Enhanced Rotation-Equivariant U-Net for Nuclear Segmentation

Benjamin Chidester[1,*], That-Vinh Ton[2,3,*], Minh-Triet Tran[2], Jian Ma[1], Minh N. Do[3]
[1]Carnegie Mellon University [2]Univ. of Science, VNU-HCM
[3]Univ. of Illinois at Urbana-Champaign
{bchidest, jianma}@cs.cmu.edu ttvinh@selab.hcmus.edu.vn
minhdo@illinois.edu tmtriet@fit.hcmus.edu.vn

Abstract

Despite recent advances in deep learning, the crucial task of nuclear segmentation for computational pathology remains challenging. Recently, deep learning, and specifically U-Nets, have shown significant improvements for this task, but there is still room for improvement by further enhancing the design and training of U-Nets for nuclear segmentation. Specifically, we consider enforcing rotation equivariance in the network, the placement of residual blocks, and applying novel data augmentation designed specifically for histopathology images, and show the relative improvement and merit of each. Incorporating all of these enhancements in the design and training of a U-Net yields significantly improved segmentation results while still maintaining a speed of inference that is sufficient for real-world applications, in particular, analyzing whole-slide images (WSIs). Code for our enhanced U-Net is available at https://github.com/thatvinhton/G-U-Net.

1. Introduction

The recent surge in interest in deep learning coupled with increasing availability of large-scale histopathical image data sets, such as The Cancer Genome Atlas [17], has resulted in significant advances in computational histological analysis [13, 19]. A crucial step in such analysis pipelines is accurate and efficient segmentation of cell nuclei [5, 24]. With the aid of large-scale training data, deep-learning-based methods for automatically segmenting nuclei have surpassed traditional approaches, such as watershed [23] and thresholding [18] algorithms, though, despite these improvements, this step remains challenging and continues to be an active area of research [9, 26]. Changes in nuclear morphology are well-studied indicators of diseases, such as cancer, which motivates the continued development of effective methods of automated segmentation.

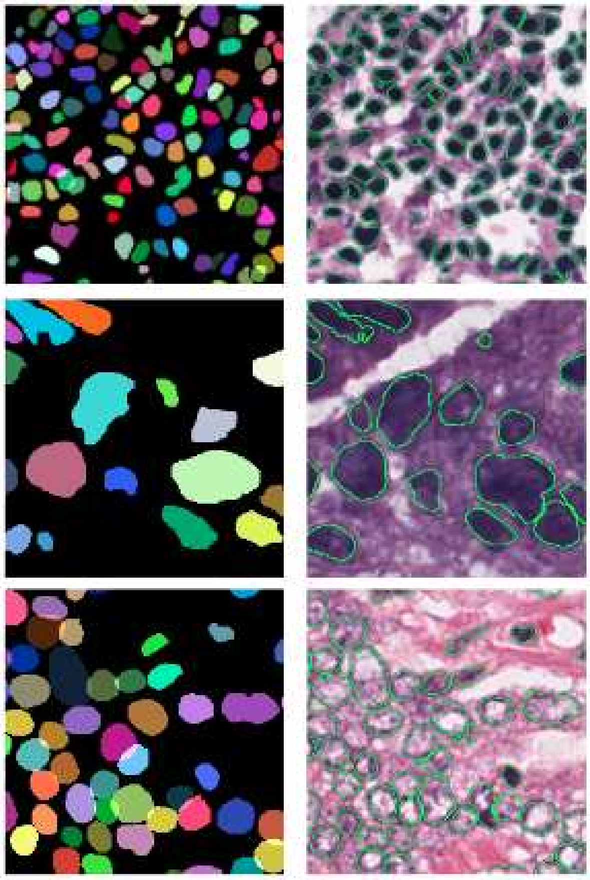


Figure 1: Segmentation results of our fully-enhanced U-Net on example test histopathology images. *Left.* Ground-truth annotations. *Right.* Our results. The estimated nuclear boundary is visualized in green. Patches are cropped for better visualization. Best view in color.

