2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)
April 3-7, 2020, Iowa City, Iowa, USA

WEAKLY-SUPERVISED DEEP STAIN DECOMPOSITION FOR MULTIPLEX IHC IMAGES

Shahira Abousamra[1], *Danielle Fassler*[2], *Le Hou*[1], *Yuwei Zhang*[3], *Rajarsi Gupta*[3], *Tahsin Kurc*[3], *Luisa F. Escobar-Hoyos*[2], *Dimitris Samaras*[1], *Beatrice Knudson*[4], *Kenneth Shroyer*[2], *Joel Saltz*[3] *and Chao Chen*[3]

[1]Stony Brook University, Department of Computer Science, USA
[2]Stony Brook University, Department of Pathology, USA
[3]Stony Brook University, Department of Biomedical Informatics, USA
[4]Cedars Sinai Medical Center, USA

ABSTRACT

Multiplex immunohistochemistry (mIHC) is an innovative and cost-effective method that simultaneously labels multiple biomarkers in the same tissue section. Current platforms support labeling six or more cell types with different colored stains that can be visualized with brightfield light microscopy. However, analyzing and interpreting multi-colored images comprised of thousands of cells is a challenging task for both pathologists and current image analysis methods. We propose a novel deep learning based method that predicts the concentration of different stains at every pixel of a whole slide image (WSI). Our method incorporates weak annotations as training data: manually placed dots labelling different cell types based on color. We compare our method with other approaches and observe favorable performance on mIHC images.

***Index Terms*—** Color decomposition; machine learning; microscopic images

1. INTRODUCTION

Multiplex immunohistochemistry (mIHC) utilizes up to six colored chromogens to label distinct cell classes within a tissue section. Digital images of mIHC-stained tissue slides can be captured in a single step with traditional brightfield light microscopy (Fig. 1) [1, 2]. It thus provides a more accessible alternative to immunofluorescence (IF) to study the complex interactions between various immune cell types and tumor cells within the tumor microenvironment [1, 2]. However, it remains a formidable challenge to analyze mIHC-stained WSIs in order to evaluate population densities and spatial distributions when more than 5 stains are present.

Correct identification and localization of different stains has been used to characterize the tumor microenvironment. The major challenge has been to distinguish multiple different cell classes based on their stains within a single image. Cells of the same class present varying levels of a given biomarker (the entity to which the stain localizes); moreover a given cell may contain multiple biomarkers simultaneously. Thus, different stains and their combinations may have significantly overlapping color spectra.

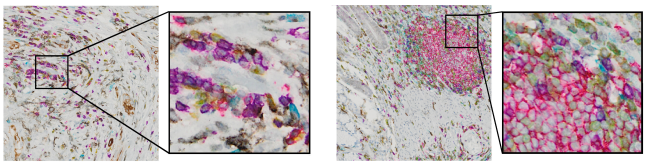

Fig. 1. Sample multiplex stained images.

Indeed, unsupervised stain unmixing has been studied in the literature. Color deconvolution, i.e., unmixing concentrations of different stains at each pixel, has been classically used to analyze images produced with different techniques, e.g., hematoxylin & eosin (H&E) and immunofluorescence (IF) images. Ruifrok and Johnston [3] recover stain concentrations by solving a linear equation system. However, when there are more than three stains, the linear equation system is underdetermined. To solve this ill-posed problem, Chen and Chefd'hotel [4] propose a lasso regression model to select a sparse set of stains at each pixel. The solution can be further improved using additional constraints based on prior information of co-occurring stains [5]. Good solutions have been developed for H&E staining [6, 7]. Duggal et al. [8] implicitly deconvolve dual-stained images for downstream classification tasks, but the deconvolution channels are not guaranteed to accurately correspond to different stains.

However, typical color deconvolution methods do not scale well with the number of simultaneous stains seen in mIHC images. Commercial tools such as HALO can unmix up to 4-colored chromogenic stains [9]. Thus arises the need for a tool that can scale with the increasing number of multiplex stains. Direct application of deep learning techniques is impractical due to lack of sufficient supervision; it is both tedious and time-consuming to manually label hundreds of thousands of cells at pixel-level. We need an innovative automatic method that leverages the intrinsic characteristics of multiplex images and requires minimal supervision.

978-1-5386-9330-8/20/$31.00 ©2020 IEEE

In this paper, we propose a novel deep autoencoder for color decomposition of multiplex images. We introduce a reconstruction loss to solve the ill-posed deconvolution problem. The autoencoder learns to predict concentrations of different stains across the image so their combination recovers the original image. Our method utilizes human annotations. However, instead of expensive per-pixel annotations, we only ask domain experts to mark dots at the approximate centers of different cell types. These weak annotations are used in a label consistency loss for the training to ensure the results are consistent with the dot label annotations. On an in-house dataset of multiplex images with 6 stains labeling 5 immune cell types, we demonstrate both qualitatively and quantitatively that our method outperforms existing ones.

