Understanding the impact of image and input resolution on deep digital pathology patch classifiers

Eu Wern Teh
School of Engineering, University of Guelph
Vector Institute for Artificial Intelligence
eteh@uoguelph.ca

Graham W. Taylor
School of Engineering, University of Guelph
Vector Institute for Artificial Intelligence
gwtaylor@uoguelph.ca

***Abstract*—We consider annotation efficient learning in Digital Pathology (DP), where expert annotations are expensive and thus scarce. We explore the impact of image and input resolution on DP patch classification performance. We use two cancer patch classification datasets PCam and CRC, to validate the results of our study. Our experiments show that patch classification performance can be improved by manipulating both the image and input resolution in annotation-scarce and annotation-rich environments. We show a positive correlation between the image and input resolution and the patch classification accuracy on both datasets. By exploiting the image and input resolution, our final model trained on $<$ 1% of data performs equally well compared to the model trained on 100% of data in the original image resolution on the PCam dataset.**

***Keywords*-Digital Pathology, Patch Classification, Annotation-efficient Learning**

I. Introduction

Digital Pathology (DP) is a medical imaging field where microscopic images are analyzed to perform Digital Pathology tasks (e.g., cancer diagnosis). The appearance of hardware for a complete DP system popularized the usage of whole slide images (WSI) [1]. These whole slide images are high-resolution images, usually gigapixel in size. Without resizing WSIs, these high-resolution images will not fit in the memory of off-the-shelf GPUs. On the other hand, resizing WSIs destroys fine-grained information crucial in DP tasks, such as cancer classification. A common practice is to divide WSIs into small patches to address this issue. Most of the DP solutions for patch classification use single-scale images, where a fixed zoom-level of WSI is used [2], [3], [4], [5]. Therefore, a standard augmentation strategy such as Random-resized-crop will not work in patch classification (more details are provided in Section IV-E).

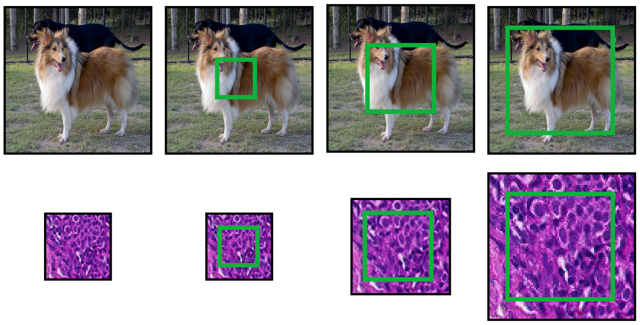

Figure 1. An illustration of the differences between input and image resolutions. In both rows of images, input resolution increases from left to right. However, only the bottom row has an increase in image resolution. In Digital Pathology, an increase in both the image and input resolution allows a model to capture fine-grained information without losing the global context of a given image.

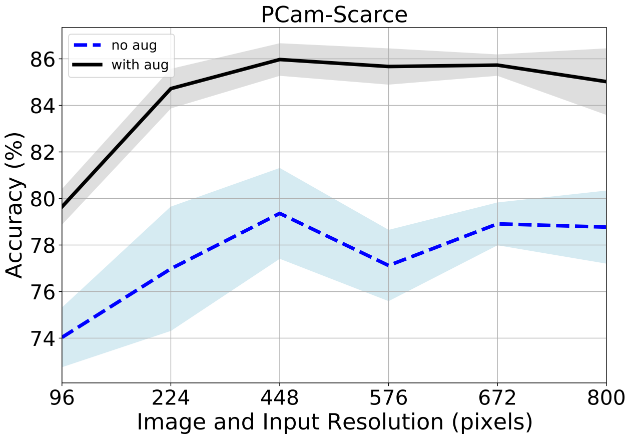

Figure 2. Image and input resolution effects on models trained with $<$ 1% of the original dataset. The shaded areas represent one standard deviation of uncertainty. The dotted line represents models trained without data augmentation.

