Assessing and Enhancing Robustness of Deep Learning Models with Corruption Emulation in Digital Pathology

Peixiang Huang[1*], Songtao Zhang[1*], Yulu Gan[1], Rui Xu[1], Rongqi Zhu[1], Wenkang Qin[1],
Limei Guo[2], Shan Jiang[3], Lin Luo[1†]

[1] College of Engineering, Peking University, Beijing, China

[2] Third Hospital, Peking University Health Science Center, Beijing, China

[3] Institute of Biomedical Engineering, Beijing Institute of Collaborative Innovation, Beijing, China

{huangpx, ganyulu, xurui, qinwk}@stu.pku.edu.cn, {songtzhang, luol}@pku.edu.cn,
rongqizhu77@gmail.com, guolimei@bjmu.edu.cn, jiangs@jingjinji.cn

Abstract—Deep learning in digital pathology brings intelligence and automation as substantial enhancements to pathological analysis, the gold standard of clinical diagnosis. However, multiple steps from tissue preparation to slide imaging introduce various image corruptions, making it difficult for deep neural network (DNN) models to achieve stable diagnostic results for clinical use. In order to assess and further enhance the robustness of the models, we analyze the physical causes of the full-stack corruptions throughout the pathological life-cycle and propose an Omni-Corruption Emulation (OmniCE) method to reproduce 21 types of corruptions quantified with 5-level severity. We then construct three OmniCE-corrupted benchmark datasets at both patch level and slide level and assess the robustness of popular DNNs in classification and segmentation tasks. Further, we explore to use the OmniCE-corrupted datasets as augmentation data for training and experiments to verify that the generalization ability of the models has been significantly enhanced.

Index Terms—digital pathology, corruption, robustness.

I. INTRODUCTION

Pathological diagnosis is the gold standard for precise diagnosis and treatment for most diseases especially tumors and cancers. Digital pathology with high-resolution scanned images of pathological slides enables the use of deep learning algorithms in helping pathologists improve diagnostic efficiency and quality [1]. However, the whole process of producing pathological slides and digital images (Fig. 1) involves various corruptions such as artifacts in specimen preparation and image processing. These corruptions challenge deep neural network (DNN) models in the diagnostic reliability under clinical circumstances. DNN models are sensitive to general image corruptions such as Gaussian noises and Exposure variance and suffer from severe performance drop. The same problems occur on medical DNN with even worse impact.To tackle the robustness issue of DNNs, several works have introduced adversarial data or corrupted samples to the training process. [2] trained a robust model for medical image classification from noisy-labelled data. [3] conduct stress-testing on diagnostic models using synthetically generated artifacts for clinical validation.

In digital pathology, some works evaluate the performance of DNN models under different corruption types. [4] applies general image processing as corruptions on pathology images to benchmark robustness of various DNNs. [3] digitally reproduces twelve types of pathological artifacts using both image processing and image style transfer methods and discovers performance loss of DNN models in prostate cancer detection. However, few method covers the full-stack pathological corruptions encountered along the digital pathology life-cycle, and most works focus on the visual similarities rather than the physical causes when reproducing the corruptions. Besides, how effective the model robustness can be improved against the corruptions are not explored in depth,

In this paper we investigate how corruptions are generated throughout the full pathological life-cycle from tissue preparation to slide imaging. 21 types of corruptions are discovered, including *Over/Under Stained with H&E/H/E*, *Residual Wax/Xylene/Alkali*, *Thick and Thin Section*, *Over/Under Exposure*, *Defocus*, *Crack*, *Venetian*, *Fold*, *Knife Line*, *Bubble*, etc, while the causes such as human operations, materials, and device setups are analyzed.

Accordingly, we propose an Omni-Corruption Emulation (OmniCE) method to simulate the physical mechanisms of the causes with mechanical engine, optical engine, chemical engine, etc, to reproduce the realistic and controllable image corruptions, with physicians' know-how in designing the scales. The OmniCE corruption benchmark datasets cover the 21 types of corruptions throughout the pathological life-cycle, each is quantified with the corruption severity of 5 levels, from shallow to deep. Three typical pathology datasets are applied with OmniCE and used to assess the influence on typical DNN models.

