# Unsupervisedly Training GANs for Segmenting Digital Pathology with Automatically Generated Annotations

**Michael Gadermayr**[1,2]
**Laxmi Gupta**[2]
[1] *Salzburg University of Applied Sciences, Austria*
[2] *Institute of Imaging and Computer Vision, RWTH Aachen University*

**Barbara M. Klinkhammer**[3]
**Peter Boor**[3]
[3] *Institute of Pathology, University Hospital Aachen, RWTH Aachen University, Germany*

**Dorit Merhof**[2]

## Abstract

Recently, generative adversarial networks exhibited excellent performances in semi-supervised image analysis scenarios. In this paper, we go even further by proposing a fully unsupervised approach for segmentation applications with prior knowledge of the objects' shapes. We propose and investigate different strategies to generate simulated label data and perform image-to-image translation between the image and the label domain using an adversarial model. For experimental evaluation, we consider the segmentation of the glomeruli, an application scenario from renal pathology. Experiments provide proof of concept and also confirm that the strategy for creating the simulated label data is of particular relevance considering the stability of GAN trainings.

## 1. Motivation

Due to the progressing dissemination of whole slide scanners generating large amounts of digital histological image data, image analysis in this field has recently gained significant importance (Hou et al., 2016; BenTaieb and Hamarneh, 2016; Gadermayr et al., 2018a; Valkonen et al., 2017; Veta et al., 2016; Gadermayr et al., 2017; Herve et al., 2011).

For segmentation applications, especially fully-convolutional networks proofed to be highly effective tools (Ronneberger et al., 2015; BenTaieb and Hamarneh, 2016; Gadermayr et al., 2017). A major challenge in the field of digital pathology is given by a large set of different application scenarios as well as changing underlying data distributions which is due to inter-subject variability, different staining protocols and/or pathological modifications (Gadermayr et al., 2018b). Each individual application scenario therefore requires large amounts of annotated training data covering the prevalent variability. The acquisition of such large amounts of labeled training data, however, is typically time-consuming and cost-intensive and thereby constitutes a burden for the deployment of automated segmentation techniques.

For training state-of-the-art machine learning approaches such as fully-convolutional networks, data augmentation proved to be a highly powerful tool (Ronneberger et al., 2015; J. Ratner et al., 2017) to keep the amount of required training data decent. A limitation of data augmentation in combination with supervised learning approaches is given by the fact that often large non-annotated data is available “for free” but is not utilized for training at all. Particularly in the fields of medicine, such as digital pathology, huge amounts of digital image data are routinely captured without any (additional) effort whereby a complete annotation of all data is definitely not feasible. In order to take advantage of non-annotated data as well, dedicated semi-supervised segmentation approaches relying on adversarial models were recently proposed (Kozí Nski et al., 2017; Isola et al., 2017; Hung et al., 2018).

Adversarial models were also developed for the field of image-to-image translation (Johnson et al., 2016; Zhu et al., 2017). Recently, the so-called cycleGAN (Zhu et al., 2017) was introduced which eliminates the restriction of corresponding image pairs for training. This architecture can also be utilized for means of unsupervised domain adaptation (Chartsias et al., 2017; Wolterink et al., 2017; Gadermayr et al., 2018a). The domain adaptation in these cases is performed on image-level, i.e. “fake” images showing similar characteristics as the target domain samples are generated. This strategy is highly flexible as it can be combined with arbitrary further segmentation or classification approaches.