Initial approaches using deep learning operated by scanning the image patch-by-patch and generating a label (e.g., nucleus or non-nucleus) for each pixel centered in each patch [10, 25, 13]. Subsequent morphological processing

---

[*] Authors contributed equally.

could be applied to help smooth and refine the generated label masks and ensure that the segmented regions are contiguous; an example is the work of [25] that uses learned nuclear shape priors to encourage consistent inference. More recently, the U-Net architecture was proposed [19], which operates on the entire image and jointly infers the label at each pixel simultaneously, leading to more spatially coherent segmentation. U-Nets have been shown to achieve improved accuracy on several bioimage segmentation tasks, even when the data set is relatively small [19].

In this work, we bring together several recent developments in bioimage analysis to enhance the current methodology for deep-learning-based nuclear segmentation and, through a thorough ablation study, show the relative improvement and merit of each. We show that combining these enhancements together can achieve significantly improved scores for several important metrics at a speed that is sufficient for real-world applications, such as analyzing whole-slide images (WSIs).

The first enhancement to the U-Net we consider is to encode equivariance to groups, specifically rotation and translation, following the work of group-equivariant CNNs (G-CNNs) [3], thereby obviating the need to learn such equivariance through extensive and time-consuming data augmentation. This enhancement helps the learned network to better generalize to such variation in unseen data. G-CNNs have recently been shown to demonstrate top performance on segmenting regions of interest in histology images [22] and biological structures in other types of bioimages [1], but have not yet been applied to the task of nuclear segmentation.

Secondly, we enhance the long-skip connections in the U-Net from the downsampling to upsamling arms of the "U" with residual blocks, an insight that has shown improved performance for other applications [2]. The motivating hypothesis of this modification is that providing richer low-level features from the downsampling arm, learned through the residual blocks, to the upsampling arm will aid in producing detailed boundaries, especially between touching nuclei.

Lastly, we propose a novel means of data augmentation specifically designed for histological images to aid in training. Although the rate of generation of H&E image data is increasing, fully-labeled training data is still scarce. Therefore, augmentation of training data is still crucial for learning robust models. In addition to standard augmentation techniques of elastic deformations, blurring, and additive noise, we generate synthetic images by slightly translating and deforming the nuclei and filling any empty pixels by inpainting [27]. This method was inspired by recent work in video object segmentation [11], which demonstrated significant improvements in performance. Others works have also proposed means of generating synthetic, realistic histological images [16], though ours is much simpler to implement.

We demonstrate the merit of each of these considerations for designing and training U-Nets for nuclear segmentation on a data set of several histology images from TCGA samples for various types of tissue [13]. Several demonstrative results of our fully-enhanced U-Net are shown in Fig. 1. We use the aggregated Jaccard index (AJI), Dice coefficient, and F1-score as metrics for evaluation. We show that the combination of these improvements yields significant gains for segmentation.

2. Method

The standard U-Net architecture consists of two arms, one for downsampling the feature maps to a lower-dimensional space and one for upsampling the feature maps back to full resolution. Each downsampling layer consists of convolution, non-linear activation, pooling, and batch normalization. Each upsampling layer consists of similar operations except that pooling is replaced by upsampling. Additionally, residual blocks [7], which are used to increase the depth of the network, can be added. In [6], some experiments evaluated the performance of different types of residual blocks. We inherit from their work the architecture producing the best reported performance. Lastly, to help with interpolating the higher-resolution feature maps, features from the downsampling arm are conveyed to the upsampling layers by long-skip connections. From this baseline architecture, we incorporate two enhancements, namely group-equivariant operations to encode equivariance to rotation and translation, and residual blocks along the long-skip connections. We describe these enhancements below.

As in other deep-learning-based approaches [13], we formulate nuclear segmentation as a pixel labeling problem with three potential labels, namely, nuclear interior, nuclear boundary, and background. The network is designed and trained to produce a probability map for each label. Creating a label especially for the boundary results in a larger contribution to the objective for boundary pixels and thereby encourages the network to produce a more accurate boundary, which is generally harder to infer by post-processing than pixels away from nuclear boundaries. Our post-processing method of morphological operations, described below, helps to refine the output of the U-Net by smoothing edges and ensuring contiguous segmentation boundaries.

Input data is first stain-normalized before being fed to the U-Net. To help train the model, we employ several means of data augmentation, including our proposed histology-specific method, which we describe below.

