Published in final edited form as:
ACM BCB. 2019 September ; 2019: 466–474. doi:10.1145/3307339.3342170.

Learning to Evaluate Color Similarity for Histopathology Images using Triplet Networks

Anirudh Choudhary[1], **Hang Wu**[2], **Li Tong**[3], **May D. Wang**[4]

[1.]Georgia Institute of Technology Atlanta, GA

[2.]Georgia Institute of Technology Atlanta, GA

[3.]Georgia Institute of Technology and Emory University Atlanta, GA

[4.]Georgia Institute of Technology and Emory University Atlanta, GA

Abstract

Stain normalization is a crucial pre-processing step for histopathological image processing, and can help improve the accuracy of downstream tasks such as segmentation and classification. To evaluate the effectiveness of stain normalization methods, various metrics based on color-perceptual similarity and stain color evaluation have been proposed. However, there still exists a huge gap between metric evaluation and human perception, given the limited explainability power of existing metrics and inability to combine color and semantic information efficiently. Inspired by the effectiveness of deep neural networks in evaluating perceptual similarity of natural images, in this paper, we propose TriNet-P, a color-perceptual similarity metric for whole slide images, based on deep metric embeddings. We evaluate the proposed approach using four publicly available breast cancer histological datasets. The benefit of our approach is its representation efficiency of the perceptual factors associated with H&E stained images with minimal human intervention. We show that our metric can capture the semantic similarities, both at subject (patient) and laboratory levels, and leads to better performance in image retrieval and clustering tasks.

Keywords

Perceptual Similarity; Whole-Slide Imaging; Representation Learning; Metric Embedding

1 INTRODUCTION

Stain normalization, which transforms the color distribution of an image into a defined reference space, is a crucial pre-processing step for histopathological image processing. Various factors such as varying staining procedures across labs, different color responses of digital scanners and inconsistent stain manufacturing processes lead to undesired color variation in microscopic images [22, 27, 45]. These variations hamper the performance of machine learning algorithms on downstream analysis tasks for automated disease diagnosis.

achoudhary46@gatech.edu.

Pre-processing with stain normalization techniques can mitigate this issue and has been shown to empirically improve classification [28] and nuclei segmentation accuracy [9, 49].

Conventional approaches rely on estimating the underlying stains by computing color deconvolution matrix [23, 31, 45] or performing histogram-matching [36]. More recently, machine learningbased methods have focused on leveraging internal morphological structures using nuclei segmentation or deep generative models Figure 2: Our training pipeline: in the a) pre-processing step, we tile a whole slide image to 512 × 512 sized patches and select ones with top filtering score (s0–1), and in the b) learning step, we first embed three images of a triplet to lower dimensional vectors, and require perceptually closer images to have a lower distance in the embedding space. This is captured by triplet loss and we train the embedding networks end-to-end on pre-processed data. like variational-autoencoders [22] and generative adversarial networks [50]. To enable automatic evaluation of the effectiveness of these methods, various metric-based approaches have been proposed. We can categorize popular metrics used in evaluating color normalization into two groups: 1) perceptual similarity-based and 2) stain evaluation-based.

In perceptual similarity-based evaluation metrics, popular approaches include comparing images using per pixel measures like peak-signal-to-noise ratio (PSNR), L2 Euclidean distance in alternate color spaces (HSV/Lαβ) or leveraging perceptually motivated distance metrics used for image quality assessment like Structural Similarity Index (SSIM) [10], Feature Similarity Index (FSIM) [53], Multi-Scale Structural Similarity Index (MS-SSIM) [47] and Quanterion Structural Similarity Index (QSSIM) [25]. PSNR and SSIM are amongst the most popular metrics in stain normalization literature [7, 40]. However, key drawbacks of these metrics include assuming pixel-wise independence or limited ability to capture higher order image structure and contextual color information for tissue components, which humans could easily perceive. Even in perceptually motivated color spaces like Lαβ, color channels are not completely independent due to dye contribution and cannot be compared independently. Another drawback of intensity-based index like SSIM is its high sensitivity to geometric and scale distortions. This becomes a big problem in stain-normalization studies wherein, the whole slide images being compared can have significant geometric differences.

Stain evaluation-based approaches rely on comparing the color distribution of hematoxylin & eosin (H&E) stained regions. Nuclei absorb hematoxylin while eosin is absorbed by stroma region. These metrics try to incorporate tissue structure by leveraging nuclei segmentation to estimate separate color distributions for hematoxylin-stained and eosin-stained regions. Normalized Median Intensity (NMI) or histogram percentiles are compared for each distribution using L2 distance [14, 16, 51]. However, accurate estimation of color distribution entails manual annotation of nuclei and stroma pixels in reference image. Also, summary statistics-based measures have limited ability to capture semantics.

There still exists a gap between metric-based evaluation and human judgment for color similarity (see an example illustration in Figure 1). Different semantic structure and tissue composition of source and reference images should be considered for developing the similarity metric. To capture both contextual and color information, inspired by the learned perceptual similarity metric (LPIPS) [54], we propose a novel learned perceptual similarity metric for H&E stained images, TriNet-P, which can be used for evaluating color normalization methods. Our metric is computed using deep embeddings, a feature vector which represents a visual concept and is extracted from a convolutional neural network. Specifically, we learn deep embeddings for images, and distances between images in the embedding space are used as our similarity metric, so that perceptually more similar images are closer to each other. The embeddings are learned via a triplet neural network structure [20] in an end-to-end fashion. In line with recent color normalization studies, the image data used in our study originates from four publicly available breast cancer histopathology datasets - MITOS [2], TUPAC [3], CAMELYON17 [4] and BACH-ICIAR [5]. For training, we leverage comprehensive image repository like The Cancer Genome Atlas (TCGA) [1], which comprises of whole-slide tissue biopsy samples with different cancer stages, and collected across multiple laboratories and patient cohorts. The main contribution of our work is three-fold:

- We propose a new pipeline to learn data-driven metrics for evaluating color similarity in histopathology images, which requires minimal supervision with the help of a state-of-theart triplet network training procedure.
- We show that compared to traditional perceptual metrics, our metric has an improved performance on identifying perceptually similar whole slide images: our metric can be used to retrieve perceptually similar image patches from multiple whole slide images.
- We also show that the embeddings can be used to define an inter cluster Euclidean distance which acts as a measure for perceptual stain similarity of whole slide images. The embeddings showed improved performance on clustering tasks and are also meaningful in low dimensional visualization tasks.

2 RELATED WORK

2.1 Image similarity metrics

Measuring perceptual similarity has been the focus of image quality assessment (IQA) methods. IQA methods have been classified as reference and no-reference based metrics. No-reference-based metrics use information contained in an image itself, such as the presence of noises and artifacts, for quality assessment, without comparing to the ground-truth image for reference [32, 33, 42]. On the other hand, reference-based metrics typically focus on evaluating the similarity between a reference image and a modified version of that image containing artifacts like blurring and distortion.

