U-Net in Histological Segmentation: Comparison of the Effect of Using Different Color Spaces and Final Activation Functions

László Körmöczi[a] and László G. Nyúl[b]
Department of Image Processing and Computer Graphics, University of Szeged, Szeged, Hungary
{kormoczi, nyul}@inf.u-szeged.hu

Keywords: Semantic Segmentation, U-Net, Machine Learning, Deep Learning, Color Space Conversion.

Abstract: Deep neural networks became widespread in numerous fields of image processing, including semantic segmentation. U-Net is a popular choice for semantic segmentation of microscopy images, e.g. histological sections. In this paper, we compare the performance of a U-Net architecture in three different color spaces: the commonly used, perceptually uniform sRGB, the perceptually uniform but device-independent CIE $L^*a^*b^*$, and linear RGB color space that is uniform in terms of light intensity. Furthermore, we investigate the network's performance on data combinations that were unseen during training.

1 INTRODUCTION

Semantic segmentation of images is a key task for numerous applications, including quantitative analysis of histological sections (Chang et al., 2017; Iizuka et al., 2020; Ahmed et al., 2022). With widespread adoption of deep neural networks, this task is mostly a matter of quantity and quality of the training data. However, in several fields, such training data are limited, either in terms of quality or quantity. By the nature of the selected task, the training dataset can be highly imbalanced (e.g. different tissue types in a histological section) or can differ from the data the model is evaluated. Among other deep learning architectures, U-Net (Ronneberger et al., 2015) became widely adopted in the field of semantic segmentation, especially for biomedical images. It is a fully convolutional network (first used for segmentation by (Shelhamer et al., 2017)) with skip connections (introduced in ResNet (He et al., 2016)). U-Net is popular for semantic segmentation tasks due to its high generalization ability and acceptable speed.

Images used for deep learning data (either for training or inference) are usually stored in popular image formats, like PNG that uses lossless compression, JPEG that uses lossy compression, or TIFF, that can use either lossy or lossless compression. These formats usually use 1 (grayscale) or 3 (RGB) channels, 8 bits per channel (16 bits per channel is common for TIFF images for scientific purposes). These image formats use the standardized but device-dependent sRGB color space and store intensities with lon-linear gamma correction, to better fit for human perception. Gamma correction ensures that the intensity values are stored in a perceptually uniform way, i.e. efficiently for displaying to human viewers. Contrary, numerous image processing tasks (e.g. color blending or even resizing images) require the intensity values to be in a linear color space, i.e. uniform in terms of physical light intensity. Image manipulation software usually linearize the opened images for editing, and apply gamma correction upon saving.

Another common color space is CIE $L^*a^*b^*$ (also referred as CIELAB), that is perceptually based but device independent (contrary to sRGB). It may be better suitable for pattern recognition and semantic segmentation tasks than an RGB representation, because it separates the lightness information (L^* value) from the color information (a^* and b^* values).

An important question is whether a deep neural network can benefit from a perceptually uniform, device independent color space like CIE L^*a^*b, or a phisically uniform, linear RGB color representation.

[a] https://orcid.org/0000-0002-3833-0609
[b] https://orcid.org/0000-0002-3826-543X

Körmöczi, L. and Nyúl, L. G.
U-Net in Histological Segmentation: Comparison of the Effect of Using Different Color Spaces and Final Activation Functions.
DOI: 10.5220/0013323300003911

In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Volume 1*, pages 386-389
ISBN: 978-989-758-731-3; ISSN: 2184-4305

2 DATA

2.1 Images

Our dataset consists of 256×256 px 3-channel 8-bit RGB images, randomly cut from high-resolution bright-field optical microscopy images of pancreas histological sections. The data is classified pixel-wise into 3 distinct classes, i.e. background, healthy tissue and diseased tissue. The images containing diseased tissue are taken from histological sections with artificially induced acute pancreatitis, where the whole section is diseased. Similarly, images of healthy tissue are taken from histological sections where the whole tissue is healthy. Considering this, each sample contains either healthy or diseased tissue, along with background. Some of the samples can be seen in 1.

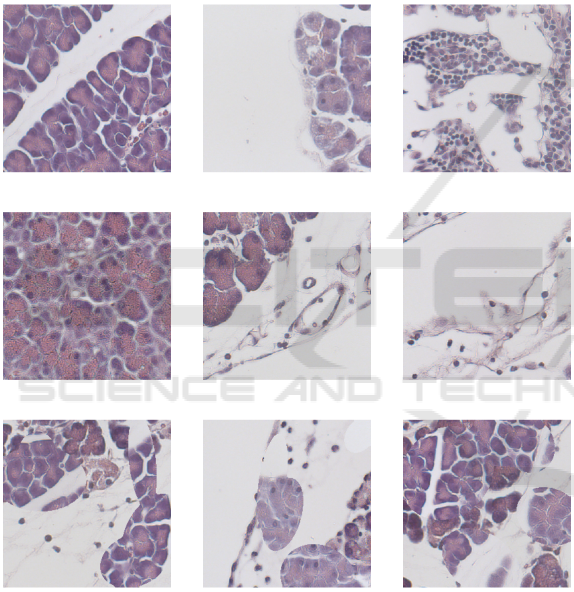

Figure 1: Sample image patches. First row: healthy, second row: diseased, third row: mixed samples.

