\uparrow}$	E-measure$^{\clubsuit\diamond\uparrow}$	SSIM$^{\spadesuit\diamond\uparrow}$
VGG-19						
EVP	**0.427** ± 0.050	0.475 ± 0.082	0.214 ± 0.047	**0.289** ± 0.025	0.503 ± 0.044	**0.270** ± 0.063
DSP	0.413 ± 0.051	0.000 ± 0.000	1.000 ± 0.000	0.293 ± 0.026	0.250 ± 0.000	0.280 ± 0.064
MD	0.394 ± 0.049	0.000 ± 0.000	1.000 ± 0.000	0.303 ± 0.025	0.250 ± 0.000	0.279 ± 0.055
EVSP	0.206 ± 0.043	0.000 ± 0.000	1.000 ± 0.000	0.397 ± 0.022	0.250 ± 0.000	0.384 ± 0.066
GSP	0.020 ± 0.004	0.000 ± 0.000	1.000† ± 0.000	**0.529** ± 0.021	0.250 ± 0.000	0.913 ± 0.014
FM	0.331 ± 0.058	0.000 ± 0.000	1.000 ± 0.000	0.335 ± 0.029	0.250 ± 0.000	0.314 ± 0.071
MISO-2	0.660 ± 0.023	0.000 ± 0.000	1.000 ± 0.000	0.170 ± 0.012	0.250 ± 0.000	0.001 ± 0.001
VGG-Att						
EVP	0.430 ± 0.050	**0.609** ± 0.083	**0.179** ± 0.056	0.285 ± 0.025	**0.581** ± 0.043	0.265 ± 0.063
DSP	**0.412** ± 0.051	**0.585*** ± 0.098	**0.170***† ± 0.068	**0.541*** ± 0.016	**0.610*** ± 0.045	**0.412*** ± 0.054
MD	**0.391** ± 0.049	**0.627** ± 0.030	**0.135***† ± 0.045	**0.306** ± 0.025	**0.657*** ± 0.042	**0.282** ± 0.055
EVSP	**0.205** ± 0.043	**0.691*** ± 0.047	**0.076*** ± 0.028	**0.446*** ± 0.013	**0.781*** ± 0.021	**0.435*** ± 0.053
GSP	0.020 ± 0.004	**0.350** ± 0.072	**0.266**† ± 0.044	0.490 ± 0.002	**0.656** ± 0.040	0.913 ± 0.014
FM	**0.313** ± 0.055	**0.664*** ± 0.046	**0.051*** ± 0.014	**0.363** ± 0.026	**0.645*** ± 0.031	**0.339** ± 0.068
MISO-2	0.660 ± 0.023	**0.390*** ± 0.078	**0.398**† ± 0.046	0.170 ± 0.012	**0.258*** ± 0.015	**0.002** ± 0.001

more challenging—regarding PPM correctness in Table 1, VGG-19 has low sensitivity (near 0.5) for all scenarios other than EVP, thus pointing to an average of near-random classification of foreground patches per image. Regarding SSM and SAM correctness (Table 2), VGG-19 often failed to identify any salient objects in scenarios other than EVP. Results from MISO-2 show us that while VGG-19, on average, tends to make more correct class predictions than VGG-Att (Table 1), VGG-Att is far superior at identifying salient objects on average (Table 2). *Are some salient objects simply latent features or spurious correlations?* Regarding latent features, VGG-19 is learning highly predictive patch representations almost all scenarios Table 1, but systemically struggles to create interpretable explanations. This discrepancy points to the architecture's potential reliance on latent features. Unfortunately, we are limited in our ability to characterize spurious correlations from this work alone. For simplicity and to probe high-supervision performance on MISO-1 and MISO-2, we conducted complete patch filtering of image backgrounds. This choice thereby limits any learning of spurious correlations and subsequently constrained learning and evaluation to foreground objects.

