Evaluating virtual staining for high-throughput screening

Tonks, Samuel; Hsu, Chih; Hood, Steve; Musso, Ryan; Hopely, Ceriden; Doan, Minh; Edwards, Erin; Krull, Alexander; Styles, Iain

DOI:
10.1109/ISBI53787.2023.10230501

Evaluating virtual staining for high-throughput screening

[1]*Samuel Tonks* [4]*Chih-Yang Hsu* [2]*Steve Hood* [4]*Ryan Musso* [4]*Ceridwen Hopely*
[4]*Steve Titus* [1,5]*Alexander Krull*⋆ [4]*Minh Doan*⋆ [1,5,6]*Iain Styles*⋆
[1]School of Computer Science, University of Birmingham, Birmingham, United Kingdom
[3]GSK Bioimaging, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, UK.
[4]GSK Bioimaging, 1250 S Collegeville Rd, Collegeville, PA 19426, United States
[5]Centre of Membrane Proteins and Receptors (COMPARE), Universities of Birmingham and Nottingham, Midlands, UK
[6]The Alan Turing Institute, London, UK.

ABSTRACT

Little is known about the feasibility of virtual staining for industry applications such as high-throughput screening (HTS). We provide a thorough analysis of the usability of image-to-image translation for the virtual staining of label-free bright-field microscopy images of live cells, using a pool of more than 1.6 million images across six lung, six ovarian and six breast cell lines consisting of paired bright-field, cytoplasm, nuclei and DNA-damage stains. To our knowledge this is the first time an analysis of virtual staining has been performed on three levels; pixel-based, biological-feature based, and determining if virtual staining can reproduce drug-effect. Our results reveal that while virtually stained nuclei and cytoplasm images often consistently and faithfully reproduce the information found in fluorescence microscopy, virtually stained images of DNA-damage are usually less accurate.

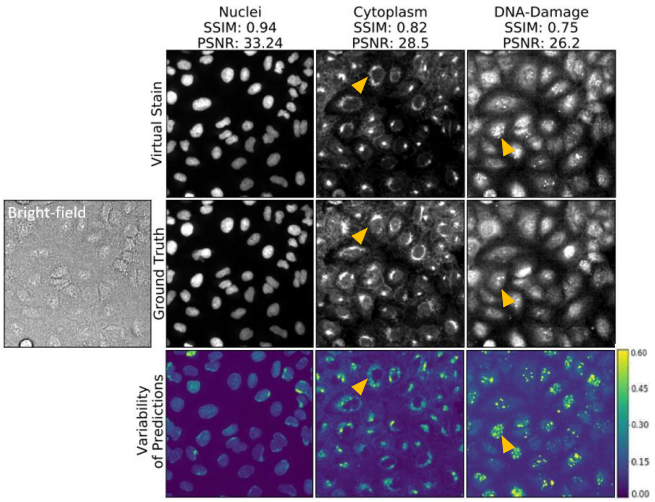

Fig. 1: **Ground truth fluorescence vs. virtual stains for a high cell density region of interest (ROI) in a lung cell line.** SSIM and PSNR for the ROI as well as prediction variability maps (VMAP), computed as the pixel-wise standard-deviation of multiple predictions generated from the same bright-field input using dropout [1]. Each VMAP is normalised by the standard-deviation of the ground truth fluorescence image. The general shape of nuclei and cells are reproduced well. Despite virtual staining appearing plausible, certain biological features; DNA-damage spots and cytoplasm intensity, show clear differences and subsequent high prediction variability. Examples shown at yellow arrow heads.

1. INTRODUCTION

High-throughput screening (HTS) is a process used in drug discovery that enables a large number of compounds to be tested simultaneously for drug effects on cell cultures. Fluorescence microscopy is the standard tool in HTS for detecting drug effects on cellular structures [2]. By covalently binding different fluorescent dyes to biomolecules (fluorescent staining), it enables biological structures to be simultaneously revealed in different parts of the optical spectrum, with each fluorescent dye captured in a separate image channel. However, fluorescent staining requires expensive, specialist machinery, the process is time-consuming and the number of parallel fluorescent dyes is restricted by spectrum saturation [3].