Furthermore, we explore the use of the OmniCE corrupted data as augmentations of training data to improve DNN perfor-

---

*∗ These authors contributed equally to this work.
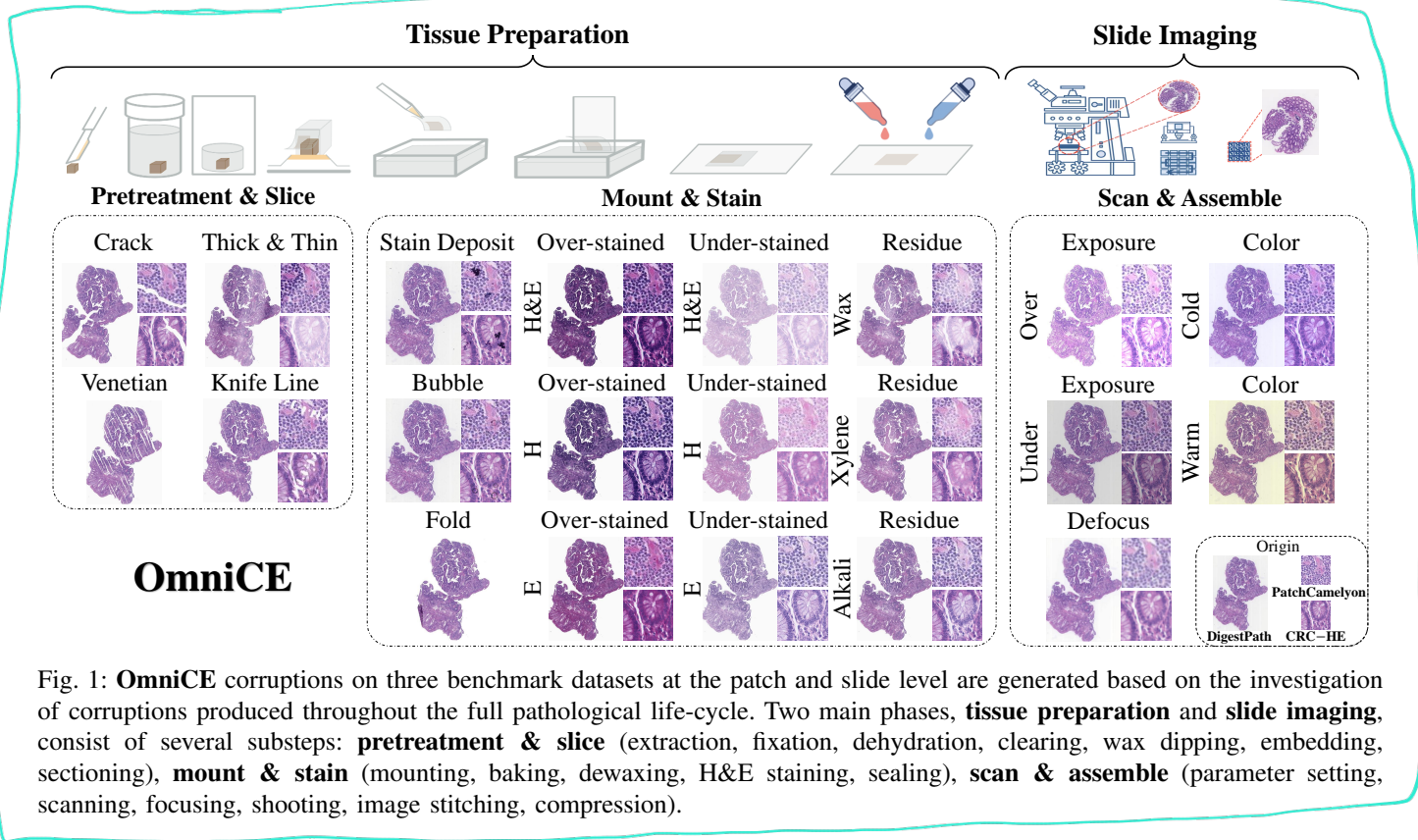
† Corresponding author.

Fig. 1: **OmniCE** corruptions on three benchmark datasets at the patch and slide level are generated based on the investigation of corruptions produced throughout the full pathological life-cycle. Two main phases, **tissue preparation** and **slide imaging**, consist of several substeps: **pretreatment & slice** (extraction, fixation, dehydration, clearing, wax dipping, embedding, sectioning), **mount & stain** (mounting, baking, dewaxing, H&E staining, sealing), **scan & assemble** (parameter setting, scanning, focusing, shooting, image stitching, compression).

mance. We compared the datasets augmented by OmniCE vs by Augmix [5] which covers the common image processing types. Experimental results show that the model trained by OmniCE-augmented datasets significantly outperforms that by Augmix-augmented ones of 8.3% and 15.3% on two different centers, and achieves the SOTA performance.

Our main contributions include: (1) We are the first to introduce the full-stack pathological corruption types that present throughout the digital pathology life-cycle. (2) We design the corruption emulator engines based on the underlying physical causes of corruptions to ensures that the corruptions reflect realistic clinical scenarios. (3) We experiment the pathology-specific corruption data as augmentation of training data and achieve a significant accuracy improvement in enhancing the model robustness compared with Augmix.

II. OMNI-CORRUPTION EMULATION

A. Physical Causes of Corruptions

As shown in Fig. 1, a full pathological lifecycle contains two main phases: tissue preparation and slide imaging, each consists of several sub-steps that involve corruption generation from human or device variance, as analyzed below.