2. METHOD

We propose an autoencoder which predicts the concentration of each stain at each pixel. Our multiplex images (Fig. 1) are of mIHC-stained pancreatic cancer tissue. Each image has various types of immune and tumor cells labelled by stains associated with cell class-specific biomarkers; they are: CD3 (yellow), CD4 (teal), CD8 (purple), CD16 (black), and CD20 (red). The tumor marker (brown) was not evaluated in this study as we were particularly interested in the immune cell infiltration. In addition, hematoxylin (blue) was employed to stain cell nuclei. See Fig. 2 for examples of different stains.

We use dot annotations as weak supervision. Our domain expert (a pathologist) marks a dot at the center of each cell (Fig. 3b). Each dot is assigned a single stain label (assumed to be the stain with the maximal concentration in this cell).

For a given image, our autoencoder jointly predicts the color concentration of all stains at every pixel. The predicted concentrations of the stains are used to derive a *stain segmentation* of the image, in which each pixel is associated with the stain with the maximal concentration. This stain segmentation is provided to domain expert for downstream analysis. See Fig. 4(A,E) for an example output concentration map and a stain segmentation, respectively.

The auto-encoder is trained on two loss terms, reconstruction loss and label consistency loss. The reconstruction loss ensures the predicted stain concentrations correctly restore the original image. The label consistency loss enforces the concentration prediction is consistent with human annotation.

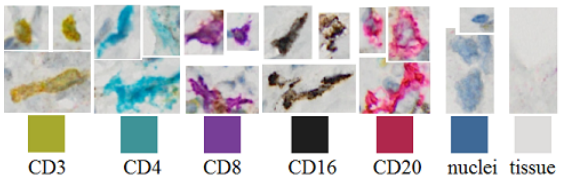

Fig. 2. Representative images of each cell class and the corresponding stain color used.

Together they ensure that the network does not overfit on the weak labels and learn the correct stain decomposition.

Reconstruction loss. To explain the reconstruction loss, we first explain how stain concentrations are related to the observed color of a pixel. Stain concentrations are linearly related to the *color optical density*, which represents the total loss of light when reflected by the material due to absorption, reflection and scattering. The optical density of the i^{th} pixel is the negative log of its normalized RGB colors:

$$y_i = [-\log(r_i/255), -\log(g_i/255), -\log(b_i/255)]^T.$$

Assume a given set of m stains, $S = [s_1, s_2, \cdots, s_m]^T$, in which s_i is a 3×1 vector representing the optical density of the i^{th} stain RGB color. In practice, we obtain these stain colors by sampling representative pixels of each stain. By Beer Lambert Law [10], the optical density and the stains satisfy a linear relationship, i.e., $y_i = S \times c_i$, in which c_i is a $m \times 1$ vector, denoting the concentrations of different stains at pixel i. Stacking the optical density of all n pixels of a given image, we have

$$Y = [y_1, \cdots, y_n] = S \times C \tag{1}$$

in which Y is a $3 \times n$ matrix and the concentration matrix $C = [c_1, \cdots, c_n]$ is of size $m \times n$.

Color decomposition is equivalent to solving the concentration matrix C given Y and S [3]. When S is 3×3 and full rank, C can be solved as $C = S^{-1} \times Y$. However, when we have more than three stains, the problem is underdetermined. There can be infinitely many solutions. We propose an autoencoder to predict C. We define the reconstruction loss as the mean squared error between the true color optical density, Y, and the reconstructed optical density based on the neural network prediction $\widehat{C}$, formally,

$$L_{recon} = \text{MSE}(Y, S \times \widehat{C}).$$

Label consistency loss. Reconstruction loss alone is insufficient; there can be infinitely many solutions satisfying $\mathcal{L}_{recon} = 0$. Our goal is to find the set of concentrations where the true marker at each pixel has the majority concentration. This enables us to predict the dominating stain at each pixel and to provide an interpretable stain segmentation.

To achieve this goal, we introduce a label consistency loss that leverages the dot annotation from domain experts. We first expand the dots to superpixels. We use SLIC [11] to partition the image into superpixels. Each superpixel that contains a dot annotation is assigned the same stain (Fig. 3c). All

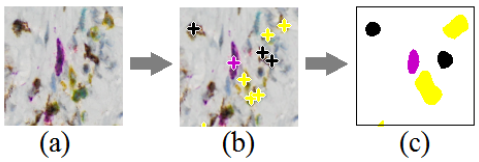

Fig. 3. Dot annotations. (a) Original patches. (b) Dot annotations (+). (c) Superpixel labels used in training.

