Comparative Analysis of Multi-Channel Feature Extraction Using a Modified K-means and PCA for PARS-to-H&E Image Translation

Marian Boktor[1,2], Paul Fieguth[1], Parsin Haji Reza[2]
[1]Vision and Image Processing Lab, University of Waterloo
[2]PhotoMedicine Labs, University of Waterloo
{marian.boktor, paul.fieguth, parsin.hajireza}@uwaterloo.ca

Abstract

Histological staining, particularly H&E staining, is essential in pathology for visualizing tissue structures, but traditional methods are time-consuming. Photon Absorption Remote Sensing (PARS), a high-resolution microscopy technique, offers a promising alternative by capturing H&E-like contrasts directly, enabling virtual staining without the need for chemical reagents. However, differentiating biological structures remains challenging for current models. We propose that channel-specific feature extraction could enhance colorization accuracy. This study investigates the effectiveness of modified K-means algorithm and Principal Component Analysis (PCA) for feature extraction in virtual staining. Results reveal that features produced by the K-means approach more effectively isolate tissue-specific structures, leading to improved labeling compared to PCA and conventional PARS channels. This advantage is demonstrated both quantitatively, through higher Structural Similarity Index (SSIM) scores, and visually, with enhanced colorization outcomes.

1 Introduction

Histological staining is essential in pathology, enabling the visualization of tissue structures for accurate diagnosis [1]. Among various techniques, Hematoxylin and Eosin (H&E) staining is the gold standard for providing high-contrast visualization of cell nuclei and extracellular components. However, conventional staining methods are time-consuming, requiring lengthy preparation and staining processes that delay diagnostic turnaround times [2, 3].

To streamline this process, researchers have pursued virtual staining methods that work with images captured by advanced, high-resolution microscopes, offering faster processing by eliminating the need for chemical reagents. A notable innovation is Photon Absorption Remote Sensing (PARS) [4, 5, 6, 7], a high-resolution, label-free microscopy technique capable of generating H&E-like images by directly capturing hematoxylin-like (non-radiative) and eosin-like (radiative) contrasts, which we refer to as conventional PARS channels. By leveraging these intrinsic tissue contrasts, PARS enables label-free imaging with high structural fidelity, facilitating the training of virtual staining models, which has proven successful [6, 7].

Despite their promise, virtual staining methods using models like Generative Adversarial Networks (GANs) [8, 9, 10] face challenges in accurately differentiating tissue types, often resulting in colorization errors. This may stem from the models' limited ability to capture and isolate distinct tissue features effectively during training, possibly due to a lack of explicit understanding of underlying structures within the data.

To address these challenges, this study introduces a novel approach that enhances the virtual staining process by first separating biological structures based on features produced by a modified K-means [11, 12] and Principal Component Analysis (PCA) [13, 14] before model training. By improving the learning process through these separation techniques, we aim to achieve more precise tissue differentiation and thus colorization in PARS images. This study builds upon the work presented in [7], with a focus on comparing two widely recognized feature extraction techniques and examining their effects on feature labeling and virtual staining.

PARS data inherently shows partial structural separation: non-radiative (NR) channels capture nuclei, while radiative (R) channels highlight cytoplasm and connective tissues [5]. Previous studies [11, 12] have identified