If we could extract the same information from unstained samples, then we could drastically reduce the resources required to image drug effects and enable new multiplex imaging combinations. Our approach to designing such a method is to explore image-to-image translation (I2I) [4] – a machine learning (ML) technique – to translate unstained bright-field microscopy images into multiple fluorescent microscopy channels, each corresponding to a structure of biological interest. Such a system could, in principle, generate virtually stained fluorescence images from fast, low cost, non-invasive, and widely available bright-field images. These could then potentially be used for HTS or other applications.

Virtual staining using I2I has been widely explored. Previous work [3, 5] used approaches based on a regression loss. This approach has been outperformed in other domain areas by other methods such as generative adversarial networks (GANs) [6, 7, 8], which have led to dramatic performance improvements on natural image datasets [9, 10] and other bioimaging applications [11]. Existing approaches to the evaluation of virtual staining have been focused on pixel-level discrepancies and have not yet determined the extent to which virtual staining accurately represents the biological information found in the fluorescence images. As a consequence, the true applicability of virtual staining for drug discovery applications such as HTS is still in doubt.

Here, we will for the first time, systematically investigate which biological features are consistently reproduced accurately in virtual staining and which are likely to be lost, opening the door to potential wide-spread future applications in drug discovery. To achieve robust quantitative results, we leverage a dataset containing more than 1.6 million individ-

ual high-resolution images from six lung, six ovarian and six breast cell lines consisting of paired bright-field and fluorescent channels showing cytoplasm, nuclei and DNA-damage.

We evaluate our virtually stained images using three levels of analysis. In level 1, we consider the established pixel-level metrics (SSIM [12], PSNR [13]) and examine the variability in virtual staining predictions. In level 2, we use the software Columbus [14] to compute 50 morphological, textural and intensity features that are frequently used in HTS. We compare the results between the real fluorescence channels and those obtained for different virtual staining. Finally, in level 3, we compute a common HTS assay quality metric, robust Z prime (RZ') factor [15] on the feature values obtained from real and virtually stained images.

We find that to determine the usability of I2I in practical applications, choosing the right quality metric is essential. Despite reasonably high performance for SSIM and PSNR across all channels, these scores do not give a good indication of whether or not the virtually stained channel could be used for HTS instead of the real fluorescence channel. Consistently the virtual nuclei and virtual cytoplasm accurately reproduce the biological feature values of the real fluorescent images and do not lead to substantial changes in RZ' factor. However, we find that the virtual DNA-damage channel, even though it is visually convincing, does not reflect the biological information contained within the real images. These results suggest that much, but not all information usually acquired by fluorescence staining can be extracted from the unstained bright-field and while virtual staining can not completely replace fluorescence microscopy, it can be a viable supplement.

2. EXPERIMENTS & RESULTS

Dataset: Our experiments are based on a pool of 1,631,232 individual images generated as part of a GSK HTS assay. The data comprises 118 16x24 well plates each of which contains negative and positive controls along with 10 GSK compounds screened for toxicity with one of three cell types; lung, ovarian and breast. Every well consists of 9 fields of view each containing a bright-field and three co-registered fluorescent stains; fluorescein (FITC) for cytoplasm, 6-diamidino-2-phenylindole (DAPI) for nuclei detection and Cyanine (Cy5) for DNA-damage. Each cell type was represented by six different cell lines.

Training: For each cell type, three models were independently trained to translate from bright-field to fluorescent stain. For each cell type 30,000 bright-field and fluorescence stain image pairs were sampled and split 70% for train, 20% for validation and 10% for test. In addition, for each cell type, a single plate, excluded from any training or validation sets, was used as a test set for levels 2 and 3 of the analysis.

This work used the Pix2PixHD [7] architecture adapted for our image resolution via rescaling. We use the same hyper-parameters as [7], after tuning we found 3 discriminators instead of the default 2 and a batch size of 16 achieved the best performance on the validation set. Each model was trained for a maximum of 100 epochs using early stopping.