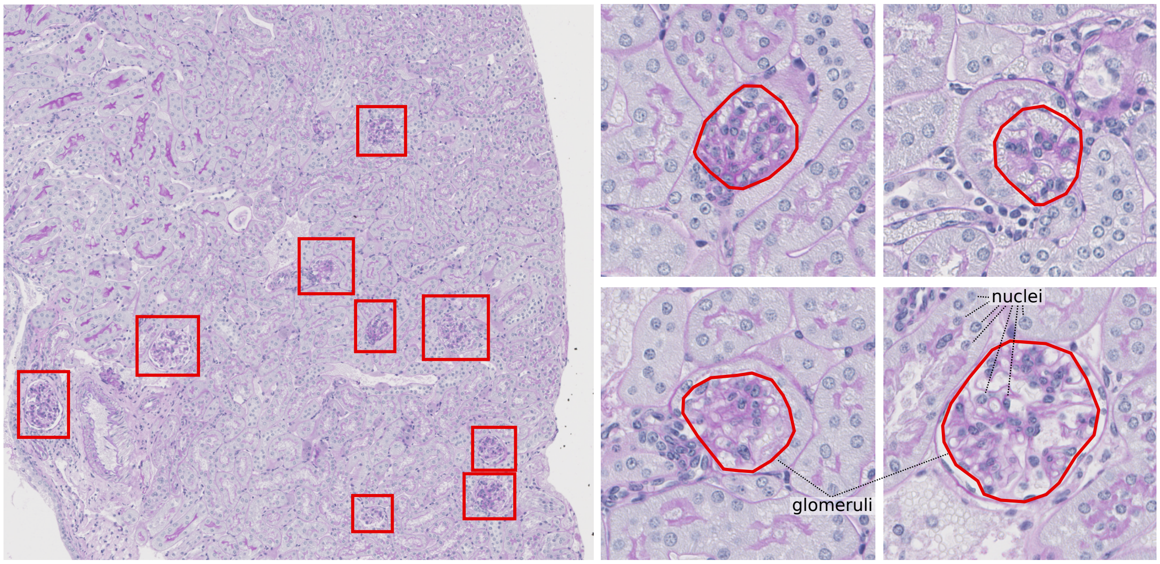


Figure 1: This illustration shows an extract of a renal whole slide image with marked glomeruli (left) as well as magnifications of single glomeruli showing precise manual annotations (right).

**Contribution:** We tackle the problem of acquiring labeled training data by proposing a framework completely bypassing the need for manually labeled objects. We focus on generating artificial annotations to perform image-to-image translation on unpaired data sets. In our experimental study, we investigate strategies for modeling the shape of the annotations and for modeling additional image information to facilitate training the translation networks. As application scenario, we consider a segmentation task from digital pathology, specifically the segmentation of the renal glomeruli (Gadermayr et al., 2017; Kato et al., 2015; Herve et al., 2011) (Fig. 1).

## 2. Methods

For the proposed method, we make use of an image-to-image translation approach. Specifically, we utilize a generative adversarial network (GAN) which facilitates training with unpaired data (Zhu et al., 2017). The four subnetworks consisting of two generators and two discriminators are optimized based on an adversarial loss as well as a cycle consistency criterion. This formulation does not require sample pairs, i.e. there is no need to obtain corresponding image samples for the two domains. Instead it is sufficient to collect a set of images, individually for each domain. If annotations are interpreted as label (e.g. binary) images, this approach can be utilized for segmentation applications as well. The architecture allows to perform training based on a set of images and a (non-corresponding) set of annotations as long as the annotations are realistic (i.e. the distribution matches the underlying distribution of real annotations).

The proposed method relies on an automated generation of realistic annotation images followed by training an image-to-image translation model which is finally able to convert original images to annotations. The procedure is based on the following assumptions: (1) we need to understand the underlying distribution of the annotation data and we need to be able to model this distribution (for details, see Sect. 2.1). (2) The unpaired image-to-image translation approach needs to be effectively applicable to translate between the image and the annotation domain. If a straight-forward translation is not effective, additional information can be added to the annotation domain to enhance the translation process (for details, see Sect. 2.2).

### 2.1. Annotation Model

In the considered application scenario (Fig. 1), the underlying distribution (assumption 1) of the objects-of-interest is rather basic and can thereby be approximately modeled quite well. The objects-of-interest show roundish shapes which are sparsely distributed over the kidney. For training we consider patches extracted from the whole slide images. We assume that the number of objects per patch can be approximated by a (quantized) Gaussian distribution $G_{\#} \sim \mathcal{N}(\mu_g, \sigma_g^2)$. The objects are uniformly distributed over the patch with one single further assumption that the objects may not overlap. For generating the annotations, we investigate two different approaches. Firstly, we consider the objects-of-interest as round objects (**Circular objects (C):**). The objects' radii $r$ are randomly sampled from a Gaussian distribution $R \sim \mathcal{N}(\mu_r, \sigma_r^2)$. In a second configuration, we incorporate the fact that the objects-of-interest often show an elliptic shape (**Elliptic objects (E):**). To incorporate this knowledge, $r_1$ is drawn from the same distribution as $r$ and $r_2 = r_1 + r_\delta$ where $r_\delta$ models the eccentricity and is drawn from $R_\delta \sim \mathcal{N}(0, \sigma_e^2)$. A further rotation parameter $\alpha$ is drawn from a uniform distribution in the interval $[0, 2\pi]$.