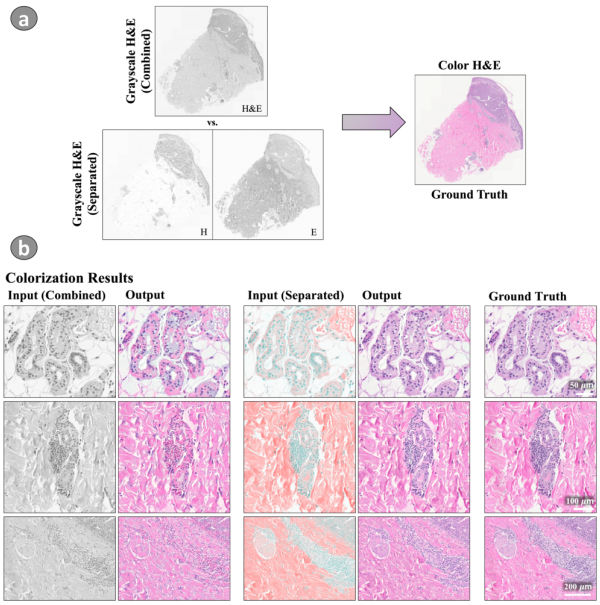

Figure 1: A visual comparison of colorization results using combined vs. separated grayscale H&E channels on human skin tissue. (a) Inputs: combined and separated grayscale H&E, with color H&E as ground truth. (b) Resulting colorizations, showing improved accuracy with separated inputs.

a relationship between PARS signals and specific structures, suggesting that enhanced labeling through channel separation could improve colorization accuracy.

To validate the potential of this approach, we conducted a proof-of-concept study with color H&E images taken by a brightfield microscope. The objective was to recolor these images with GAN models by providing inputs in the form of separated H&E channels. Using Ruifrok's method [15] with the sci-kit-image library [16], we isolated the hematoxylin and eosin channels, converted them to grayscale, and recombined them into RGB images as inputs for the virtual staining GAN model.

For comparison, we also used grayscale versions of the color H&E images as a baseline input alongside the separated channels to assess how different input formats affect colorization. Results (Figure 1 (b)) show that channel-separated inputs significantly enhance colorization accuracy, reducing blending between structures like nuclei and connective tissues. These findings highlight that separating contrasts helps the model learn relationships between input and target domains, resulting in improved colorization outcomes.

2 Methodology

Building on these findings, our methodology utilizes non-radiative time-domain (TD) signals from PARS to investigate how feature extraction can enhance virtual staining. These TD signals have the potential to convey valuable multimodal information about the observed targets. To capture this diversity, we selected modified K-means and PCA due to their strengths in feature extraction and differentiation, enabling us to isolate key structural information for input into the staining model.

In combining feature channels, we recognize that including all possible elements could introduce redundancy and potentially reduce model effectiveness. To address this, we later conduct a combinatorial analysis of feature sets to identify the optimal subset of extracted features that maximizes colorization accuracy and avoids overloading the model with repetitive information.

2.1 Feature Extraction Approaches

2.1.1 Modified K-means

To capture tissue-specific information from non-radiative TD signals, we applied a tailored version of K-means (K*-means) approach based on Pellegrino et al. [12]. This method identifies clusters by analyzing signal shape, and is both robust to noise and signal inversion—critical for accurately labeling distinct biological structures in virtual staining. Each TD signal is treated as a vector in $\mathbb{R}^n$, where n is the number of samples per signal. The angle between vectors is used to measure similarity, with orthogonal signals considered maximally distant. Cluster centroids (i.e., the learned features) are computed as the principal component of each cluster and its negative, ensuring centroids are resilient to noise.

After clustering, each TD signal is represented by a weighted sum of K feature vectors, $\mathcal{F} = \{\vec{f}_i\}$, with K chosen to balance structural detail and redundancy. These feature vectors were then used to generate feature images, which highlight different tissue components, such as nuclei and connective tissues. This approach yielded K feature images per sample, arranged as channels for the virtual staining model. We selected K based on visual clarity, testing values from 2 to 6 to avoid redundant or indistinct clusters. For detailed information about the methodology, refer to [7].

2.1.2 Principal Component Analysis (PCA)

To complement K*-means, we used PCA to reduce data dimensionality while retaining major variance in the

TD signals [12]. Unlike K*-means, PCA doesn't directly capture specific biological features but instead highlights overall variance. Each TD signal was transformed into a set of principal components (PCs), preserving data variance in fewer dimensions while minimizing information loss. This reduced-dimension representation served as an alternative input for virtual staining, allowing the model to leverage broad signal patterns without focusing on specific signal shapes.