Level 1 - Visual Inspection & Pixel-level Performance: Upon visual inspection (Figure 1), we find the general shape of the nuclei and cytoplasm are represented well in the virtual stains. However, aspects such as the position of DNA-damage spots or the intensity of the cytoplasm are visually plausible but incorrect. We believe the required information is simply not contained in the bright-field channel. One way to determine the level of confidence a GAN has in its prediction is to sample multiple solutions via dropout [16] during inference. We used this technique to produce 400 solutions on which the pixel-wise standard-deviation was calculated to visualise prediction variability, i.e., lack of confidence. We find (Figure 1) the highest variability in the DNA-Damage spots and areas of high intensity in the cytoplasm channel suggesting high levels of uncertainty.

Our first step to quantitatively evaluate our virtual staining is to use pixel-level metrics. In contrast to previous virtual staining work [3, 5] that evaluated its predictions using the Pearson correlation between the real and virtual pixel intensities, we used two well-established image quality metrics to compare our virtually stained images to the real fluorescence stains. Peak signal-to-noise ratio (PSNR) [13] was selected as it is sensitive to shift and scale of pixel intensities and structural similarity index measure (SSIM) [12] because it is invariant in that respect.

Across all cell types, virtual nuclei stain and virtual cytoplasm achieved very similar average SSIM scores of 0.96 ± 0.02 and 0.95 ± 0.02 while virtual DNA-damage scored 0.87 ± 0.10. Similarly for PSNR; virtual nuclei and cytoplasm achieved on average 35.2 ± 1.6 and 35.1 ± 2.8 respectively while virtual DNA-damage achieved 34.9 ± 3.1. Compared to the results from image restoration [17] these numbers indicate good quality. However, microscopy images contain large areas of uniform background leading to overstated pixel-level quality. Both metrics dramatically drop when considering a smaller region of interest with less background (Figure 1).

These results show that virtual nuclei and virtual cytoplasm are reproduced more accurately compared to the virtual DNA-damage. Despite appearing visually plausible and achieving relatively high performance for SSIM and PSNR, we identify certain biological structures in this channel that are reproduced with low confidence and differ from the real fluorescence channels.

Level 2 - Biological Feature Representation: In HTS, the toxicity of compounds is often evaluated by computing quantitative morphological, textual and intensity features for each cell and comparing their statistics to positive and negative controls. For I2I to be useful for HTS it must faithfully translate biological features of interest. We utilised a Columbus [14] feature extraction pipeline, designed for the original

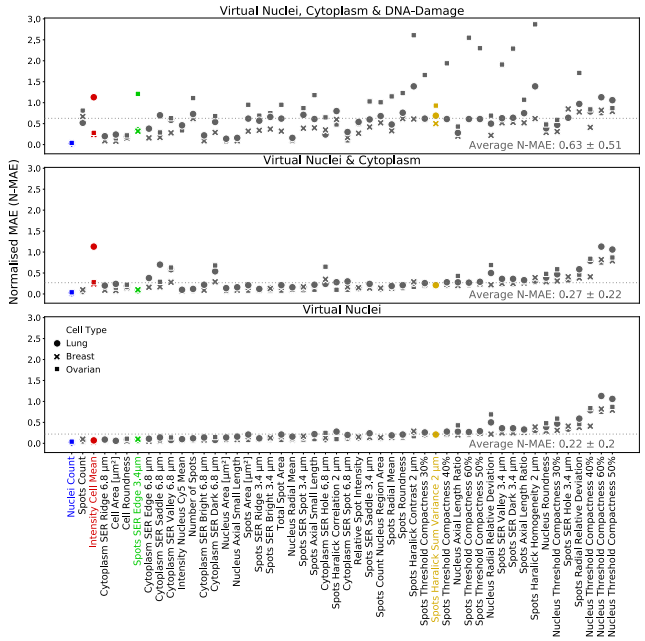

Fig. 2: **Normalised MAE (N-MAE) between fluorescence feature scores and each virtual staining stage for 50 biological features across all three cell types**. On average the virtual nuclei and virtual cytoplasm achieved lower N-MAE compared to virtual DNA-damage. One nuclei, one cytoplasm and two DNA-damage spot features are highlighted.