2.1. Encoding group-equivariance

As noted in [3], it is helpful to think of the input image to a neural network as a function $f\colon \mathbb{Z}^2 \to \mathbb{R}^K$ that maps 2D

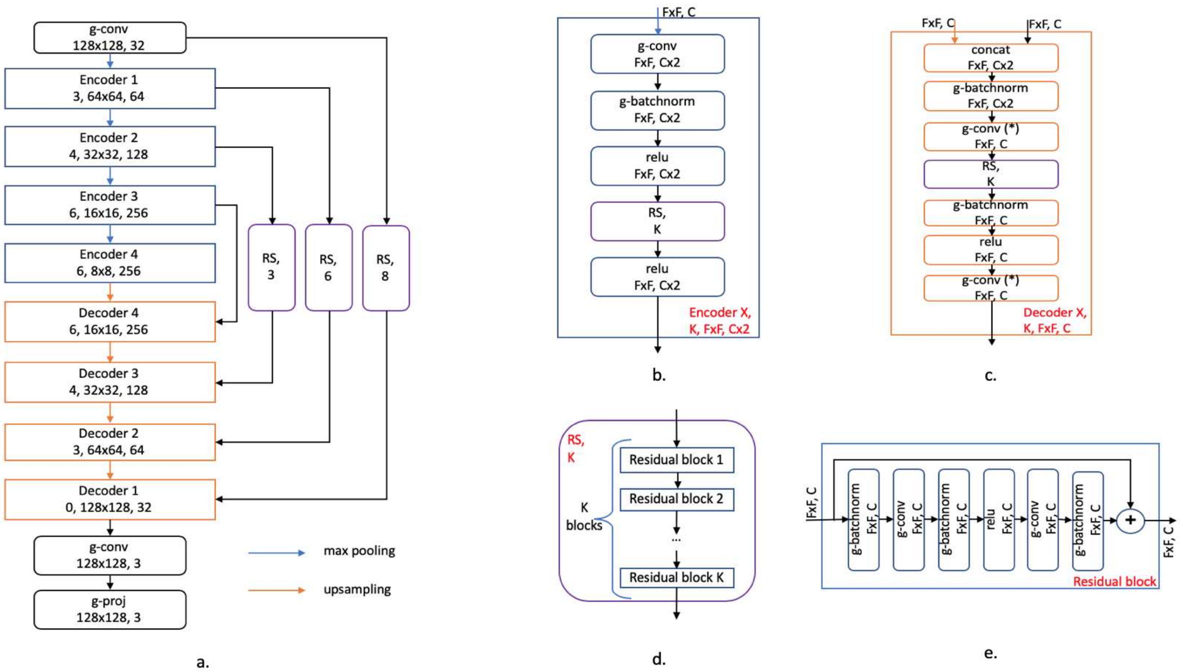


Figure 2: Our proposed rotation and translation-equivariant U-Net architecture. (a) Highest-level view, showing sequential encoder and decoder blocks. Our residual blocks along the long-skip connections of low-level layers are also shown. (b) Details of encoder blocks. (c) Details of decoder blocks. (d) Details of RS blocks, with several residual blocks in series, which are found on long-skip connections and within decoder and encoder blocks. (e) Details of residual blocks. (*) Because of limited space, we omit the batch normalization layers and ReLu activation layers after g-conv blocks.

space to pixel intensities. The insight of [3] for CNNs was to generalize equivariance to translation, which is inherent to standard convolution, to other transformations by defining convolution for *groups* in general, of which $\mathbb{Z}^2$ with translations is a specific example. An important group for histology images is the $p4$ group, which consists of translations and rotations about the origin by 90 degrees of elements in $\mathbb{Z}^2$. A group-equivariant neural network can be created by the composition of group-equivariant convolution with several other group operations, given below, which preserve equivariance to such transformations throughout the network. The placement of these operations in the network can be seen in Fig. 2.

Group-equivariant convolution

Group-equivariant convolution is the generalization of convolution to functions on groups, the set $\mathbb{Z}^2$ with translation, on which convolution is normally defined, being a specific type of group. For a group G, the convolution of a filter $\psi\colon G \to \mathbb{R}^K$ with a feature map $f\colon G \to \mathbb{R}^K$ is defined to be the sum, over all elements in G, of their inner product:

$$(f * \psi)(g) = \sum_k \sum_{h \in G} f_k(h)\psi_k(g^{-1}h). \tag{1}$$

Here, the action of element g on $h \in G$ is expressed by gh, and $g^{-1}h$ is the action of the inverse of g. For example, if the group is the translation of elements $x \in \mathbb{Z}^2$, then $gx = x + g$ and $g^{-1}x = x - g$ and we would have standard convolution. Since the output function is a function of G, which indexes not only pixel locations, but also rotations, this information can be preserved throughout the network thereby preserving equivariance to such transformations.