The Deep Learning field has focused on innovating architectures ranging from AlexNet to ResNet to Transformers [6], [7], [8]. “The bigger the model, the better” approach works when a massive quantity of human annotations is available. However, in an annotation-scarce environment, models with higher capacity often perform worse due to overfitting [9]. We take a step back from designing complex methods and explore image and input resolution factors, which are often ignored as part of an input processing pipeline. We show that we can improve patch classification with a fixed zoom level by increasing the image and input resolution of image patches in DP (Figure 2). By increasing the image and input resolution, we allow our model to focus on fine-grained information without losing coarse-grained information, resulting in an increase of 6.33% (PCam-Scarce) and 9.83% (CRC-Scarce) in patch classification accuracy.

II. Related work

Annotation-efficient learning: In the Digital Pathology (DP) domain, the availability of human labels is often scarce due to the high annotation costs [5]. This data scarcity is a challenge for supervised deep learning models as these models often require tremendous amounts of human labels to be effective [9]. An effective way of combating label scarcity is via transfer learning in the form of pre-training [10]. Transfer learning from a pre-trained ImageNet model is shown to be effective in the DP domain, especially in the low-data regime [11], [12], [13].

The effect of input resolution on model performance: There is an important distinction between image resolution and input resolution (Figure 1). Image resolution refers to the width and height of an image in pixels, while input resolution refers to the input width and height being fed to a model. Random-resized-crop augmentation is a standard input augmentation technique used in the Natural Image (NI) domain for image classification [14], [15], [7]. During Random-resized-crop augmentation, the image resolution remains the same throughout training, while the crop area changes [16]. The crop area is randomly scaled between 0.08 to 1.00 of the image resolution, followed by another random re-scaling between 0.75 to 1.33. Finally, the crop area is resized to the given input resolution.

In the NI domain, classification performance can be improved by increasing the input resolution (crop size) [14], [15]. Touvron et al. discover that different training and testing input resolution affects model performance [14]. The most optimal train and test input resolution for the ImageNet dataset are 384×384 and 448×448. Tang et al. push the limits of their EfficientNet model by increasing the input resolution of the images [15]. Teh et al. show that input resolution has a large influence on image retrieval performance [17]. It is reasonable to use Random-resized-crop in the NI domain, as NI images come with a wide range of different image resolutions (Figure 8). However, in the DP domain, Random-resized-crop is ineffective for patch classification because image patches have the same image resolution (fixed zoom-level) throughout the dataset (Section IV-E).

III. Methodology and Control Factors

Supervised-training: A ResNet-34 architecture is used as our model in all of our experiments[1] where we trained them with cross-entropy loss as described in Equation 1 [7]. B denotes the batch size, x_i denotes a single image with a corresponding image label, y_i, and K denotes the number of classes in the respective dataset. $f(x_i)$ represents a deep model that accepts an image, x_i and returns a vector of size K denoting the classification scores.

$$L = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp(f(x_i)[y_i])}{\sum_{k=1}^{K} \exp(f(x_i)[k])} \tag{1}$$

Model initialization: In Section IV-A and IV-B, we use randomly initialized weights in our experiments. In Section IV-C, we use both randomly initialized and ImageNet pre-trained ResNet-34 model to explore the limits of annotation-efficient learning for patch classification in DP.

Data augmentation: We perform data augmentation via random rotation, random cropping, random horizontal flipping, and color jittering following Teh et al. [12] (Algorithm 1). Each image is resized with matching height, h and width, w via bilinear interpolation. As image resolution increases, the receptive field of each convolution layer in the model increases, resulting in model awareness towards finer-grained details in the images (Figure 3). Additionally, we also perform experiments without data augmentation in Section IV-A (Algorithm 3).

Dataset demography: We use the Patch Camelyon (PCam) and the Colorectal cancer (CRC) dataset in our experiments. For each dataset, we create two subsets: scarce and full. The scarce subset consists of <1 % of the dataset, while the full subset consists of 100% of the dataset.

The PCam dataset consists of 262,144 training images, 32,768 validation images, and 32,768 test images [2]. It comprises two classes: normal tissue and tumor tissue. Each image has a dimension of 96×96 pixels with a resolution of 0.972μm per pixel. On the scarce subset of PCam, we randomly extract 2000 images from the training set and 2000 images from the validation set. The 2000 validation images are used in both scarce and full experiments to determine early stopping. There is approximately 0.76% of images in PCam-Scarce training set when compared to the full training set.

