Transferability of Deep Learning Models for Focus Quality Assessment in Digital Pathology

Adyn Miles
University of Toronto
Mahdi S. Hosseini (mahdi.hosseini@unb.ca)
University of New Brunswick
Sheyang Tang
University of Waterloo
Zhou Wang
University of Waterloo
Savvas Damaskinos
Huron Digital Pathology Inc
Konstantinos N. Plataniotis
University of Toronto

Research Article

Keywords: Transferability, Deep Learning, Quality Assessment, Digital Pathology

Posted Date: December 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1120682/v1

Transferability of Deep Learning Models for Focus Quality Assessment in Digital Pathology

Adyn Miles[1,*], Mahdi S. Hosseini[2,*], Sheyang Tang[3], Zhou Wang[3], Savvas Damaskinos[4], and Konstantinos N. Plataniotis[1]

[1]Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON MS5 3G4, Canada
[2]Department of Electrical and Computer Engineering, University of New Brunswick, Fredericton, NB E3B 5A3, Canada
[3]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[4]Huron Digital Pathology Inc., Waterloo, ON N2L 5V4, Canada
[*]adyn.miles@mail.utoronto.ca
[*]mahdi.hosseini@unb.ca

ABSTRACT

Out-of-focus sections of whole slide images are a significant source of false positives and other systematic errors in clinical diagnoses. As a result, focus quality assessment (FQA) methods must be able to quickly and accurately differentiate between focus levels in a scan. Recently, deep learning methods using convolutional neural networks (CNNs) have been adopted for FQA. However, the biggest obstacles impeding their wide usage in clinical workflows are their generalizability across different test conditions and their potentially high computational cost. In this study, we focus on the transferability and scalability of CNN-based FQA approaches. We carry out an investigation on ten architecturally diverse networks using five datasets with stain and tissue diversity. We evaluate the computational complexity of each network and scale this to realistic applications involving hundreds of whole slide images. We assess how well each full model transfers to a separate, unseen dataset without fine-tuning. We show that shallower networks transfer well when used on small input patch sizes, while deeper networks work more effectively on larger inputs. Furthermore, we introduce neural architecture search (NAS) to the field and learn an automatically designed low-complexity CNN architecture using differentiable architecture search which achieved competitive performance relative to established CNNs.

Introduction

Digital pathology is an expanding field focused on using Whole Slide Image (WSI) scans to facilitate the clinical workflow[1,2]. One critical issue with this field is reliable and efficient quality control (QC) for the scanned images. With a global shortage of trained pathologists, automated QC methods are an attractive option for digital pathology[3–5]. Making effective diagnoses from whole slides requires high quality images, which can be affected by lighting conditions, the optical system, and the scanner's sensor itself[6–9]. In this context, QC refers to Focus Quality Assessment (FQA), which differentiates between varying degrees of in-focus and out-of-focus sections of an image.

Recently, deep learning models based on convolutional neural networks (CNNs) have emerged as viable FQA methods[2,10–23]. Open source platforms such as HistoQC[13], CellProfiler 3.0[17] and ImageJ[24,25] also leverage deep learning models for FQA. Moreover, there is advancement in artificial intelligence (AI) for medical diagnosis purposes in digital pathology[26–34]. Out-of-focus regions in an image are a major contributor to systematic errors in these diagnoses[2,35,36], highlighting the importance of reliable FQA methods to accompany these diagnosis tools. However, two major barriers have been slowing down the adoption of deep learning methods in clinical workflows. The first is their undertested transferablity to diverse imaging conditions, and the second is their potentially high computational cost and scalability to extremely high scanning throughput in practical clinical workflows[3,10].

Dynamic imaging conditions mean that FQA methods must be generalizable to different datasets. This requires that a variety of tissue types, stain types, and resolutions be used to train the model[13,37,38]. Unfortunately, there has until recently been a shortage of large and diverse datasets for FQA purposes[39], which increases the risk of overfitting data-driven models[40]. Additionally, a principal advantage of digital pathology technology is that scanners can process images much faster than human pathologists, with some able to scan hundreds of images at a time[6,26,40]. In clinical settings, it is ideal for scans to be completed during night-time hours, so that they are ready for diagnosis the following day[10]. The high throughput of digital scans therefore

requires QC pipelines that can handle this high throughput. Computational complexity of the FQA method should not be a restricting factor, and is therefore equally important in the evaluation of the method as its performance[10,40,41].

A drawback of deep learning models for FQA is that they are not as easily applicable as knowledge-based methods[42–48], which have low computational complexity and can be easily applied without adjustment or tailoring[10]. Conversely, when transferring CNNs to other computer vision applications it is usually recommended to use a process called fine-tuning[49–53], where the network already has a majority of parameters set and is then trained to adjust the remaining parameters. This is used when the datasets and computational resources for training are limited[49]. However, this increases the resources spent transferring the network to different scanners, and would ideally not be necessary for high performance. While foregoing fine-tuning would have efficiency advantages, it could potentially produce some concerns regarding the explainability of the model and the trust in AI-based decision making[3,27].

Figure 1 shows our process for the evaluation of each deep learning model. The images are first normalized[27,39], then used to train a CNN. The trained model, including the fully connected layer, is then tested on the same dataset on which it was trained to validate the success of the training. Afterwards, the full model is tested on a separate dataset it has not yet seen to evaluate the transfer process[13,54,55], without the use of fine-tuning. Numerous metrics are used to assess each deep learning model, including computational complexity, layer probing quality metrics, and a spatial focus quality distribution.

Table 1 shows an overview of existing deep learning models used for FQA purposes. The models are split into two categories based on the architecture of the CNN being used. This section explains the two categories in more detail:

1. *Lightweight CNNs*: these refer to CNNs that are optimized towards efficiency by containing only one convolutional layer[10–15,22]. Some of these networks are able to perform FQA even more quickly than knowledge-based methods[14,15]. They also directly address the scalability problem, as their low computational complexity allows gigabytes of WSI data to be assessed for focus quality in a matter of hours. Shallow CNNs can also be well trained using a limited number of samples[14,22], which reduces the effort needed for training. These networks generally have worse performance than deeper CNNs, but can still provide reasonable FQA performance because blur is assumed to be encoded in low-level features in a well-controlled environment[10,22].

2. *Deep CNNs*: these CNNs have multiple layers and are optimized towards obtaining an accurate result using a robust framework[2,16–21]. These networks focus on achieving high levels of accuracy and being highly generalizable to different tissue types, stain types, and resolutions. These relatively deeper CNNs have more layers which enable the network to capture fine features that may be missed by a shallow network[56]. A drawback of deep CNNs is that they are prone to overfitting when training on small datasets, which means more effort must be put into dataset creation[14]. After a critical depth, CNNs can be overparametrized which worsens their ability to generalize[57].