For a fair comparison, we varied the number of PCs from 2 to 6 to match the range used in K*-means. The chosen PCs were arranged as multi-channel inputs for the GAN model, similar to the K*-means features, enabling the model to explore the impact of different data representations on colorization accuracy.

2.2 Multi-Channel GAN (MC-GAN)

In this study, we apply CycleGAN [8], a variant of GAN model [17], to translate label-free PARS images (source domain) into virtually stained images that resemble H&E-stained samples (target domain).

The Multi-Channel GAN (MC-GAN), introduced in [7], extends the CycleGAN framework by enabling multi-channel input, allowing it to process data with richer features. Traditional CycleGAN models typically handle single-channel (grayscale) or three-channel (RGB) images [18, 8, 19]. Prior studies [6, 9, 10] replaced RGB channels with NR and R channels, which proved effective when using three or fewer channels. However, to enhance virtual staining, we leverage additional structural information from PARS NR signals, requiring a model capable of handling more than three input channels.

MC-GAN accommodates this by expanding the allowable input channels to integrate multiple feature layers, capturing a broader range of structural details during training. Other than this channel expansion, the MC-GAN architecture retains the core CycleGAN design [8]. This multi-channel capability enables MC-GAN to leverage richer data, potentially improving colorization accuracy and overall model effectiveness.

2.3 Dataset and Training Settings

We used a human skin dataset collected by the research team at the Photomedicine Labs, University of Waterloo. In this study, the dataset was collected using two excitation wavelengths, 266 nm and 532 nm, to selectively target nuclei and red blood cells (RBCs), respectively. We denote the non-radiative channels as NR_{266} and NR_{532}, while the radiative channel is labeled as R_{266}.

From this dataset, we extracted overlapping 256×256, resulting in approximately 500 patches. We split the data into 70% training, 10% validation, and 20% testing sets. The GAN model was trained with a learning rate of 0.0002 for a maximum of 200 epochs, with early stopping applied if the generator loss did not improve. This setup ensured model stability and prevented overfitting.

Empirically, we found that setting the patch overlap to ∼50% minimized boundary artifacts, allowing seamless reconstruction of the final images from processed patches. All training was implemented in Python 3.10.6 using PyTorch 2.0.0 with CUDA 12 support, ensuring computational efficiency.

3 Results and Discussion

3.1 Feature Extraction

To determine a suitable number of features, a preliminary study was conducted using the K*-means algorithm. Feature extraction was performed for $K \in \{2, \ldots, 6\}$, generating feature sets M_f^K for each value of K, as shown in Figure 2, top. The results indicate that when $K < 3$, key structures like connective tissue and cell nuclei are not effectively distinguished in the feature images. For instance, with $K = 2$, these structures merge into single feature images. However, at $K = 3$, they become clearly separated. Increasing K beyond 3 results in redundancy, with similar structures forced into multiple images. Further experiments with these features as input to the virtual staining model confirmed that $K = 3$ achieves maximum separation without redundancy, enhancing the virtual staining process.

Similarly, PCA was applied to the PARS signals, generating six PC images, as seen in Figure 2 (bottom). Unlike K*-means, PCA captures data variance but lacks alignment with specific biological structures, resulting in feature images without clear biological separation. For example, PC_1 combined multiple structures at different intensities, while PC_5 and PC_6 appeared nearly redundant. These observations emphasize the limitations of PCA when applied to PARS image data, as it falls short in isolating and emphasizing relevant structures, which can be beneficial for improving virtual staining.