Can model explanations reveal learning mechanisms and behaviors? What mechanisms are desirable for megapixel imagery? While model explanations on synthetic datasets do not directly reveal learning mechanisms, they can shed light on resulting emergent behaviors and capabilities that certain architectures afford us. Firstly, VGG-Att seems to have a stronger ability to learn ranges of pixel values between classes and to identify relatively small salient morphologies, respectively supported by DSP and FM scenarios (*e.g.*, near-doubled PPM



(a) VGG-19 **(b)** VGG-Att

Fig. 4. Example Report Card for morphological differences (MD). Test-set examples, PPMs, PCMs, and SSMs (and SAMs if applicable) are displayed in rows 1–4, respectively.

sensitivity and F_β-measure in Table 1). Additionally, while VGG-Att has lower patch-level classification performance scores for GSP (Appendix A.5 Table 3), the scores actually reflect the model's ability to work beyond the confines of fuzzy (*i.e.*, ILI) patch labeling and classify "guilty" regions as the sole class-1 patches. VGG-Att's under-reliance on fuzzy patch labeling is also apparent in its explanations. Curiously, it identified salient "guilty" ROIs in larger morphologies (Appendix A.6 Fig. 10, Table 2 ScagDist and *E*-measure) despite flawed modeling assumptions: despite only having access to IID-assumed patches and ILI labels, solely computing Attention scores within each patch, and without any form of global Attention across patches. This finding hints at superior contextual awareness, such as capabilities for Gestalt Closure [56] and foreground-background edge detection. VGG-Att's heightened edge detection is supported by properly modulated patch prediction probabilities in the MD scenario's PCMs (Fig. 4b, Table 4)—by assigning lower prediction probabilities to interior foreground patches, it appears to use edges as the primary discernment between large-scale morphologies. Interestingly, MISO-2's explanation results (Table 2) could point to VGG-19's heightened global textural awareness (a potential bias for CNNs [34]), despite overall lower SSM performance than VGG-Att (Table 1). The relative complexity of MISO-2 as a dataset may point to VGG-Att's limited bias toward textures, but could also reflect the need for increased training iterations or sample size. Even so, VGG-Att's dominance in interpretability points to its ability to construct representations without solely relying on latent features. VGG-Att's overall modeling capabilities are most likely granted by its model-intrinsic explanations [29], but don't necessarily imply greater utility of SAMs over SSMs—both were generally accurate, but also excelled in different scenarios (*e.g.*, SSMs outperformed SAMs in all statistics in the EVP scenario). These findings reiterate Attention-based models' (*e.g.*, *Vision Transformers*) starkly different learning mechanisms for visual recognition than traditional CNNs, their fine-grained attentiveness, and their subsequent human-interpretability in standard image domains [14,19,21,80].

How should we assess interpretability? To quantify and assess interpretability, we recommend measuring explanation plausibility. Specifically for megapixel imagery, classifiers should be assessed via wsSOD to investigate their abilities to explain predictions with long-range dependencies. Even though our evaluation framework only tests plausibility, it allows us to interrogate model behaviors, hypothesize learning mechanisms, and even sufficiently differentiate between two comparable, but architecturally distinct baselines. Because this form of analysis requires complete salient object annotations, synthetic controls (with differentially expressed properties) provides automated evaluation in deployment settings where exhaustive annotations are infeasible or impossible. Furthermore, because predictive performance and explanation plausibility are not necessarily correlated, we support similar frameworks for architectural (*i.e.*, model family) evaluation, benchmarking, debugging and testing workflows [118], and selection prior to model deployment. In order to push models toward XAI and IML, we must promote interpretability as a quantifiable criterion in the design process to ultimately build inherently interpretable architectures and trained models.

While MISO-1 lacks realism, its scenarios independently and systematically test megapixel imagery's common properties to reveal architectural behaviors. Even with MISO-2's real-world origins and combined elements of MISO-1 scenarios (*i.e.*, MD, GSP, FM), VGG-19's low-plausibility explanations for MISO-1 recur for MISO-2 (Table 2). While this consistency hints at generalized behavior, generalizability should be assessed further in settings with greater data complexity (as discussed in Limitations section Appendix A.7). For these reasons, we recommend that MISO-1, MISO-2, and any other custom MISO-generated datasets be used *in tandem* with domain-specific segmentation datasets (*i.e.*, where annotated ROIs are withheld during training and evaluated against). This multi-pronged strategy can respectively test (A) a model's full set of plausible explanations in settings without exhaustive ground truths and (B) true positive salient objects in real-world environments. We believe the proposed benchmarks can act as a community resource similar to MNIST [15], but customized for megapixel imagery and assessing interpretability. These datasets act as debugging sanity checks that *any* interpretable model should be able to pass—especially before deployment in low-annotation settings.

5 Conclusion

In summary, our primary goals are to both provide synthetic datasets that reflect one or more common, differentially expressed data properties, as well as systematically probe the interpretability of megapixel image classifiers. We show the utility of this approach for evaluating classifiers and their explanation plausibility via their propensities to perform wsSOD. Through experimentation, we also put commonplace PatchCNN architectures into question. While current modeling paradigms lack interpretability, extensions toward context-aware, Attention-based architectures have great potential as salient object detectors aligned with human interpretation. This work has widespread applicability for megapixel image and application domains, and can even provide an groundwork for interpretability evaluation in standard image domains.