HTS assay to see if these feature's statistics are reproduced in our virtually stained images.

The Columbus pipeline takes a complete HTS plate as input. Here, we use the held out plate of each cell type that was not used for training or validation. The feature extraction begins with the segmentation of the nuclei, which then enables the segmentation of the cytoplasm and DNA-damage spots. Based on these segmentation outputs the scores for 50 features are then computed from their respective channel, see Figure 2. The feature score is computed per well by first quantifying the score across all cells per field of view and then averaging across all nine fields of view for all features.

To evaluate our virtual staining, we first generated feature scores from the real fluorescence channels and then introduced the virtual staining in 3 stages; stage 1 involved virtual nuclei only, stage 2 virtual nuclei and virtual cytoplasm and in stage 3 all virtually stained channels were used. For each cell type, feature and stage, the mean absolute error (MAE) between the real fluorescence and virtual staining feature scores was computed over all wells. To enable the comparison and interpretation of MAE across different feature ranges the resulting MAE for each feature was normalised (N-MAE) by the standard-deviation of the real fluorescence feature score.

Stage 1 with virtual nuclei only, achieved an average N-MAE score of 0.22 ± 0.2. Remarkably, a large proportion of features have produced stable N-MAE across the different cell types. As highlighted in Figure 2 the "Nuclei Count" feature from the nuclei channel, consistently scores the lowest N-MAE across all cell types, verifying the qualitative observations of the virtual nuclei channel in Figure 1. The largest errors are observed for the nuclei channel features associated with "Nucleus Threshold Compactness %", which measure nuclei shape at different intensity thresholds. This finding aligns with the uncertainty in shape shown for the nuclei channel in Figure 1.

In stage 2, as the virtual cytoplasm is introduced the average N-MAE rises to 0.27 ± 0.22. The increase in error is predominantly driven by features that rely on cytoplasm intensity or cell shape. Figure 2 shows an intensity-based feature, "Intensity Cell Mean" which, is not reproduced well for all cell types and in particular for lung. The problem of incorrect cell intensity was also visible in Figure 1.

Finally, in stage 3, when all virtually stained channels are used we observed a dramatic increase in error for many features extracted from the DNA-damage channel resulting in the average N-MAE increasing to 0.63 ± 0.51. In Figure 2, we highlight two examples of this; "Spots Haralick Sum Variance 2µm" and "Spots SER Edge 3.4µm" which are both measurements for the texture of DNA-damage spots.

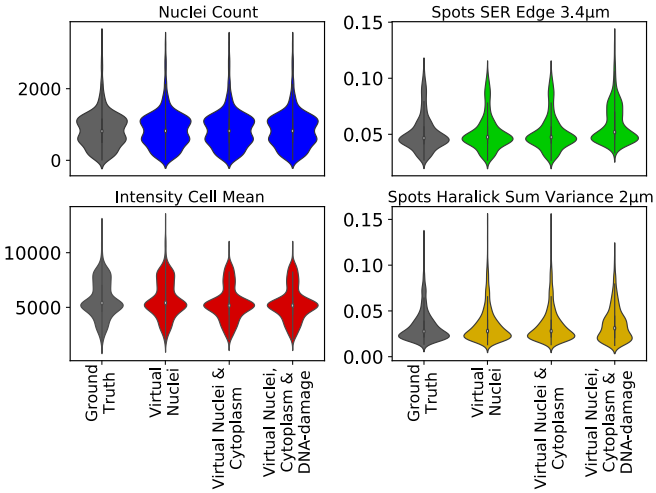

Fig. 3: **Distribution of feature scores highlighted for fluorescence and virtual staining stages**. The nuclei channel feature distributions appear very similar, the cytoplasm feature distributions are moderately different when the virtual cytoplasm is introduced meanwhile the two remaining DNA-damage spot features show considerable differences when virtual DNA-damage channel is used.