3.2 Virtual Staining

After extracting features with K*-means, the MC-GAN model was trained using the feature set M_f^K for each K from 2 to 6, along with the radiative (R) channel. Since the K features are derived from non-radiative (NR) signals and are independent of the R channel, the R channel was consistently included in the virtual staining phase for fair comparison. Model performance was then assessed using both visual evaluation and SSIM [20, 21],

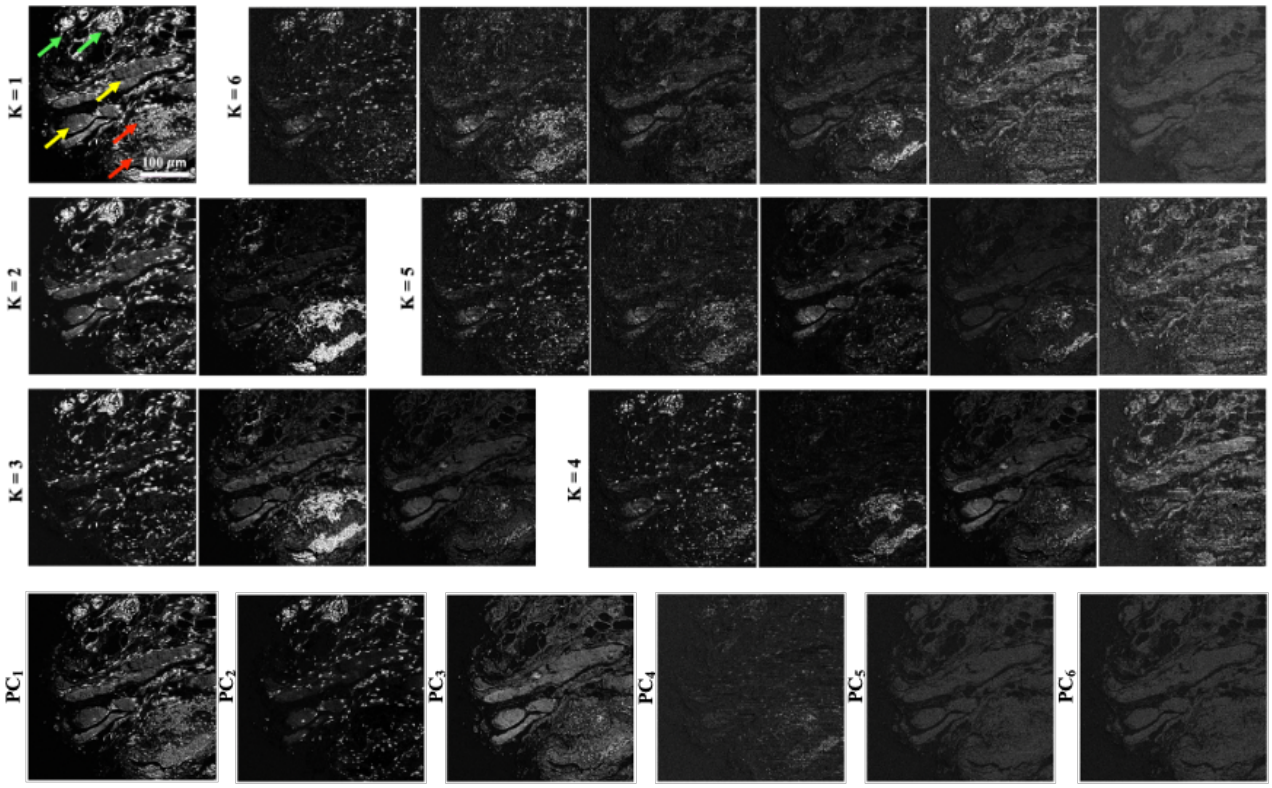

Figure 2: Comparison of feature extraction results of the K*-means (top) and PCA (bottom) using human skin data. K*-means: When $K < 3$, major structures like RBCs (red arrows) and connective tissue (yellow arrows) are not effectively separated. When $K > 3$, the features start to become redundant and are forcefully separated like in the case of RBCs and cell nuclei (green arrows). PCA: Results for the first X components for each $X \in \{2, \ldots, 6\}$. It can be noted that the PCA features do not align with identifiable biological structures.

comparing colorized images to their ground truth. The human skin dataset achieved optimal results at $K = 3$.