References

1. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: Salient region detection and segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 66–75. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79547-6_7
2. Adadi, A., Berrada, M.: Peeking inside the Black-Box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
3. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps (2020)
4. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I.: Precise4Q consortium: explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inform. Decis. Mak. **20**(1), 310 (2020)

5. Amores, J.: Multiple instance classification: review, taxonomy and comparative study. Artif. Intell. **201**, 81–105 (2013)
6. Anonymous: Patches are all you need? In: Submitted to The Tenth International Conference on Learning Representations (2022). https://openreview.net/forum?id=TVHS5Y4dNvM. under review
7. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. arXiv:1908.02983 [cs], June 2020. arXiv: 1908.02983
8. Arun, N., et al.: Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging. bioRxiv, July 2020
9. Bartol, K., Bojanić, D., Pribanić, T., Petković, T., Donoso, Y.D., Mas, J.S.: On the comparison of classic and deep keypoint detector and descriptor methods. arXiv, July 2020
10. Berry, S., et al.: Analysis of multispectral imaging with the AstroPath platform informs efficacy of PD-1 blockade. Science **372**(6547) (2021)
11. Black, S., et al.: CODEX multiplexed tissue imaging with DNA-conjugated antibodies. Nat. Protoc. **16**, 3802–3835 (2021)
12. Bommasani, R., et al.: On the opportunities and risks of foundation models. arXiv, August 2021
13. Borji, A., Cheng, M.-M., Hou, Q., Jiang, H., Li, J.: Salient object detection: a survey. Comput. Vis. Media **5**(2), 117–150 (2019). https://doi.org/10.1007/s41095-019-0149-9
14. Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., Wattenhofer, R.: On identifiability in transformers. arXiv, August 2019
15. Burges, C.J.C.: MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. https://yann.lecun.com/exdb/mnist/. Accessed 20 July 2022
16. Bándi, P., et al.: From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. IEEE Trans. Med. Imaging **38**(2), 550–560 (2019). https://doi.org/10.1109/TMI.2018.2867350
17. Campanella, G., et al.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. **25**(8), 1301 (2019)
18. Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: a survey of problem characteristics and applications. arXiv, December 2016
19. Caron, M., et al.: Emerging properties in self-supervised vision transformers. arXiv, April 2021
20. Chan, L., Hosseini, M., Rowsell, C., Plataniotis, K., Damaskinos, S.: HistoSeg-Net: semantic segmentation of histological tissue type in whole slide images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, October 2019
21. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. arXiv, December 2020
22. Chevrier, S., et al.: An immune atlas of clear cell renal cell carcinoma. Cell **169**(4), 736-749.e18 (2017). https://doi.org/10.1016/j.cell.2017.04.016
23. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. arXiv, January 2020
24. Coffey, V.C.: Multispectral imaging moves into the mainstream. Opt. Photonics News **23**(4), 18 (2012)
25. Cohen, T.S., Welling, M.: Group equivariant convolutional networks. arXiv, February 2016