Since the input images to the network are functions on $\mathbb{Z}^2$, the output of the first group-equivariant convolutional layer is a special case, given by

$$(f * \psi)(g) = \sum_k \sum_{z \in \mathbb{Z}^2} f_k(z)\psi_k(g^{-1}z). \tag{2}$$

Group-equivariant upsampling

For the upsampling arm of the U-Net, before each layer, we first upsample the feature map from the layer below by 2 using nearest-neighbor interpolation, which preserves equivariance to translations and rotations of 90 degress. This method of upsampling is the same as *deconvolution* or *transpose* convolution with a 2×2 filter of all ones [15] and helps to keep the number of trainable filters in the network manageable.

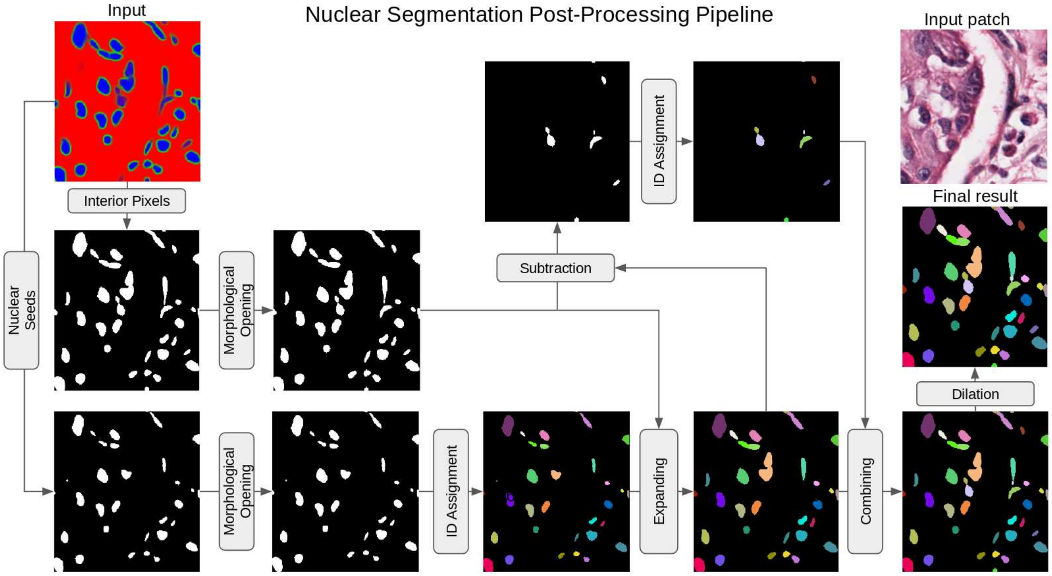


Figure 3: Our proposed post-processing method, consisting of a series of several morphological operations. Each step is visualized by its result.

Group-equivariant pooling and residual blocks

As noted in [3], for the $p4$ group, max-pooling with a stride of 2 preserves the equivariance properties of the network, and this is the pooling operation we use in the downsampling arm of the U-Net. Also noted in [3], since residual blocks are simply the addition of two group-equivariant feature maps, the output will also be group-equivariant.

Group projection

To create the final segmentation image, we must transform the domain of the feature maps in the U-Net from G back to $\mathbb{Z}^2$. To do so, we average the feature map for each filter over rotations. This is called the group *projection* layer, as in other works [14].