For PCA, MC-GAN models were similarly trained with the R channel and the first X PCs for X from 2 to 6. Visual assessment and SSIM were used to determine the optimal number of PCs, which was three. However, these first three PCs captured only 60% of data variance, while 137 PCs would be needed to explain 90%, limiting PCA's effectiveness in capturing critical structural details.

Overall, features from K*-means outperformed those from PCA in capturing meaningful biological features for virtual staining, as shown in Figure 3. The misalignment of PCA-derived features with distinct structures led to color blending between regions (see Figure 3, bottom). For example, PCA failed to specifically highlight RBCs, leading to low-contrast RBC colorization. In contrast, K*-means at $K = 3$ provided superior contrast, isolating RBCs in a bright pink color in the colorized output (Figure 3, top). RBCs are labeled with red arrows in the feature images in Figure 2, top, for reference. K*-means also effectively segmented nuclei and connective tissue, resulting in clearer colorization. These findings support our hypothesis that channel separation enhances learning and improves colorization accuracy.

Both methods produced comparable SSIM metrics; however, the K*-means features consistently outperformed PCA in terms of visual clarity and contrast [12], underlining the practical advantages of using K*-means for virtual staining. This improved performance has significant implications for applications in medical diagnostics, where clear and accurate representation of tissue structures is crucial.

After feature extraction, an analysis was performed to identify the optimal subset of K*-means features for virtual staining. Using the optimal value of K from the initial study, a comprehensive feature set, M_f^{opt}, was created and combined with the conventional NR_{532}, NR_{266}, and R_{266} channels to form an array, $A = [NR_{532}, NR_{266}, R_{266}, M_f^{opt}]$. Since including all elements of A could introduce redundancy, an exhaustive search was conducted across all possible feature combinations to identify the most effective set for training. Model performance was evaluated using multiple quantitative metrics, including SSIM and PSNR, by comparing colorized images with the true H&E counterparts.

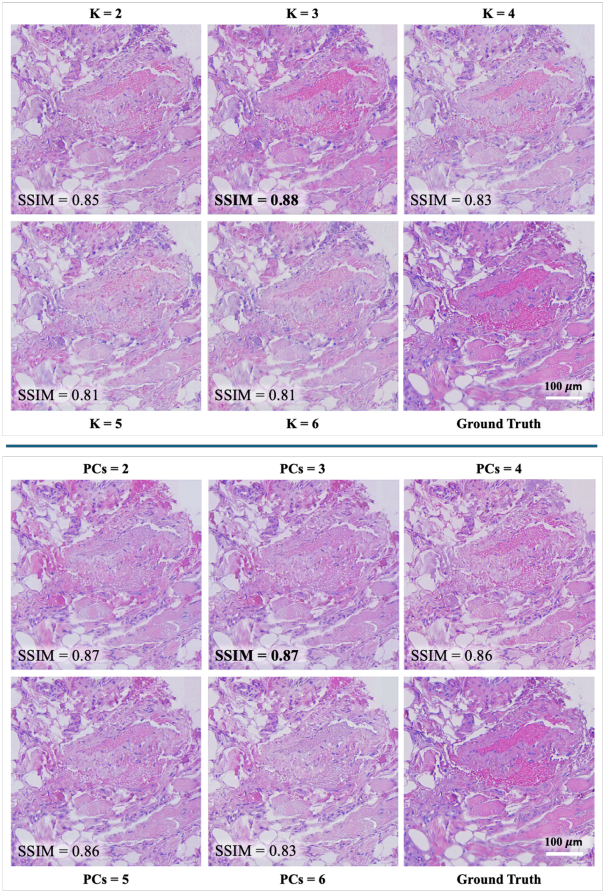

Figure 3: Comparison of virtual staining results using K*-means and PCA features using human skin data. K*-means (top): Results for each $K \in \{2, \ldots, 6\}$. The findings indicate that the optimal colorization is produced when $K = 3$. PCA (bottom): Results for the first X components for each $X \in \{2, \ldots, 6\}$. The findings indicate that the optimal colorization is produced when $X = 3$.