2.2 Training, Validation and Test Dataset

We partition the dataset into distinct training, validation and test subsets. The trainig dataset contains 4500 images, with a total of 180 833 184 px of background, 25 999 369 px of normal (healthy) tissue and 88 079 441 px of diseased tissue. The validation dataset comprises 3016 images, with a total of 114 602 379 px of background, 20 262 843 px of normal (healthy) tissue and 62 791 354 px of diseased tissue. Finally, the test set includes 3787 images, consisting of 148 533 175 px of background, 22 332 040 px of normal (healthy) tissue and 77 319 617 px of diseased tissue.

2.3 Mixed Samples

Deep neural networks generally perform well on data similar to the training data. It is interesting though, how they perform on data variations that were not present during training. To evaluate the model, since there were no samples that contained both healthy and diseased pixels, we artificially generated mixed samples by combining samples taken from the healthy and diseased subset of the validation dataset. Mixing was performed using a randomly generated binary mask that contained 5 potentially overlapping ellipsoids of random size, orientation and location. The masks were blurred and then thresholded, to have a more natural appearance in the corners resulting from overlapping ellipsoids. The new image sample is generated by taking parts from a diseased sample where the mask is 0 and from a healthy sample where the mask is 1. These new images may contain both healthy and diseased tissue, along with background.

This new mixed dataset consists of 1497 images, with a total of 52 876 272 px of background, 7 895 774 px of normal (healthy) tissue and 37 335 346 px of diseased tissue.

3 ARCHITECTURE

Using U-Net for semantic segmentation is popular and well-documented (Du et al., 2020). We used a U-Net architecture of 5 encoder-decoder block pairs, with 1 px padding at the convolutional layers to keep original image dimensions at the output segmentation. Although softmax is the preferred activation function when having mutually exclusive classes in a semantic segmentation task, we also trained the model using sigmoid instead of softmax and examined the differences.

4 TRAINING THE MODEL

The images of our dataset were originally stored in 3-channel PNG image format, in sRGB color space, gamma-encoded. For comparison of the effect of the color space (i.e. gamma-encoded sRGB, linear RGB and CIE *L*a*b**), we trained and evaluated different models of the same architecture (described in Section 3), with both the training and validation dataset converted to the 3 different color spaces. Furthermore, to investigate the effect of changing the activation function of the output layer, we trained a model with softmax and another with sigmoid on all three color spaces, so as a result, we have 6 models to compare.

For training the models, we used Adam optimizer with an initial learning rate of 0.001 and a learning rate scheduler was utilized for better convergence. We used Categorical Cross-Entropy loss, weighted to compensate for the imbalanced dataset. The training was run for 160 epochs, with a batch size of 8. Training was performed on a desktop computer, having an NVIDIA RTX 3060 Ti GPU with 8GB VRAM.

5 RESULTS

5.1 Evaluation Metrics

We evaluated the 6 different models on both the original validation dataset and on the generated, artificially mixed images. For performance metrics, we used confusion matrix, per-class precision, recall, F1 score and micro-averaged F1 score, that equals to micro-averaged precision and micro-averaged recall. While per-class metrics are sensitive to class imbalance, micro-averaged F1 score is implicitly compensated for that effect and gives an impression on the overall performance of the model. This note is important, because one of the 3 classes (healthy tissue) has weak support in the mixed dataset (8 million pixels compared to 53 million pixels background and 37 million pixels diseased tissue), as detailed in Section 2.3.

For each trained model, regardless of the final activation function used in that model, we considered the class with the highest predicted probability for a pixel as the predicted class for that pixel.

The confusion matrix provides detailed information for each class, including the number of data samples (pixels, in this case) that were correctly predicted as that class (true positives, denoted as TP), as well as those misclassified as other classes. This includes false positives (samples incorrectly predicted as the given class, denoted as FP) and false negatives (samples from the given class incorrectly predicted as another class, denoted as FN).

Precision is the ratio of true positives and all predictions for a given class, while recall is the ratio of true positives and all pixels originally labeled as the given class. F1 score is calculated as

$$\mathrm{F1}_i = 2 \times \frac{\mathrm{precision}_i \times \mathrm{recall}_i}{\mathrm{precision}_i + \mathrm{recall}_i}$$

for the ith class.

Micro-averaged precision, recall and F1 scores take all samples into account, so the true positives are the diagonals of the confusion matrix, and all other values are false negatives and also false positives. Considering this, micro-averaged precision, recall and F1 score is equal for any confusion matrix, so we only display the F1 score from the micro-averaged metrics.