26. Cruz-Roa, A., Arévalo, J., Judkins, A., Madabhushi, A., González, F.: A method for medulloblastoma tumor differentiation based on convolutional neural networks and transfer learning. In: 11th International Symposium on Medical Information Processing and Analysis, vol. 9681, p. 968103. International Society for Optics and Photonics, December 2015. https://doi.org/10.1117/12.2208825. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9681/968103/A-method-for-medulloblastoma-tumor-differentiation-based-on-convolutional-neural/10.1117/12.2208825.short
27. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 886–893 (2005). https://doi.org/10.1109/CVPR.2005.177
28. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (XAI). A survey. arXiv, June 2020
29. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. arXiv, July 2018
30. Echle, A., Rindtorff, N.T., Brinker, T.J., Luedde, T., Pearson, A.T., Kather, J.N.: Deep learning in cancer pathology: a new generation of clinical biomarkers. Br. J. Cancer **124**(4), 686–696 (2020)
31. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: a new way to evaluate foreground maps. arXiv, August 2017
32. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv, May 2018
33. Frintrop, S., García, G.M., Cremers, A.B.: A cognitive approach for object discovery. In: 2014 22nd International Conference on Pattern Recognition, pp. 2329–2334, August 2014
34. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv, November 2018
35. Ghafoorian, M., et al.: Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Sci. Rep. **7**(1), 5110 (2017)
36. Giesen, C., et al.: Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat. Methods **11**(4), 417–422 (2014). https://doi.org/10.1038/nmeth.2869
37. Goltsev, Y., et al.: Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. Cell **174**(4), 968–981.e15 (2018)
38. Gupta, A.K., Seal, A., Prasad, M., Khanna, P.: Salient object detection techniques in computer vision-a survey. Entropy **22**(10) (2020)
39. Gurcan, M.N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., Yener, B.: Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. **2**, 147–171 (2009). https://doi.org/10.1109/RBME.2009.2034865. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2910932/
40. Halicek, M., et al.: Head and neck cancer detection in digitized Whole-Slide histology using convolutional neural networks. Sci. Rep. **9**(1), 14043 (2019)
41. Harris, C.R., et al.: Quantifying and correcting slide-to-slide variation in multiplexed immunofluorescence images, July 2021
42. Hickey, J.W., et al.: Spatial mapping of protein composition and tissue organization: a primer for multiplexed antibody-based imaging. arXiv, July 2021
43. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. arXiv:1504.07947 [cs], March 2016. arXiv: 1504.07947

44. Huang, H., Chen, Z., Rudin, C.: SegDiscover: visual concept discovery via unsupervised semantic segmentation. arXiv, April 2022
45. Huttenlocher, D., Klanderman, G., Rucklidge, W.: Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. **15**(9), 850–863 (1993). https://doi.org/10.1109/34.232073
46. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. arXiv, February 2018
47. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? arXiv, April 2020
48. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. arXiv:1506.02025 [cs], February 2016. arXiv: 1506.02025
49. Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S.: Combining satellite imagery and machine learning to predict poverty. Science **353**(6301), 790–794 (2016)
50. Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., Ermon, S.: Tile2Vec: unsupervised representation learning for spatially distributed data. arXiv, May 2018
51. Jetley, S., Lord, N.A., Lee, N., Torr, P.H.S.: Learn to pay attention. arXiv, April 2018
52. Jiang, Y., Yang, M., Wang, S., Li, X., Sun, Y.: Emerging role of deep learning-based artificial intelligence in tumor pathology. Cancer Commun. **40**(4), 154–166 (2020)
53. Kandemir, M., Hamprecht, F.A.: Computer-aided diagnosis from weak supervision: a benchmarking study. Comput. Med. Imaging Graph. **42**, 44–50 (2015)
54. Kao, P.Y., et al.: Improving patch-based convolutional neural networks for MRI brain tumor segmentation by leveraging location information. Front. Neurosci. **13**, 1449 (2019)
55. Kawamura, Y., et al.: Histological and immunohistochemical evaluation of stroma variations and their correlation with the KI-67 index and expressions of glucose transporter 1 and monocarboxylate transporter 1 in canine thyroid C-cell carcinomas. J. Vet. Med. Sci. **78**(4), 607–612 (2016)
56. Kim, B., Reif, E., Wattenberg, M., Bengio, S., Mozer, M.C.: Neural networks trained on natural scenes exhibit gestalt closure. arXiv, March 2019
57. Kim, B., Seo, J., Jeon, S., Koo, J., Choe, J., Jeon, T.: Why are saliency maps noisy? Cause of and solution to noisy saliency maps. arXiv, February 2019
58. Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: the efficient transformer. arXiv:2001.04451 [cs, stat], February 2020. arXiv: 2001.04451
59. Klevesath, M.B., Bobrow, L.G., Pinder, S.E., Purushotham, A.D.: The value of immunohistochemistry in sentinel lymph node histopathology in breast cancer. Br. J. Cancer **92**(12), 2201–2205 (2005)
60. van der Laak, J., Ciompi, F., Litjens, G.: No pixel-level annotations needed. Nat. Biomed. Eng. **3**(11), 855–856 (2019)
61. Lähnemann, D., et al.: Eleven grand challenges in single-cell data science. Genome Biol. **21**(1), 31 (2020)
62. LeCun, Y., Bottou, L., Bengio, Y., Ha, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**, 46 (1998)
63. Lerousseau, M., Vakalopoulou, M., Deutsch, E., Paragios, N.: SparseconvMIL: sparse convolutional context-aware multiple instance learning for whole slide image classification. In: COMPAY 2021: The Third MICCAI Workshop on Computational Pathology (2021). https://openreview.net/forum?id=3byhkJb8FUj
64. Lewis, S.M., et al.: Spatial omics and multiplexed imaging to explore cancer biology. Nat. Methods **18**, 997–1012 ((2021)