Traditional reference-based metrics which incorporate structural information include SSIM [10], Feature Similarity Index (FSIM) [53], Visual Information Fidelity (VIF) [35] and HaarPSI [37]. These metrics were not designed to compare images with spatial ambiguities and rely on sliding window based convolutions to measure local similarity within images. This makes them unsuitable for comparing images from two domains with different internal composition. SSIM is widely used for comparing natural images due to its simplicity and effectiveness in measuring perceptual changes by taking pixel inter-dependencies into account. SSIM performs structural comparison after normalizing for luminance & contrast changes across images and combines the structure comparison index with luminance & contrast comparison indices. It can be extended for multi-channel images by averaging the indices across all channels.Another line of research [44] has attempted to define new color spaces and corresponding distance metric that map colors with better perceptual uniformity. Recently, Zolotarev and Kaarna [55] leveraged a triplet network to learn a color metric using spectral data, which aligns better with human perception compared to Lαβ space. However, these methods are not widely popular for image similarity assessment due to their limitation in linking color with structural information.

In recent years, internal activations of deep convolutional networks trained for high-level image classification tasks have been shown to correspond well with perceptual sensitivity in humans, thus yielding an optimum space to measure image distance. Distance metric based on activations derived from simple unsupervised network initialization beats traditional image quality metrics significantly [54]. Our formulation is similar in spirit to the Learned Perceptual Similarity (LPIPS) metric proposed by Zhang et al. [54], LPIPS uses the hidden unit activations of the convolutional layers from image classification network such as AlexNet [29] and combines them with feature-specific weight to compute weighted activations. The final image distance metric d (·, ·) is computed by summing the average squared L2 distances between weighted activations over all layers. LPIPS optimizes the weights by training a fully connected network (G) with modified ranking loss (L), such that the distance metric best agrees with human judgment (h) derived from two-alternative forced choice (2AFC) tests. Given the pairwise distances in a image triplet (I, I1, I2) the loss function is given by:

$$L(l, l1, l2, h) = -h\log G(d(l, l1), d(l, l2)) - (1-h)\log(1 - G(d(l, l1), d(l, l2))) \quad (1)$$

where I is the reference image and h ∈ [0, 1].

2.2 Deep Metric Embeddings

Metric embedding aims to learn a function, fθ , which maps semantically similar samples of the data onto closer points in metric space while dissimilar samples are pushed apart. The function fθ is parameterized by θ and can range from a linear transform to complex non-linear mappings usually represented by deep neural networks. Deep metric learning has been applied to learn embeddings in various tasks including person classification [19], face recognition [41], visual product search [8], and object retrieval [52].

Learning perceptual similarity generally focuses on comparing example-to-example distances to train deep embeddings. This is achieved with pairwise distance based loss function using structures like pairs or triplets. Two widely popular formulations are contrastive loss and triplet loss. Contrastive loss based network tries to minimize the distance between matching pair (same class) while maximizing the distance between non-matching pair. Siamese networks [11] is a key milestone utilizing contrastive loss for signature recognition and face identification. A drawback of contrastive embedding is that it requires real-valued accurate pair-wise similarities for training, which is difficult to obtain in pathology. Natural image perceptual similarity metrics rely on human judgement studies to generate class labels for training. To address this, Triplet network was proposed by Ding and Tao [13] which focuses on relative dissimilarities between image pairs. Triplet network samples a triplet (i.e. a positive, a negative, and an anchor) and tries to learn the embeddings such that positive sample is pulled closer to the anchor by a pre-defined margin compared to the negative sample. Positive sample belongs to the same class as the anchor while negative sample belongs to a different class.

The performance of triplet network depends on selecting informative training triplets. Since the overall number of triplets is O(N3), where N is the size of the dataset, covering all combinations is inefficient and leads to slower training. Various approaches for triplet selection include pair-wise relevance based [46], semi-hard mining [39] and hard mining [20]. Recently, online hard mining based on selecting hardest triplets on a mini-batch level has been shown to generate higher accuracy for person re-identification [19]. Deep Embedding Learning in Medical Imaging. Deep metric embedding methods have also been applied in medical imaging applications. For example, Thammasorn et al. [43] used triplet loss to extract meaningful features that result in improved classification performance on gamma images during radiotherapy deliveries, and Yan et al. [48] used triplet network to learn lesion image embeddings that be used for abnormality detection.

3 METHOD

3.1 Problem Statement

We aim to learn a distance metric d that measures the distance between two images I, I1 as d(I, I1). The intuition is that the qualitative evaluation results matches human judgment of similarity:assume the oracle judgment can be represented as a function h, we'd like for any triplets (I, I1, I2), pairs deemed closer by human h should also be deemed closer by our metric d:

$$d(l, l1) > d(l, l2),\ \forall\, l, l1, l2,\ \text{for}\, h(l, l1) > h(l, l2) \tag{2}$$

To obtain a parametric form for d(I, I1), we adopt embedding based approach: we learn an embedding function fθ parameterized by θ, and the distance between two images is represented by the Euclidean distance between their embedding vectors in the embedding space as d(I, I1) = L2(fθ (I), fθ (I1)). To learn a distance metric is then to learn an embedding function that captures the semantic relationships between images, specifically in our case, the majority stain component from whole slide images.

3.2 Learning fθ with Triplet Loss

We use the triplet loss proposed by Hoffer and Ailon [20] to capture the intuition of Eq. (2). More specifically, for an anchor image I, we choose a positive image I+ that is from the same class of I (i.e. two patches from the same whole slide image), and also a negative image I− that is from a different class of I (i.e. two patches from different whole slide images). We provide a detailed description of how we assign the class of each image in later sections.

To match our intuition on the property of the metric, we require that the distance between the anchor and the positive sample d+ = L2(fθ(I), fθ(I+)) is smaller than the distance between the anchor and the negative sample d− = L2(fθ (I), fθ (I−)), by at least a margin γ . Mathematically, the loss function is

$$\mathrm{loss}(\mathrm{l}, \mathrm{l}+, \mathrm{l}-) = (\mathrm{y} + \mathrm{d}+ \; - \; \mathrm{d}-)+ \tag{3}$$

where the function (t)+ = max(t, 0) is the hinge loss function. We utilize the numerically more stable soft-margin approximation given by softplus function : ln(1 + exp(d+ − d−)). Minimizing the sum of softplus loss over valid triplets, for which d+ > d−, is our overall training objective.

Triplet Mining.—The key to learning meaningful embeddings is to sample triplets which effectively contribute to the loss function (i.e., generate larger gradients). To this end, we evaluate batch-hard and batch-all-based triplet mining proposed by Hermans et al. [19]. In batch-hard sampling, for each anchor image I, we select I+ that is from the same class of I, but is farthest from I in terms of perceptual distance, and I− that is not from the same class but is closest to I. Batch-all sampling considers all combinations of valid triplets (semi-hard and hard) in a mini-batch.