The CRC dataset consists of 100,000 images, and we split the dataset into 70% training set, 15% validation set, and 15% test set following Kather et al. [3]. It consists of nine classes: adipose tissue, background, debris, lymphocytes, mucus, smooth muscle, normal colon tissue, cancer-associated stroma, and colorectal adenocarcinoma epithelium. Each image has a dimension of 224×224 pixels with a resolution of 0.5μm per pixel. On the scarce subset of CRC, we randomly extract 693 images (77 images per class) from the training set and 693 images (77 images per class) from the validation set. The 693 validation images

[1] With the exception of experiments in Table I where we explore ResNet-152 and ResNeXt-101 architectures [18].

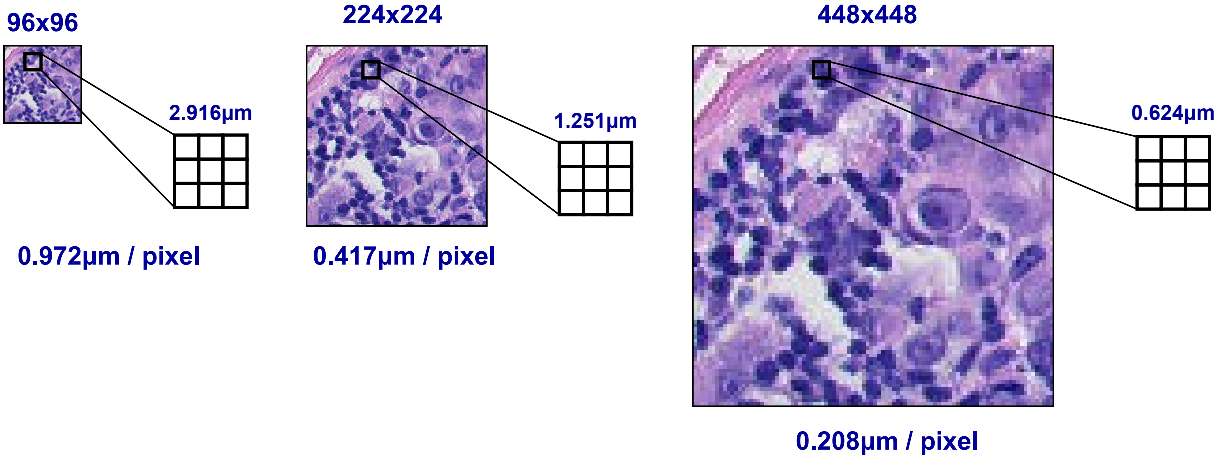

Figure 3. An illustration of image resolution and the corresponding receptive fields of a 3×3 convolution layer at various image resolutions. The leftmost image is resized twice (224×224 and 448×448) via bilinear interpolation from an original image at 96×96 resolution with 0.972μm/pixel. As image resolution increases, the receptive field decreases, allowing finer-grained features to be captured by the convolution layer.

are used in both scarce and full experiments to find the best epoch at which to early stop. There is approximately 0.99% of images in CRC-Scarce training set when compared to the full training set.

Additional Experimental Settings: We train all our models on the scarce subset for 100 epochs and 20 epochs for the full subset of the corresponding dataset. We use a validation set to select the best epoch and a separate test set is used for model evaluation. All models are trained with the Adam optimizer with a learning rate of $1e^{-4}$, a weight decay of $5e^{-3}$, and a batch size of 32. We train all models five times with different seeds [2] and report the mean accuracy with one standard deviation of uncertainty. Furthermore, we also use the SciPy Pearson library to compute the Pearson Correlation coefficient between test accuracy and the input and image resolution.

IV. Experiments

We examine the effects of image and input resolution in three different settings: Annotation-scarce environments (§ IV-A), Annotation-rich environments (§ IV-B), and Transfer Learning settings (§ IV-C).