1) Corruptions in Tissue Preparation: include *Crack*, *Venetian*, *Knife Line*, *Thick and Thin Section*, *Fold*, *Bubble*, *Stain Deposit*, *Stain Variance*, *Residue*. During sectioning, a defective blade may cause *Crack* (tearing of the tissue structure) or many fine lines called *Knife Lines*. And when disposable blades are not properly supported in the knife holder, tiny

vibrations in the knife edge will bring fine parallel cracking like *Venetian*. Besides, a loosely attached knife may form *Thick and Thin Sections* presenting banded areas of different staining levels adjacent to each other. In addition, there are some common impurities, such as the *Stain Deposit* comes with incomplete dissolution of the stains. And because of the unflattened slide, some small *Bubbles* and *Folds* are often be seen. When it comes to the staining, there may be irregular staining during the staining process, such as a change in the concentration or staining time of hematoxylin or eosin, then leads to *Over-staining* or *Under-staining*. Besides, some *Residues* can also lead to locally uneven staining. Sometimes slide dewaxing is incomplete, then *Residual Wax* results in unevenly H&E stained areas. And inefficient washing after "blueing" will leave *Residual Alkali* resulting in uneven eosin staining. *Residual Xylene* appears when hematoxylin solution rapidly, which causes uneven hematoxylin staining.

2) Corruptions in Slide Imaging: include *Color Cast*, *Exposure*, *Defocus*. *Color Cast* includes *Cold Color* and *Warm Color* caused by the color temperature of the microscope illumination. Besides, uneven illumination will introduce *Under-exposure* or *Overexposure*. And *Defocus* is a blur phenomenon due to inaccurate focusing.

B. Emulation of Corruptions

For emulation, we classify the lifecycle corruptions according to three types of their physical causes and the corresponding engines : Stain (*Stain Variance*, *Residue*, *Thick and Thin Section*) generated by **Chemical Engine**, Deformation
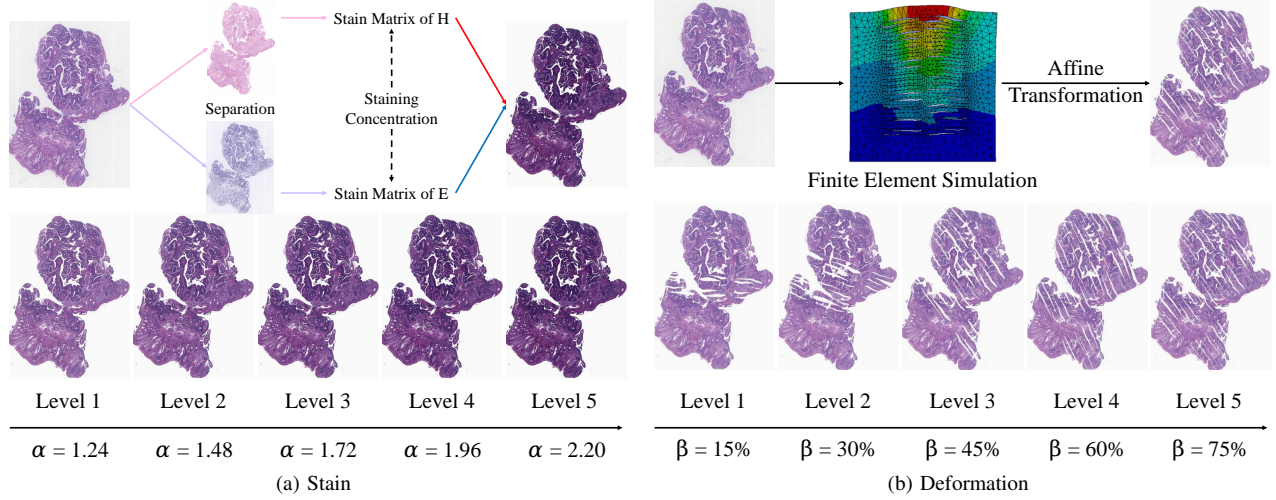
Fig. 2: Corruption emulations of 5-level severity. Take *Over-stained with H&E* corruption and *Venetian* corruption for example. α is used for controlling staining concentration and β denotes the area scale covered by venetian.

(*Crack*, *Venetian*, *Fold*) generated by **Mechanical Engine**, Color (*Exposure*, *Color Cast*), Defocus and Coverage (*Stain Deposit*, *Bubble*, *Knife Line*) generated by **Optical Engine**.

All of these corruptions are emulated in five levels. And indicators for each level have been discussed and confirmed with different senior pathologists to make ensure that OmniCE-corrupted images within a certain level range are possible in realistic clinical scenarios. After confirming the most severe level, even parameter intervals are assigned to different levels (Fig. 2).