3.3 Class Assignment and Image Selection

To generate training data for our network, we partition each whole slide image (WSI) into 512 × 512 sized tiles and sample relevant patches. Tiles belonging to a WSI are assigned to a particular class. Since a large portion of WSI contains background, thresholding is required to extract tissue regions. Random sampling from WSI would lead to high intra-class variation in training data, primarily due to color variations caused by different absorption of H & E stains, and presence of artifacts like tissue folds, shadows, smudges & pen markings [27]. Hard mining-based triplet loss is sensitive to noise and as such, it would make the training difficult. Hence, at a class level, we need to ensure that there is relative color homogeneity. Moreover, tissue regions containing at least a certain number of nuclei have been leveraged in color normalization studies [14] implying that prioritizing nuclei containing areas can improve our metric's accuracy. To ensure this, we incorporate weak supervision by leveraging the scoring function proposed by Eriksson and Hu.

$$\mathrm{s} = \mathrm{t}\,2\,\mathrm{p} * \ln(1 + \mathrm{sf} * \mathrm{qf} * \mathrm{cf}), \; \mathrm{scaledscore}(\mathrm{s}0 - 1) = 1 - 10\;10 + \mathrm{s}\,. \tag{4}$$

Here, tp represents tissue percentage, sf measures how broadly saturation channel values are distributed, qf represents tissue quantity (more the better) and cf represents colorfactor which differentiates purple shades from pink shades and measures the relative deviation of each pixel's hue value from pink and purple vectors in HSV (Hue, Saturation, Value) color space.

The scoring function focuses on tiles with higher tissue percentage and weighs hematoxylin staining over eosin staining. This ensures that extracted images within a class have sufficient nuclei component and fully represent the semantic color dependencies between various tissue components.

3.4 Network Architecture for fθ

The core architecture of our model is based on the standard deep learning classification network and contains a convolutional neural network initialized with pre-trained weights. We utilize ResNet50- v1 [18] architecture pre-trained on ImageNet [38] database as our backbone architecture. It allows us to accommodate a larger batch size which leads to better optimization of triplet loss due to increased possibility of finding harder triplets. We also replace the last classification layer of the original ResNet network with two fully connected layers, the first one has 1,024 units, followed by ReLU and Batch Normalization [21] while the second layer is the feature embedding of dimension 128. We use global average pooling to reduce the spatial dimension of last convolutional layer. During each training step, the image triplet is fed to three convolution blocks with weight sharing to generate three embeddings for loss computation and back-propagation.

3.5 Training Configurations

Using pre-trained ResNet within our architecture leads to more effective generalization, in line with previous studies [6, 19]. Due to GPU memory constraints, we utilize 256 × 256 dimension images to train the network. Sampled tiles are re-sized to 256 × 256 dimension while inputting to the network and standard data augmentation using random cropping and flipping is applied. We apply data augmentation during training only, no test time augmentation is performed. Training is performed using both batch-hard and batch-all sampling approaches with mini-batch size of 80 images (P= 4 classes and K = 20 images per class). We use the Adam optimizer [24] available in TensorFlow with default hyperparameter values (ϵ = 10–8, β1 = 0.9, β2 = 0.999). We train the network for 30,000 iterations, starting with a learning rate of 3 × 10–4 and decay the learning rate exponentially after 20,000 iterations to close to zero.

4 EXPERIMENT AND RESULTS

4.1 Datasets

Our experiments were conducted using four breast cancer histological datasets with training and testing data derived as follows:

Training dataset. : Tumor Proliferation Assessment Challengeb (TUPAC) and The Breast Cancer Histology (BACH) datasets were used for training our model. TUPAC dataset has been derived from The Cancer Genome Atlas (TCGA) and consists of H&E-stained invasive cancer whole slide images (WSI). It comprises of 500 training WSIs with multiple levels of cancer proliferation and scanned at different laboratories. The WSIs are stored as multi-resolution pyramid structures and contain multiple downsampled versions of the original tissue scanned at 40x magnification and a spatial resolution of 0.25 μm/pixel. BACH dataset was released as a part of ICIAR-2018 challenge and consists of 30 high resolution H&Estained breast histology WSI comprising four carcinoma stages. We merged TUPAC and BACH datasets, and randomly sampled 210 WSI for training. From each WSI, we sample a maximum of 1,000 tiles at highest magnification level, generating 200,000 image patches, each of size 512 × 512.

Test dataset. : Test images were drawn from MITOS and CAMELYON17 datasets, both of which are distinct from the training dataset. The class labels for each test image were defined based on its respective subject and laboratory ids. MITOS dataset is a publicly available dataset released as a part of MITOS-ATYPIA ICPR'14 challenge. The dataset comprises of 16 WSI with multiple 10x frames per case, scanned using Aperio scanner and re-scanned using Hamamatsu scanner. It allows us to evaluate our metric on inter-microscope variability. We consider 11 subject cases in MITOS and sample 2,700 non-overlapping patches of dimension 512 × 512 across each microscope at 10x magnification. Given the paired images from two microscopes, extracted patches have overlapping image content but different staining. CAMELYON17 dataset was collected across five medical centers in the Netherlands and released as part of CAMELYON17 Grand Challenge. It comprises of 1,000 WSI cases with 5 slides per subject, each having lymph-node metastasis condition. Since the centers might have employed different slide scanners or tissue staining procedures, we leverage the dataset to evaluate our metric on laboratory-related color variations. From CAMELYON17, we randomly select three subjects at each laboratory and sample 500 image patches for each subject at 40x magnification, generating 7,500 images. For quantitative benchmarking (Experiments 1 and 2), we create a smaller dataset by subsampling 20% images from the original test dataset. This is due to computational challenges involved in calculating SSIM and LPIPS metrics across image pairs. For visualizing the embeddings (Experiment 3), we consider the complete test dataset (12,900 images).

4.2 Pre-processing

Pre-processing has been an important step in histopathology analysis for effective color normalization [14] or malignancy prediction [17, 34]. In line with the standard practice, we pre-process the whole slide image to improve the quality of our training dataset.

Tissue Segmentation.

Typically, whole slide images comprise of large amount of non-useful regions including background which need to be removed. We sub-sample each whole slide image by a factor of 1/16 to generate lower resolution image for tissue segmentation. Hysteresis thresholding is applied to remove background regions followed by color-based filtering across R,G, and B channels respectively to remove tissue artifacts. This is followed by morphological operations (closing & dilation) and removal of small objects to generate the final segmentation mask.

Tile Scoring.

Post tissue extraction, slides are split into contiguous image tiles of 512 × 512 pixels in size and a corresponding score (s0–1 defined in (4)) is generated for each tile. Moreover, only tiles with more than 90% tissue are considered. Tiles are sampled in decreasing order of scores and maximum of 1,000 tiles, with more than 90% tissue, are extracted for each WSI for training. We found that sampling 1,000 tiles gives a balanced representation of various tissue components.

4.3 Experiment Results

We evaluate the performance of our proposed similarity metric in terms of its ability to retrieve and cluster similar image tiles from WSIs, both quantitatively and qualitatively. We also evaluate as to how effectively the performance of two popular perceptual similarity assessment metrics, SSIM and LPIPS translates to pathology images. We utilize AlexNet architecture for LPIPS, with weights calibrated to human judgment using 2AFC data. For SSIM, we evaluate the structural similarity index independently for each channel and subsequently, average the indices across channels. Our experimental evaluation is split up into two sections. The first part highlights the performance on class-wise image retrieval task with much better performance than perception-based image quality assessment metrics. In the second part, we evaluate the 128 dimensional embedding vector on its ability to effectively capture perceptual differences, in images acquired via different microscopes or at different laboratories, using a clustering-based approach. We also highlight our embedding's effectiveness in assessing the quality of color normalization.