A. The effects of image and input resolution on annotation-scarce dataset

We analyze the impact of image and input resolution on two annotation-scarce datasets: PCam-Scarce and CRC-Scarce with and without data augmentation (Figures 2 and 4). On models with augmentation, there is a positive correlation of 0.703 (PCam-Scarce) and 0.944 (CRC-Scarce) between the image and input resolution and patch accuracy. Similarly, there is also a correlation of 0.771 (PCam-Scarce) and 0.938 (CRC-Scarce) on models without augmentation. Models' performance suffers in general without data augmentation on the PCam-Scarce dataset. Nevertheless, on the CRC-Scarce dataset, models without data augmentation surpass models with data augmentation as image and input resolution approach 576×576 and beyond. This result shows that an increase of image and input resolution alone may be sufficient to regularize the network on the CRC-Scarce dataset. Additionally, this result also shows that too much regularization could hurt generalization on test set.

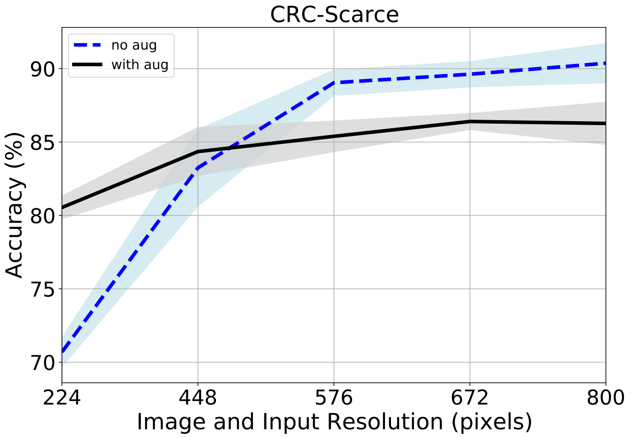

Figure 4. Image and input resolution effects on models trained with $< 1\%$ of the original dataset. The shaded areas represent one standard deviation of uncertainty. The dotted line represents models trained without data augmentation.

Table I shows our experimental results on the PCam-

[2] Random seeds affect the weight initialization of our models.

			PCam-Scarce	CRC-Scarce
Models	Resolution	GFLOPs	Accuracy (%)	Accuracy (%)
ResNet-34	96×96	0.68	79.64±0.74	72.17±1.65
ResNet-34	224×224	3.68	84.72±0.83	80.54±0.78
ResNet-34	448×448	14.73	**85.97±0.68**	84.35±1.64
ResNet-34	576×576	24.35	85.67±0.76	85.39±1.05
ResNet-34	672×672	33.14	85.73±0.44	**86.40±0.55**
ResNet-34	800×800	46.97	85.02±1.41	86.27±1.44
ResNet-152	96×96	2.14	77.47±1.57	55.68±1.00
ResNet-152	224×224	11.63	79.97±1.10	65.34±1.44
ResNeXt-101	96×96	3.04	79.34±0.60	61.53±0.89
ResNeXt-101	224×224	16.55	80.06±0.77	71.21±1.67

Table I

Patch classification accuracy of our models trained on the PCam-Scarce dataset and CRC-Scarce dataset. All models are trained with data augmentation. Column 2 denotes both the image and input resolution. The shaded cells represent the default configurations on both datasets.

Scarce and CRC-Scarce dataset at various image and input resolutions. In addition to the ResNet-34 architecture, we also experiment with ResNet-152 and ResNeXt-101 (32×8d) architectures. ResNet-152 is a direct upgrade of ResNet-34, with more convolution layers. ResNeXt is an enhanced version of the corresponding ResNet architecture with increased parallel residual blocks and more channels [18]. ResNeXt-101 (32×8d) has 32 parallel residual blocks with 8× more channels compared to the vanilla ResNet-101 architecture. Additionally, we also report the number of floating-point operations (GFLOPS) for each model at various image and input resolutions[3].

Model performance decreases when larger and more complex architectures (ResNet-152 and ResNeXt-101) are used on the annotation-scarce dataset (Table I). This result concurs with Goodfellow et al., where higher capacity models can overfit the dataset [9]. On the other hand, increasing the image and input resolution yields a performance gain, while models with higher capacity failed to do so. This result suggests that increasing the image and input resolution may indeed be regularizing high capacity models in the low-annotation setting. However, increasing the image and input resolution also leads to increased computational cost, which grows quadratically with those factors.

Figure 5 shows that varying the image resolution alone is not sufficient. There is a negative correlation of 0.835 (Pcam-Scarce) and 0.970 (CRC-Scarce) between image resolution and patch classification accuracy. As the image resolution increases, a model can capture finer-grained information. Nonetheless, if we do not increase the input resolution proportionally, the model loses global information, causing a significant drop in performance.

