



Which Way Round? A Study on the Performance of Stain-Translation for Segmenting Arbitrarily Dyed Histological Images

Michael Gadermayr¹(✉), Vitus Appel¹, Barbara M. Klinkhammer²,
Peter Boor², and Dorit Merhof¹

¹ Institute of Imaging and Computer Vision, RWTH Aachen University,
Aachen, Germany

`Michael.Gadermayr@lfb.rwth-aachen.de`

² Institute Pathology, University Hospital Aachen, RWTH Aachen University,
Aachen, Germany

Abstract. Image-to-image translation based on convolutional neural networks recently gained popularity. Especially approaches relying on generative adversarial networks facilitating unpaired training open new opportunities for image analysis. Making use of an unpaired image-to-image translation approach, we propose a methodology to perform stain-independent segmentation of histological whole slide images requiring annotated training data for one single stain only. In this experimental study, we propose and investigate two different pipelines for performing stain-independent segmentation, which are evaluated with three different stain combinations showing different degrees of difficulty. Whereas one pipeline directly translates the images to be evaluated and uses a segmentation model trained on original data, the other “way round” translates the training data in order to finally segment the original images. The results exhibit good performance especially for the first approach and provide evidence that the direction of translation plays a crucial role considering the final segmentation accuracy.

Keywords: Histology · Adversarial network · Segmentation · Kidney

1 Motivation

Image-to-image translation gained popularity during the last few years generating highly attractive and realistic output [8, 9, 14]. The majority of approaches require image pairs for training the image-to-image transformation models and make use of single fully-convolutional networks (FCNs) [9] or adversarial networks (ANs) [8]. Recently, the so-called cycleGAN [14] was introduced which eliminates the restriction of corresponding image pairs for training the network.

The authors proposed a generative adversarial network (GAN) relying on a cycle-consistency loss which is combined with the discriminator loss (GAN loss) to perform circular trainings. That means, translations from domain A to domain B and back to domain A and the same vice versa are conducted. This GAN architecture exhibits excellent performance in image-to-image translation applications, based on unpaired training. As image pairs often cannot be obtained (or are at least difficult and/or expensive to achieve), this architecture opens up entirely new opportunities especially in the field of biomedical image analysis.

In this work, we investigate the applicability of image-to-image translation for image-level domain adaptation showing the following advantages: (1) image-to-image translation allows completely unsupervised domain adaptation. (2) The domain adaptation model can be trained independently of the underlying segmentation or classification problem which increases flexibility and saves computation time in case of more than one segmentation and domain adaptation tasks (compared to other methods incorporating both steps into one architecture [10]). (3) Domain adaptation is completely transparent as the intermediate representation is an image and (4) domain adaptation is typically utilized to adapt quite similar domains [3]. Domain pairs considered in image-to-image translation on the other hand are often highly divergent considering color as well as texture. Despite all of these advantages, one problem of the cycleGAN formulation is that there is no guarantee that the objects' outline is kept stable during the adaptation process. Problems can especially occur if the underlying distribution of the objects' shapes are dissimilar between the domains. In this case, it is very likely that GAN training leads to changed shapes as otherwise the discriminator could easily distinguish between real and fake images. If the objects' shapes are changed during GAN training, a segmentation of the fake data and subsequently a transfer of the segmentation mask to the real image cannot be conducted without losing segmentation accuracy.

Due to the dissemination of digital whole slide scanners generating large amounts of digital histological image data, image analysis in this field has recently gained significant importance. Considered applications mostly consist of segmentation [2, 4], classification [1, 7, 12] and regression tasks [13]. For segmentation, especially FCNs [2, 11] yielded excellent performances. However, problems arise if the underlying distribution between training and testing data is dissimilar, which could be introduced by various aspects, such as inter-subject variability, dye variations, different staining protocols or pathological modifications [5]. Although FCNs are capable to learn variability if sufficient (and the right) training data is available, annotating whole slide images (WSIs) for all potential combination of characteristics is definitely not feasible due to the large number of degrees-of-freedom. The authors of previous work [6] proposed a pipeline to perform stain-independent segmentation by registering an arbitrarily dyed WSI with a differently stained WSI for which a trained model exists in order to directly transfer the obtained segmentation mask. Although this strategy allows a segmentation of arbitrarily stained WSIs, it requires for consecutive slices (which show similar image content but are in general not available).

The authors also showed that the registration step and the fact that consecutive slices do not show exactly the same content constitute limiting factors considering segmentation accuracy.