Table 1

Author	Year	Method Type	Organ Variety	Stain Variety	Transferability	Scalability	Method Description
Wang[10]	2020	C	•	•	•	•	Built a high-efficiency, light-weight CNN and compared performance to other CNNs and knowledge-based methods.
Wang[11]	2019	C				•	Developed a simple CNN architecture called EONSS for superior IQA performance at lower cost.
Pinkard[12]	2019	C	•			•	Developed a fully connected Fourier neural network to make accurate predictions with less memory usage.
Yu[14]	2017	C				•	Trained a shallow single-layer CNN for cost-efficient predictions.
Yu[22]	2017	C			•		Shallow CNN combined with general regression neural network for higher prediction accuracy.
Kang[15]	2014	P			•	•	Trained a single layer CNN to predict image quality without reference.
Kohlberger[2]	2019	P	•	•	•		Trained a CNN ConvFocus on synthetically blurred image patches.
Yang[16]	2018	C			•		Deep neural network model trained on synthetically blurred images for an absolute focus quality measure.
Senaras[18]	2018	C		•	•	•	Deep CNN DeepFocus used to assess image focus quality.
Bosse[19]	2017	C			•		Robust ten-layer CNN DeepIQA which shows high ability to generalize.
Campanella[20]	2017	P	•	•	•		Trained a residual network (ResNet18) to assess focus quality.
Ma[21]	2017	C			•		Multi-task learning framework with GDN activation to improve efficiency.

Table 1 also shows which studies focused on scalability and transferability, the two key pillars of a successful FQA method[3,10]. The majority of studies using lightweight CNNs do not examine the transferability of the model, while the majority of studies using deep CNNs do not examine the scalability of their model. Three studies[10,15,18] examined both the transferability and the scalability of the method, which is not a large enough representation to draw conclusions about the characteristics of a model and method that translate to successful FQA. Additionally, transferability cannot effectively be compared between studies because of major differences in the methodologies[41].

Additionally, of the papers that examined deep CNN methods for transferability, three studies[2,16,18] used a variety of tissue and stain types. A variety of WSIs is necessary to confirm that the methods can be generalized to FQA in digital pathology. To perform a more systematic analysis of how well a model can generalize, diverse datasets should be used in the training process, before transferring the models to a separate dataset that it has not yet seen for testing[13,54,55].

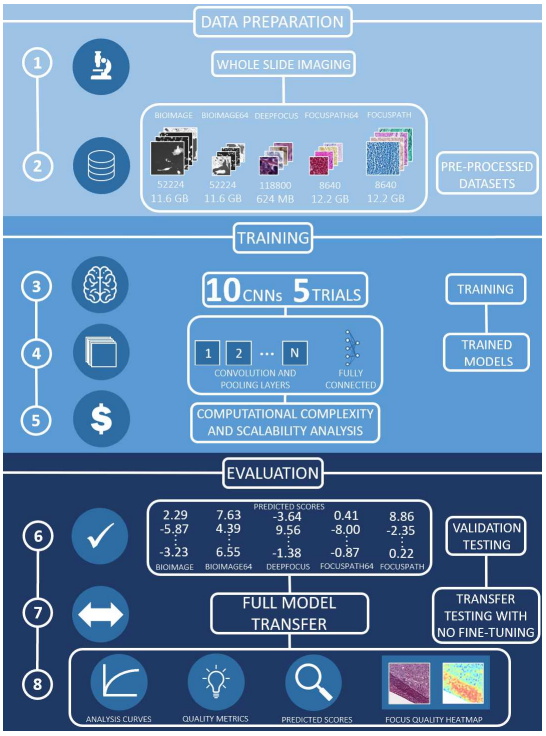


Figure 1

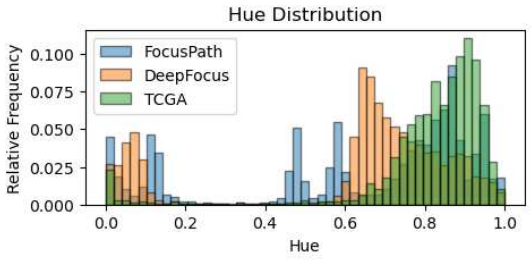


Figure 2

While many of the models in Table 1 were custom-made for the application, none of them use neural architecture search (NAS) to arrive at an optimal architecture. Automating the architecture design process requires less effort on the part of human researchers, while still performing well compared to human-designed CNNs[58–62]. Networks that are tailored for digital pathology datasets may result in higher performance FQA.

A variety of metrics are used across these works to evaluate accuracy and transferability. These include Area under the Receiver Operating Characteristics (ROC) and Precision Recall (PR) Curves[10,18], F-Scores[16,17], Pearson's Linear Correlation Coefficient (PLCC) and Spearman's Rank Correlation Coefficient (SRCC)[2,10,11,14,15,19–22], Root Mean Square Error (RMSE)[12,20], post-FQA qualitative assessment[13,20], and Rand Index[17]. The most popular metrics of this list, including PLCC, SRCC, ROC, and PR, are therefore evaluated in this paper to clearly compare the network performance.

Heatmaps are used in four studies[2,10,18,20] to spatially represent the FQA of a deep learning model. Spatial representations are important for better visualizing which features the model is able to capture well and which features it misses. They are also important for understanding the characteristics of a dataset that may make it more or less transferable to other applications[2,10,20].

In this paper, we make the following contributions to improving FQA in digital pathology:

Experiment Design: We train ten architecturally diverse CNNs on five datasets of different stain types, tissue types, resolutions, and input patch sizes. We evalute the computational complexity of each CNN, and scale this to realistic scanning applications. We use these methods to draw conclusions about the effect of input patch size, tissue diversity, and stain diversity on the transferability of a deep learning model to other datasets.

Architecture Design: We develop an automatically designed architecture using differentiable architecture search[58] on the same diverse datasets to evaluate the performance of searched architectures relative to conventional CNNs.

Validation Methods: We use the knowledge gain and mappign condition metrics[63] to evaluate how well the model is learning as well as its degree of stability. We use ROC and PR metrics to evaluate the performance of each network when transferred to another dataset. We use focus quality heatmap representations to understand the spatial distribution of focus quality in an image.

Dataset Selection

This section describes the datasets used for this experiment, with information about each dataset summarized in Table 2. Datasets with the suffix "64", such as FocusPath64[43], are copies of the original dataset with a 64 x 64 input patch size. Figure 3 shows examples of patches used for each dataset. Further information about each dataset can be found in the Supplementary

Materials. The tissue scans used in this work are deemed exempt from University of Toronto Research Ethics Board regulations. The datasets consist of anonymized tissue scans and metadata relevant only to the focus quality of the image.