65. Li, X.H., et al.: Quantitative evaluations on saliency methods: an experimental study. arXiv, December 2020
66. Liebel, L., Körner, M.: Auxiliary tasks in multi-task learning. arXiv, May 2018
67. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1) (2020)
68. Lipton, Z.C.: The mythos of model interpretability. arXiv, June 2016
69. Litjens, G., et al.: 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. Gigascience **7**(6) (2018)
70. Liu, Y., Zhuang, B., Shen, C., Chen, H., Yin, W.: Auxiliary learning for deep multi-task learning. arXiv, September 2019
71. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. arXiv, March 2021
72. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. **55**, 555–570 (2021)
73. Marcos, D., Volpi, M., Tuia, D.: Learning rotation invariant convolutional filters for texture classification. arXiv, April 2016
74. Matos, L.L.D., Trufelli, D.C., de Matos, M.G.L., da Silva Pinhal, M.A.: Immunohistochemistry as an important tool in biomarkers detection and clinical practice. Biomark. Insights **5**, 9–20 (2010)
75. Molnar, C., Casalicchio, G., Bischl, B.: Interpretable machine learning - a brief history, State-of-the-Art and challenges. arXiv, October 2020
76. Nathanson, S.D.: Insights into the mechanisms of lymph node metastasis. Cancer **98**(2), 413–423 (2003). https://doi.org/10.1002/cncr.11464
77. Nazeri, K., Aminpour, A., Ebrahimi, M.: Two-stage convolutional neural network for breast cancer histology image classification. arXiv:1803.04054 [cs] 10882, pp. 717–726 (2018). https://doi.org/10.1007/978-3-319-93000-881. arXiv: 1803.04054
78. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 427–436. IEEE, Boston, June 2015. https://doi.org/10.1109/CVPR.2015.7298640. https://ieeexplore.ieee.org/document/7298640/
79. Pesenson, M.Z., Pesenson, I.Z., McCollum, B.: The data big bang and the expanding digital universe: high-dimensional, complex and massive data sets in an inflationary epoch. Adv. Astron. **2010**, 1–16 (2010). https://doi.org/10.1155/2010/350891. arXiv: 1003.0879
80. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? arXiv, August 2021
81. Rahwan, I., et al.: Machine behaviour. Nature **568**(7753), 477–486 (2019)
82. Ras, G., Xie, N., van Gerven, M., Doran, D.: Explainable deep learning: a field guide for the uninitiated. arXiv, April 2020
83. Ratner, A., De Sa, C., Wu, S., Selsam, D., Ré, C.: Data programming: creating large training sets, quickly. Adv. Neural. Inf. Process. Syst. **29**, 3567–3575 (2016)
84. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. arXiv, February 2016
85. Robinson, J., Jegelka, S., Sra, S.: Strength from weakness: fast learning using weak supervision. arXiv, February 2020
86. Rost, S., Giltnane, J., Bordeaux, J.M., Hitzman, C., Koeppen, H., Liu, S.D.: Multiplexed ion beam imaging analysis for quantitation of protein expression in cancer tissue sections. Lab. Invest. **97**(8), 992–1003 (2017)

