Domain-specific Cycle-GAN Augmentation Improves Domain Generalizability for Mitosis Detection

Rutger H.J. Fick, Alireza Moshayedi, Gauthier Roy, Jules Dedieu, Stéphanie Petit, and Saima Ben Hadj

Tribun Health, Paris, France

Abstract. As the third-place winning method for the MIDOG mitosis detection challenge, we created a cascade algorithm consisting of a Mask-RCNN detector, followed by a classification ensemble consisting of ResNet50 and DenseNet201 to refine detected mitotic candidates. The MIDOG training data consists of 200 frames originating from four scanners, three of which are annotated for mitotic instances with centroid annotations. Our main algorithmic choices are as follows: first, to enhance the generalizability of our detector and classification networks, we use a state-of-the-art Residual Cycle-GAN to transform each scanner domain to every other scanner domain. During training, we then randomly load, for each image, one of the four domains. In this way, our networks can learn from the fourth non-annotated scanner domain even if we don't have annotations for it. Second, for training the detector network, rather than using centroid-based fixed-size bounding boxes, we create mitosis-specific bounding boxes. We do this by manually annotating a small selection of mitoses, training a Mask-RCNN on this small dataset, and applying it to the rest of the data to obtain full annotations. We trained the follow-up classification ensemble using only the challenge-provided positive and hard-negative examples. On the preliminary and final test set, the algorithm scores an F1 score of 0.7578 and 0.7361, respectively, putting us as the preliminary second-place and final third-place team on the leaderboard.

Keywords: MIDOG Challenge · Mitosis Detection · Instance Segmentation.

1 Introduction

Mitosis detection is a highly challenging task in pathology due to the rarity of the events and the highly variable morphological appearance of a cell undergoing mitosis - some being very clear and others highly ambiguous [17]. While several mitosis detection challenges have been organized over the past years (MITOS12 [13], AMIDA13 [14], MITOS14 [11], TUPAC16 [15]), none of them focused on testing the effect of domain shift on the robustness of a mitosis detection method. The MIDOG challenge [1, 2] specifically addresses this by providing training data originating from four different scanners but making the unseen test set (partially) consist of images that are not from these same scanners.

1.1 Dataset

Following the challenge description: the MIDOG training subset consists of 200 whole slide images (WSI) from human breast cancer tissue samples stained with routine H&E dye. The samples were digitized with four slide scanning systems: the Hamamatsu XR NanoZoomer 2.0, the Hamamatsu S360, the Aperio ScanScope CS2 and the Leica GT450, resulting in 50 WSI per scanner. For the slides of three scanners, a selected field of interest sized approximately 2mm^2 (equivalent to ten high power fields) was annotated for mitotic figures and hard negative look-alikes. These annotations were collected in a multi-expert blinded set-up, but with the help of computer augmentation, similar to [4]. For the Leica GT450, no annotations were available. The preliminary and final test set consist of four (at the time undisclosed) scanners, only two of which were also part of the training set, namely the Hamamatsu XR, Leica GT450, 3DHistech P1000 and Hamamatsu RS. The preliminary test set consists of only five WSI from each for four test scanners. This preliminary test set was used for evaluating the algorithms prior to submission and publishing preliminary results on a leaderboard. The final test set consists of 20 additional WSI from the same four test scanners. The evaluation through a Docker-based submission system ensured that the participants had no access to the (preliminary) test images during method development.

2 Material and Methods

We base our method around a classic cascade approach to detect mitotic instances in H&E-stained images. We first use a Mask-RCNN [8] to detect mitotic candidates in an image. These candidates are then extracted as small patches and given to a classifier ensemble of a ResNet50 [7] and DenseNet201 [9]. The predictions are merged via weighted average and the final score is returned.

To improve the generalizability of the method - which is the main purpose of the challenge - we used a Residual Cycle-GAN [3] to transform each image of the training images into all other available domains. In this way, each mitotic annotation is available in all 4 scanner domains. This differs from standard data augmentation (color, hue, brightness, etc.), in that these are not random shifts in appearance for the training process, but specifically towards domains that we *know* are in the testing set. In Figure 1 we show a 4×4 grid of images of the 4 domains that we transformed to all other domains.

To improve the information present in the data for training a detector, we use Mask-RCNN to create pixel-wise annotations for all annotated mitotic instances. Since we know where all mitoses are, we use Inkscape to manually annotate the first 100 or so, train a pretrained Mask-RCNN model on this small dataset, and apply it specifically around other known mitoses. We use test-time augmentation (8 rotations and flips) and average the predicted masks for each mitosis, resulting in clean masks for most annotations. The remaining ”difficult” cases were manually completed, providing us mitosis-specific bounding boxes for all mitotic instances. The average bounding box diagonal in the dataset is 28.8±7.9 pixels, which is consistent with the MITOS12 dataset [10].