2.2. Long-skip connections with residual blocks

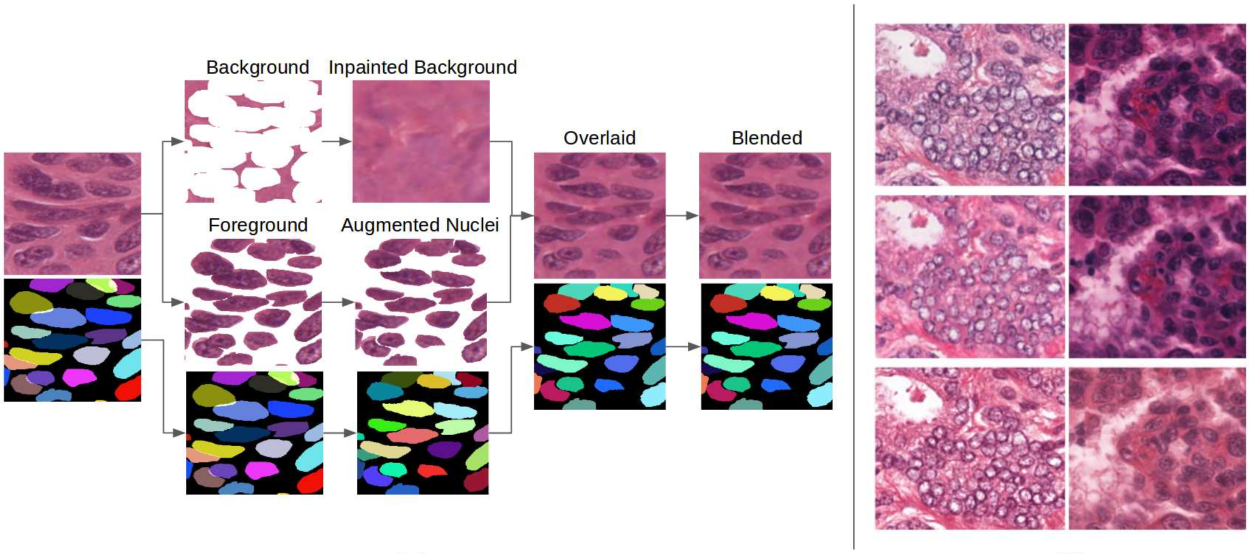
In the typical U-Net [19] architecture, the number of convolution blocks at low-level layers is small, which limits the effective local field-of-view and thereby decreases the quality of features which the network can use to delineate the output boundary. To provide richer low-level features to the final layers of the upsampling arm of the U-Net, we enhanced the baseline U-Net by adding residual blocks on long-skip connections. Our long-skip connections enhanced with residual blocks are visualized in Fig. 2a, and a detailed view of the long-skip connections and residual blocks are shown in Fig. 2d and Fig. 2e, respectively.

2.3. Morphological post-processing

Even though the U-Net deep learning architecture produces a full segmentation mask, unlike patch-based deep learning methods, post-processing, specifically morphological operations, are still essential to yielding contiguous regions and accurate nuclear boundaries. We designed a post-processing pipeline, shown in Fig. 3, to accomplish this. It consists of the following steps. A mask of confident interior pixels is created by identifying pixels for which the inside probability is greater than other labels, followed by morphological opening. A map of nuclear seeds is created by thresholding ($thres = 0.85$) the probability of the interior class of these pixels, followed by opening. Each seed is assigned a unique index, which is propagated to all connected interior pixels. Regions not covered will be assigned new index and combined with the previous result. Then, we apply binary dilation to create the final result. The parameters, such as window size, for these various morphological operations were optimized on the validation set in our experiments.

2.4. Data augmentation by nuclear deformation

To aid in training a robust network that minimizes overfitting, we implemented a novel approach designed specific for histology image analysis to generate augmented training images. Given the ground-truth annotations, each nucleus in an image is extracted and then deformed slightly by affine, spline, and elastic transformations as well as small random translation, yielding a new orientation, position,



(a)

(b)

Figure 4: (a) Our proposed process generating new synthetic images with their corresponding annotations. (b) Data augmentation and color normalization examples. The first row shows example training images from [13]. The second row shows synthetic images generated by our proposed data-augmentation method. The last row shows the original training images normalized by the method proposed in [21].

and shape. Extracting the nucleus from the image produces empty holes in the background, which are filled by the image inpainting method proposed in [27]. Some elastic transformations and slight Gaussian blurring are applied at this point, along with some added speckle noise. The transformed nuclei are then overlaid back onto the reconstructed background image. Finally Gaussian blurring is applied to help smooth any harsh or unnatural edges between the augmented nuclei and the background. Fig. 4 diagrams this process and shows examples of resulting augmented training images.