### 2.2. Image-to-Label Translation

The straightforward approach consists of adding either circles or ellipses as binary objects into two dimensional matrices which are interpreted as single channel images. However, for training the image-to-image translation approach, this setting can be highly challenging due to the loss criteria:

For training the GAN (Zhu et al., 2017), two generative models, $F : \mathcal{X} \rightarrow \mathcal{Y}$ and $G : \mathcal{Y} \rightarrow \mathcal{X}$ and two discriminators $D_X$ and $D_Y$ are trained optimizing the cycle consistency loss $\mathcal{L}_c$

$$\begin{aligned} \mathcal{L}_c = \mathbb{E}_{x \sim p_{data}(x)}[||G(F(x)) - x||_1] + \\ \mathbb{E}_{y \sim p_{data}(y)}[||F(G(y)) - y||_1] \end{aligned} \tag{1}$$

and the adversarial loss $\mathscr{L}_d$

$$\begin{aligned}\mathscr{L}_d = \mathbb{E}_{x \sim p_{data}(x)}[\log(D_X(x)) + \log(1 - D_Y(F(x)))] + \\ \mathbb{E}_{y \sim p_{data}(y)}[\log(1 - D_X(G(y))) + \log(D_Y(y))] \ .\end{aligned} \tag{2}$$

$F$ and $G$ try to generate fake images that look similar to real images, while $D_X$ and $D_Y$ aim to distinguish between translated and real samples. The generators aim to minimize this adversarial objective against the discriminators that try to maximize it.

Let $X$ be the domain referring to the original images and let $Y$ be the label domain. The cycle criterion requires that an annotation mask can be translated to an image by the generator $G$. The generator $F$, however, hides all low-level image details, such as nuclei and tubuli (Fig. 1) and only preserves the high-level shapes of the glomeruli. Based on these shapes only, it will not be able to reconstruct e.g. the nuclei at the right (i.e. the same) positions leading to a high cycle-consistency loss even though the images might look realistic. To take this into account, we propose and investigate a second setting simulating the nuclei exhibiting low-level information as well. As for the glomeruli, the number of nuclei is drawn from a (quantized) normal distribution $N_{\#} \sim \mathscr{N}(\mu_n, \sigma_n^2)$. They are uniformly distributed over the whole patch with the restriction that they may not overlap. Diameter is fixed to $d_n$. The additional binary matrix containing the nuclei is added as further image channel to the annotation image. This channel is only needed to train the GAN. For testing, this further channel is simply ignored. Whereas the setting incorporating only the target labels (i.e. the glomeruli) is referred to as single-class scenario, the setting also incorporating further low-level information is referred to as multi-class scenario. Finally, we identified four different settings: single-class circular objects (SC), single-class elliptical objects (SE), multi-class circular objects (MC) and multi-class elliptical objects (ME).

To facilitate learning, Gaussian random noise ($\sigma_{f_n}$) is added to the annotation maps followed by the application of a Gaussian filter ($\sigma_{f_s}$) to smooth the objects' borders in all settings.

### 2.3. Experimental Setting

Paraffin sections ($1\mu m$) are stained with periodic acid-Schiff (PAS) reagent and counterstained with hematoxylin. Whole slides are digitalized with a Hamamatsu NanoZoomer 2.0HT digital slide scanner and a 20× objective lens. From each of the 23 WSIs overall, 200 patches with a size of $500 \times 500$ pixels are randomly extracted (resulting in 4600 patches overall). For evaluation purpose, the WSIs are manually annotated under the supervision of a medical expert. Learning is performed in a transductive setting, i.e. both training and testing is executed on all patches. This does not introduce bias in this case, as no label data is used during training.

As large context is required to assess whether segmentations are realistic, a (rather low) resolution corresponding to a 2.5× magnification is utilized (original images downscaled by factor eight).