The best-performing feature subsets from A were selected for final model evaluation [7].

3.3 Challenges and Potential Applications

While the results indicate that $K = 3$ provides the optimal balance for the tissue structures in this study, challenges remain in determining the best value for K for different tissue types and stains. For tissues with more complex or heterogeneous structures, it may be necessary to fine-tune the value of K to achieve optimal feature separation. Future work should explore the impact of varying K for different tissue types (e.g., skin vs. kidney tissue) and staining protocols (e.g., H&E vs. PAS stains), which may require adjustments to the feature extraction process.

Integrating K*-means with existing machine learning models has the potential to significantly enhance diagnostic tools in pathology labs. By providing more accurate and biologically meaningful feature extraction, it could improve the model's ability to distinguish subtle tissue variations and detect early-stage abnormalities, such as tumors. This, in turn, could lead to more reliable and precise diagnoses, reducing the likelihood of misdiagnoses and supporting pathologists in making data-driven decisions. Ultimately, this integration could expedite the diagnostic process, enabling quicker treatment decisions and improving patient outcomes, especially in time-sensitive clinical scenarios.

4 Conclusion

This study demonstrates that the modified K-means algorithm outperforms PCA in both visual contrast and structural alignment, making it a more effective method for separating biological structures in PARS data. The improved feature separation achieved by K*-means significantly benefits virtual staining models. Its ability to enhance virtual staining highlights its potential as a valuable tool for medical diagnostics, especially in clinical and pathology lab settings. While determining the optimal number of features for various tissue types and stains remains a challenge, the findings suggest that K*-means offers an accurate and practical solution for virtual staining with promising real-world applications. Future research could explore advanced feature extraction methods and test the approach on a wider range of tissues and stains to broaden its applicability.

Acknowledgments

The authors would like to thank James E.D. Tweel and Benjamin R. Ecclestone for helping with data collection, and Jennifer Ai Ye for helping with MC-GAN implementation. The authors would also like to thank Dr. Marie Abi Daoud at the Alberta Precision Laboratories in Calgary, Canada for providing the human skin tissue samples and Dr. Deepak Dinakaran and Dr. Kevin Camphausen from the radiation oncology branch at the National Cancer Institute, NIH, Bethesda, MD, USA for providing the mouse brain samples. Additionally, the authors would like to acknowledge Hager Gaouda for their valuable assistance in staining the tissue samples used in this study. The authors gratefully acknowledge the financial support provided by the following funding sources throughout the duration of this project: Natural Sciences and Engineering Research Council of Canada (DGECR-2019-00143, RG-

PIN201906134, DH-2023-00371), Canada Foundation for Innovation (JELF 38000), Mitacs Accelerate (IT13594), University of Waterloo Startup funds, Centre for Bioengineering and Biotechnology (CBB Seed fund), illumiSonics Inc (SRA 083181), New frontiers in research fund – exploration (NFRFE-2019-01012), and The Canadian Institutes of Health Research (CIHR PJT 185984).

References

[1] V. Baxi, R. Edwards, M. Montalto, and S. Saha, "Digital pathology and artificial intelligence in translational medicine and clinical practice," *Modern Pathology*, vol. 35, no. 1, pp. 23–32, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893395222003477

[2] C. E. Day, Ed., *Histopathology: Methods and Protocols*, ser. Methods in Molecular Biology. New York, NY: Springer New York, 2014, vol. 1180. [Online]. Available: http://link.springer.com/10.1007/978-1-4939-1050-2

[3] L. Kang, X. Li, Y. Zhang, and T. T. Wong, "Deep learning enables ultraviolet photoacoustic microscopy based histological imaging with near real-time virtual staining," *Photoacoustics*, vol. 25, p. 100308, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213597921000689

[4] P. Hajireza, W. Shi, K. Bell, R. J. Paproski, and R. J. Zemp, "Non-interferometric photoacoustic remote sensing microscopy," *Light: Science & Applications*, vol. 6, no. 6, pp. e16 278–e16 278, Jun. 2017. [Online]. Available: http://www.nature.com/articles/lsa2016278

[5] B. R. Ecclestone, K. Bell, S. Sparkes, D. Dinakaran, J. R. Mackey, and P. H. Reza, "Label-free virtual Hematoxylin and Eosin (H&E) staining using second generation Photoacoustic Remote Sensing (PARS)," p. 38.