B. The effects of image and input resolution on an annotation-rich dataset

We examine the impact of image and input resolution on two annotation-rich datasets: PCam-Full and CRC-Full. Figure 6 shows that as we increase the image and input resolution, the models' performance generally increases. There is a positive correlation of 0.649 (PCam-Full) and 0.961 (CRC-Full) between the image and input resolution and patch classification accuracy. For the PCam-Full dataset, peak performance occurs at 576×576 resolution. The model achieves peak performance at 800×800 resolution for the CRC-Full dataset. There is a gain of 2.52% in accuracy for the PCam-Full dataset and a gain of 0.62% in accuracy for the CRC-Full dataset when compared to the accuracy of models trained with the original image resolution of the respective datasets (PCam-Full: 96×96, CRC-Full: 224×224).

C. Annotation-efficient learning in the transfer learning setting

Transfer learning in the form of pre-training is a common strategy to tackle annotation efficient learning (§ II). In the previous experiments we avoided transfer learning to isolate the contribution of image and input resolution on performance. In this experiment, we study the interaction between image and input resolutions and pre-trained features (Table II and Figure 7). Previously, we observed that on the PCam-Scarce and CRC-Scarce datasets, there is a gain of up to 6.33% and 9.83% in patch classification accuracy by increasing the image and input resolution. Here, we observe an additional gain of 3.30% and 5.42% in patch classification accuracy when we initialize the weights of the ResNet-34 model by supervised pre-training on the ImageNet 2012 dataset. On the PCam-Scarce dataset, our final model performs equally well compared to the model trained on 100% of the dataset in the original image resolution (96×96). For the CRC-Scarce dataset, our final model is only 2.22% away from the model trained on 100% of the dataset in the original image resolution (224×224). Additionally, there is a positive correlation of 0.810 (PCam-Scarce) and 0.875 (CRC-Scarce) between the image and input resolution and

[3] We exclude the final linear classifier in this calculation as it is held constant across architectures.

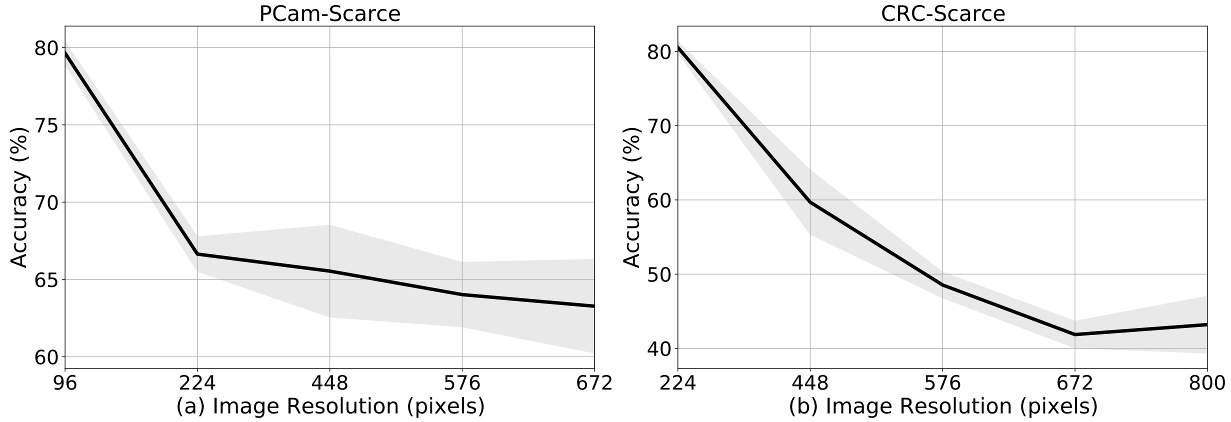

Figure 5. Effect of image resolution on models trained with $< 1\%$ of the original dataset. Input resolution is held constant in this experiment. The shaded areas represent one standard deviation of uncertainty. All models train with data augmentation.