4.4 Experiment 1: Effectiveness of Learned Metrics in Retrieving Similar Images

The setup of this experiment is an image retrieval task, where for each query image, we retrieve the closest images from a candidate set (search pool) based on our similarity metric. Due to laboratory specific protocols, images taken at a particular laboratory would be perceptually closer in terms of color distribution compared to images captured at different laboratories. Hence, an ideal metric should be able to retrieve images belonging to the same laboratory with high accuracy. Moreover, since our metric has been trained on inter-slide perceptual variations, it should be able to retrieve patches belonging to the same subject with reasonable accuracy, even if two different subjects from the same laboratory have similar stain composition. We use the standard approach in retrieval tasks i.e. evaluating the recall at k scores (Rank-1, Rank-5 and Rank-10).

$$\text{Recall@k} = \#\,\text{of retrieved items @k that are relevant}/\ \text{total}\,\#\,\text{of relevant items} \tag{5}$$

We subsample 20% test dataset and split it equally into query and candidate sets comprising 1,300 images each. Each image is assigned a class label based on its subject (patient) id and laboratory/microscope scanner id. For each query image, we retrieve the closest candidate image using our metric and compare their classes. We evaluate the retrieval accuracy using both subject and laboratory as class ids. In Tables 1 and 2, we can see that our metric yields the best Rank-1 scores for both datasets, highlighting the limitations of existing metrics in comparing pathological images with varying semantics. Batch-all generates slightly better results than batch-hard mining which can be attributed to the fact that batch-hard mining is more sensitive to the presence of color variation and artifacts within the same slide. Qualitative analysis of subject-based retrieval (Figure 7) indicates our metric's ability to capture color-perceptual similarity across different subjects and laboratories. Our embedding not only takes into account the color similarity but also prioritizes tissue patches with similar nuclei structure and distribution. Due to the similar stain composition of tissues scanned at labs 1,2 and 4 in CAMELYON17 dataset, our metric's retrieval performance at subject level is sub-optimal, although in terms of patch perceptual similarity, it performs quite well.

4.5 Experiment 2: Effectiveness of Learned Metrics in Clustering

We then utilize clustering-based approach to evaluate the effectiveness of embeddings in capturing color-based perceptual nuances across different laboratories. We perform hierarchical clustering on embeddings generated for the subsampled test dataset. For SSIM and LPIPS, clustering is performed using pairwise distance matrix across all image pairs. The number of clusters (K) is set to the number of imaging sites across which the dataset was acquired (K = 2 for MITOS & K = 5 for CAMELYON17). The clustering performance is evaluated using normalized mutual information (NMI) score (NMI ranges in [0, 1] with 1 being perfect clustering in line with the ground truth clusters) as shown in Table 3. Our metric successfully captures the perceptual differences due to different scanners in MITOS dataset with perfect NMI score. For CAMELYON dataset, although the NMI score is lower, our performance is much better than existing metrics. The lower NMI score is due to similarity in stain color appearance of slides prepared at labs 2 & 4 (Figure 7).

4.6 Experiment 3: Visualization of Learned Embeddings

We visualize our embeddings for test samples using t-Distributed Stochastic Neighbor Embedding (t-SNE) [30] plots (Figure 3 & Figure 4). The plots clearly highlight the separation between different laboratories and imaging equipment captured by our embedding. To evaluate our embedding's effectiveness on capturing color shift during stain normalization, we consider 2,700 paired image samples from Aperio (A) & Hamamatsu (H) scanners in MITOS dataset. Using Structure-Preserving Color Normalization [45], we colornormalize domain 'A' images with respect to domain 'H', generating stain normalized image A*. In Figure 5, we visualize the embeddings for A, A* & H using t-SNE plot. As evident, the embeddings of A* and domain 'H' images overlap very well, indicating our embedding's efficacy in capturing color related perceptual changes. We also present examples of patient-wise WSI samples in Figure 6 highlighting that better stain normalization leads to higher overlap between the embeddings of color-normalized and reference images.

5 CONCLUSIONS AND FUTURE WORK

In this paper, inspired by the effectiveness of deep neural networks in extracting representations and evaluating natural image perceptual similarity, we proposed a pipeline to learn a metric for evaluating the color-based perceptual similarity of whole slide images. The key method of our approach is to embed whole slide images such that in the embedding space, perceptually similar images are closer than dissimilar images. As shown by the evaluation results, our learned metric can capture the perceptual similarity of whole slide images and has shown improved performance over popular perceptual similarity metrics in image retrieval and clustering tasks.

In the next phase of this project, we plan to collect physicians' ratings for similar whole slide image pairs, so that we can learn a metric that is even more aligned with human perception in the same fashion. We also aim to improve our pre-processing pipeline by developing a more sophisticated tile selection algorithm which identifies informative regions using a combination of histopathological image features with tile scores, in line with Kothari et al. [26]. Since stain normalization leads to improved performance on downstream analysis tasks, we also plan to incorporate our method as a quality control step in the pre-processing pipeline of real-world pathological image tasks such as diagnosis of heart rejection. Moreover, it will also be interesting to examine how alternative metric learning methods can be leveraged in our learning, for example, histogram distance-based loss, quadruplet network [12], and how this metric can be used in other biomedical imaging analysis settings.

ACKNOWLEDGMENTS

This work is supported by funding from American Heart Association, Giglio Cancer Research fund, Petit Institute Faculty Fellow Fund, and Carol Ann and David Flanagan Faculty Fellow Fund to Professor May D. Wang.

REFERENCES

[1]. [n. d.]. TCGA Research Network. https://www.cancer.gov/tcga/.

[2]. 2014 MITOS-ATYPIA-14. https://mitos-atypia-14.grand-challenge.org/.

[3]. 2016 Tumor Proliferation Assessment Challenge. http://tupac.tue-image.nl/.

[4]. 2017 CAMELYON17 Challenge. https://camelyon17.grand-challenge.org/Data/.

[5]. 2018 BACH ICIAR Challenge. https://iciar2018-challenge.grand-challenge.org/.

[6]. Almazan Jon, Gajic Bojana, Murray Naila, and Larlus Diane. 2018 Re-ID done right: towards good practices for person re-identification. arXiv:cs.CV/1801.05339

[7]. Bayramoglu N, Kaakinen M, Eklund L, and HeikkilÃď J. 2017 Towards Virtual H E Staining of Hyperspectral Lung Histology Images Using Conditional Generative Adversarial Networks. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) 64–71. 10.1109/ICCVW.2017.15

[8]. Bell Sean and Bala Kavita. 2015. Learning visual similarity for product design with convolutional neural networks. ACM Transactions on Graphics (TOG) 34, 4 (2015), 98.