[6] M. Boktor, B. R. Ecclestone, V. Pekar, D. Dinakaran, J. R. Mackey, P. Fieguth, and P. Haji Reza, "Virtual histological staining of label-free total absorption photoacoustic remote sensing (TA-PARS)," *Scientific Reports*, vol. 12, no. 1, p. 10296, Jun. 2022. [Online]. Available: https://www.nature.com/articles/s41598-022-14042-y

[7] M. Boktor, J. E. Tweel, B. R. Ecclestone, J. A. Ye, P. Fieguth, and P. Haji Reza, "Multi-channel feature extraction for virtual histological staining of photon absorption remote sensing images," *Scientific Reports*, vol. 14, no. 1, p. 2009, 2024.

[8] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2242–2251, iSSN: 2380-7504.

[9] J. E. D. Tweel, B. R. Ecclestone, M. Boktor, J. A. T. Simmons, P. Fieguth, and P. H. Reza, "Virtual Histology with Photon Absorption Remote Sensing using a Cycle-Consistent Generative Adversarial Network with Weakly Registered Pairs," 2023, publisher: arXiv Version Number: 1. [Online]. Available: https://arxiv.org/abs/2306.08583

[10] J. E. Tweel, B. R. Ecclestone, M. Boktor, D. Dinakaran, J. R. Mackey, and P. H. Reza, "Automated whole slide imaging for label-free histology using photon absorption remote sensing microscopy," *IEEE Transactions on Biomedical Engineering*, vol. 71, no. 6, pp. 1901–1912, 2024.

[11] N. Pellegrino, B. R. Ecclestone, D. Dinakaran, F. Van Landeghem, P. Fieguth, and P. Haji Reza, "Time-domain feature extraction for target specificity in photoacoustic remote sensing microscopy," *Optics Letters*, vol. 47, no. 15, p. 3952, Aug. 2022. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=ol-47-15-3952

[12] N. Pellegrino, P. W. Fieguth, and P. Haji Reza, "K-Means for noise-insensitive multi-dimensional feature learning," *Pattern Recognition Letters*, vol. 170, pp. 113–120, Jun. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167865523001150

[13] J. Shlens, "A tutorial on principal component analysis," 2014.

[14] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology*, vol. 24, no. 6, p. 417–441, 1933.

[15] A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal Quant Cytol Histol*, vol. 23, Jan. 2001.

[16] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "scikit-image: image processing in python," *PeerJ*, vol. 2, p. e453, 2014.

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *arXiv:1406.2661 [cs, stat]*, Jun. 2014, arXiv: 1406.2661. [Online]. Available: http://arxiv.org/abs/1406.2661

[18] B. Bai, X. Yang, Y. Li, Y. Zhang, N. Pillar, and A. Ozcan, "Deep learning-enabled virtual histological staining of biological samples," *Light: Science & Applications*, vol. 12, no. 1, p. 57, Mar. 2023. [Online]. Available: https://www.nature.com/articles/s41377-023-01104-7

[19] Y. Liang, D. Lee, Y. Li, and B.-S. Shin, "Unpaired medical image colorization using generative adversarial network," *Multimedia Tools and Applications*, Jan. 2021. [Online]. Available: https://doi.org/10.1007/s11042-020-10468-6

[20] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004. [Online]. Available: http://ieeexplore.ieee.org/document/1284395/

[21] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 193–201, 2009.