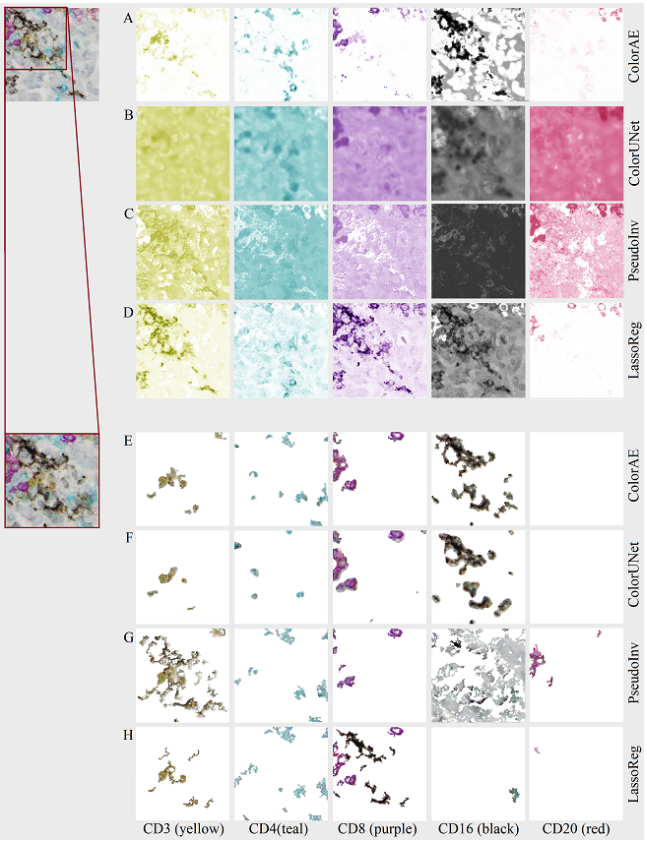

Fig. 4. Prediction results. A-D: stain concentration maps. E-H: stain segmentation on the magnified region.

remaining superpixels are assigned background stain. To get proper concentration maps for pixels belonging to the background tissue it is important to have in S a background stain vector that represents the plain tissue color, Fig. 2

Using this superpixel-based stain labeling, we now assign a single stain to each pixel, called its *label stain*. The intuition is that the label stain should have the majority concentration. In other words, using the label stain alone should be able to reconstruct the input image fairly well. To this end, we design the label consistency loss as the quality of another image reconstructed using only the label stain of each pixel and its corresponding concentration. Denote by ℓ_i the one hot vector of the label stain of pixel i, and let $L = [\ell_1, \cdots, \ell_n]$. The label consistency loss is

$$\mathcal{L}_{label} = \text{MSE}\left(Y, S \times (L \odot \widehat{C})\right),$$

in which " $\odot$ " is the Hadamard product

Our training loss is a weighted sum of the reconstruction loss and the label consistency loss. The weights are tuned empirically. The architecture of the network is shown in Fig. 5. The autoencoder starts with a 1×1 convolution that transforms the input into 128 channels. We find this first layer is important to capture the color information. The number of channels decreases through the encoder and increases back through the decoder. The final layer is a 1×1 convolution followed by a squaring operation to ensure the concentration map is all positive. We find the training converges better with squaring than with other alternatives such as a ReLU layer. The network is trained with dropout to avoid overfitting.

3. EXPERIMENTS AND RESULTS

We train our method and baselines on randomly extracted patches from the tumor area in 6 whole slide images of pancreatic cancer tissue. We test all models on patches from 4 different whole slide images. All patches have a resolution of 0.174 microns per pixel. For training, we sample 300 patches of size 400×400. For testing, we also randomly sample patches, but patch size varies depending on the evaluation strategy. The reference stain color vectors were obtained by averaging random samples from multiple locations (Fig. 2).

We compare our autoencoder (**ColorAE**) with several baselines: (1) a color decomposition UNet (**ColorUNet**): we train a UNet [12] using the same superpixel-based stain labels that were used for our label consistency loss. We train the network using weighted cross entropy loss and dropout at the end of each block in the concatenating path. We use the softmax output of UNet as the prediction of color concentration. (2) **PseudoInv**: we extend the method in [3] for multiplex stains by solving $C = S^{-1} \times Y$ using the pseudo inverse of S. (3) **LassoReg** [4]: we solve the color deconvolution problem with lasso regression. We empirically select the $L1$ norm weight $\lambda = 10^{-3}$. The regression is applied to every pixel in the input image.