Dataset Name	Number of Organs	Number of Stains	Pixel Resolution [um/pixel]	Optical Zoom	Real/Synthetic Blur	Number of Classes	Percentage In-Focus	Original Patch Size [pixels]	Training/Testing Patch Size	Number of Patches	Training/Testing Split	Mean Hue
BioImage[16]	1	1	Unspecified	20X	Synthetic	12	38%	696 x 520	64 x 64 235 x 235	52224	90% / 10%	N/A
DeepFocus[18]	4	4	0.2461	40X	Synthetic	6	9%	64 x 64	64 x 64	118800	80% / 10%, 10% Validation	$0.618_{0.285}$
FocusPath[43]	9	9	0.25	40X	Real	15	57%	1024 x 1024	64 x 64 235 x 235	8640	60% / 20%, 20% Validation	$0.632_{0.303}$
TCGA[64]	52	Unspecified	Variable	40X	Real	2	79%	1025 x 1025	64 x 64 235 x 235	14371	0% / 100%	$0.812_{0.179}$

Table 2

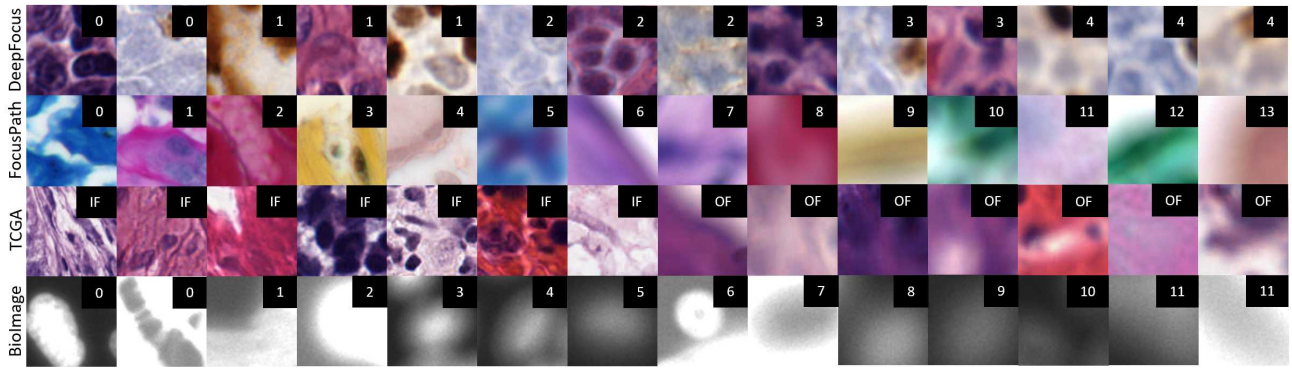


Figure 3

FocusPath[43]: The FocusPath[43] dataset contains 8640 patches of size 1024 x 1024 extracted from nine different stained slides. This dataset is useful for the development of CNNs geared towards FQA methods due to its diverse distribution of colors and stains relative to other datasets. A stain distribution can be found in the Supplementary Materials, with a sample hue distribution for this dataset shown in Figure 2.

DeepFocus[18]: The DeepFocus[18] dataset contains 118800 patches of size 64 x 64, consisting of 16 different slides with 4 types of stains. This dataset, alongside the FocusPath64[43] and BioImage64[16] datasets, were useful for determining the effect that a varying patch size can have on the quality of training and transferability to other datasets. This dataset has less stain diversity than FocusPath[43], and slightly less hue diversity as well, with a standard deviation 1.8% smaller than that of FocusPath[43].

BioImage[16]: The Broad BioImage Dataset[16] consists of 52224 patches of size 696 x 520. BioImage[16] was useful for investigating the effect that grayscale images have on the quality of training and transferability to other datasets. It has been observed that color information can positively enhance the FQA performance of a CNN[14]. Successful FQA methods should have the ability to distinguish focus levels in an image regardless of the colors in the image. This study therefore also seeks to further investigate the effect of color information on transfer performance.

TCGA[64]: The TCGA@Focus dataset contains 14371 image patches in total, with 11328 patches labelled in-focus and 3043 patches labelled out-of-focus. This dataset was chosen due to its wide spectrum of tissue textures and colors.

CNN Analysis

This experiment trained ten architecturally diverse CNN models on the five training datasets. The first is FocusLiteNN[10], a light-weight CNN built for FQA, which was trained using 1, 2, and 10 input channels. Other architectures used include EONSS[11], DenseNet13[65], MobileNetv2[66], variations of ResNet[67] (ResNet18, ResNet50, and ResNet101), and DARTS[58]. A summary of each network's parameters and convolutional layers, as well as the floating point operations (FLOPs) cost and GPU latency are shown in Table 3.

Model	# Layers	# Param	# FLOPs	Latency
FocusLiteNN-1[10]	1	0.15K	0.99M	0.25ms
FocusLiteNN-2[10]	1	0.30K	1.99M	0.31ms
FocusLiteNN-10[10]	1	1.49K	9.92M	0.32ms
EONSS[11]	4	0.11M	1.79M	1.06ms
DenseNet-13[65]	8	0.19M	52.4M	2.24ms
MobileNetv2[66]	21	2.23M	48.9M	7.39ms
ResNet-18[67]	16	4.91M	145M	2.26ms
ResNet-50[67]	48	23.5M	667M	9.29ms
ResNet-101[67]	99	42.5M	1273M	18.7ms
DARTS-FQA	3	35.1K	78.8M	3.47ms

Table 3

	Input Patch Size			
Model	64	128	235	300
FocusLiteNN (1 kernel)[10]	3.12	0.58	0.18	0.10
FocusLiteNN (2 kernel)[10]	3.90	0.72	0.21	0.12
FocusLiteNN (10 kernel)[10]	4.08	0.85	0.30	0.15
EONSS[11]	13.53	3.30	1.15	0.68
DenseNet-13[65]	28.51	8.46	2.69	1.75
MobileNetv2[66]	93.98	24.17	7.61	5.94
ResNet-18[67]	28.67	7.41	2.43	1.94
ResNet-50[67]	118.12	31.79	11.80	11.55
ResNet-101[67]	237.49	60.55	21.68	21.30
DARTS-FQA	44.10	12.34	7.15	6.73

Table 4

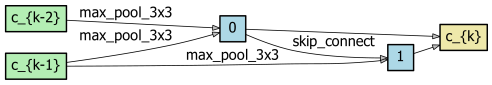


Figure 4

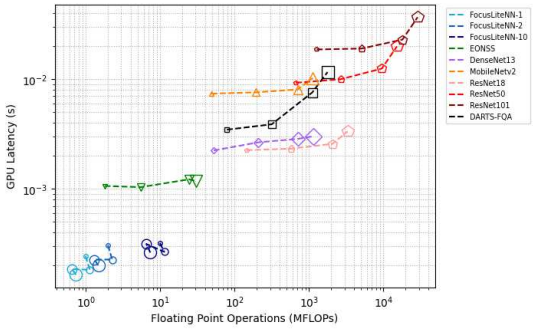


Figure 5

Differentiable Architecture Search

This experiment involves a NAS on the FocusPath[43], BioImage[16], and DeepFocus[18] datasets. Typical NAS algorithms are extremely computationally expensive due to the number of architecture evaluations required[58]. We apply the differentiable architecture search (DARTS)[58] due to its convolutional architecture applications and scalability advantages. We refer to our searched architecture as DARTS-FQA.