For image-to-image translation, we make use of the cycle GAN (Zhu et al., 2017). We rely on the provided pytorch reference implementation. Apart from the following changes, we use the proposed standard settings. As generator model, a residual network consisting of four blocks is utilized. As discriminator, we rely on the suggested patch-wise CNN with three layers (Zhu et al., 2017). Learning rate is fixed to $10^{-6}$, number of training epochs is set to 15 and batch size is set to one. The losses are equally weighted. For data augmentation, flipping, rotation and random cropping ($384 \times 384$ pixel sub-patches) is performed.

The annotations are generated based on the following visually assessed parameters (we did not incorporate statistical information of the data set to avoid introducing significant supervision): $\mu_g = 7$, $\sigma_g = 2$, $\mu_r = 18$, $\sigma_r = 2$, $\sigma_e = 2$, $d_n = 4$, $\mu_n = 5000$, $\sigma_n = 50$, $\sigma_{f_n} = 5$ and $\sigma_{f_s} = 2$.

For evaluation, we investigate two optimization strategies. The first strategy does not incorporate any optimization and we basically report the obtained segmentation performance after training for all 15 epochs. As GAN training is, in general, often unstable, we also optimize the epoch by separating the testing data set into one patch for optimization and the others for testing. We use only one patch for optimization because the approach is intended to be unsupervised.

Apart from pixel-level scores ($F_1$-score (F), precision (P), recall(R)), we also report the corresponding object-level scores ($F_o$, $P_o$, $R_o$). That means, we distinguish between true positive objects (i.e. the distance between the center of a detected object and a real object is smaller than 10 pixels), objects which were missed and false positively detected objects.

All experiments are repeated four times. The obtained performances are compared with the a supervised fully-convolutional network (Gadermayr et al., 2017).

## 3. Results

Fig. 2 shows quantitative results for each of the four different settings. We investigate pixel-level as well as object-level scores. The left two columns show the testing pixel-level and object-level $F_1$-scores for different numbers of training epochs. The third column shows the scores obtained with cross validation (i.e. the epoch is optimized) and the last column shows the rates corresponding to training for 15 epochs without any further optimization.

Considering these results, we notice that the single-class settings (SC, SE) do not show any useful results. In case of elliptical shapes (SE), at least the best configuration exhibits acceptable outcomes, however, GAN training is highly unstable in this scenario. In case of the multi-class settings (MC, ME), we notice a more stable behavior, as in each repetition good scores are obtained after few training epochs. Mean pixel-level $F_1$-scores of 0.63 (MC) and 0.62 (ME) as well as mean object-level F-scores of 0.74 (MC and ME) are achieved. Convergence is obtained approximately after six epochs for both settings. We notice slightly higher precision than recall, especially on object-level. A further optimization of the number of the training epoch does not show a high influence.

The baseline results of the supervised approach are provided in Fig. 3. We obtain $F_1$-scores of 0.49, 0.65 and 0.71 on pixel level and 0.52, 0.68 and 0.76 on object-level for training with 2, 4 and 8 WSIs. We notice that the break-even point of the supervised approach is reached with approximately four fully-annotated training WSIs corresponding to roughly 500 annotated glomeruli. Considering the object-level scenario, the proposed method exhibits increased performances (comparable with the supervised method trained on eight WSIs).

We further investigated the annotation images with respect to the shape of the masks. Comparing the best fitting circle with the mask showed a mean $F_1$-score of 0.92 while for the best fitting ellipse, a $F_1$-score of 0.95 is obtained. The difference is statistically significant ($p < 0.001$) and indicates that ellipses provide better approximations for the objects-of-interest.

Example output of the image-to-image translation process is provided in Fig. 4. With the single-class setting ((a)–(b)), we notice a tendency to segment vessel structures instead of the target objects. This is not the case if making use of the multi-class settings ((c)–(d)).