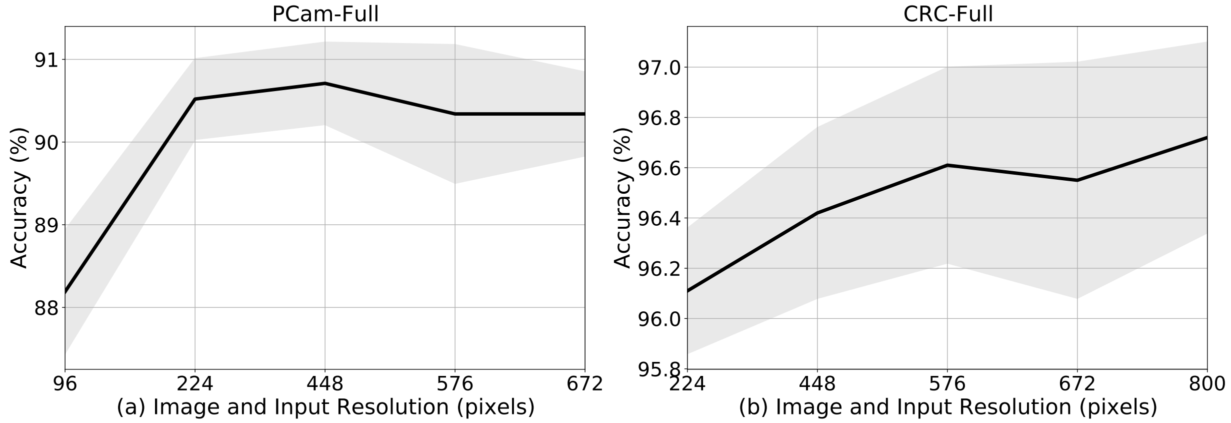

Figure 6. Image and input resolution effects on models trained with 100% of the original dataset. The shaded areas represent one standard deviation of uncertainty. All models are trained with data augmentation.

the patch classification accuracy on models initialized with ImageNet pre-trained weights (Figure 7). This result shows that pre-trained models work well with the image and input resolution factors, yielding an additive gain in performance.

D. ImageNet 2012 statistics

Figure 8 shows the image resolution distribution of the ImageNet 2012 dataset. The average image width and height are 472 and 405, with a standard deviation of 208 and 179. 54% of the images have an image width of 500 to 550, and 52% have an image height of 300 to 400. The large variation in image resolution makes Random-Resized Crop a suitable image augmentation strategy for the ImageNet 2012 dataset. The aim of the Random-Resized Crop algorithm is to generalize a model to unseen images of different image resolutions.

E. Data augmentation for Digital Pathology

Algorithm 1 shows the data augmentation strategy specifically designed for patch classification in Digital Pathology, following Teh et al. [12]. Table III shows the image and input resolution and the corresponding pad size used in our experiments. The pad size is approximately 12.5% to 14.3% of the corresponding image resolution.

Algorithm 2 shows a typical data augmentation strategy used in the Natural Image domain. We compare our data augmentation strategy (Algorithm 1) with respect to Algorithm 2 on PCam-Scarce dataset with the image resolution of 96×96 and a pad size of 12. The mean accuracy of Algorithm 1 and Algorithm 2 are 79.64 ± 0.74 and 75.25 ± 0.91. There is a drop of 4.39% of mean accuracy by switching from Algorithm 1 to Algorithm 2, showing the ineffectiveness of Random-resized-crop on patch classification in the Digital Pathology domain.

V. Conclusion

Input and image resolution are important meta-parameters that have been overlooked to-date in DP classification re-

PCam dataset					CRC dataset				
Subset	Initialization	Augmentation	Resolution	Accuracy (%)	Subset	Initialization	Augmentation	Resolution	Accuracy (%)
Scarce	Random	✔	96×96	79.64±0.74	Scarce	Random	✔	224×224	80.54±0.78
Scarce	Random	✔	448×448	85.97±0.68	Scarce	Random	✕	800×800	90.37±1.34
Scarce	ImageNet	✔	448×448	**89.27±0.68**	Scarce	ImageNet	✕	800×800	95.79±0.31
Full	ImageNet	✔	96×96	88.88±0.92	Full	ImageNet	✔	224×224	**98.01±0.14**

Table II
ACCURACY OF OUR MODELS TRAINED ON THE PCAM DATASET AND CRC DATASET. WE ALSO SHOW DATASET SUBSET (COLUMN 1 AND 6), MODEL INITIALIZATION (COLUMNS 2 AND 7), THE USE OF DATA AUGMENTATION DURING TRAINING (COLUMNS 3 AND 8) AS WELL AS THE IMAGE AND INPUT RESOLUTION (COLUMNS 4 AND 9) IN THIS TABLE.