For each of the features highlighted in Figure 2, we show in Figure 3 the distribution of values at each virtual staining stage compared to the distribution of the real florescence values. For the nuclei channel feature, the virtual nuclei distribution is almost indistinguishable from the real fluorescence distribution and remains unchanged throughout. The distributions of feature scores for the cytoplasm and DNA-damage features are visually impacted when the corresponding vir-

tual channel is introduced suggesting that any measurements based on them are no longer reliable.

Level 3 - RZ' Factor: In HTS assay screening pipelines for cell damage, after biological features of each plate have been extracted, the plate quality is checked using the RZ' factor [15].

The method compares the feature distributions in negative and positive control sections of the plate. The RZ' factor is a value in the range of $(-\infty \leq RZ' \leq 1)$ that is computed for each feature. A value close to 1 indicates the feature distributions are very dissimilar and that this feature provides a potentially strong measurement for cell damage. Often an HTS study will focus on the features with the highest RZ' factors. A value close to 0 or negative suggests that the feature distributions are highly similar and therefore not suitable for measuring cell damage.

We compute the RZ' factors and select the ten features with the highest scores from the real fluorescence images to compare with the RZ' factors from all stages (see section 2) of virtual staining. For the breast plate analysed, we found all RZ' factors to be negative, indicating that they are not suitable for HTS and as such we exclude it from further analysis.

For the lung and ovarian plates the highest 10 RZ' factors all rely on the DNA-damage channel. Figure 4 shows remarkably small changes occur for many features when the virtual nuclei and virtual cytoplasm channels are used. However, the introduction of virtual DNA-damage leads to a substantial reduction in the majority of RZ' factors, indicating that they are no longer usable in the toxicity study. We again highlight "Spots Haralick Sum Variance 2µm" in Figure 4 as an example, where the drop in RZ' factor aligns with the increases in N-MAE shown in Figure 2 and the distorted distribution shown in Figure 3. We see the same pattern for the "Spots SER Edge 3.4µm" feature in the lung cell type. However, unexpectedly, we see an increased RZ' factor in the ovarian cell line as the virtual DNA-damage is introduced. One might naively interpret this an improvement introduced by virtual staining. However, considering the increased error for ovarian and the deformed distribution in Figure 3 we believe that this simply reflects inaccuracies in the translation that are systematically different in the positive and negative controls.

3. DISCUSSION AND CONCLUSION

Here, we have investigated I2I as a technique to extract information from the bright-field usually requiring fluorescence staining in HTS. Remarkably, across all three levels of our analysis both virtual nuclei and virtual cytoplasm have consistently performed well; achieving high pixel-level quality, producing similar scores for the majority of features found in the real fluorescence channels and leading to relatively small changes in RZ' factor outcomes compared to the fluorescence. As such, we believe that the use of virtual nuclei and virtual cytoplasm stains as a non-invasive imaging method could become a viable alternative to nuclei and cytoplasm

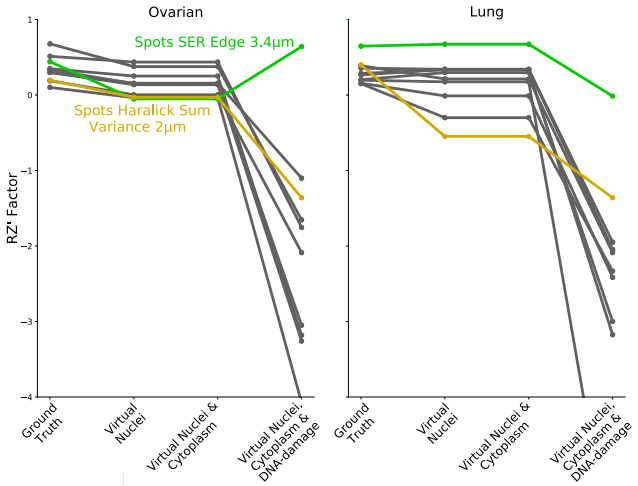

Fig. 4: **Highest 10 RZ' factor obtained with fluorescence and the change in RZ' factor with the progressive virtually staining stages for lung and ovarian plates.** For virtual nuclei and cytoplasm we observe small changes in RZ' factor. However, as the virtual DNA-damage is introduced the RZ' factor for the majority of features plummets indicating that these features are no longer usable for HTS.

fluorescence staining. These findings could lead to a reduction in the time needed to image these two labels and could enable new multiplex imaging combinations.