[9]. Bentaieb A and Hamarneh G. 2018 Adversarial Stain Transfer for Histopathology Image Analysis. IEEE Transactions on Medical Imaging 37, 3 (3 2018), 792–802. 10.1109/TMI.2017.2781228 [PubMed: 29533895]

[10]. Bovik AC, Sheikh HR, and Simoncelli EP. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13, 4 (4 2004), 600–612. 10.1109/TIP.2003.819861 [PubMed: 15376593]

[11]. Bromley Jane, Guyon Isabelle, LeCun Yann, Eduard Säckinger, and Shah Roopak. 1994 Signature verification using a" siamese" time delay neural network. In Advances in neural information processing systems. 737–744.

[12]. Chen Weihua, Chen Xiaotang, Zhang Jianguo, and Huang Kaiqi. 2017 Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (7 2017). 10.1109/cvpr.2017.145

[13]. Ding Changxing and Tao Dacheng. 2017 Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE transactions on pattern analysis and machine intelligence 40, 4 (2017), 1002–1014. [PubMed: 28475048]

[14]. Ehteshami Bejnordi B, Litjens G, Timofeeva N, Otte-HÃűller I, Homeyer A, Karssemeijer N, and van der Laak JA. 2016 Stain Specific Standardization of Whole-Slide Histopathological Images. IEEE Transactions on Medical Imaging 35, 2 (2 2016), 404–415. 10.1109/TMI.2015.2476509 [PubMed: 26353368]

[15]. Eriksson Deron and Hu Fei. 2018 Whole-Slide Image Preprocessing in Python - Top Tile Retrieval. https://developer.ibm.com/articles/an-automatic-method-toidentify-tissues-from-big-whole-slide-images-pt4/ (2018).

[16]. Zanjani Farhad Ghazvinian, Zinger Svitlana, de With PHN, Bejnordi Babak E., and van der Laak Jeroen A.W.M.. 2018 Histopathology stain-color normalization using deep generative models. In 1st Conference on Medical Imaging with Deep Learning (MIDL 2018) 1–11. (a) A03 (b) A05 (c) A11 (d) A12

[17]. Ertosun Mehmet GÃijnhan and Rubin Daniel. 2015 Automated Grading of Gliomas using Deep Learning in Digital Pathology Images: A modular approach with ensemble of convolutional neural networks. AMIA Annu Symp Proc 2015 (11 2015), 1899–1908. [PubMed: 26958289]

[18]. He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. 2016 Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (6 2016). 10.1109/cvpr.2016.90

[19]. Hermans Alexander, Beyer Lucas, and Leibe Bastian. 2017 In Defense of the Triplet Loss for Person Re-Identification. arXiv:cs.CV/1703.07737

[20]. Hoffer Elad and Ailon Nir. 2015 Deep Metric Learning Using Triplet Network. Lecture Notes in Computer Science (2015), 84–92. 10.1007/978-3-319-24261-3_7

[21]. Ioffe Sergey and Szegedy Christian. 2015 Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:cs.LG/1502.03167

[22]. Janowczyk Andrew, Basavanhally Ajay, and Madabhushi Anant. 2017 Stain Normalization using Sparse AutoEncoders (StaNoSA): Application to digital pathology. Computerized Medical Imaging and Graphics 57 (2017), 50 – 61. 10.1016/j.compmedimag.2016.05.003 Recent Developments in Machine Learning for Medical Imaging Applications. [PubMed: 27373749]

[23]. Khan AM, Rajpoot N, Treanor D, and Magee D. 2014 A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using ImageSpecific Color Deconvolution. IEEE Transactions on Biomedical Engineering 61, 6 (6 2014), 1729–1738. 10.1109/TBME.2014.2303294 [PubMed: 24845283]

[24]. Kingma Diederik P. and Ba Jimmy. 2014 Adam: A Method for Stochastic Optimization. arXiv:cs.LG/1412.6980

[25]. Kolaman A and Yadid-Pecht O. 2012 Quaternion Structural Similarity: A New Quality Index for Color Images. IEEE Transactions on Image Processing 21, 4 (4 2012), 1526–1536. 10.1109/TIP.2011.2181522 [PubMed: 22203713]

[26]. Kothari Sonal, Osunkoya Adeboye O., Phan John H., and Wang May D.. 2012 Biological interpretation of morphological patterns in histopathological wholeslide images. ACM-BCB : ACM Conference on Bioinformatics, Computational Biology and Biomedicine 2012 (2012), 218–225.

[27]. Kothari Sonal, Phan John, Stokes Todd H, and Wang May. 2013 Pathology imaging informatics for quantitative analysis of whole-slide images. Journal of the American Medical Informatics Association : JAMIA 20 (8 2013). 10.1136/amiajnl-2012-001540

[28]. Kothari Sonal, Phan John, Moffitt Richard, Stokes Todd H, Hassberger Shelby E, Chaudry Qaiser, Young Andrew, and Wang May. 2011 Automatic batch-invariant color segmentation of histological cancer images. Proceedings / IEEE International Symposium on Biomedical Imaging: from nano to macro. IEEE International Symposium on Biomedical Imaging 2011, 657–660. 10.1109/ISBI.2011.5872492

[29]. Krizhevsky Alex, Sutskever Ilya, and Hinton Geoffrey E.. 2012 ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12) Curran Associates Inc., USA, 1097–1105. http://dl.acm.org/citation.cfm?id=2999134.2999257

[30]. van der Maaten Laurens and Hinton Geoffrey. 2008 Visualizing data using t-SNE. Journal of machine learning research 9, 11 (2008), 2579–2605.

[31]. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, , Schmitt C, and Thomas NE. 2009 A method for normalizing histology slides for quantitative analysis. In 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro 1107–1110. 10.1109/ISBI.2009.5193250

[32]. Moorthy Anush Krishna and Bovik Alan Conrad. 2010 A two-step framework for constructing blind image quality indices. IEEE Signal processing letters 17, 5 (2010), 513–516.

[33]. Moorthy Anush Krishna and Bovik Alan Conrad. 2011 Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE transactions on Image Processing 20, 12 (2011), 3350–3364. [PubMed: 21521667]

[34]. Paeng Kyunghyun, Hwang Sangheum, Park Sunggyun, and Kim Minsoo. 2017 A Unified Framework for Tumor Proliferation Score Prediction in Breast Histopathology. Lecture Notes in Computer Science (2017), 231–239. 10.1007/978-3-319-67558-9_27

[35]. Sheikh Hamid R and Bovik Alan. 2005 A visual information fidelity approach to video quality assessment. (1 2005).

[36]. Reinhard E, Adhikhmin M, Gooch B, and Shirley P. 2001 Color transfer between images. IEEE Computer Graphics and Applications 21, 5 (7 2001), 34–41. 10.1109/38.946629

[37]. Reisenhofer Rafael, Bosse Sebastian, Kutyniok Gitta, and Wiegand Thomas. 2018 A Haar wavelet-based perceptual similarity index for image quality assessment. Signal Processing: Image Communication 61 (2 2018), 33–43. 10.1016/j.image.2017.11.001

[38]. Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, Huang Zhiheng, Karpathy Andrej, Khosla Aditya, Bernstein Michael, Berg Alexander C., and Fei-Fei Li. 2015 ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115, 3 (2015), 211–252. 10.1007/s11263-015-0816-y

[39]. Schroff Florian, Kalenichenko Dmitry, and Philbin James. 2015 FaceNet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (6 2015). 10.1109/cvpr.2015.7298682

[40]. M Tarek Shaban Christoph Baur, Navab Nassir, and Albarqouni Shadi. 2018 StainGAN: Stain Style Transfer for Digital Histological Images. arXiv preprint arXiv:1804.01601 (2018).