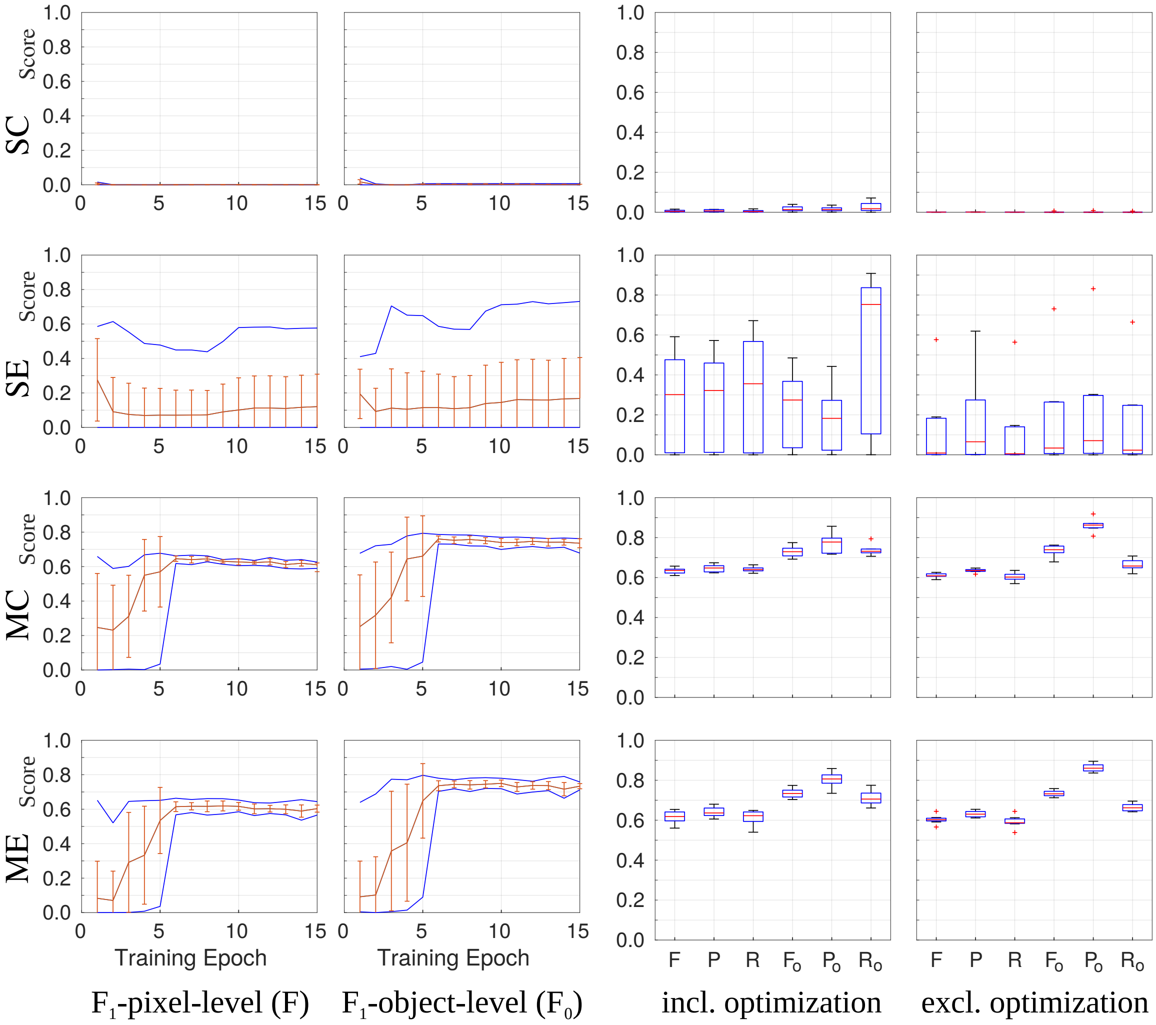
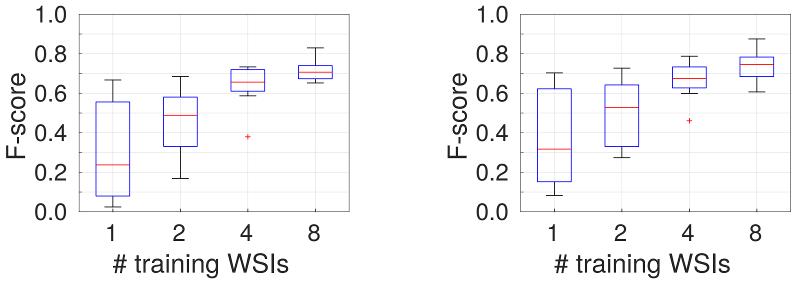


Figure 2: Experimental results for the four different settings (row 1 to row 4). The left columns show pixel- and object-level $F_1$-scores for testing after varying number of training epochs. The right columns provide $F_1$-scores (F), precision (P) and recall (R) as well as object-level measures ($F_o$, $P_o$, $R_o$) for training for 15 epochs (excl. optimization) and for optimizing the epoch (incl. optimization).