2.5. Stain normalization

To reduce the uninformative and possibly confusing color variation inherent to H&E stained images, we use the structure-preserving color normalization method proposed in [21]. One image from the training data set is used as the target and all other images, including the new augmented data, are converted to its color space. We use $\lambda = 0.1$ as recommended in [21]. The last row in Fig. 4b shows example normalized images.

3. Experiments and Results

3.1. Data set and evaluation metrics

We evaluated our enhanced U-Net architecture and proposed training data augmentation approach on the data set curated by [13], which consists of 30 histopathology images with accompanying full segmentation masks. The images are 1000×1000 pixels in size and were extracted from WSIs from unique TCGA samples, from a variety of different organs. To comprise the training set, four images were randomly selected from breast, liver, kidney and prostate samples. The remaining 14 images were from 7 different organs and were evenly split into validation and testing sets. To aid in training, for each image, we generate 25 new synthetic images and their corresponding masks by our proposed method, yielding a total of 750 new synthetic images.

For evaluation, we focused on the AJI metric, proposed in [13], which balances detection accuracy with the accuracy of the delineated boundaries of nuclei, though we also considered the F1-score & Dice coefficient to shed more light upon the performance of the various architectures.

3.2. Experimental setup

Our U-Net architectures operate on patches of size 128×128 (due to size limitations of the GPU), so to generate a set of training patches, we extracted random patches from each image during training. Before extracting the patch, the original image was rotated and scaled by random amounts. The network was trained to minimize the cross entropy loss plus the generalized Dice loss [20], which helps specifically to learn sharper boundaries by addressing the class imbalance problem of boundary pixels. We used the Python library NiftyNet [4] for an implementation of the Dice loss. We adopt the weight initialization proposed in [8]. In our experiments, we found that our enhanced U-Net works better without dropout layers, so we removed them, since they only increased the number of training steps required to converge. We used the Adam [12] optimizer for learning. The learning rate was set constant as $5e-5$ during the training process because of the adaptive property of Adam optimization [12]. Our batch size was restricted to be four, due to memory limitations on the GPU. We trained each network for roughly 300,000 steps. We used the validation set to fine-tune parameters.

During test-time inference, to generate a complete segmentation map, since the images are larger than 128×128 pixels, we perform inference on overlapping patches, with an overlap of 62 pixels, and then merge the results. We use the reflection transformation to pad patches on the boundaries of the original image.

3.3. Experimental results

Fig. 1 shows the results of our fully-enhanced U-Net, trained with augmented images, on several example histopathology images from the held-out test set, and the first row of Table 1 shows its performance according to the aforementioned metrics. As can be seen in Fig. 1, the boundary between overlapping nuclei can be reasonably separated by our method. Moreover, our method is able to generalize well to other types of organs, even those for which it was not trained. The first two lines in Fig. 1 are from stomach and colon tissue, respectively, which tissue types were not in the training set, yet our method still produces strong results. On the entire test set, our method achieved an AJI of 0.629. For comparison, the method proposed in [13] achieved an AJI of 0.508 on the same data set, although this comparison is not definitive, since 7 fewer images from the data set were used for training in their experiments.

Processing an entire image of size 1000×1000 on a single TITAN V 12GB GPU with our enhanced U-Net architecture takes only about 17 seconds. The subsequent post-processing incurs an additional 2 seconds on a CPU to create the final result.

3.4. Ablation study

We further evaluated the effect of each proposed component to the performance of our method through an ablation study. Example resulting segmentations from each of the following experiments are shown in Fig. 5 and the performance according to the aforementioned metrics are given in Table 1.

Residual blocks on long-skip connections

In this experiment, we removed all residual blocks on the long-skip connections of the U-Net architecture. To make a fair comparison, we compensated by adding more residual blocks in the encoder and decoder components to maintain the number of trainable parameters. All of other hyper-parameters were kept constant. The first and third line in Table 1 show the comparison between the two different models. Without residual blocks on the long-skip connection, the AJI value decreased by 1.4%, and the Dice score and F1-score decreased by 1.3% and 1.2%, respectively. This strengthens the credibility of our hypothesis that residual blocks on the long-skip connections help the network extract richer low-level features and thereby aid the network in delineating nuclear boundaries of touching nuclei. Fig. 5c and 5d further visualize the results of the two architectures. Some nuclei in Fig. 5d evidence a difficulty in separating their overlap, while our architecture with long-skip residual blocks successfully separates them.