5.2 Results on Original Data

Table 1 shows the micro-averaged F1 score on the original validation dataset. Softmax activation and *L*a*b** color space model performs best, but all values are within 0.01 difference.

Table 1: Micro-averaged F1 score of the different models for the original validation data.

Class	sRGB	linear RGB	CIE L*a*b*
Sigmoid	0.9778	0.9739	0.9782
Softmax	0.9753	0.9713	0.9817

Table 2 shows the micro-averaged F1 score on the original test dataset. Softmax activation and *L*a*b** color space model performs best again, but all values are within 0.01 difference. The results match those on the validation data.

Table 2: Micro-averaged F1 score of the different models for the original test data.

Class	sRGB	linear RGB	CIE L*a*b*
Sigmoid	0.9804	0.9761	0.9803
Softmax	0.9781	0.9742	0.9830

5.3 Results on Mixed Samples

When dealing with previously unseen combinations of data (i.e. healthy and diseased tissue in the same sample), performance degrades compared to the previous results. Micro-averaged F1 score for the mixed dataset can be seen in Table 3. *L*a*b** performs best, and sRGB is superior to linear RGB. For *L*a*b** and linear RGB, softmax performs slightly better than sigmoid.

Table 3: Micro-averaged F1 score of the different models for the generated mixed data.

	sRGB	linear RGB	CIE L*a*b*
Sigmoid	0.7274	0.7078	0.7531
Softmax	0.7362	0.6971	0.7677

We aimed to investigate the performance of the models in more details, by focusing on areas around the artificial borders were the parts from healthy and diseased images meet. We selected parts based on which layer can be affected by the mixed data. When a pixel is far enough from the meeting parts, all layers see data similar to the training dataset. When we are getting closer to the meeting border, the deepest layer can see more of the mixed data. Going further, the mixed data can potentially affect more and more layers. In Table 4 we show the micro-averaged F1 score to the selected bands. Here, *BM* denotes the areas around the meeting border that can only affect the deepest, 5th (bottleneck) level. *L4* denotes the areas that can affect the bottleneck and one higher level, but where the higher levels are unaffected. *L1* denotes the areas where all layers can potentially see mixed data. We marked the areas where all layers see only original type of data as *out*.

Table 4: Micro-averaged F1 score for the different models on different parts of the mixed dataset, by affected layers.

		L*a*b*	RGB	sRGB
Sigmoid	out	0.8765	0.8476	0.8623
	BN	0.7487	0.6993	0.7304
	L4	0.6926	0.6406	0.6689
	L3	0.6532	0.6061	0.6170
	L2	0.6221	0.5634	0.5494
	L1	0.6073	0.5158	0.5076
Softmax	out	0.8776	0.8308	0.8608
	BN	0.7636	0.6904	0.7377
	L4	0.7154	0.6208	0.6823
	L3	0.6818	0.5995	0.6301
	L2	0.6482	0.5727	0.5749
	L1	0.6214	0.5556	0.5548

All models show similar performance at the outer regions, but going closer to the meeting border of healthy and diseased parts, we see performance degradation as more layers are affected by mixed data. The advantage of using CIE *L*a*b** color space is higher than on the original dataset. For this color space, using softmax is preferred over sigmoid as final activation function. Interestingly, for the sRGB and linear RGB color spaces, the models using sigmoid perform better for outer regions, but they lose this advantage near the mixed data.

6 CONCLUSIONS

In this paper, we compared the performance of U-Net models of the same architecture and structure, trained on the same datased but using different color spaces and output activation functions. We also investigated the performance on a dataset that differs significantly from the training data. Deeper examination of the results show how each layer affects the prediction.

Experimental results show that a perceptually uniform, device-independent color space, CIE *L*a*b**, that separates the lightness and color information, has advantage over the traditionnaly used, gamma-encoded sRGB color space and also over phisically uniform RGB representation that is used in image processing.

ACKNOWLEDGEMENTS

We thank Dr. József Maléth from Department of Medicine, Albert Szent-Györgyi Medical School, University of Szeged, Szeged, Hungary for generously providing the microscopy images as source data that was essential for this study.

This research was supported by project TKP2021-NVA-09. Project no TKP2021-NVA-09 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme.

REFERENCES

Ahmed, A. A., Abouzid, M., and Kaczmarek, E. (2022). Deep learning approaches in histopathology. *Cancers*, 14(21):5264.

Chang, Y. H., Thibault, G., Madin, O., Azimi, V., Meyers, C., Johnson, B., Link, J., Margolin, A., and Gray, J. W. (2017). Deep learning based nucleus classification in pancreas histological images. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 672–675.

Du, G., Cao, X., Liang, J., Chen, X., and Zhan, Y. (2020). Medical image segmentation based on u-net: A review. *Journal of Imaging Science and Technology*, 64(2):020508–1–020508–12.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Iizuka, O., Kanavati, F., Kato, K., Rambeau, M., Arihiro, K., and Tsuneki, M. (2020). Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific Reports*, 10(1).

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Shelhamer, E., Long, J., and Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651.