87. Sakamoto, T., et al.: A narrative review of digital pathology and artificial intelligence: focusing on lung cancer. Transl. Lung Cancer Res. **9**(5), 2255–2276 (2020)
88. Salvi, M., Acharya, U.R., Molinari, F., Meiburger, K.M.: The impact of pre- and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis. Comput. Biol. Med. **128**, 104129 (2021)
89. Schaumberg, A.J., et al.: Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. Mod. Pathol. **33**(11), 2169–2185 (2020)
90. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. arXiv, October 2016
91. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. **128**(2), 336–359 (2020). https://doi.org/10.1007/s11263-019-01228-7. arXiv:1610.02391
92. Shaban, M., et al.: Context-aware convolutional neural network for grading of colorectal cancer histology images. IEEE Trans. Med. Imaging **39**(7), 2395–2405 (2020)
93. Shamir, R.R., Duchin, Y., Kim, J., Sapiro, G., Harel, N.: Continuous dice coefficient: a method for evaluating probabilistic segmentations. arXiv, June 2019
94. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. arXiv:1511.04119 [cs], February 2016. arXiv: 1511.04119
95. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034 [cs], April 2014. arXiv: 1312.6034
96. Simonyan, K., Zisserman, A.: Very deep convolutional networks for Large-Scale image recognition. arXiv, September 2014
97. Sooriakumaran, P., Lovell, D.P., Henderson, A., Denham, P., Langley, S.E.M., Laing, R.W.: Gleason scoring varies among pathologists and this affects clinical risk in patients with prostate cancer. Clin. Oncol. **17**(8), 655–658 (2005)
98. Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M., Costa da Silva, E.: Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors (Basel, Switzerland) **19**(13) (2019). https://doi.org/10.3390/s19132969. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6651753/
99. Stamey, T.A., McNeal, J.E., Yemoto, C.M., Sigal, B.M., Johnstone, I.M.: Biological determinants of cancer progression in men with prostate cancer. JAMA **281**(15), 1395–1400 (1999). https://doi.org/10.1001/jama.281.15.1395. http://jamanetwork.com/journals/jama/fullarticle/189523 Association 7
100. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. arXiv:1512.00567 [cs], December 2015. arXiv: 1512.00567
101. Thompson, G.Z., Maitra, R.: CatSIM: a categorical image similarity metric. arXiv, April 2020
102. Tourniaire, P., Ilie, M., Hofman, P., Ayache, N., Delingette, H.: Attention-based multiple instance learning with mixed supervision on the camelyon16 dataset. In: COMPAY 2021: the third MICCAI Workshop on Computational Pathology (2021). https://openreview.net/forum?id=Z_L9j0HW3QM
103. Vaswani, A., et al.: Attention is all you need. arXiv, June 2017

104. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology. arXiv:1806.03962 [cs, stat], June 2018. arXiv: 1806.03962
105. Vinuesa, R., Sirmacek, B.: Interpretable deep-learning models to help achieve the sustainable development goals. Nat. Mach. Intell. **3**(11), 926–926 (2021)
106. Wagner, J., et al.: A single-cell atlas of the tumor and immune ecosystem of human breast cancer. Cell **177**(5), 1330-1345.e18 (2019). https://doi.org/10.1016/j.cell.2019.03.005
107. Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. arXiv:1606.05718 [cs, q-bio], June 2016. arXiv: 1606.05718
108. Wang, L., et al.: Learning to detect salient objects with image-level supervision. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, July 2017
109. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: an in-depth survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 3239–3259 (2021)
110. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
111. Wen, S., et al.: Comparison of different classifiers with active learning to support quality control in nucleus segmentation in pathology images. AMIA Jt Summits Transl. Sci. Proc. **2017**, 227–236 (2018)
112. Wilkinson, L., Anand, A., Grossman, R.: Graph-theoretic scagnostics. In: IEEE Symposium on Information Visualization 2005, INFOVIS 2005, pp. 157–164, October 2005
113. Xie, J., Xu, K., Li, Z., Bi, Q., Qin, K.: Building scene recognition based on deep multiple instance learning convolutional neural network using high resolution remote sensing image. In: Proceedings of the 2019 International Conference on Video, Signal and Image Processing, VSIP 2019, pp. 60–63. Association for Computing Machinery, New York, October 2019
114. Xu, H., Jiang, C., Liang, X., Li, Z.: Spatial-aware graph relation network for large-scale object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2019
115. Yeh, C., et al.: Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. Nat. Commun. **11**(1), 2583 (2020)
116. Yildirim, G., Sen, D., Kankanhalli, M., Süsstrunk, S.: Evaluating salient object detection in natural images with multiple objects having multi-level saliency. arXiv, March 2020
117. Yurtsever, E., Lambert, J., Carballo, A., Takeda, K.: A survey of autonomous driving: common practices and emerging technologies. IEEE Access **8**, 58443–58469 (2020)
118. Zhang, J.M., Harman, M., Ma, L., Liu, Y.: Machine learning testing: survey, landscapes and horizons. arXiv, June 2019
119. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-Supervised salient object detection via scribble annotations. arXiv, March 2020
120. Zhang, M., Sohoni, N.S., Zhang, H.R., Finn, C., Ré, C.: Correct-N-contrast: a contrastive approach for improving robustness to spurious correlations (2021)
121. Zhao, W., Du, S.: Learning multiscale and deep representations for classifying remotely sensed imagery. ISPRS J. Photogramm. Remote. Sens. **113**, 155–165 (2016)
122. Zhou, Z.H.: Multi-instance learning: a survey (2016)