Method	AJI	F1-score	Dice's coef.	#Params
Ours	**0.6291**	0.8469	**0.7980**	102M
Ours - D.A	0.6019	0.8006	0.7796	102M
Ours - R.S	0.6151	0.8349	0.7846	101M
Ours - G.E.	0.6125	**0.8490**	0.7893	101M

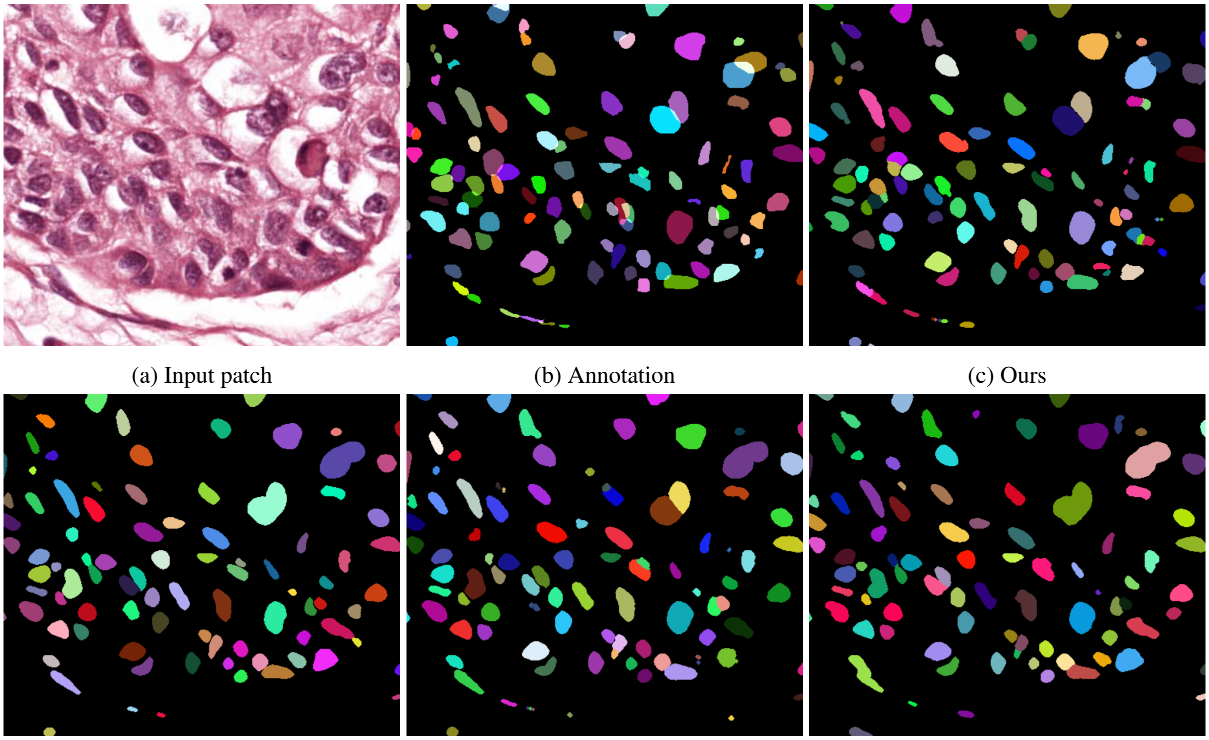
Table 1: Quantitative comparison of each proposed component of the U-Net architecture and training procedure. -D.A: without our proposed synthetic data; -R.S: without residual blocks on the long-skip connections; -G.E.: the U-Net without group-equivariant operations.

New synthetic data

To see the effect of the new synthetic data we proposed, we withheld these synthetic images from the training set and trained a separate network on this reduced training set. The architecture of the network and all of other hyper-parameters were kept the same. As can be seen in Table 1, training with these new synthetic images can improve the model's performance by approximately 2.7% in AJI, showing the most profound impact upon performance compared to the other two contributions. This suggests that this method of data augmentation further offsets the problem of sample scarcity for histopathological analysis, even beyond standard augmentation techniques. This is significant since labeling this type of data is extremely labor intensive.