Qualitative results. Fig. 4 shows sample predictions on a raw image. In the top block, we show the predicted concentration map for different stains. The darker a pixel is, the higher the concentration is. A pixel is white if its concentration is near zero. We show results of the four methods in four rows: ColorAE, ColorUNet, PseudoInv and LassoReg. The five columns correspond to the five different stains. Recall for ColorUNet we use the softmax layer output as the concentration map. In the bottom block, we show the derived stain segmentation, i.e., each pixel takes the stain with the maximal concentration. A pixel is white if it does not take the stain. This stain segmentation can be delivered to domain expert for downstream analysis.

Our method (ColorAE) produces the best results in both

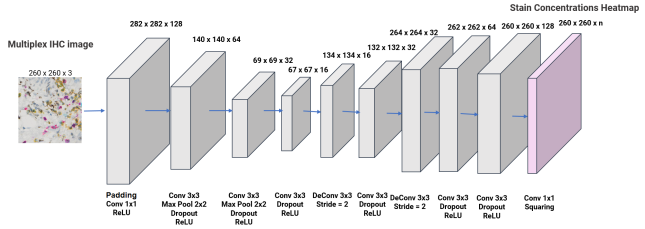

Fig. 5. Network architecture for an input size 260×260.

Methods	CD3	CD4	CD8	CD16	CD20
ColorAE	**0.62**	**0.61**	**0.63**	**0.67**	0.22
ColorUNet	0.50	0.52	0.57	0.64	0.19
PseudoInv	0.17	0.29	0.43	0.25	**0.23**
LassoReg	0.34	0.27	0.28	0.01	0.18

Table 1. F-score of stain segmentation on dots annotations.

Metric	CD3	CD4	CD8	CD16	CD20
SSim (Concentration)	0.9	0.87	0.90	0.54	0.87
Dice (Stain Seg.)	0.87	0.81	0.86	0.59	0.34

Table 2. Evaluation of ColorAE compared with automatic annotation. First row: SSim of the concentration maps. Second row: DICE of the stain segmentation.

concentration map and stain segmentation. ColorUNet predicts poor concentration maps; they tend to get uniformly nonzero, except for a few locations (in which some stains take very high concentration). The main underlying issue is that the model is trained on cross-entropy loss, which only ensures the stain with the highest probability to be consistent with the superpixel-based stain labeling. The prediction of other stains are not penalized even if they are all nonzero. On the other hand, the stain segmentation of ColorUNet is good, except for a few small stain components (in yellow and teal).

A common issue of both PseudoInv and LassoReg is that different stains tend to share concentration equally on the same pixel. Furthermore, black is often confused with other stains. For PseudoInv, Purple and red are mixed. The sparsity constraint in LassoReg leads to sparser and higher confidence. But some colors are still mixed (e.g., in yellow and purple). In terms of stain segmentation results, LassoReg is significantly better than PseudoInv, but is still worse than ColorAE and ColorUNet.

Quantitative evaluation. We use two evaluation strategies. First, we use **dot labels** (Fig. 3b) to evaluate the results. This is the most scalable way to obtain annotations. These dots are weak labels as they do not provide exact stain area. Second, we propose a strategy to evaluate the predicted stain concentration maps. It is impossible to obtain human labels for the continuous-valued concentration maps. Instead, we resort to restricted single stain slides on which existing automatic color decomposition methods can produce reliable color concentrations. Using these **automatic concentration labels**, we can evaluate the quality of our concentration map.

Table 1 reports the F-score of the predicted stain segmentation evaluated on dot labels. We test on 19 patches of size 1200×1920 from the 4 test slides. For each stain we define the true/false positive as the number of connected components of stain segmentation that are overlapping/disjoint with the dot labels. The false negative is the number of dot labels that do not overlap with a component of the stain segmentation.

Previous evaluation focuses on stain segmentations. To evaluate the concentration quality, we use two stacks of whole slide images. Each slide in a stack is stained with only one chromogen plus blue counterstain (hematoxylin). On these slides, we use automatic decomposition results of state-of-the-art method [6]. Note this method only works on 3-colored images. We consider these automatic results reliable and evaluate our results against them. See Fig. 6 for the original image, the automatic results generated for comparison, and our results. Our results generally look similar to the automatic labels. The only exception is the black stain. Our method tends to mix it with the background (gray).