The DARTS-FQA search space used a three-cell system, all of which are classified as reduction cells. Each cell is a directed acyclic graph which is built from four nodes, which each represent an ordered sequence of feature maps. Figure 4 shows the DARTS-FQA reduction cell architecture. The reduction cell architecture is one where all the operations adjacent to the input nodes have a stride of two, halving the pixel resolution of the image. The algorithm takes 60 epochs to learn the operations on the edges of these acyclic graphs from a few candidate operations in the search space, which it does using the highest validation accuracy score[58]. After the model search operations have been completed, the model is frozen and transferred to evaluation. A 3-layer, 20 input channel model based on the searched architecture is trained for 120 epochs using the Adam optimizer, with full details found in the Supplementary Materials.

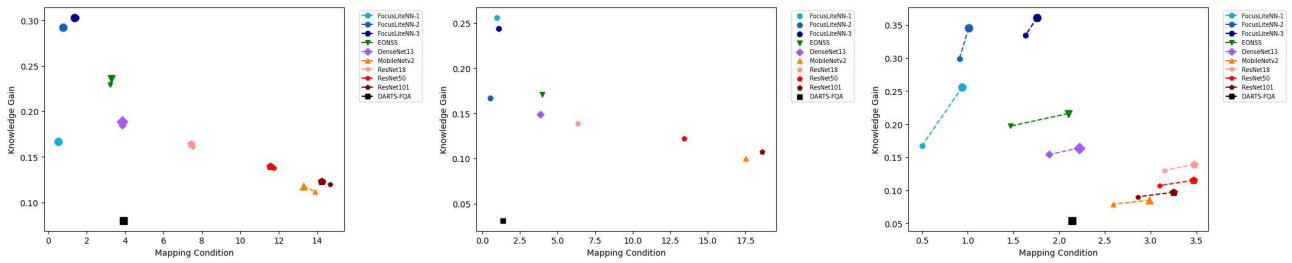
CNN Complexity

FLOPs and GPU latency were the two parameters used to investigate the computational complexity of these CNNs for 4 different randomly cropped input patch sizes: 64, 128, 235, and 300. To fairly compare complexity, all models were evaluated on a Windows station using an Intel Core i7-10875H CPU @ 2.30GHz, and NVIDIA GeForce GTX 1650 Ti. For latency measurements, the GPU was given some time to initialize, and the latency was calculated as the average of 100 trials.

Figure 5 shows that the GPU Latency and the FLOPs cost increase with the input patch size when evaluating a single image, especially for the deepest networks such as ResNet[67]. With shallower networks, the change is less noticeable because the time frames are much shorter and have more stochasticity. Figure 5 also confirms that the FocusLiteNN[10] networks have the lowest GPU latency per image, with a reduction in time per image of 88.7% from DenseNet-13[65], and 98.7% from ResNet101[67] for a single 64 x 64 input patch. These FocusLiteNN[10] models as well as EONSS[11] also save on FLOPs per image, with a reduction of 99.2% between FocusLiteNN (10-channel)[10] and ResNet101[67] for a single 64 x 64 input patch.

Scalability for High Throughput Scanning

These complexity metrics are most relevant when scaled to larger processes, to see how they can affect digital pathology systems on a daily basis. A smaller latency can scale to days saved in diagnosis processes. Assuming a slide at 0.5μm/pixel@20X



(a) BioImage[16] **(b)** DeepFocus[18] **(c)** FocusPath[43]

Figure 6

magnification, containing an approximately 1cm x 1cm tissue which translates to a pixel size of 25000 x 25000 for each WSI, Table 4 shows how GPU latency scales for each network.

The two smaller patch sizes are worse at scale than larger patch sizes, even though they require less GPU inference time and FLOPs when considering a single input image. For a network such as EONSS[11], decreasing the input patch size from 128 to 64 increases the time by a factor of 4.36. This is the difference between a successful overnight QC session, and one that carries over into the next working day. Smaller networks such as FocusLiteNN[10] can achieve high throughput scanning in only 0.3% of the time spent processing the WSIs using ResNet101[67], and is able to complete it in just over 3 hours. If the performance is not severely impacted, there is an advantage to using larger input patch sizes, and especially to using shallower CNNs.

Training Performance Metrics

Assessing the training performance is important to understanding which models will train and transfer better when no fine-tuning is applied. The metrics chosen for this purpose are the Knowledge Gain and the Mapping Condition. The Knowledge Gain[63] quantitatively encodes the useful information carried over each convolution layer, which serves as a representation of how well a network is learning or gaining knowledge. The Mapping Condition[63] quantitatively encodes the sensitivity of the convolution mapping between a layer's inputs and outputs. A lower mapping condition is valuable when paired with a high knowledge gain, which shows good stability and a good ability to map input features to output features. If the mapping condition is high and the knowledge gain is low, the layers are very sensitive to input perturbations, and do not show desirable mapping capabilities[63]. For the ten CNNs used in the experiment, these metrics were averaged between the input and output channels, then averaged across each layer, then finally over the five trials used in the experiment.

Figures 6a, 6b, and 6c plot the Knowledge Gain against the Mapping Condition for each network. A point near the top left of this plot is desired, as this means the network has a strong mapping condition and a high knowledge gain. For BioImage[16], a change in patch size does not affect the knowledge gain and mapping condition significantly, shown by the proximity of each pair of points. For the FocusPath[43] dataset, the input patch size has a noticeable effect on the knowledge gain and mapping condition. Generally, the larger patch sizes have a worse mapping condition performance. This could be because smaller input patches have fewer features, meaning that input perturbations are less likely to impact the output.

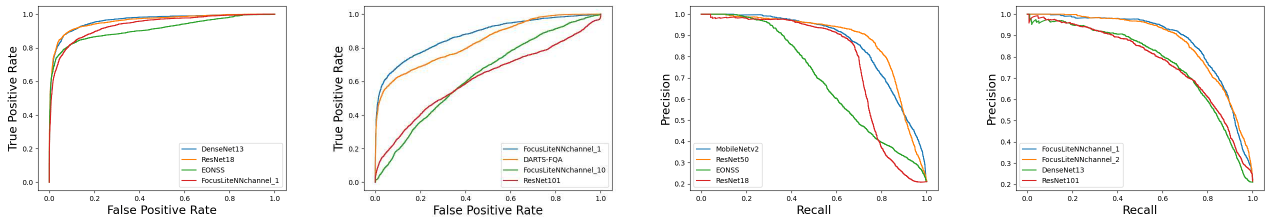
Interestingly, across all three datasets, deeper networks have a worse average knowledge gain and a worse mapping condition performance than shallower networks. For BioImage[16] as an example, FocusLiteNN (10 channel)[10] shows 84.7% knowledge gain improvement and 87.8% mapping condition improvement over DenseNet-13[65] and 146% and 97.4% respective improvements over ResNet-101[67]. DARTS-FQA[58] falls outside of this trend with a lower knowledge gain but a more stable mapping condition relative to networks of similar complexity, likely because tailoring the network to a specific dataset makes the network much more stable in response to that dataset. Overall, these deeper networks would benefit from fine-tuning before transferring to other datasets, which is less necessary for shallower networks.

Results

The training and testing sets were randomly shuffled and distributed before each trial, and each model was trained five times on each dataset. Each model was then tested on a testing set from the same dataset on which they were trained before being transferred to TCGA@Focus[64]. The input patches were normalized by color before being passed into the CNN, for both the training and testing cases. A full description of training hyperparameters, evaluation metrics, and validation results can be found in the Supplementary Materials.