[41]. Tadmor Oren, Rosenwein Tal, Shai Shalev-Shwartz Yonatan Wexler, and Shashua Amnon. 2016 Learning a metric embedding for face recognition using the multibatch method. In Proceedings of the 30th International Conference on Neural Information Processing Systems Curran Associates Inc., 1396–1397.

[42]. Tang Huixuan, Joshi Neel, and Kapoor Ashish. 2011 Learning a blind measure of perceptual image quality In CVPR 2011. IEEE, 305–312.

[43]. Thammasorn Phawis, Wootton Landon, Ford Eric, and Nyflot Matthew. 2017 Deep convolutional triplet network for quantitative medical image analysis with comparative case study of gamma image classification. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) IEEE, 1119–1122.

[44]. Tkalcic M and Tasic JF. 2003 Colour spaces: perceptual, historical and applicational background In The IEEE Region 8 EUROCON 2003. Computer as a Tool., Vol. 1. 304–308 vol.1. 10.1109/EURCON.2003.1248032

[45]. Vahadane A, Peng T, Sethi A, Albarqouni S, Wang L, Baust M, Steiger K, Schlitter AM, Esposito I, and Navab N. 2016 Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. IEEE Transactions on Medical Imaging 35, 8 (8 2016), 1962–1971. 10.1109/TMI.2016.2529665 [PubMed: 27164577]

[46]. Wang Jiang, Song Yang, Leung Thomas, Rosenberg Chuck, Wang Jingbin, Philbin James, Chen Bo, and Wu Ying. 2014 Learning Fine-Grained Image Similarity with Deep Ranking. 2014 IEEE Conference on Computer Vision and Pattern Recognition (6 2014). 10.1109/cvpr.2014.180

[47]. Wang Z, Simoncelli EP, and Bovik AC. 2003 Multiscale structural similarity for image quality assessment. In The Thirty-Seventh Asilomar Conference on Signals, Systems Computers, 2003, Vol. 2 1398–1402 Vol.2 10.1109/ACSSC.2003.1292216

[48]. Yan Ke, Wang Xiaosong, Lu Le, Zhang Ling, Adam P Harrison Mohammadhadi Bagheri, and Summers Ronald M. 2018 Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 9261–9270.

[49]. Yuan Edwin and Suh Junkyo. 2018 Neural Stain Normalization and Unsupervised Classification of Cell Nuclei in Histopathological Breast Cancer Images. arXiv:cs.CV/1811.03815

[50]. Zanjani FG, Zinger S, Bejnordi BE, van der Laak JAWM, and de With PHN. 2018 Stain normalization of histopathology images using generative adversarial networks. In 2018 IEEE

15th International Symposium on Biomedical Imaging (ISBI 2018) 573–577. 10.1109/ISBI.2018.8363641

[51]. Zarella Mark D., Yeoh Chan, Breen David E., and Garcia Fernando U.. 2017 An alternative reference space for H&E color normalization. PLOS ONE 12, 3 (3 2017), 1–14. 10.1371/journal.pone.0174489

[52]. Zezula Pavel, Amato Giuseppe, Dohnal Vlastislav, and Batko Michal. 2006 Similarity search: the metric space approach. Vol. 32 Springer Science & Business Media.

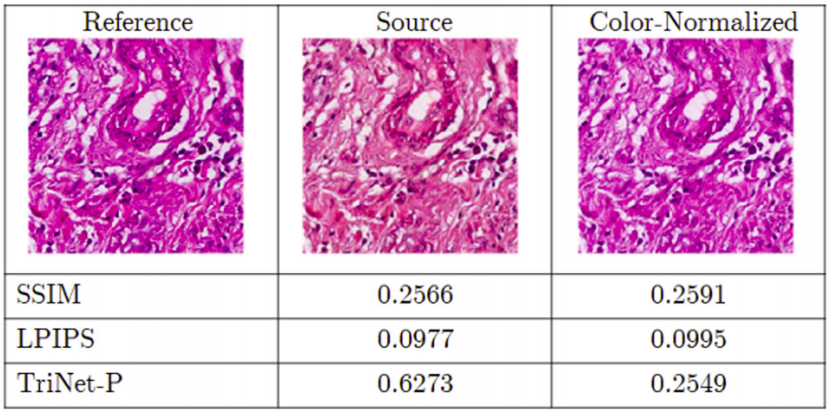
[53]. Zhang L, Zhang L, Mou X, and Zhang D. 2011 FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Transactions on Image Processing 20, 8 (8 2011), 2378–2386. 10.1109/TIP.2011.2109730 [PubMed: 21292594]

[54]. Zhang Richard, Isola Phillip, Efros Alexei A, Shechtman Eli, and Wang Oliver. 2018 The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In CVPR.

[55]. Zolotarev F and Kaarna A. 2018 Deep Metric Learning for Color Differences. In 2018 7th European Workshop on Visual Information Processing (EUVIP) 1–6.

CCS CONCEPTS

Applied computing → Imaging; • Computing methodologies → Image representations; Neural networks.



Reference	Source	Color-Normalized
SSIM	0.2566	0.2591
LPIPS	0.0977	0.0995
TriNet-P	0.6273	0.2549

Figure 1:

Perceptual Distances Scores: Efficacy of our perceptual distance (TriNet-P) over existing metrics for evaluating stain-normalization: structural similarity (SSIM) and learned perceptual similarity (LPIPS): in terms of color perception, for the first image, the third image should be closer compared to the second one. Our metric successful captures this perception distance while existing metrics cannot distinguish the two.