To quantitatively compare our concentration map and the automatic annotation for each stain, we use the mean structural similarity index (SSim) on 20 randomly selected patches of size 1440×1440. SSim [13] is designed to measure the quality of an image prediction compared to groundtruth by taking into account similarity in luminance, contrast, and structure over sliding windows. This makes it more robust than traditional image quality measurement methods such as peak signal to noise ratio (PSNR) and mean squared error (MSE). We also compare the derived stain segmentation with the automatic results, i.e., stain segmentation derived from the automatic concentration maps. We compare the stain segmentations using dice score. Results presented in Table 2 confirm that our method is consistent with the automatic annotation in general. The noticeable drop in CD16 (black) is consistent with what we observe in Fig. 6. CD20 (red) have a generally very low distribution and so small mistakes are heavily penalized by the dice score as observed in Table 2

4. CONCLUSION

We have proposed a deep autoencoder for multiplex stain decomposition using weak labels. We use a reconstruction loss and a label consistency loss to train our network. The loss function and the weak labels allow our method to scale to a larger number of stains and achieve favorable performance.

Acknowledgement. This research was partially supported by NIH U24CA180924, the National Pancreas Foundation and the Pancreatic Cancer Action Network (PanCAN) 18-65-SHRO.

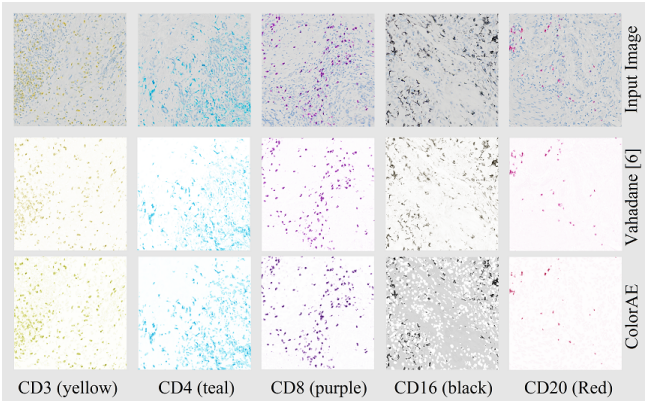

Fig. 6. Concentration maps on single-stained slides.

5. REFERENCES

[1] P. Hofman, C. Badoual, F. Henderson, L. Berland, M. Hamila, E. Long-Mira, S. Lassalle, H. Roussel, V. Hofman, E. Tartour, and M. Ilie, "Multiplexed Immunohistochemistry for Molecular and Immune Profiling in Lung Cancer-Just About Ready for Prime-Time?," *Cancers (Basel)*, vol. 11, no. 3, Feb 2019.

[2] A. Dixon, C. Bathany, M. Tsuei, J. White, K. Barald, and S. Takayama, "Recent developments in multiplexing techniques for immunohistochemistry," *Expert Rev. Mol. Diagn.*, vol. 15, no. 9, pp. 1171–1186, 2015.

[3] A. Ruifrok and D. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal Quant Cytol Histol.*, vol. 23, 01 2001.

[4] T. Chen and C. Chefd'hotel, "Deep learning based automatic immune cell detection for immunohistochemistry images," in *Machine Learning in Medical Imaging*, 2014.

[5] T. Chen and C. Srinivas, "Group sparsity model for stain unmixing in brightfield multiplex immunohistochemistry images," *Computerized Medical Imaging and Graphics*, vol. 46, pp. 30 – 39, 2015.

[6] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 8, pp. 1962–1971, Aug 2016.

[7] M. Macenko, M. Niethammer, J. Marron, D. Borland, J. Woosley, X. Guan, C. Schmitt, and N. Thomas, "A method for normalizing histology slides for quantitative analysis," in *ISBI*, 2009.

[8] R. Duggal, A. Gupta, R. Gupta, and P. Mallick, "SD-layer: Stain deconvolutional layer for cnns in medical microscopic imaging," in *MICCAI*, 2017.

[9] D. S. Thommen, V. H. Koelzer, P. Herzig, A. Roller, M. Trefny, S. Dimeloe, A. Kiialainen, J. Hanhart, C. Schill, C. Hess, S. Savic Prince, M. Wiese, D. Lardinois, P. C. Ho, C. Klein, V. Karanikas, K. D. Mertz, T. N. Schumacher, and A. Zippelius, "A transcriptionally and functionally distinct PD-1+ CD8+ T cell pool with predictive potential in non-small-cell lung cancer treated with PD-1 blockade," *Nat. Med.*, vol. 24, no. 7, pp. 994–1004, 07 2018.

[10] JH Lambert, "Photometria sive de mensura et gradibus luminis colorum et umbrae augsburg," *Detleffsen for the widow of Eberhard Klett*, 1760.

[11] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, Nov 2012.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[13] 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, 2004.