The Stain corruption refers to the uneven staining concentrations of different stains, or even serious deviation from normal staining. The emulation uses Macenko's method [6] to separate the staining concentration of the pathological image and the color matrices of the two stains as shown in Fig. 2. Then the color matrices of the two stains are multiplied by different coefficients or assign different coefficients to different random areas to vary the staining concentration of different stains in the local area.

The Deformation corruption involves a large area of the slide. Specially to emulate *Crack* and *Venetian* more realistically, we use ANSYS WORKBENCH [7] to emulate the mechanical properties of polymer materials to approximate the deformation of tissue sections when being stretched by various external forces. Firstly, irregular cracks are preset on the model of the thin section. After setting the material and Young's modulus on the software, we emulate the applied force and obtain the templates before and after deformation. Then templates are randomly rotated and used for affine transformation of image tiles within each mesh on the templates to get the whole slide image (WSI) after the deformation corruption as shown in Fig. 2. *Fold* is also achieved by affine transformation of some preset folding templates. The overlapping regions uses sum-up in the optical density space to get a more reasonable visual effect.

For Color corruptions, we adjust the proportion of red and blue channels for *Color Cast* or scale pixel values for *Exposure*. Then *Defocus* corruption is emulated by setting parameters, i.e. RGB illumination wavelength, refractive index and objective NA, to quantitatively generate point spread functions (PSF) of RGB center wavelengths at different locations, and then convolving images of corresponding color channels with defocus PSFs. And for Coverage corruptions, we overlay with preset templates in the optical density space.

Due to the limited number of pages, we will organize and open-source all the detailed formulas, related quantitative parameters and corruptions generating codes involved here after publication as soon as possible.

III. EXPERIMENT

A. OmniCE Corruption Benchmark Datasets

We collect two different types of pathology datasets, which are patch level and slide level, for omni-corruption emulation and robustness evaluation. Then we train on clean images and test on these benchmark datasets with corruption.

1) Patch-level Dataset: Given consideration to different tissue types, we select two datasets derived from lymph node and colon tissue respectively, i.e. Patchcamelyon [8] and CRC-HE dataset [9], for benchmarking.

In PatchCamelyon, patches with the size of 96×96 are extracted from slides of potentially metastatic breast cancer, the label is whether the patch contains tumor. Here, we have removed duplicates and normalize the staining style of images from different centers (1,2,3) for more rational grading of staining corruptions. Finally, we obtain 208,401 training examples and 10,000 remaining examples for omni-corruption emulation as synthetic distribution shift to benchmark the robustness. The data from Center 4, 5 are used to benchmark the natural distribution shift. The CRC-HE dataset contains 100,000 patches with 224×224 pixels divided into 9 classes for training. And for the validation, there are 7180 patches