Contributions: To tackle the problem of a large range of different stains, (1) we propose two stain-independent segmentation approaches (P1, P2) for the analysis of histological image data (Fig. 1). We consider a scenario where annotated training data is available for one staining protocol only (S_T). Both pipelines consist of completely separate segmentation and GAN-based stain-translation stages which learn to convert between an arbitrary stain (S_U) and the stain for which annotated data is available (S_T). In case of P1, the input images to be segmented are adapted to match the stain of the training data, whereas in case of P2, the training data is adapted in order to train a segmentation model which fits the images to be segmented. (2) We investigate if stain-translation based on image-to-image translation can be performed effectively for stain-independently segmenting WSIs. (3) As the characteristics and the segmentation-difficulty of the individual stains differ, we expect dissimilar performances on the two pipelines and therefore pose the question for the “best way round”. There exists only one related publication which focusses on stain-independent segmentation of WSIs [6]. Compared to this work, the proposed method does not require consecutively cut slices which are in general not available. Evaluation is performed based on a segmentation task in renal histopathology. Particularly, we segment the so-called glomeruli exhibiting probably the most relevant renal structure (Fig. 2).

2 Methods

We propose two stain-independent segmentation pipelines (P1, P2) consisting of a separate stain-translation and a segmentation stage. Supposed we have annotated training WSIs available for a domain S_T where the domain corresponds to a specific staining protocol. For another domain S_U , there are only non-annotated WSIs available. In the following, focus is on obtaining segmentation masks for new images of the domain without available annotations (S_U , Fig. 1) by making use of two different pipelines. For both pipelines, first a stain-translation GAN (cycleGAN [14]) is trained (Fig. 1, right) consisting, inter alia, of the two generators G_U and G_T converting from S_T to S_U and vice versa.

Pipeline 1 (P1): For P1, the segmentation model M_T is trained with original (S_T) training data. The input images to be segmented are first stain-adapted using model G_T , then segmented based on model M_T and finally the output masks obtained for the stain-adapted fake images are directly transferred to the original images (as shown in Fig. 1, P1). An advantage of P1 is, that the segmentation model can be trained independently of the stain-translation model which improved efficiency in case of more than one adaptation and segmentation tasks.

Pipeline 2 (P2): The training data is stain-adapted utilizing model G_U to translate it from S_T to S_U before training the segmentation model M_U based on fake S_U image data. This model is directly utilized to segment the original (S_U) images without the need for adapting them (Fig. 1, P2). An advantage of P2 is, that during testing only one network is needed (and no stain-translation is performed) improving segmentation efficiency.

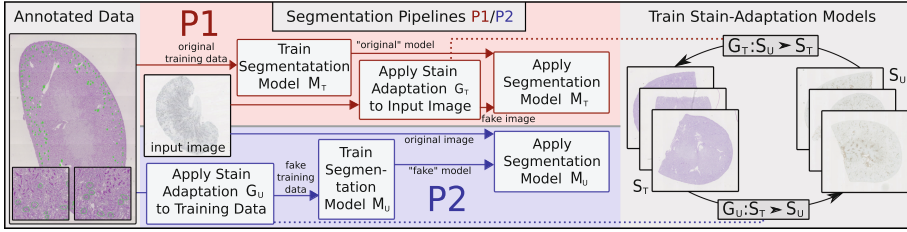


Fig. 1. Outline of the proposed stain-independent segmentation approaches (P1 & P2): In case of P1, the input image is adapted and finally segmented with the model M_T trained on original data (S_T). In case of P2, the training data is adapted before training the segmentation model M_U which is finally used to segment the original data (S_U).

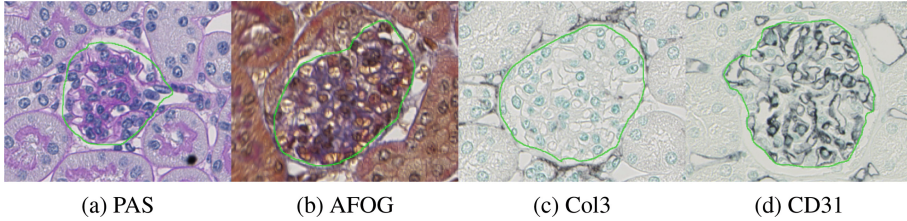


Fig. 2. Example patches from renal tissue showing a glomerulus dyed with four different staining protocols.