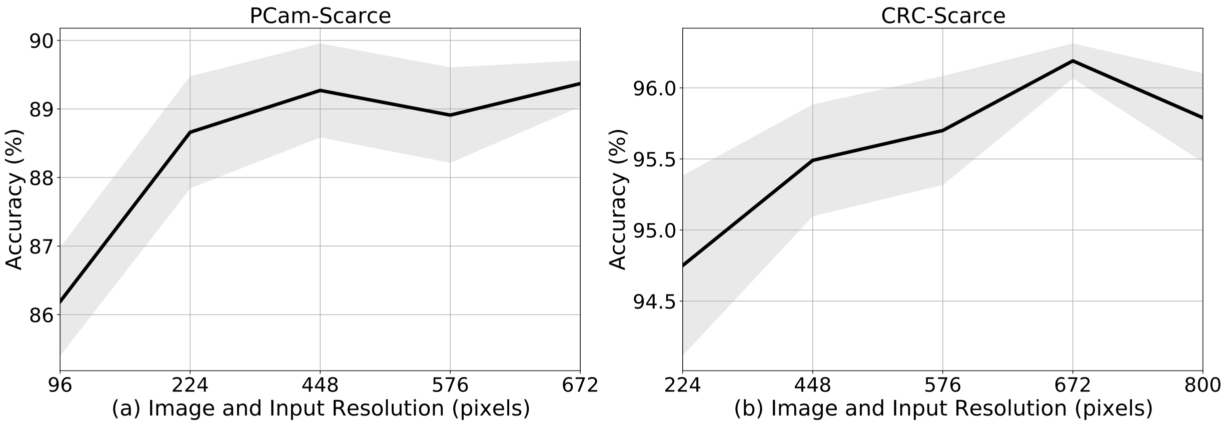

Figure 7. Image and input resolution effects on models trained with $< 1\%$ of the original dataset. The shaded areas represent one standard deviation of uncertainty. PCam-Scarce models are trained with data augmentation, but CRC-Scarce models are trained without data augmentation. All models are initialized with ImageNet pre-trained weights.

Algorithm 1 Data Augmentation, PyTorch-like

```
# h,w: height, width; p: padding size
import torchvision.transforms as t

transform = {’train’:t.Compose([
                  t.Resize((h, w)),
                  t.Pad(p, padding_mode=’reflect’),
                  t.RandomRotation([0, 360]),
                  t.RandomCrop((h,w)),
                  t.RandomHorizontalFlip(0.5),
                  t.ColorJitter(
                     hue= 0.4,
                     saturation=0.4,
                     brightness=0.4,
                     contrast=0.4),
                  t.ToTensor(),
                  ]),
             ’test’:t.Compose([
                  t.Resize((h, w)),
                  t.ToTensor(),
                  ])}
```

Image Resolution (pixels)	96	224	448	576	672	800
Image Padding (pixels)	12	32	64	80	96	114

Table III
IMAGE RESOLUTION AND THE CORRESPONDING PAD SIZE.

Algorithm 2 Data Augmentation, PyTorch-like

```
# h,w: height, width; p: padding size
import torchvision.transforms as t

transform = {’train’:t.Compose([
                  t.RandomResizedCrop((h, w)),
                  t.RandomHorizontalFlip(0.5),
                  t.ToTensor(),
                  ]),
             ’test’:t.Compose([
                  t.Resize((h+p, w+p)),
                  t.CenterCrop((h, w)),
                  t.ToTensor(),
                  ])}
```