OmniCE	AlexNet mCE	rCE	VGG16 mCE	rCE	ResNet18 mCE	rCE	ResNet50 mCE	rCE	ResNet101 mCE	rCE	DenseNet121 mCE	rCE	AlexNet mCE	rCE	VGG16 mCE	rCE	ResNet18 mCE	rCE	ResNet50 mCE	rCE	ResNet101 mCE	rCE	DenseNet121 mCE	rCE
						PatchCamelyon											CRC-HE							
Under-stained H&E	100.0	**2.3**	55.8	3.9	**45.3**	3.1	55.6	3.7	55.1	3.7	62.6	5.1	100.0	4.6	**48.0**	4.8	75.4	4.6	71.1	**4.1**	<u>111.6</u>	<u>7.4</u>	67.1	5.3
Over-stained H&E	<u>100.0</u>	2.0	**31.3**	1.9	37.6	2.3	55.6	<u>3.3</u>	44.1	2.6	46.1	<u>3.3</u>	100.0	**3.3**	54.6	3.9	<u>140.7</u>	6.1	104.5	4.4	114.8	5.5	113.4	<u>6.4</u>
Under-stained H	<u>100.0</u>	2.3	59.7	4.2	**52.9**	3.7	57.6	3.8	53.4	3.6	53.6	4.4	100.0	4.6	**81.5**	8.1	97.5	5.9	108.9	6.3	<u>128.4</u>	<u>8.5</u>	88.0	6.8
Over-stained H	<u>100.0</u>	1.9	32.0	1.8	36.6	2.1	44.1	<u>2.4</u>	28.9	**1.6**	27.3	1.8	100.0	2.7	**64.9**	3.8	141.0	5.1	151.9	5.2	<u>154.3</u>	6.1	102.7	4.8
Under-stained E	<u>100.0</u>	**1.5**	42.2	1.9	**34.2**	1.6	39.6	1.7	35.0	**1.5**	36.8	2.0	100.0	2.4	47.4	<u>2.4</u>	66.7	2.1	54.8	**1.6**	49.7	1.7	**44.3**	1.8
Over-stained E	<u>100.0</u>	2.3	**43.4**	3.0	61.3	4.2	78.6	5.2	51.3	3.4	66.4	<u>5.4</u>	100.0	3.2	67.6	4.7	<u>151.5</u>	6.4	122.8	5.0	110.3	5.1	126.5	<u>6.9</u>
Residual Wax	<u>100.0</u>	1.7	37.4	1.9	**33.1**	1.7	37.2	1.8	36.2	1.8	40.6	2.4	100.0	1.6	37.4	1.3	58.0	**1.2**	67.0	1.4	94.4	<u>2.2</u>	53.3	1.5
Residual Xylene	<u>100.0</u>	1.7	34.2	1.8	**30.7**	**1.6**	35.6	1.7	35.9	1.8	31.1	1.9	100.0	1.5	38.5	1.2	58.7	**1.1**	73.4	1.4	<u>113.8</u>	<u>2.4</u>	68.4	1.7
Residual Alkali	<u>100.0</u>	**1.3**	36.3	1.4	32.8	**1.3**	37.8	1.4	36.1	**1.3**	32.3	1.5	100.0	**1.3**	43.2	1.2	56.2	**1.0**	62.0	**1.0**	60.2	1.1	50.1	1.1
Thick and Thin	<u>100.0</u>	1.8	35.2	1.9	**33.9**	1.8	42.3	2.1	39.5	2.0	42.8	2.7	100.0	2.4	**39.4**	**2.0**	105.0	3.3	86.3	2.6	<u>112.7</u>	<u>3.9</u>	91.0	3.7
Stain Deposit	<u>100.0</u>	<u>1.4</u>	25.6	1.1	23.5	**1.0**	26.9	1.1	26.1	1.1	**22.5**	1.1	100.0	1.2	42.8	1.1	65.1	**1.0**	75.2	<u>1.2</u>	61.7	1.1	53.3	1.1
Bubble	<u>100.0</u>	**1.1**	42.9	1.4	46.5	1.5	53.7	<u>1.6</u>	51.0	<u>1.6</u>	**37.4**	1.4	100.0	1.1	66.2	<u>1.6</u>	78.5	1.2	71.2	**1.0**	64.3	**1.0**	**63.4**	1.2
Knife Line	<u>100.0</u>	**1.1**	35.5	<u>1.1</u>	31.8	**1.0**	34.6	<u>1.1</u>	33.6	**1.0**	29.1	1.1	100.0	1.5	**31.5**	**1.0**	49.7	**1.0**	54.8	1.1	44.9	**1.0**	39.9	**1.0**
Crack	<u>100.0</u>	**1.2**	53.6	2.0	56.3	2.1	54.5	1.9	50.7	1.8	49.3	2.2	100.0	2.5	40.2	2.2	38.5	**1.3**	49.0	1.5	50.0	1.8	**32.7**	1.4
Cold Color	100.0	**2.0**	88.6	5.3	**71.3**	4.3	84.6	4.8	<u>106.3</u>	<u>6.1</u>	79.0	5.6	100.0	3.2	**43.6**	3.1	102.9	4.4	113.4	4.6	153.0	7.1	<u>191.8</u>	<u>10.6</u>
Warm Color	100.0	**3.0**	83.1	7.6	89.9	8.2	91.0	7.9	85.5	7.5	84.5	9.1	100.0	**7.5**	77.8	<u>12.6</u>	93.6	9.2	93.2	8.8	95.5	10.3	82.6	10.5
Overexposure	**100.0**	**2.0**	<u>127.3</u>	7.7	112.7	6.8	125.9	7.2	<u>142.2</u>	8.2	121.1	8.6	**100.0**	**3.0**	168.9	11.0	251.9	9.9	226.9	8.6	<u>280.1</u>	12.1	263.5	<u>13.5</u>
Underexposure	100.0	**1.7**	90.9	4.8	<u>117.3</u>	6.2	114.3	5.7	<u>136.1</u>	6.9	107.3	6.6	100.0	2.0	78.8	3.4	79.4	2.1	<u>154.6</u>	<u>3.9</u>	124.8	3.6	94.5	3.2
Defocus	100.0	**2.4**	97.0	7.0	97.4	7.0	<u>100.6</u>	6.9	**81.7**	5.7	90.4	<u>7.7</u>	**100.0**	2.7	100.9	5.9	150.8	5.4	128.3	4.4	<u>206.7</u>	<u>8.1</u>	164.2	7.6
Average	100.0	**1.8**	55.4	3.3	**55.0**	3.2	61.6	3.4	59.4	3.3	55.8	3.9	100.0	**2.8**	61.7	4.0	98.0	3.8	98.4	3.6	<u>112.2</u>	<u>4.7</u>	94.3	<u>4.7</u>
Original Error	<u>12.9</u>		4.2		4.2		4.5		4.4		**3.6**		<u>8.4</u>		**3.9**		6.4		6.7		5.9		5.0	

with colorectal adenocarcinoma which are used to emulate 19 different types of corruptions.