Considerations: Due to the final segmentation task, we do not only need to generate realistic images, but also corresponding image pairs (i.e. the objects' masks need to be similar). For example, if the generator creates images with displaced objects, they could look realistic and could potentially also be inversely translated to satisfy the cycle consistency. Such data, however, would be useless for our segmentation task. For this purpose, in case of both pipelines, the following two assumptions need to hold. (a) Firstly, the objects' shapes need to be stain-invariant, i.e. the outline of the objects-of-interest must not depend on the staining protocol. This is because in case of changing shapes between the stains, the GAN would need to change the objects' shape as well. As a result, the unchanged corresponding annotations which are reused either for obtaining the final mask (P1) or for training the segmentation model (P2) would no longer

be adequate. (b) Secondly, obviously information on the outline of the objects-of-interest need to be available in both stainings to facilitate image-to-image translation.

For the considered image data sets, both conditions hold true [6], as can also be assessed based on Fig. 2. Therefore, we expect that cycleGAN is capable to maintain the underlying objects’ shape and to perform appropriate stain-translation. Evaluation is performed by assessing the finally obtained segmentation scores.

2.1 Stain-Translation Model and Sampling Strategies

For training the stain-translation cycleGAN model, first patches are extracted from source domain S_U as well as from target domain S_T WSIs. The target domain corresponds to the stain which should be finally segmented whereas the source domain corresponds to the stain for which training data for segmentation is available. A patch extraction is required, because due to the large size of the WSIs in the range of gigapixels, a holistic processing of complete images is not feasible. Training patches with a size of 512×512 pixels are extracted from the original WSIs. For each data set, we extract 1500 of these patches. To account for the sparsity of the glomeruli, we consider uniform sampling of training patches as well as an equally weighted mixture of uniformly sampled patches (750) and patches containing glomeruli (750). Uniform sampling in both domains is referred to as T_{rand} , 50%/50% sampling in the PAS domain combined with uniform sampling in the other as $T_{50/rand}$ and 50%/50% sampling in both domains is referred to as T_{50} . The first two scenarios are completely unsupervised (as non-PAS domain data is uniformly sampled) whereas the last is not completely unsupervised. With these patches, a cycleGAN based on the GAN-loss \mathcal{L}_{GAN} , the cycle-loss \mathcal{L}_{cyc} as well as the identity loss \mathcal{L}_{id} is trained [14] (with corresponding weights $w_{id} = 1, w_{cyc} = 1, w_{GAN} = 1$). Apart from a U-Net based generator network [11] and an initial identity loss only (w_{id} is only used to stabilize training at the beginning of training and is set to zero after five epochs), the standard configuration based on the patch-wise CNN discriminator is utilized [14)]¹.

2.2 Segmentation Model and Evaluation Details

For segmentation, we rely on an established fully-convolutional network architecture, specifically the so-called U-Net [11] which was successfully applied for segmenting kidney pathology [4]. For taking the distribution of objects into account (the glomeruli are small, sparse objects covering only approximately 2% of the renal tissue area) training patches are not randomly extracted. Instead, as suggested in [4], 50% of the patches are extracted in object-containing-area (to obtain class balance) whereas the other 50% are randomly extracted (to include regions far away from the objects-of-interest).

¹ We use the provided PyTorch reference implementation [14].

The experimental study investigates WSIs showing renal tissue of mouse kidney. Images are captured by the whole slide scanner model C9600-12, by Hamamatsu with a $40\times$ objective lens. As suggested in previous work [4], the second highest resolution ($20\times$ magnification) is used for both segmentation and stain-translation. We consider a scenario where manually annotated WSIs dyed with periodic acid Schiff (PAS) are available for training the segmentation model. For adaptation and finally for stain-independent segmentation, we consider WSIs dyed with Acid Fuchsin Orange G (AFOG), a cluster-of-differentiation stain (CD31) and a stain focused on highlighting Collagen III (Col3). The overall data set consists of 23 PAS, 12 AFOG, 12 Col3 and 12 CD31 WSIs, respectively. 10 of the 12 AFOG, Col3 and CD31 images are used for training the stain-translation model and two are employed for evaluation. All 23 PAS WSIs are utilized for training the segmentation network. For segmentation, we rely on the original U-Net architecture [11]. Batch-size is set to one and L2-normalization is applied. Besides standard data augmentation (rotation, flipping), moderate non-linear deformations are applied similar to [4]. Training is conducted with 4,566 patches (492×492 pixels) extracted from all 23 PAS-stained WSIs. For evaluating the final segmentation performance, the evaluation WSIs (which are not used for training) for each of the stainings AFOG, CD31 and Col3 are manually annotated.