Algorithm 3 Without Data Augmentation, PyTorch-like

```
# h,w: height, width;
import torchvision.transforms as t

transform = {’train’:t.Compose([
                  t.Resize((h, w)),
                  t.ToTensor(),
                  ]),
             ’test’:t.Compose([
                  t.Resize((h, w)),
                  t.ToTensor(),
                  ])}
```

search. In this paper, we show that tuning the image and input resolution can yield impressive gains in performance. Across annotation-scarce and annotation-rich environments, we demonstrate a positive correlation between the image and input resolution and the patch classification accuracy. By increasing the image and input resolution, our models can capture finer-grained information without losing coarse-grained

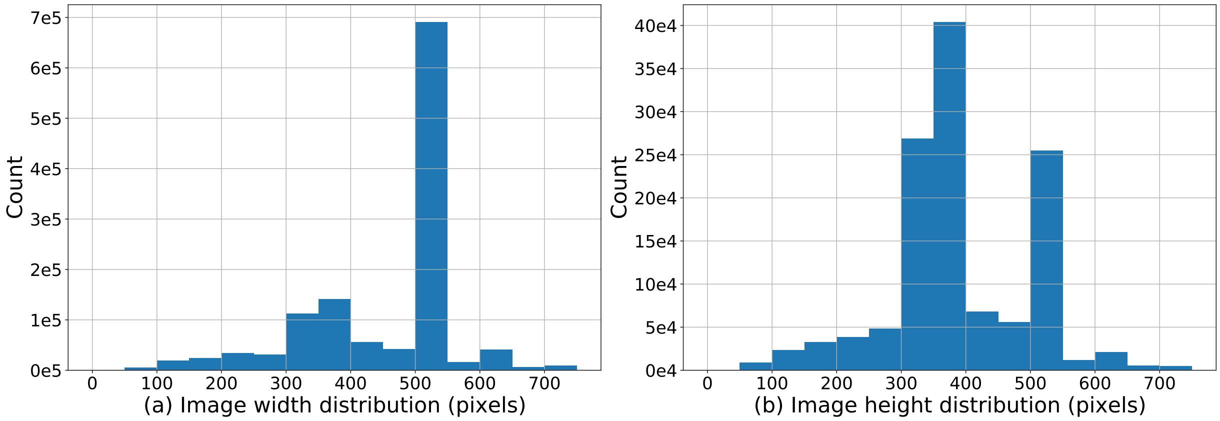

Figure 8. Image resolution distribution of ImageNet 2012 dataset. The average image width and height are 472 and 405, with a standard deviation of 208 and 179.

information, yielding a gain of 6.33% (PCam-Scarce) and 9.83% (CRC-Scarce) in patch classification accuracy. We highlighted other important practical considerations such as the interaction of the data augmentation strategy with resolution and choice of dataset.

References

[1] L. Pantanowitz, "Digital images and the future of digital pathology," *Journal of pathology informatics*, vol. 1, 2010.

[2] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant cnns for digital pathology," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2018, pp. 210–218.

[3] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine*, vol. 16, no. 1, p. e1002730, 2019.

[4] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan *et al.*, "Bach: Grand challenge on breast cancer histology images," *Medical image analysis*, vol. 56, pp. 122–139, 2019.

[5] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Medical Image Analysis*, p. 101813, 2020.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[7] 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.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[10] M. Huh, P. Agrawal, and A. A. Efros, "What makes imagenet good for transfer learning?" *arXiv preprint arXiv:1608.08614*, 2016.

[11] N. Hegde, J. D. Hipp, Y. Liu, M. Emmert-Buck, E. Reif, D. Smilkov, M. Terry, C. J. Cai, M. B. Amin, C. H. Mermel *et al.*, "Similar image search for histopathology: Smily," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[12] E. W. Teh and G. W. Taylor, "Learning with less data via weakly labeled patch classification in digital pathology," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 471–475.

[13] K. L. Kupferschmidt, E. W. Teh, and G. W. Taylor, "Strength in diversity: Understanding the impacts of diverse training sets in self-supervised pre-training for histology images," 2021.

[14] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy," *arXiv preprint arXiv:1906.06423*, 2019.

[15] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114.

[16] "Random-resized crop algorithm - pytorch (source code)," https://github.com/pytorch/vision/blob/main/torchvision/transforms/transforms.py#L847, accessed: 2022-02-14.

[17] E. W. Teh, T. DeVries, and G. W. Taylor, “Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 448–464.

[18] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.