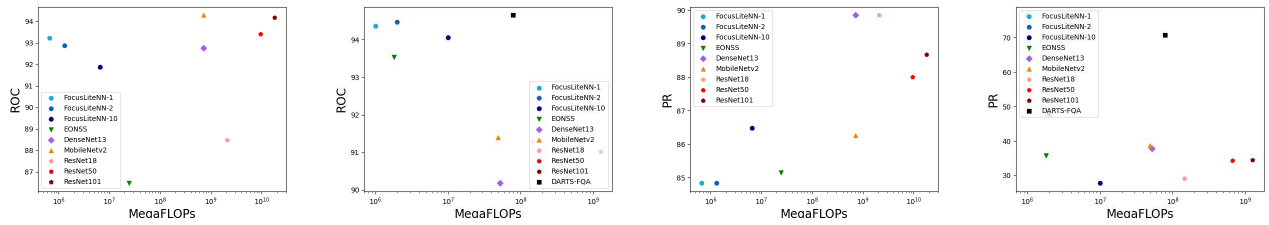
Network	FocusPath[43]		FocusPath64[43]		BioImage[16]		BioImage64[16]		DeepFocus[18]	
	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR
FocusLiteNN (1 kernel)[10]	$93.35_{0.00}$	$84.83_{0.00}$	$93.30_{0.06}$	$86.78_{0.15}$	$93.22_{0.07}$	$84.78_{0.09}$	$94.36_{0.14}$	$87.97_{0.23}$	$87.29_{0.07}$	$76.49_{0.14}$
FocusLiteNN (2 kernel)[10]	$93.14_{0.03}$	$84.83_{0.02}$	$94.51_{0.23}$	$88.41_{0.40}$	$92.88_{0.24}$	$84.36_{0.62}$	$94.47_{0.03}$	$87.28_{0.07}$	$73.48_{1.95}$	$47.81_{3.34}$
FocusLiteNN (10 kernel)[10]	$93.63_{0.25}$	$86.48_{0.32}$	$93.86_{0.17}$	$87.86_{0.28}$	$91.87_{0.92}$	$83.66_{1.31}$	$94.06_{0.31}$	$87.16_{0.57}$	$60.15_{3.85}$	$27.76_{3.02}$
EONSS[11]	$91.03_{0.09}$	$85.15_{0.10}$	$90.68_{0.34}$	$83.11_{0.31}$	$86.47_{3.29}$	$73.80_{6.38}$	$93.54_{0.21}$	$86.27_{0.52}$	$67.06_{4.78}$	$35.84_{7.56}$
DenseNet-13[65]	$95.64_{0.30}$	$89.85_{0.40}$	$93.28_{0.30}$	$85.28_{0.40}$	$92.76_{1.26}$	$84.37_{2.50}$	$90.18_{1.19}$	$78.87_{1.79}$	$61.73_{8.08}$	$37.88_{11.5}$
MobileNetv2[66]	$94.34_{0.52}$	$86.26_{1.45}$	$92.93_{0.63}$	$83.83_{0.98}$	$94.29_{0.55}$	$87.16_{1.00}$	$91.39_{0.37}$	$80.72_{1.16}$	$70.69_{2.31}$	$38.63_{2.69}$
ResNet-18[67]	$94.83_{1.04}$	$89.85_{1.11}$	$93.28_{0.51}$	$86.73_{0.48}$	$88.48_{2.71}$	$80.97_{2.67}$	$91.29_{0.59}$	$79.26_{0.81}$	$61.68_{5.41}$	$29.17_{5.84}$
ResNet-50[67]	$94.85_{0.59}$	$88.01_{0.73}$	$93.14_{0.60}$	$86.08_{0.37}$	$93.42_{0.14}$	$87.33_{0.69}$	$91.31_{0.25}$	$80.02_{0.84}$	$63.32_{7.29}$	$34.32_{8.49}$
ResNet-101[67]	$95.01_{0.46}$	$88.67_{0.86}$	$92.80_{0.73}$	$85.80_{0.82}$	$94.18_{0.65}$	$86.65_{0.78}$	$91.03_{0.50}$	$79.91_{1.28}$	$60.71_{6.24}$	$34.49_{10.4}$
DARTS-FQA	–	–	$92.34_{0.80}$	$85.09_{0.74}$	–	–	$94.65_{1.01}$	$83.97_{2.45}$	$82.58_{0.81}$	$70.68_{0.55}$

Table 5



(a) FocusPath[43] **(b)** DeepFocus[18] **(c)** BioImage[16] **(d)** BioImage64[16]

Figure 7



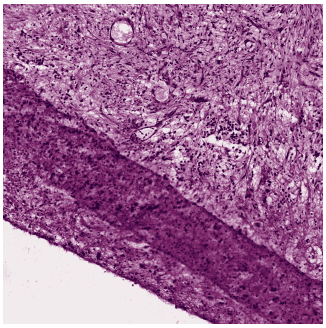
(a) BioImage[16] **(b)** BioImage64[16] **(c)** FocusPath[43] **(d)** DeepFocus[18]

Figure 8

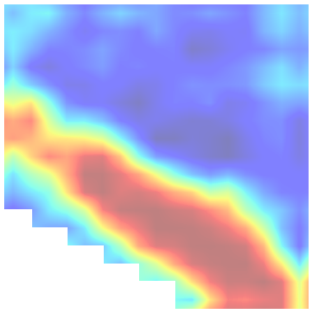
Discussion

Table 5 shows the ROC and PR performance for each network and dataset when transferred to the TCGA[64] dataset, and Figure 7 shows the best and worst two ROC and PR curves for selected datasets. More ROC and PR Curves can be found in the Supplementary Materials. In general, the FocusPath[43] dataset has the highest transferability to the TCGA[64] dataset, with an average ROC of 93.98% and PR of 87.10%. DeepFocus[18] has the worst with respective averages of 68.87% and 43.31%, which is unacceptable for pathology applications when compared to ROC results from other studies[10,18]. Table 2 shows that FocusPath[43] has a higher organ and stain diversity than any of the other training datasets. This would suggest that the organ and stain diversity does have an impact on the transferability of the model. The main issue with DeepFocus[18] is the distribution of in-focus and out-of-focus classes, where only 9% of the patches are labelled in-focus. Since there are so few positive cases, it does not take many false positives to heavily impact the precision. This highlights the importance of a dataset with balanced focus levels as well.