## 4. Discussion

In this work, we investigated a concept of fully-unsupervised learning for segmentation applications by making use of a GAN in combination with simulated annotation data.

We obtained highly divergent results for the four different settings. One substantial finding is that a simulation of the annotations of the objects-of-interest only (referred to as single-class scenario) is not sufficient to obtain proper segmentations of the glomeruli in the investigated unpaired image-to-image translation scenario. In the majority of attempts, an unwanted translation between the image and the label domain is observed. A major problem here is that a translation from the



(a) $F_1$-pixel-level (F)

(b) $F_1$-object-level ($F_o$)

Figure 3: Baseline pixel-level (a) and object-level (b) $F_1$-scores indicating the segmentation performance of the supervised U-Net-based approach (Gadermayr et al., 2017) with variable numbers of fully-annotated training WSIs. One single training WSI contains on average 120 single objects.

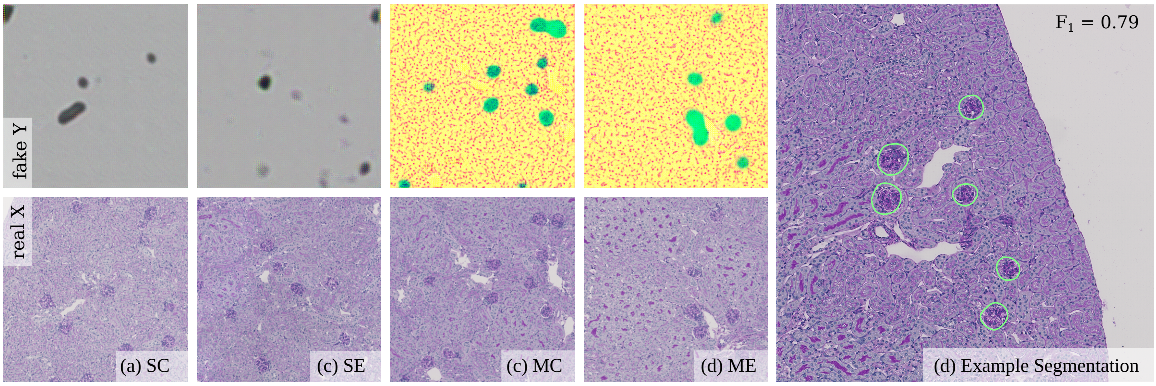


Figure 4: Qualitative results of the image translation process for the four different settings (a) – (d). While the setting SC and SE do not show any good segmentations, MC and ME perform similarly well. Subfigure (e) shows an example segmentation extracted from a fake image generated with setting MC.

label to the image domain cannot be performed which complicates the GAN training. The generator $G$ in this case has no chance to place the low-level objects (here the nuclei) in a way that the cycle consistency loss can become small as the position of the nuclei cannot be effectively derived from the annotation image. This behavior can also be seen in the example reconstructed images where nuclei cannot be clearly detected (Fig. 4, third column). In the multi-class scenarios with added simulated nuclei during GAN training, these objects are maintained during the training cycles. That means, the nuclei are segmented during translation to the label domain followed by a reconstruction of the nuclei based on the label domain in case of the inverse mapping.

A further interesting finding is that the distribution of the shapes of the simulated objects does not have a major impact on final segmentation performance. We do not consider the single-class scenarios here as they showed either completely wrong or highly unstable performance. The multi-class scenarios show similar performances for the setting based on circles and for the setting based on ellipses.

Considering the multi-class settings MC and ME, we assess the obtained segmentation performance as good and applicable for medical applications although the scores seem to be rather low. We need to mention here that this is on the one hand due to the fact that small objects are often not identified as glomeruli in the ground-truth but are detected by our approach. On the other hand, there are also small objects which are in the ground-truth but are not detected. Anyway, these objects are neglected by the medical experts and are thereby excluded from further analysis.

A comparison with a state-of-the-art supervised approach showed that the novel method is highly competitive. Especially the detection performance (indicated by the object-level $F_1$-scores) is outperformed by the supervised technique only if training is performed with a large amount of annotated data (specifically with eight WSIs corresponding to approx. 1000 single objects). Due to the stable training process, a “slightly-supervised” optimization of the training epoch is not required as the results are only marginally improved (Fig. 2, column 3 vs. column 4).