3 Results

The mean Dice similarity coefficients (DSCs) including standard deviations, precision as well as recall are provided in Fig. 3. We notice that P1 generally exhibits higher DSCs compared to P2. P1 also shows stable DSCs with lower standard deviations and rather balanced recall and precision. Considering the different

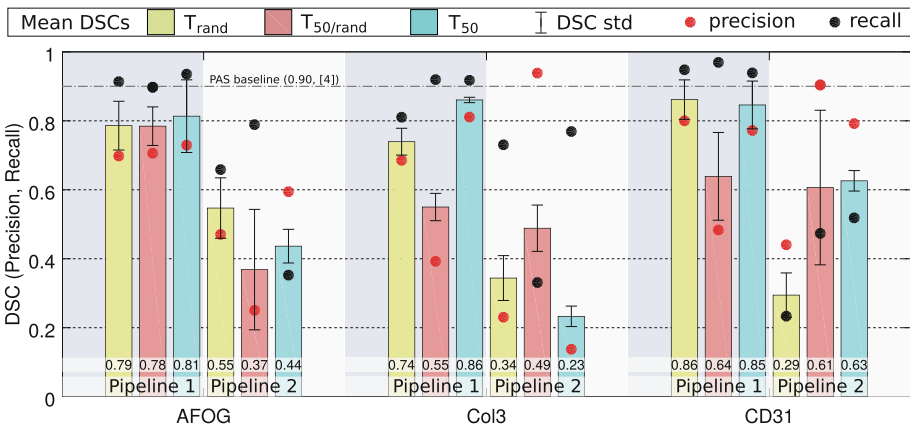


Fig. 3. Segmentation results (DSC, Precision, Recall) individually shown for the two pipelines, the three training configurations and the three stain modalities. PAS baseline indicates the DSC obtained for segmenting original PAS stained images [4].

training strategies, either T_{50} or T_{rand} shows the best rates. The strategy based on different sampling strategies in the two domains $T_{50/rand}$ performs worst. Regarding the three stain combinations, we notice similar rates (between 0.81 and 0.86) in case of T_{50} compared to more divergent DSCs between 0.74 and 0.86 in case of T_{rand} . The overall best DSC is obtained for CD31 in combination with T_{rand} . Figure 4 shows example images after the stain-translation process. We notice that the translation process generally results in highly realistic fake images. We also do not notice any significant changes of the shape of the glomeruli which would automatically lead to degraded final segmentations (in Fig. 3).

4 Discussion

Making use of unpaired image-to-image translation, we propose a methodology to facilitate stain independent segmentation of WSIs relying on unsupervised image-level domain adaptation. A crucial outcome is given by the divergent segmentation performances considering the two proposed and investigated pipelines. It proved to be highly advantageous to translate the WSIs to the PAS staining before segmenting the images and not translating the training images to the target stain (i.e. the stain to be segmented). A reason for this behavior could be given by weakly translated images in case of converted PAS patches. Visual assessment (Fig. 4) indicates that PAS-to-any translation leads to even visually indistinguishable fake images. Thus we are confident that this is not the limiting factor here. Therefore, we assume that this is because the PAS stained images are easier to segment (there is mostly a distinct change in color distribution in

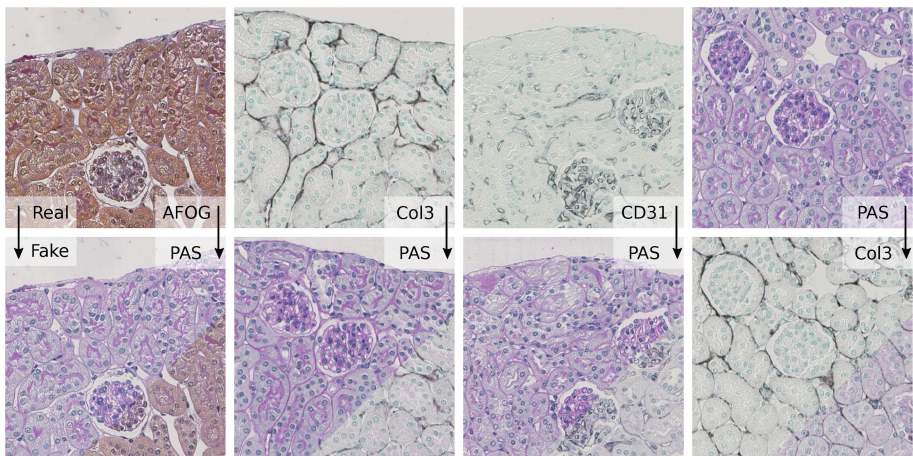


Fig. 4. Example translations as well as overlays of the real and the corresponding fake images (see bottom-right corners). The fake images look highly realistic and do not show any significant changes in objects’ morphology.