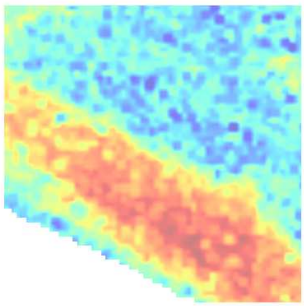
As suggested by the training quality metrics, deep CNNs do not perform best on every dataset. Rather, it is clear that complex networks such as ResNet[67] exhibit better transferability when training on datasets with large input patch sizes. ResNet-101[67] shows 3.15% better ROC, and 6.74% better PR on BioImage[16] compared to BioImage64[16], and similar performance on FocusPath[43] compared to FocusPath64[43]. Conversely, lightweight networks such as FocusLiteNN[10] perform better when working on datasets with small input patch sizes, with 1.47% better ROC and 3.58% better PR for FocusPath64[43] when



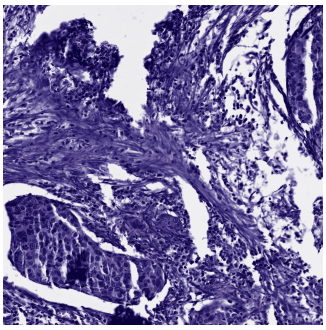
(a) Original WSI



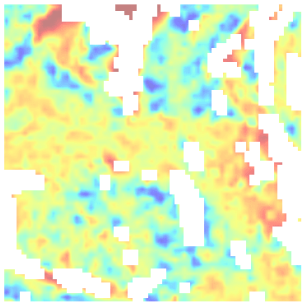
(b) FocusPath, DenseNet13



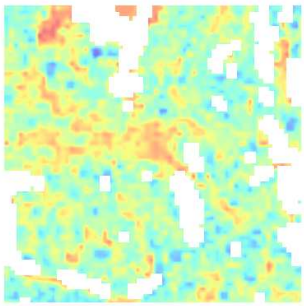
(c) FocusPath64, FocusLiteNN-2



(d) Original WSI



(e) DeepFocus, DARTS-FQA



(f) DeepFocus, FocusLiteNN-1

Figure 9

compared to FocusPath[43]. Figure 8 reinforces this trend, as datasets with small input patch sizes display negative trends, while datasets with larger input patch sizes have positive trends.

While the average transfer performance is worse for the FocusLiteNN[10] networks, the best performance that can be achieved from using a shallower network, using FocusLiteNN (2-channel)[10] on FocusPath64[43], is only 1.13% lower than the best performance measured in this study, using DenseNet-13[65] and FocusPath[43]. Networks such as FocusLiteNN[10] and EONSS[11] scale so well to high-throughput scanning that using a smaller patch size is acceptable. Using FocusLiteNN (1 channel)[10] with a 64 x 64 input patch size still processes 300 WSIs in just over three hours, while largely outperforming deeper networks on FocusPath64[43], BioImage64[16], and DeepFocus[18] as shown in Table 5.

Table 5 shows that FocusPath[43] has an advantage on average over BioImage[16] in ROC (2.3%) and PR (3.4%) performance, as does FocusPath64[43] over BioImage64[16]. This does support the notion that color information is helpful for network performance and transferability[14].

The DARTS-FQA[58] architecture performed competitively compared to other CNNs, even scoring the best ROC performance by 0.18% for BioImage64[16] and the second best ROC and PR performance for DeepFocus[18]. DARTS-FQA[58] performed the best relative to networks of similar complexity (DenseNet-13[65], MobileNetv2[66]) on BioImage64[16] and DeepFocus[18], but performed slightly worse on FocusPath64[43]. This suggests that architecture search can have clear advantages over other networks for transferability for certain datasets, though the effort involved in optimizing these networks make simpler, more generic networks such as FocusLiteNN[10] very valuable for efficient and reliable transferability performance as well.

Figure 9 shows an example of a visual FQA representation of a whole slide image. Using the models trained on each dataset, a heatmap was produced using the predicted focus scores from each network. The background was filtered out of the image, and then each heatmap was normalized. The heatmaps were then interpolated, making them visually smoother. The same patch size and stride were used as were applied to testing each network.

Figures 9e and 9f show some differences between shallow and deep networks trained on the same dataset, DeepFocus[18]. DARTS-FQA[58] and FocusLiteNN (1 channel)[10] show similar focus distributions of Figure 9d. From Figure 9e and the Supplementary Materials, DARTS-FQA[58] is generally more aggressive with classifying focus regions in the image, but it does identify the correct regions upon visual inspection. For example, the right side of the image was correctly labelled out-of-focus by DARTS-FQA[58], while FocusLiteNN (1 channel)[10] misses this region.

From the heatmaps analyzed in this paper, there is no pattern concerning how these deeper networks and coarsely sampled datasets will predict the focus quality distributions, which does affect their overall explainability. While Figure 9b is somewhat aggressive in its evaluation, the Supplementary Materials display numerous examples that are not aggressive enough. In

contrast, shallower networks and smaller input patch sizes such as the one shown in Figure 9c seem to spatially represent the focus quality distribution more closely to expected human perception.

From the samples analyzed in this study, models trained on datasets with smaller input patch sizes generally have spatial representations that match human perception more closely, which gives them an explainability advantage. If a pathologist wants to understand how the model was able to reach a classification result, the spatial distributions should match what the pathologist would expect. If they do not match, there should be patterns that pathologists could identify to explain why a network's spatial distribution is not what they expected. Further work would need to be done to discover patterns in spatial distributions for deep models with coarse sampling such as BioImage[16] and FocusPath[43].

Conclusions

This paper sought to identify the characteristics of datasets and CNN architectures that make them more transferable and more scalable for FQA applications in digital pathology. The FocusPath dataset[43] showed the best transferability performance with an average ROC of 94.0% and PR of 87.1%. FocusPath's[43] tissue, stain, and hue diversity were major contributors to this success. Color information in the dataset also had a small transferability advantage of 2.03% when compared to the grayscale BioImage[16] dataset. DeepFocus[18] showed the worst transfer performance with an average ROC of 68.9% and PR of 43.3%, which can be attributed to a heavy imbalance between out-of-focus and in-focus patches.

The DARTS-FQA network performed quite well against the other nine CNNs, achieving performance in the intermediate and above-average range for each of the datasets. This is encouraging performance for a low-cost algorithm, and further optimizations of these networks would allow it to outperform other networks of similar complexity significantly.

Shallower networks showed excellent scalability performance, with the FocusLiteNN[10] networks running 300 WSIs in just over 4 hours for the slowest case. They also show acceptable transferability performance, and perform even better than deep networks and searched architectures when trained on datasets with small input patch sizes. Furthermore, shallower networks also have much better training quality metrics results than their deeper counterparts, which suggests that fine-tuning would not be as essential. Shallower networks are more easily explainable when showing focus quality heatmaps as they align better with human perception, which can help build confidence in AI-based solutions for QC in digital pathology. Overall, using diverse datasets to train shallow networks on small input patches can help optimize scalability without sacrificing significant transferability performance.

Data Availability

The FocusPath[43] dataset used in this study can be found in an open Zenodo repository (https://zenodo.org/record/3926181#.YWXyJmLMJOQ). The BioImage[16] dataset can be found at the Broad BioImage Benchmark Collection (https://bbbc.broadinstitute.org/BBBC006). The DeepFocus[18] dataset can be found in an open Zenodo repository (https://zenodo.org/record/1134848#.YWXyoGLMJOR). TCGA@Focus can also be found in an open Zenodo repository (https://zenodo.org/record/3910757#.YWXywmLMJOQ).