We use AlexNet [10], VGG16 [11], ResNet18/50/101 [12] and DenseNet121 [13] for training and testing. As for metrics, error rate $Error^f$ of each model f is computed for original images and $CE^f_{s,c}$ for each corruption type c and severity level s, we can then get the mean corruption error rate mCE^f_c for a single corruption type c by taking the average across all severity levels for that corruption. As presumably not all corruptions are equally difficult, we adjust by a baseline which in our case is AlexNet's corruption error rate $CE^{AlexNet}_{s,c}$. Thus, we get:

$$mCE^f_c = (\sum_{s=1}^{5} CE^f_{s,c})/(\sum_{s=1}^{5} CE^{AlexNet}_{s,c}) \quad (1)$$

And we also introduce:

$$rCE^f_c = (\frac{1}{5}\sum_{s=1}^{5} CE^f_{s,c})/(Error^f) \quad (2)$$

It measures how gracefully a classifier degrades in the presence of corruptions.

2) Slide-level Dataset: The DigestPath dataset [14] is used for slide level colorectal tissue segmentation, including 872 tissue sections with an average size of 3000 × 3000. We take 172 slide images color-normalized globally for OmniCE. In addition, two extra corruptions (*Fold* and *Venetian*) are emulated specific to slide images. The two corruptions are also common in clinical settings, they are more suitable for application and quantification on slide images and has great significance in studying the impact of corruptions on segmentation. It is worth noting that both original images and masks

are deformed in the same way. The following segmentation models, UNet [15], Deep Contour-aware Network (DCAN) [16], Global Convolutional Network (GCN) [17] and Dense-UNet [18] are chosen to train and test, and the dice score is used as the main metric.

B. Robustness Evaluation

1) Patch-level Benchmark: As shown in Table I, our experiment investigates effects of corruptions on model performance in the patch classification task. For corruptions in the top six rows of the table, which emulated by the staining engine, we observe that under-stained with H&E patches result in lower model performance compared to over-stained with H&E patches, since what is inline with our common sense, the under-stained patches may reduce contrast information that is critical for discrimination. We also find that over-stained with H patches result in higher model performance compared to under-stained with H patches, which is expected given more distinguishable nuclei. Conversely, over staining with E will indirectly affect the contrast of nuclei, which has a negative impact on model performance. For other 4 stain-related corruptions, both the mCE and rCE scores are lower than the under-stained ones due to smaller stain-corrupted regions, while the staining shift of the whole patch severely degrades model discrimination. For the next 4 deformation and coverage corruptions, we find that they have a more subtle effect on model performance compared to stain-related corruptions. And the deformation corruption has a relatively strong effect on model performance compared to coverage corruptions due to the fact that the deformation engine can
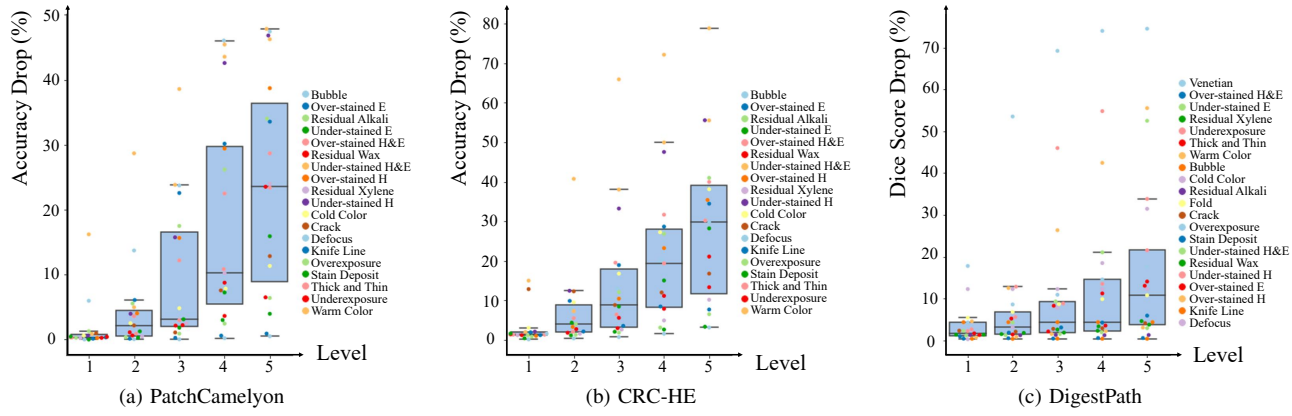
Fig. 3: Box plots of dropped metric values with different severity levels on different benchmarks. Dropped metric values are calculated by subtracting metric values of corrupted images from metric values of clean images.