Group equivariance

To see the effect of group-equivariant operations, we replaced these operations with standard operations for a CNN and retrained the network. Following the work in [3], to preserve the same number of trainable parameters, we doubled the number of filters in each convolution layer in the ordinary network, while keeping the same architecture oth-



(a) Input patch (b) Annotation (c) Ours

(d) Ours - R.S (e) Ours - D.A (f) Ours - G.E.

Figure 5: The visualization of example results of different methods in the ablation study. The segmented region for each distinct detected nucleus is shown by a unique color. The background is visualized in black. Patches are cropped for better visualization. Best view in color. -D.A: without our proposed synthetic data; -R.S: without residual blocks on the long-skip connections. -G.E.: the U-Net without group-equivariant operations.

erwise. This new model was trained with a similar set of hyper-parameters, except that we increased the learning rate to $1e-4$ and we were able to use a batch size of eight, since the network did not require as much storage on the GPU as the group-equivariant version. Since there were no group-equivariant operations, we added dropout to the model to help with regularization. As shown in Table 1, both the AJI and Dice's coefficient drop by a significant margin, roughly 1.7% and 0.9% respectively, without encoded group equivariance. This implies that integrating group-equivariant convolution and operations into the current U-Net architecture can indeed enable the network to learn better parameters that generalize well to simple transformations, namely, translations and rotations. Although, since it is only equivariant to rotations of 90 degrees, the network can still benefit from data augmentation of rotations of finer, arbitrary angles.

4. Conclusion

In this work, we have shown the value of several enhancements to the standard U-Net architecture, namely, encoding rotation and translation equivariance and adding additional residual blocks, and our novel data augmentation method for automated nuclear segmentation in histology images. This work can be considered as a parallel work to the many other current developments in deep-learning-based nuclear segmentation, which also could be incorporated to further improve performance. By contributing to improved performance for this crucial step for many computational pathology pipelines, without adding significant computational overhead, we believe these enhancements will help to enable future discoveries in pathology, with the hope of increasing the clinical impact of computational pathology.

References

[1] E. J. Bekkers, M. W. Lafarge, M. Veta, K. A. Eppenhof, J. P. Pluim, and R. Duits. Roto-translation covariant convolutional networks for medical image analysis. In *MICCAI*, pages 440–448. Springer, 2018.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016.

[3] T. Cohen and M. Welling. Group equivariant convolutional networks. In *ICML*, pages 2990–2999, 2016.

[4] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren. Niftynet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 2018.

[5] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147, 2009.

[6] D. Han, J. Kim, and J. Kim. Deep pyramidal residual networks. *CoRR*, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, 2015.

[9] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review – current status and future potential. *IEEE reviews in biomedical engineering*, 7:97–114, 2014.

[10] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

[11] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. *CoRR*, 2017.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, 2014.

[13] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, July 2017.

[14] J. Linmans, J. Winkens, B. S. Veeling, T. S. Cohen, and M. Welling. Sample efficient semantic segmentation using rotation equivariant convolutional networks. *arXiv:1807.00583*, 2018.

[15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[16] F. Mahmood, D. Borders, R. Chen, G. N. McKay, K. J. Salimian, A. Baras, and N. J. Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *arXiv:1810.00236*, 2018.

[17] T. C. G. A. Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.

[18] N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[19] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, 2015.

[20] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

[21] A. Vahadane, T. Peng, S. Albarqouni, M. Baust, K. Steiger, A. M. Schlitter, A. Sethi, I. Esposito, and N. Navab. Structure-preserved color normalization for histological images. In *ISBI*, pages 1012–1015, April 2015.

[22] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, pages 210–218. Springer, 2018.

[23] M. Veta, A. Huisman, M. A. Viergever, P. J. van Diest, and J. P. Pluim. Marker-controlled watershed segmentation of nuclei in h&e stained breast cancer biopsy images. In *ISBI*, pages 618–621. IEEE, 2011.

[24] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever. Breast cancer histopathology image analysis: A review. *IEEE transactions on biomedical engineering*, 61(5):1400–1411, 2014.

[25] F. Xing, Y. Xie, and L. Yang. An automatic learning-based framework for robust nucleus segmentation. *IEEE transactions on medical imaging*, 35(2):550–566, 2016.

[26] F. Xing and L. Yang. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE reviews in biomedical engineering*, 9:234–263, 2016.

[27] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. *arXiv:1801.07892*, 2018.