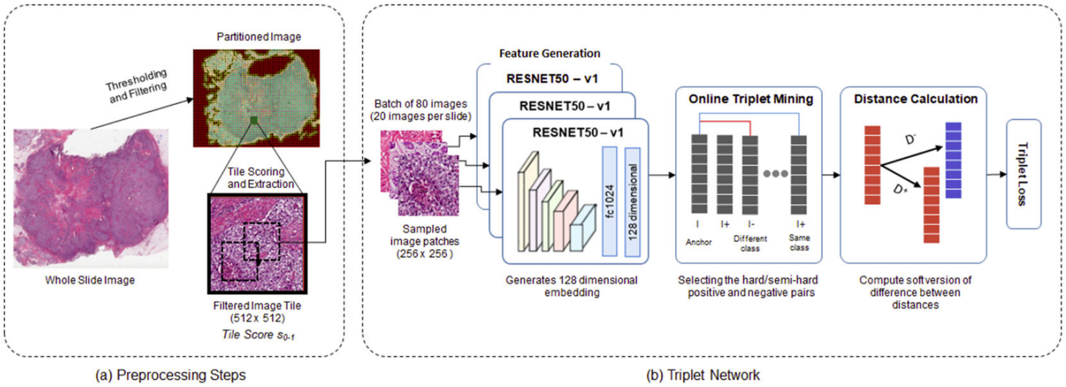

Thresholding and Filtering
Partitioned Image
Whole Slide Image
Tile Scoring and Extraction
Filtered Image Tile (512 x 512)
Tile Score $s_{0\text{-}1}$
(a) Preprocessing Steps
Feature Generation
RESNET50 – v1
RESNET50 – v1
RESNET50 – v1
Batch of 80 images (20 images per slide)
Sampled image patches (256 x 256)
fc1024
128 dimensional
Generates 128 dimensional embedding
Online Triplet Mining
I
I+
I-
I+
Anchor
Different class
Same class
Selecting the hard/semi-hard positive and negative pairs
Distance Calculation
D-
D+
Compute soft version of difference between distances
Triplet Loss
(b) Triplet Network

Figure 2:

Our training pipeline: in the a) pre-processing step, we tile a whole slide image to 512 × 512 sized patches and select ones with top filtering score (s_{0-1}), and in the b) learning step, we first embed three images of a triplet to lower dimensional vectors, and require perceptually closer images to have a lower distance in the embedding space. This is captured by triplet loss and we train the embedding networks end-to-end on pre-processed data.

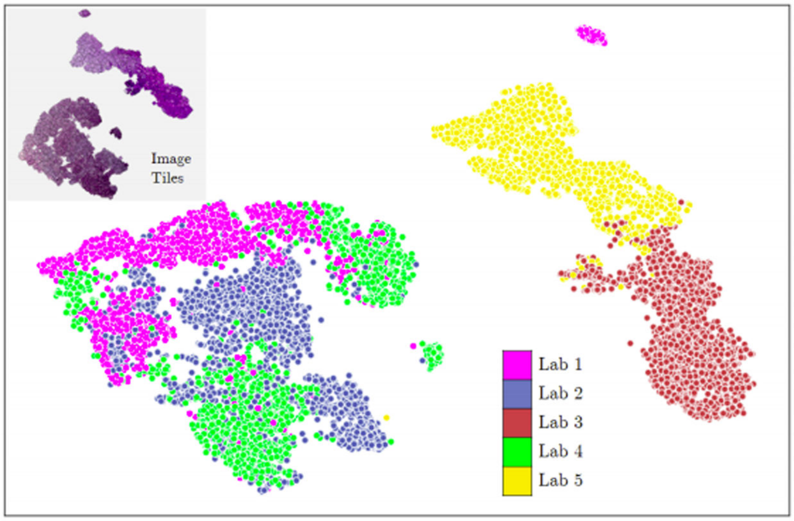


Figure 3:

t-SNE plot of embeddings for whole slide samples across five pathology labs in CAMELYON17

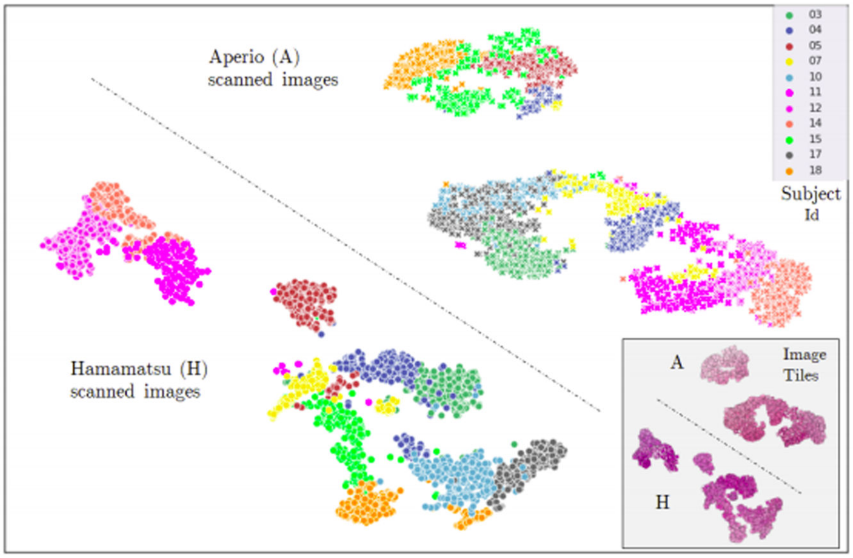


Figure 4:

t-SNE plot of embeddings for image samples generated from 11 subject case slides scanned using two micro scopes - Aperio (A) and Hamamatsu (H) (MITOS dataset)

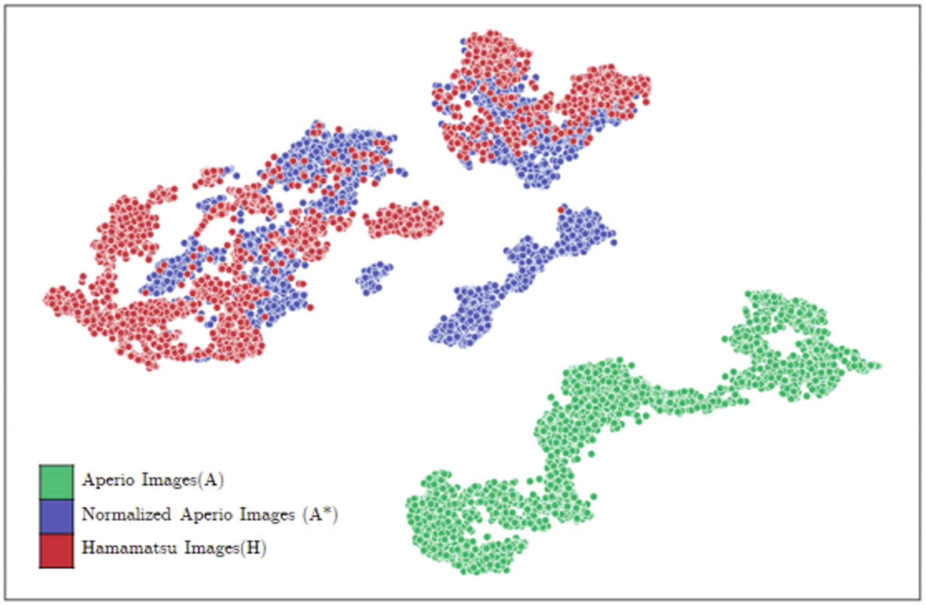
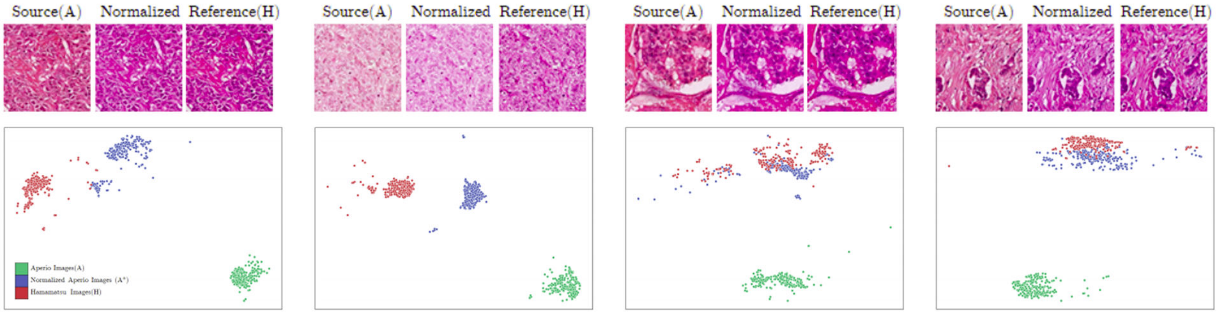