TABLE II: Dice score results (↑) of the slide level OmniCE corruption benchmark dataset. The best and worst results are marked in the same way as in Table I.

OmniCE	UNet	GCN	DCAN	Dense-UNet
Under-stained H&E	**0.6138**	0.5668	<u>0.3968</u>	0.5724
Over-stained H&E	**0.7438**	0.7137	<u>0.6088</u>	0.7194
Under-stained H	0.3450	0.4454	0.2415	**0.4496**
Over-stained H	**0.7435**	0.7291	<u>0.6378</u>	0.7416
Under-stained E	**0.7620**	0.7283	<u>0.6183</u>	0.7349
Over-stained E	0.6312	0.6649	<u>0.4729</u>	**0.6768**
Residual Wax	**0.7690**	0.7243	<u>0.6560</u>	0.6950
Residual Xylene	**0.7620**	0.7184	0.6482	0.6882
Residual Alkali	**0.7736**	0.7332	0.6635	0.7404
Thick and Thin	**0.7535**	0.7030	<u>0.6281</u>	0.7091
Stain Deposit	**0.7749**	<u>0.7408</u>	**0.7749**	0.7504
Bubble	0.7750	<u>0.7419</u>	**0.7751**	0.7539
Knife Line	0.6912	**0.7023**	<u>0.6721</u>	0.7015
Crack	**0.7651**	0.7174	<u>0.6690</u>	0.7110
Fold	**0.7253**	0.6626	<u>0.6254</u>	0.6499
Venetian	**0.6813**	0.6340	<u>0.6183</u>	0.6512
Cold Color	**0.7727**	0.6234	<u>0.2383</u>	0.6337
Warm Color	**0.6745**	0.4629	<u>0.1168</u>	0.3573
Overexposure	**0.3937**	0.1673	<u>0.1100</u>	0.3614
Underexposure	**0.6622**	0.6424	<u>0.3888</u>	0.6098
Defocus	**0.7181**	0.6231	<u>0.612</u>	0.6731
Average	**0.6920**	0.6402	<u>0.5320</u>	0.6467
Original	0.7812	<u>0.7457</u>	**0.7831**	0.7583

change the shape of some areas in the image, while coverage corruptions mainly result in the loss of some pixels.

We also explore the effects of optical imaging corruptions and our results indicate that these corruptions have the greatest impact on model performance. The color temperature degradation introduces the domain shift, which has more devastating effects on models that are trained and fitted better on the original image domain, such as DenseNet121. Furthermore, overexposure and defocus corruptions are found to be more detrimental to model performance, likely due to the greater amount of information loss.

Regarding the above conclusion, it can also be seen more intuitively from Fig. 3.

2) Slide-level Benchmark: As for the segmentation benchmark shown in Table II, the trend of most corruptions on model performance is consistent with the previous analysis of Table I. But for coverage corruptions, they may block the local features for discriminating the image category at the patch level. But at the slide level, semantic segmentation tends to capture the global feature for structural discrimination, hence, the local coverage will cause a lower impact on the segmentation task. By the same token, deformation corruptions disrupt the original morphology of the tissue slice and bring more serious impact on the global feature extraction than individual patches.

3) Corruption Level and Performance Drop : What's more, we explore the model's robustness to image features under different levels of corruption in Fig. 3. For different benchmark datasets, the performance drop of the model shows a strong consistency with the level of corruption, that is, images with higher corruption levels are more difficult to be classified by the model.

In summary, performances of models on these corrupted images conform to common sense and pathological prior knowledge, proving that our OmniCE can effectively emulate corruptions of different causes in real scenes, and the level of corruption accurately reflects the quality of image features.

C. Augmentation with OmniCE-Corrupted Dataset

In addition, we have explored data augmentation in conjunction with our corruptions and the training dataset of PatchCamelyon [8], which is obtained from three hospital centers, is chosen for model training, then the data from two other different centers is used for testing. We train with SE-ResNeXt101 [19] without pretrained models as our baseline and then replace original corruptions on natural images with OmniCE corruptions proposed in this paper for the operation pool of Augmix [5], which effectively applies corruptions for data augmentation, to leverage support of the OmniCE-corrupted dataset.

As shown in Table III, our method further enhances the model generalization compared to common Augmix and sur-

TABLE III: Performance comparison on Camelyon val set and test set.

Algorithm	Backbone	Val Acc(Center 4)	Test Acc (Center 5)
CORAL [20]	DenseNet121	86.2	59.5
IRM [20]	DenseNet121	86.2	64.2
CGD [21]	DenseNet121	86.8	69.4
Fish [22]	DenseNet121	83.9	74.1
LISA [23]	DenseNet121	81.8	77.1
ERM w/ data aug [24]	DenseNet121	90.6	82.0
ERM w/ targeted aug [25]	DenseNet121	92.7	92.1
ERM w/ H&E jitter [26]	SE-ResNeXt101	88.0	91.6
Normal Training	SE-ResNeXt101	75.3	57.2
Augmix [5]	SE-ResNeXt101	86.6	76.9
Ours (+OmniCE)	**SE-ResNeXt101**	**94.9**	**92.2**

passes the existing methods including different augmentations (e.g., normal H&E jitter) on the leaderboard.