case of the glomeruli) and a translation from PAS to a more difficult-to-segment staining leads to a loss of discriminative information. In opposite, if the difficult-to-segment image is converted to an easier-to-segment image, the GAN visually makes a segmentation task even easier. This hypothesis is also supported by the fact that Col3, which is visually most difficult to segment and which exhibits the lowest average DCSs, shows the most significant decrease with P2. Consequently, P1 could be considered as a multi-stage segmentation approach first facilitating the segmentation task using GAN-based stain-conversion followed by segmenting the easy-to-segment image data. In case of P1, we observe that the DSCs (at least in case of T_{rand} and T_{50}) are similar for all stainings whereas strong differences are observed in case of P2. The GAN is obviously able to perform stain-translation similarly well for all stain combinations (see P1 results), although the segmentation networks show divergent outcomes for the different stains (see P2 results). This again demonstrates the high effectiveness of the image translation stage indicating that the limiting factor is rather given by the segmentation network. Considering the different training set strategies, we notice that the approaches considering similar distributions in both domains perform best (T_{rand} and T_{50}). A dissimilar distribution ($T_{50/rand}$) partly leads to a transformation of glomerulus-like samples to fake-glomeruli which are finally also segmented as glomerulus tissue in case of P1. In previous work on registration-based segmentation [6], WSIs stained with CD31 were investigated. While this reference approach reaches 0.83 and also exhibits higher variance, here we obtain DSCs of 0.85 and 0.86, respectively. With our novel method, the inconvenient requirement of consecutive slides can also be circumvented.

To conclude, we introduced two pipelines to enable a stain-independent segmentation of histopathological image data requiring for annotated data for one single stain only. The pipeline based on translating the image to be segmented showed excellent performance and distinctly outperformed the other way round for all configurations. Fortunately, “the best way round” not only delivers the most accurate results, but also constitutes the more flexible method as it allows to arbitrarily combine pre-trained segmentation and translation models. Extended analysis indicates that actually segmentation and not translation is the limiting factor here. Therefore we expect that a pre-selection of special high-quality (and thereby easy-to-segment) slides for training the stain-translation model can boost the overall performance even further by facilitating the segmentation task.

Acknowledgement. This work was supported by the German Research Foundation (DFG) under grant no. ME3737/3-1.

References

1. Barker, J., Hoogi, A., Depeursinge, A., Rubin, D.L.: Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. *Med. Image Anal.* **30**, 60–71 (2016)
2. BenTaieb, A., Hamarneh, G.: Topology aware fully convolutional networks for histology gland segmentation. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 460–468 (2016)
3. Chopra, S., Balakrishnan, S., Gopalan, R.: DLID: deep learning for domain adaptation by interpolating between domains. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2013)
4. Gadermayr, M., Dombrowski, A., Klinkhammer, B.M., Boor, P., Merhof, D.: CNN cascades for segmenting whole slide images of the kidney. *CoRR*, <https://arxiv.org/abs/1708.00251> (2017)
5. Gadermayr, M., Strauch, M., Klinkhammer, B.M., Djudjaj, S., Boor, P., Merhof, D.: Domain adaptive classification for compensating variability in histopathological whole slide images. In: Campilho, A., Karray, F. (eds.) *ICIAR 2016. LNCS*, vol. 9730, pp. 616–622. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41501-7_69
6. Gupta, L., Klinkhammer, B.M., Boor, P., Merhof, D., Gadermayr, M.: Stain independent segmentation of whole slide images: a case study in renal histology. In: *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)* (2018)
7. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: *Proceedings of the International Conference on Computer Vision (CVPR)* (2016)
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)
9. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2016)
10. Kamnitsas, K., et al: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: *Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI)*, pp. 597–609 (2017)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015. LNCS*, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. Sertel, O., Kong, J., Shimada, H., Catalyurek, U., Saltz, J., Gurcan, M.: Computer-aided prognosis of neuroblastoma on whole-slide images: classification of stromal development. *Pattern Recognit.* **42**(6), 1093–1103 (2009)
13. Veta, M., van Diest, P.J., Pluim, J.P.W.: Cutting out the middleman: measuring nuclear area in histopathology slides without segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) *MICCAI 2016. LNCS*, vol. 9901, pp. 632–639. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_73
14. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the International Conference on Computer Vision (ICCV)* (2017)