Code Availability

DARTS-FQA pretrained models and codes are found at https://github.com/mahdihosseini/DARTS-FQA. FocusLiteNN models and codes are found at https://github.com/icbcbicc/FocusLiteNN.

References

1. Pantanowitz, L. Digital images and the future of digital pathology. *J. pathology informatics* **1** (2010).
2. Kohlberger, T. *et al.* Whole-Slide Image Focus Quality: Automatic Assessment and Impact on AI Cancer Detection. *J. Pathol. Informatics* **10**, 39, DOI: 10.4103/jpi.jpi_11_19 (2019).
3. Steiner, D. F., Chen, P.-H. C. & Mermel, C. H. Closing the translation gap: Ai applications in digital pathology. *Biochimica et Biophys. Acta (BBA)-Reviews on Cancer* 188452 (2020).
4. Wilson, M. L. *et al.* Access to pathology and laboratory medicine services: a crucial gap. *The Lancet* **391**, 1927–1938 (2018).
5. Lundberg, G. D. How many pathologists does the united states need? *JAMA network open* **2**, e194308–e194308 (2019).
6. Hosseini, M. S. *et al.* Focus quality assessment of high-throughput whole slide imaging in digital pathology. *IEEE Transactions on Med. Imaging* **39**, 62–74, DOI: 10.1109/TMI.2019.2919722 (2020).
7. Pantanowitz, L. *et al.* Review of the current state of whole slide imaging in pathology. *J. Pathol. Informatics* **2**, 36, DOI: 10.4103/2153-3539.83746 (2011). https://www.jpathinformatics.org/article.asp?issn=2153-3539;year=2011;volume=2;issue=1;spage=36;epage=36;aulast=Pantanowitz;t=6.

8. Farahani, N., Parwani, A. V. & Pantanowitz, L. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathol. Lab. Medicine Int.* **7**, 23–33 (2015).
9. Bueno, G., Fernández-Carrobles, M. M., Deniz, O. & García-Rojo, M. New trends of emerging technologies in digital pathology. *Pathobiology* **83**, 61–69 (2016).
10. Wang, Z., Hosseini, M., Miles, A., Plataniotis, K. & Wang, Z. Focuslitenn: High efficiency focus quality assessment for digital pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (Springer International Publishing, 2020).
11. Wang, Z., Athar, S. & Wang, Z. Blind quality assessment of multiply distorted images using deep neural networks. In *International Conference on Image Analysis and Recognition*, 89–101 (Springer, 2019).
12. Pinkard, H., Phillips, Z., Babakhani, A., Fletcher, D. A. & Waller, L. Deep learning for single-shot autofocus microscopy. *Optica* **6**, 794–797, DOI: 10.1364/OPTICA.6.000794 (2019).
13. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. Histoqc: An open-source quality control tool for digital pathology slides. *JCO Clin. Cancer Informatics* 1–7, DOI: 10.1200/CCI.18.00157 (2019). PMID: 30990737, https://doi.org/10.1200/CCI.18.00157.
14. Yu, S. *et al.* A shallow convolutional neural network for blind image sharpness assessment. *PLOS ONE* **12**, 1–16, DOI: 10.1371/journal.pone.0176632 (2017).
15. Kang, L., Ye, P., Li, Y. & Doermann, D. Convolutional neural networks for no-reference image quality assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1733–1740, DOI: 10.1109/CVPR.2014.224 (2014).
16. Yang, S. J. *et al.* Assessing microscope image focus quality with deep learning. *BMC bioinformatics* **19**, 77 (2018).
17. McQuin, C. *et al.* Cellprofiler 3.0: Next-generation image processing for biology. *PLOS Biol.* **16**, 1–17, DOI: 10.1371/journal.pbio.2005970 (2018).
18. Senaras, C., Niazi, M. K. K., Lozanski, G. & Gurcan, M. N. Deepfocus: detection of out-of-focus regions in whole slide digital images using deep learning. *PloS one* **13** (2018).
19. Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T. & Samek, W. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing* **27**, 206–219 (2017).
20. Campanella, G. *et al.* Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. *Comput. Med. Imaging Graph.* **65**, 142–151, DOI: https://doi.org/10.1016/j.compmedimag.2017.09.001 (2018). Advances in Biomedical Image Processing.
21. Ma, K. *et al.* End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Process.* **27**, 1202–1213 (2017).
22. Yu, S., Jiang, F., Li, L. & Xie, Y. Cnn-grnn for image sharpness assessment. In *Computer Vision - ACCV 2016 Workshops*, 50–61, DOI: 10.1007/978-3-319-54407-6_4 (2017).
23. Sangeetha, S., Dhaya, R., Shah, D. T., Dharanidharan, R. & Reddy, K. P. S. An empirical analysis of machine learning frameworks for digital pathology in medical science. In *Journal of Physics: Conference Series*, vol. 1767, 012031 (IOP Publishing, 2021).
24. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. Nih image to imagej: 25 years of image analysis. *Nat. methods* **9**, 671–675 (2012).
25. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. methods* **9**, 676–682 (2012).
26. Bian, Z. *et al.* Autofocusing technologies for whole slide imaging and automated microscopy. *J. Biophotonics* **13**, e202000227 (2020).
27. Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *The lancet oncology* **20**, e253–e261 (2019).
28. Radakovich, N., Nagy, M. & Nazha, A. Artificial intelligence in hematology: current challenges and opportunities. *Curr. hematologic malignancy reports* **15**, 203–210 (2020).
29. Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. pathology informatics* **7** (2016).
30. Fakoor, R., Ladhak, F., Nazi, A. & Huber, M. Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the international conference on machine learning*, vol. 28 (ACM, New York, USA, 2013).
31. Xu, J., Luo, X., Wang, G., Gilmore, H. & Madabhushi, A. A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing* **191**, 214–223 (2016).
32. Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. reports* **6**, 1–11 (2016).
33. Niazi, M. K. K., Beamer, G. & Gurcan, M. N. A computational framework to detect normal and tuberculosis infected lung from h and e-stained whole slide images. In *Medical Imaging 2017: Digital Pathology*, vol. 10140, 101400J (International Society for Optics and Photonics, 2017).
34. Parwani, A. V. Digital pathology as a platform for primary diagnosis and augmentation via deep learning. In *Artificial Intelligence and Deep Learning in Pathology*, 93–118 (Elsevier, 2021).