In contrast, the virtual DNA-damage channel has consistently shown relatively higher errors across all levels of analysis and cell types. In particular, the DNA-damage spots showed high variability and considerable losses of information for DNA-damage features such as "Spots SER Edge 3.4µm" and large changes in RZ' factor compared to the other two virtual channels. We believe it is not possible to predict the precise spot location because this information is not present in the bright-field image. Despite this, features such as "Spot Count" still achieved a low N-MAE for certain cell types as they are invariant to the precise spot location. One limitation of this component of our analysis is its inability to evaluate N-MAE per instance instead of averaging over all instances per well which would help improve the accuracy with which we evaluate virtual staining.

We saw that the use of the virtual DNA-damage channel could lead to an improved RZ' factor for individual features even when the error of the feature values had increased. This is unexpected because an increased RZ' factor indicates a potentially improved ability to distinguish between negative and positive controls which should be negatively correlated with increased errors of feature values. Ultimately, the positive correlation arises from incorrect I2I and its applicability should be subject to further scrutiny. We hope, future work will explore the ability of I2I models to generalise to unseen cell types and explore uncertainty-based approaches to improve virtual staining performance.

4. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required. The data used was originally generated for other GSK purposes and was re-used for this study.

5. ACKNOWLEDGMENTS

This work was funded by GSK and a studentship from the Engineering and Physical Sciences Research Council (project reference 2457518)

6. REFERENCES

[1] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[2] Jyrki Selinummi, Pekka Ruusuvuori, Irina Podolsky, Adrian Ozinsky, Elizabeth Gold, Olli Yli-Harja, Alan Aderem, and Ilya Shmulevich, "Bright field microscopy as an alternative to whole cell fluorescence in automated analysis of macrophage images," *PloS one*, vol. 4, no. 10, pp. e7497, 2009.

[3] Chawin Ounkomol, Sharmishtaa Seshamani, Mary M Maleckar, Forrest Collman, and Gregory R Johnson, "Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy," *Nature methods*, vol. 15, no. 11, pp. 917–920, 2018.

[4] Pang Yingxue, Lin Jianxin, Qin Tao, and Chen Zhibo, "Image-to-image translation: Methods and applications," *arXiv preprint arXiv:2101.08629*, 2021.

[5] Eric M Christiansen, Samuel J Yang, D Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison O'neil, Kevan Shah, Alicia K Lee, et al., "In silico labeling: predicting fluorescent labels in unlabeled images," *Cell*, vol. 173, no. 3, pp. 792–803, 2018.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[7] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[10] Radim Tyleček and Radim Šára, "Spatial pattern templates for recognition of objects with regular structure," in *German conference on pattern recognition*. Springer, 2013, pp. 364–374.

[11] Jacob C Reinhold, Yufan He, Shizhong Han, Yunqiang Chen, Dashan Gao, Junghoon Lee, Jerry L Prince, and Aaron Carass, "Validating uncertainty in medical image translation," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 95–98.

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[13] Osama S Faragallah, Heba El-Hoseny, Walid El-Shafai, Wael Abd El-Rahman, Hala S El-Sayed, El-Sayed M El-Rabaie, Fathi E Abd El-Samie, and Gamal GN Geweid, "A comprehensive survey analysis for present solutions of medical image fusion and future directions," *IEEE Access*, vol. 9, pp. 11358–11371, 2020.

[14] "Columbus image data storage and analysis system," https://www.perkinelmer.com/uk/product/image-data-storage-and-analysis-system-columbus, Accessed: 10/10/2022.

[15] Ji-Hu Zhang, Thomas DY Chung, and Kevin R Oldenburg, "A simple statistical parameter for use in evaluation and validation of high throughput screening assays," *Journal of biomolecular screening*, vol. 4, no. 2, pp. 67–73, 1999.

[16] Alex Kendall and Yarin Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila, "Noise2noise: Learning image restoration without clean data," *arXiv preprint arXiv:1803.04189*, 2018.