The most notable advantage, however, does not consist in high scores, but in a very high flexibility. The method can be easily adapted e.g. to other stains without a need for collecting novel annotated training data. An intrinsic limitation is certainly given regarding the shape of the objects-of-interest. While rather basic shapes can be easily modeled, complex or irregular shapes are either difficult or even impossible to model.

To conclude, we proposed and investigated a concept of fully-unsupervised learning for segmentation applications by making use of a GAN trained with real images and simulated annotations. The experimental results, in general highly promising, indicate that it is not crucial to accurately model the underlying shape as long as a good approximation is available. This is a highly relevant finding as the shapes of the objects-of-interest are often too complex to be modeled accurately. It is clearly more relevant to support the GAN to fulfill the cycle consistency criterion. Adding additional information to the label domain proved to be an effective way to facilitate the unpaired training process. A comparison with a state-of-the-art supervised segmentation approach shows that the novel method is only outperformed if a large amount of labeled training data is available.

## Acknowledgments

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

## References

Aicha BenTaieb and Ghassan Hamarneh. Topology aware fully convolutional networks for histology gland segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'16)*, pages 460–468, 2016.

Agisilaos Chartsias, Thomas Joyce, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *Proceedings of the International MICCAI Workshop Simulation and Synthesis in Medical Imaging (SASHIMI'17)*, pages 3–13, 2017.

Michael Gadermayr, Ann-Kathrin Dombrowski, Barbara Mara Klinkhammer, Peter Boor, and Dorit Merhof. CNN cascades for segmenting whole slide images of the kidney. *CoRR, https://arxiv.org/abs/1708.00251*, 2017.

Michael Gadermayr, Vitus Appel, Barbara M. Klinkhammer, Peter Boor, and Dorit Merhof. Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'18)*, pages 165–173, 2018a.

Michael Gadermayr, Dennis Eschweiler, Barbara Mara Klinkhammer, Peter Boor, and Dorit Merhof. Gradual domain adaptation for segmenting whole slide images showing pathological variability. In *International Conference on Image and Signal Processing (ICISP'18)*, Springer LNCS, pages 461–469, 2018b.

N. Herve, A. Servais, E. Thervet, J.-C. Olivo-Marin, and V. Meas-Yedid. Statistical color texture descriptors for histological images analysis. In *Proceedings of ISBI'11*, pages 724–727, 2011.

Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the International Conference on Computer Vision (CVPR'16)*, 2016.

Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *CoRR, https://arxiv.org/abs/1802.07934*, 2018. URL http://arxiv.org/abs/1802.07934.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017.

Alexander J. Ratner, Henry R. Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Re. Learning to compose domain-specific transformations for data augmentation. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS'17)*, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV'16)*, 2016.

Tsuyoshi Kato, Raissa Relator, Hayliang Ngouv, Yoshihiro Hirohashi, Osamu Takaki, Tetsuhiro Kakimoto, and Kinya Okada. Segmental HOG: new descriptor for glomerulus detection in kidney microscopy image. *BMC Bioinformatics*, 16(1), 2015.

Mateusz Kozí Nski, Loïc Simon, and Frédéric Jurie. An Adversarial Regularisation for Semi-Supervised Training of Structured Output Neural Networks. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS'17)*, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Aided Interventions (MICCAI'15)*, pages 234–241, 2015.

Mira Valkonen, Kimmo Kartasalo, Kaisa Liimatainen, Matti Nykter, Leena Latonen, and Pekka Ruusuvuori. Dual structured convolutional neural network with feature augmentation for quantitative characterization of tissue histology. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*, 2017.

Mitko Veta, Paul J. van Diest, and Josien P. W. Pluim. Cutting out the middleman: Measuring nuclear area in histopathology slides without segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'16)*, pages 632–639, 2016.

Jelmer M. Wolterink, Anna M. Dinkla, Mark H. F. Savenije, Peter R. Seevinck, Cornelis A. T. van den Berg, and Ivana Išgum. Deep MR to CT synthesis using unpaired data. In *Proceedings of the International MICCAI Workshop Simulation and Synthesis in Medical Imaging (SASHIMI'17)*, pages 14–23, 2017.

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 International Conference on Computer Vision (ICCV'17)*, 2017.