IV. CONCLUSION

In this paper, we firstly analyse physical causes of 21 types of corruptions throughout the pathological life-cycle and propose an OmniCE emulator to construct benchmark datasets for evaluating the robustness of typical DNNs in digital pathology. Furthermore, the OmniCE-corrupted dataset is used for data augmentation during model training, which is validated on the multicenter data of Camelyon and obtains a significant improvement in generalisation capability.

ACKNOWLEDGMENT

REFERENCES

[1] W. Qin, R. Xu, S. Jiang, T. Jiang, and L. Luo, "Pathtr: Context-aware memory transformer for tumor localization in gigapixel pathology images," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3603–3619. 1

[2] C. Xue, L. Yu, P. Chen, Q. Dou, and P.-A. Heng, "Robust medical image classification from noisy labeled data with global and local representation guided co-training," *IEEE Transactions on Medical Imaging*, vol. 41, no. 6, pp. 1371–1382, 2022. 1

[3] B. Schömig-Markiefka, A. Pryalukhin, W. Hulla, A. Bychkov, J. Fukuoka, A. Madabhushi, V. Achter, L. Nieroda, R. Büttner, A. Quaas, and Y. Tolkach, "Quality control stress test for deep learning-based diagnostic model in digital pathology." *Modern Pathology*, 2021. 1

[4] Y. Zhang, Y. Sun, H. Li, S. Zheng, C. Zhu, and L. Yang, "Benchmarking the robustness of deep neural networks to common corruptions in digital pathology," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 242–252. 1

[5] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv preprint arXiv:1912.02781*, 2019. 2, 5, 6

[6] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009, pp. 1107–1110. 3

[7] H.-H. Lee, *Finite element simulations with ANSYS Workbench 18*. SDC publications, 2018. 3

[8] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 210–218. 3, 5

[9] J. N. Kather, N. Halama, and A. Marx, "100,000 histological images of human colorectal cancer and healthy tissue," Apr. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1214456 3

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. 4

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 4

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 4

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708. 4

[14] Q. Da, X. Huang, Z. Li, Y. Zuo, C. Zhang, J. Liu, W. Chen, J. Li, D. Xu, Z. Hu, H. Yi, Y. Guo, Z. Wang, L. Chen, L. Zhang, X. He, X. Zhang, K. Mei, C. Zhu, W. Lu, L. Shen, J. Shi, J. Li, S. S, G. Krishnamurthi, J. Yang, T. Lin, Q. Song, X. Liu, S. Graham, R. Bashir, C. Yang, S. Qin, X. Tian, B. Yin, J. Zhao, D. Metaxas, H. Li, C. Wang, and S. Zhang, "Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system," *Medical Image Analysis*, vol. 80, Aug. 2022, publisher Copyright: © 2022 Elsevier B.V. 4

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241. 4

[16] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "Dcan: deep contour-aware networks for accurate gland segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 2487–2496. 4

[17] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4353–4361. 4

[18] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network," *Quantitative imaging in medicine and surgery*, vol. 10, no. 6, p. 1275, 2020. 4

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. 5

[20] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5637–5664. 6

[21] V. Piratla, P. Netrapalli, and S. Sarawagi, "Focus on the common good: Group distributional robustness follows," *arXiv preprint arXiv:2110.02619*, 2021. 6

[22] Y. Shi, J. Seely, P. H. Torr, N. Siddharth, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," *arXiv preprint arXiv:2104.09937*, 2021. 6

[23] H. Yao, Y. Wang, S. Li, L. Zhang, W. Liang, J. Zou, and C. Finn, "Improving out-of-distribution robustness via selective augmentation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 407–25 437. 6

[24] S. Sagawa, P. W. Koh, T. Lee, I. Gao, S. M. Xie, K. Shen, A. Kumar, W. Hu, M. Yasunaga, H. Marklund *et al.*, "Extending the wilds benchmark for unsupervised adaptation," *arXiv preprint arXiv:2112.05090*, 2021. 6

[25] I. Gao, S. Sagawa, P. W. Koh, T. Hashimoto, and P. Liang, "Out-of-distribution robustness via targeted augmentations," in *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*. 6

[26] D. Tellez, M. Balkenhol, I. Otte-Höller, R. van de Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, G. Litjens, J. van der Laak, and F. Ciompi, "Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks," *IEEE Transactions on Medical Imaging*, vol. 37, no. 9, pp. 2126–2136, 2018. 6