35. Liu, Y. *et al.* Artificial intelligence–based breast cancer nodal metastasis detection: Insights into the black box for pathologists. *Arch. pathology & laboratory medicine* **143**, 859–868 (2019).
36. Stathonikos, N., Veta, M., Huisman, A. & van Diest, P. J. Going fully digital: Perspective of a dutch academic pathology lab. *J. pathology informatics* **4** (2013).
37. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med. image analysis* **33**, 170–175 (2016).
38. Bhargava, R. & Madabhushi, A. Emerging themes in image informatics and molecular analysis for digital pathology. *Annu. review biomedical engineering* **18**, 387–412 (2016).
39. Komura, D. & Ishikawa, S. Machine learning methods for histopathological image analysis. *Comput. structural biotechnology journal* **16**, 34–42 (2018).
40. Gupta, A. *et al.* Deep learning in image cytometry: a review. *Cytom. Part A* **95**, 366–380 (2019).
41. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. medicine* **25**, 44–56 (2019).
42. Hosseini, M. S. & Plataniotis, K. N. Image sharpness metric based on maxpol convolution kernels. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 296–300, DOI: 10.1109/ICIP.2018.8451488 (2018).
43. Hosseini, M. S., Zhang, Y. & Plataniotis, K. N. Encoding visual sensitivity by maxpol convolution filters for image sharpness assessment. *IEEE Transactions on Image Process.* **28**, 4510–4525 (2019).
44. Hosseini, M. S. *et al.* Focus quality assessment of high-throughput whole slide imaging in digital pathology. *IEEE transactions on medical imaging* **39**, 62–74 (2019).
45. Leclaire, A. & Moisan, L. No-reference image quality assessment and blind deblurring with sharpness metrics exploiting fourier phase information. *J. Math. Imaging Vis.* **52**, 145–172 (2015).
46. Hassen, R., Wang, Z. & Salama, M. M. A. Image sharpness assessment based on local phase coherence. *IEEE Transactions on Image Process.* **22**, 2798–2810 (2013).
47. Bahrami, K. & Kot, A. C. A fast approach for no-reference image sharpness assessment based on maximum local variation. *IEEE Signal Process. Lett.* **21**, 751–755 (2014).
48. Li, L. *et al.* Image sharpness assessment by sparse representation. *IEEE Transactions on Multimed.* **18**, 1085–1097 (2016).
49. Tajbakhsh, N. *et al.* Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging* **35**, 1299–1312 (2016).
50. Zhou, Z. *et al.* Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7340–7351 (2017).
51. Kumar, A., Kim, J., Lyndon, D., Fulham, M. & Feng, D. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE journal biomedical health informatics* **21**, 31–40 (2016).
52. Kandel, I. & Castelli, M. How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset. *Appl. Sci.* **10**, 3359 (2020).
53. Tajbakhsh, N. *et al.* On the necessity of fine-tuned convolutional neural networks for medical imaging. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, 181–193 (Springer, 2017).
54. Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. medicine* **24**, 1559–1567 (2018).
55. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* **15**, e1002683 (2018).
56. Rawat, W. & Wang, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation* **29**, 2352–2449 (2017).
57. Nichani, E., Radhakrishnan, A. & Uhler, C. Do deeper convolutional networks perform better? *arXiv preprint arXiv:2010.09610* (2020).
58. Liu, H., Simonyan, K. & Yang, Y. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055* (2018).
59. Zoph, B. & Le, Q. V. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578* (2016).
60. Zoph, B., Vasudevan, V., Shlens, J. & Le, Q. V. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8697–8710 (2018).
61. Liu, H., Simonyan, K., Vinyals, O., Fernando, C. & Kavukcuoglu, K. Hierarchical representations for efficient architecture search. *arXiv preprint arXiv:1711.00436* (2017).
62. Real, E., Aggarwal, A., Huang, Y. & Le, Q. V. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, vol. 33, 4780–4789 (2019).
63. Hosseini, M. S. & Plataniotis, K. N. Adas: Adaptive scheduling of stochastic gradients (2020). 2006.06587.
64. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemp. oncology* **19**, A68 (2015).
65. Huang, G., Liu, Z., v. d. Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269, DOI: 10.1109/CVPR.2017.243 (2017).
66. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks (2019). 1801.04381.
67. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, DOI: 10.1109/CVPR.2016.90 (2016).

Acknowledgements

The authors would like to greatly thank Huron Digital Pathology (St. Jacobs, ON N0B 2N0, Canada) for the support of this research.

Author contributions statement

M.S.H. conceived the experiment(s), A.M. conducted the experiments, with feedback from S.T. and M.S.H. A.M., S.T. and M.S.H. analysed the results. All authors reviewed the manuscript.

Competing Interests Statement

The authors declare no competing interests in this study.

Figure Legends

Table 1: List of existing deep learning FQA methods. The Method Type column refers to whether the model was custom made for the application [C], or if it was a pre-existing architecture with slight modifications [P].

Table 2: WSI Dataset Information, listed in alphabetical order. Number of classes refers to the number of focus levels used in the dataset. The subscript for the Hue Mean column represents the standard deviation.

Table 3: CNN architecture and complexity summary. The #FLOPs and latency are listed for an input patch size of 64.

Table 4: Hours taken per GPU to process 300 WSIs. Each WSI has a 25k x 25k pixel size. The shortest time is in green.

Table 5: Performance Metrics for Transfer to TCGA@Focus[64]. The best PR and ROC result for each network is in green. The subscript indicates the standard deviation over the 5 trials.

Figure 1: An overview of the data preparation, training, and evaluation pipeline for FQA.

Figure 2: Hue relative frequency distribution for all colored datasets using a 1% sampling of pixels.

Figure 3: Examples of patches for each dataset with various focus levels. The number in the top right corner of each image corresponds to the focus level. For TCGA, "IF" stands for "in-focus", and "OF" stands for "out-of-focus".

Figure 4: Visualization of a reduction cell in the DARTS-FQA[58] architecture, trained on the FocusPath[43] dataset.

Figure 5: Relationship between FLOPs and Latency. The increasing marker size corresponds to the increasing patch sizes.

Figure 6: The knowledge gain is plotted against the mapping condition for each dataset. The 64 and 235 pixel input size versions are plotted on the same axes if applicable, with the larger points corresponding to the 235 pixel input size. Each image corresponds to a different trained dataset: **a)** BioImage[16] **b)** DeepFocus[18], and **c)** FocusPath[43].

Figure 7: ROC and PR curves that plot the two best and two worst performing networks for each dataset. The full set of curves can be found in the Supplementary Materials. Each plot corresponds to a different trained dataset: **a)** FocusPath[43] **b)** DeepFocus[18] **c)** BioImage[16] , and **d)** BioImage64[16].

Figure 8: The FLOPs cost (MegaFLOPs) is plotted on a logarithmic scale against the area under the ROC and PR curves for each dataset. The full set of plots can be found in the Supplementary Materials. Each plot corresponds to a different trained dataset: **a)** BioImage[16] **b)** BioImage64[16] **c)** FocusPath[43], and **d)** DeepFocus[18].

Figure 9: Focus Quality heat maps for a WSI from the TCGA[64] dataset. Red corresponds to a lower focus quality meaning out-of-focus, while blue is a higher focus quality. FocusPath[43] and BioImage[16] have a larger input patch size, explaining their coarser heatmap distribution. More heatmaps can be found in the Supplementary Materials. Figures **a)** and **d)** are the original Whole Slide Images. Heatmap representations with the appropriate training dataset and network are given in the remaining figures: **b)** FocusPath, DenseNet13, **c)** FocusPath64, FocusLiteNN-2, **e)** DeepFocus, DARTS-FQA, **f)** DeepFocus, FocusLiteNN-1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NatureFQA2021supp.pdf