Figure 5:

t-SNE plot of embeddings for color-normalized Aperio scanner images in MITOS dataset



(a) A03

(b) A05

(c) A11

(d) A12

Figure 6:

tSNE plots of embeddings for Aperio scanned images color-normalized with respect to Hamamatsu for 4 subjects. First row highlights color normalization results. Good normalization in A11 & A12 leads to overlapping embeddings.

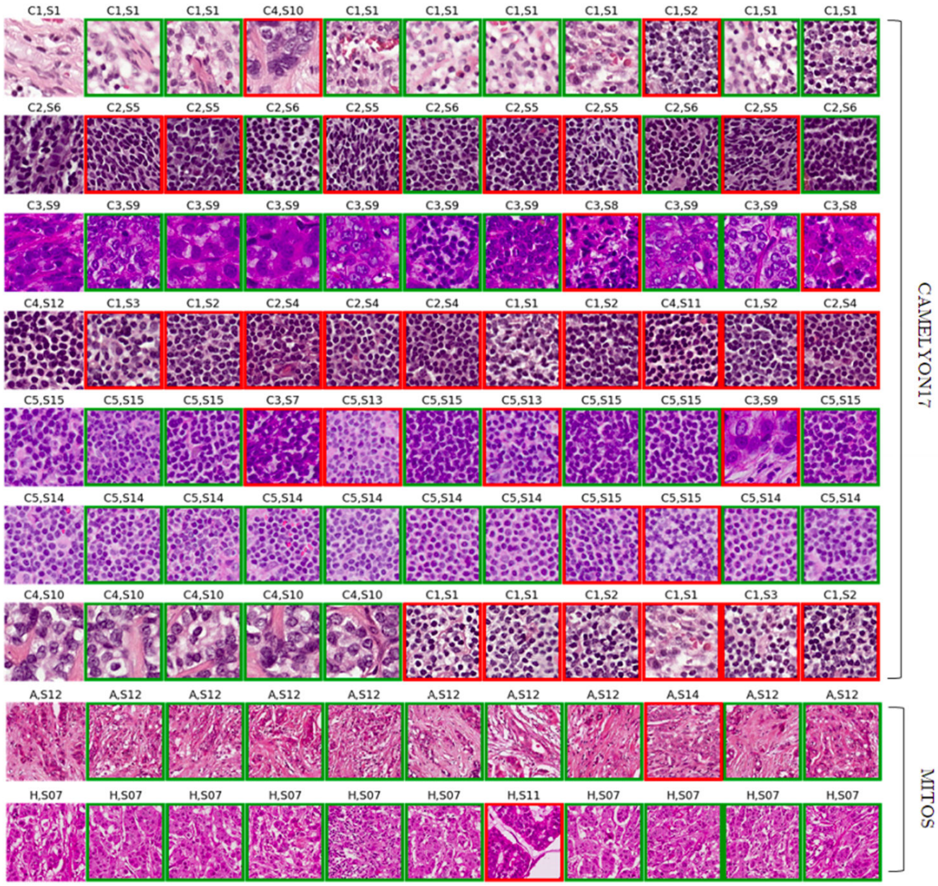

C1,S1 C1,S1 C1,S1 C4,S10 C1,S1 C1,S1 C1,S1 C1,S1 C1,S2 C1,S1 C1,S1
C2,S6 C2,S5 C2,S5 C2,S6 C2,S5 C2,S6 C2,S5 C2,S5 C2,S6 C2,S5 C2,S6
C3,S9 C3,S9 C3,S9 C3,S9 C3,S9 C3,S9 C3,S9 C3,S8 C3,S9 C3,S9 C3,S8
C4,S12 C1,S3 C1,S2 C2,S4 C2,S4 C2,S4 C1,S1 C1,S2 C4,S11 C1,S2 C2,S4
C5,S15 C5,S15 C5,S15 C3,S7 C5,S13 C5,S15 C5,S13 C5,S15 C5,S15 C3,S9 C5,S15
C5,S14 C5,S14 C5,S14 C5,S14 C5,S14 C5,S14 C5,S14 C5,S15 C5,S15 C5,S14 C5,S14
C4,S10 C4,S10 C4,S10 C4,S10 C4,S10 C1,S1 C1,S1 C1,S2 C1,S1 C1,S3 C1,S2
A,S12 A,S12 A,S12 A,S12 A,S12 A,S12 A,S12 A,S12 A,S14 A,S12 A,S12
H,S07 H,S07 H,S07 H,S07 H,S07 H,S07 H,S11 H,S07 H,S07 H,S07 H,S07
CAMELYON17
MITOS

Figure 7:

Subject-based retrieval results with 10-nearest neighbours (batch-all). The heading of each image indicates the laboratory/scanner id & subject id. First column contains query image; ‘Red’ border indicates subject mismatch, ‘Green’ border indicates same subject.

Table 1:

Laboratory-based retrieval scores for MITOS and CAMELYON17 (CAM) datasets

Metric	Rank-1		Rank-5		Rank-10	
	MITOS	CAM	MITOS	CAM	MITOS	CAM
SSIM [10]	51.71	19.40	84.03	25.39	92.78	32.42
LPIPS [54]	47.72	20.44	72.81	54.69	80.99	72.14
TriNet-P (batch-hard)	100.00	85.02	100.00	97.13	100.00	98.18
TriNet-P (batch-all)	100.00	88.02	100.00	96.48	100.00	97.92

Table 2:

Subject-based retrieval scores for MITOS and CAME-LYON17 (CAM) datasets

Metric	Rank-1		Rank-5		Rank-10	
	MITOS	**CAM**	**MITOS**	**CAM**	**MITOS**	**CAM**
SSIM [10]	4.37	7.16	12.74	9.64	19.77	13.28
LPIPS [54]	4.56	6.38	12.36	23.44	19.58	35.94
TriNet-P (batch-hard)	76.81	48.57	95.06	85.42	97.91	94.01
TriNet-P (batch-all)	84.03	58.59	97.53	87.63	98.67	94.66

Table 3:

NMI scores for clustering on MITOS and CAME-LYON17 (CAM) datasets using laboratory/scanner as class

Metric	MITOS	CAM
SSIM [10]	0.003	0.016
LPIPS [54]	0.005	0.007
TriNet-P (batch-hard)	1.000	0.456
TriNet-P (batch-all)	1.000	0.536