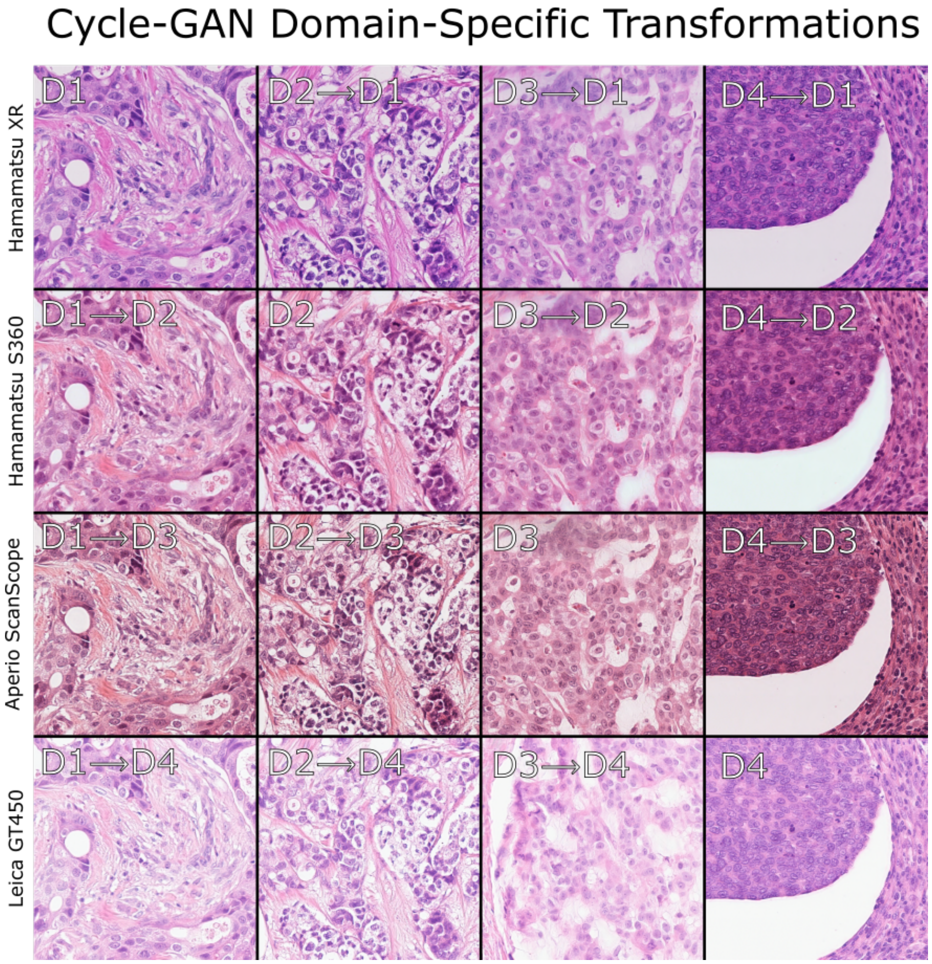

Fig. 1. Residual Cycle-GAN transformed patches. The diagonal are original patches, off-diagonal patches are domain-transformed.

2.1 Domain-Specific Residual Cycle-Gan Augmentation

For the Residual Cycle-GAN [3] we followed the reference model architecture of two sets of generators and discriminators. The Residual Cycle-GAN follows the same principle as a regular one, with the difference that the input image has a direct skip connection with the generated output image. In this way, the generator does not need to reconstruct the image from a set of filter outputs, but only needs to add a "residual", i.e. a color change in the input image so that it resembles a target domain. As it reduces the computational load on the generator, this approach requires fewer data and converges more quickly.

We train six Cycle-GANs to obtain domain transformation functions from all four scanner domains to each other. We train each cycle-GAN for 150000 iterations with generator learning rate $1e-4$, discriminator learning rate $1e-2$, batch size 4, cycle-consistency loss weight 1 and adversarial loss weight 5. We then use the trained generators of these models to create four complete "scanner" copies of the training data, where each copy corresponds to one of four scanners. This means that each "scanner" data set consists of 25% real data and 75% GAN-transformed data, which will be equally sampled from during training. We show an illustration of the 4 data sets in Figure 1.

2.2 Final Submission Network Training

We split our training data into 45 training slides and 5 validation slides per scanner, ensuring that the validation set had both highly mitotic and non-mitotic slides. The Torchvision implementation of Mask-RCNN with ResNet50 backbone was pretrained on the public COCO2017 dataset, and both the ResNet50 and DenseNet201 classifiers were pretrained on ImageNet. Note that this is the same model architecture as we used for creating the mitosis masks, but now trained on in For Mask-RCNN training, we used a patch size of 3000×3000 pixels and a batch size of 1. We did not train on patches that did not contain any mitoses. We found that using a larger patch size improves the validation performance, and did not improve when adding negative patches. We augmented Mask-RCNN training using skewing, 8 random flips/mirroring, and the domain-specific Cycle-Gan augmentation stated before. We used SGD with a plateau-reduction learning rate scheduler starting at 0.002 and reducing by a factor of 2 if the PR-AUC does not improve after 5 epochs. We warmed up the optimizer during the first epoch and only unfroze the last two convolutional blocks of the Mask-RCNN network. We ran the algorithm until convergence after around 200 epochs.

The classification networks were only trained with the positive and negative instances provided by the challenge organizers - we found that adding hard negatives detected by the detector did not improve leaderboard performance. We used a batch size of 32, and trained for 100 epochs, and kept the model with the best F1 score. We used ADAM with standard parameters and a Cosine annealing learning rate scheduler starting at 2×10^{-5} with a focal loss. For both networks, we only unfroze the backbone after 5 epochs. We used a patch size of 80×80, which we resized to 224×224 to conform with ImageNet pretraining.

We used our GAN-based domain augmentation, together with H&E specific data augmentation [6], with parameters $n = 3$, $m = 7$. The classification head consists of 3 blocks of convolutions with Relu, batch normalization, and dropout set to 0.5, followed by a fully connected layer to the output.

For both the detector and classifiers, many variations of optimization parameters were tried and the model with the best PR-AUC on validation was selected.

2.3 Ablation Study on Instance Segmentation and Domain-Specific GAN Augmentation

The outcome of the challenge indicated that using label enhancement (i.e. adding instance segmentation masks) for mitosis annotations was a winning ingredient for all winning MIDOG approaches. For this reason, we conducted an ablation study on the aforementioned training strategy to understand what aspect of our algorithm most contributed to our success. Note that we perform this study *on the detection algorithm only*, without the subsequent cascade classifiers.

To this end, we started our experiment with a basic Faster-RCNN network [12] with ResNet50 as a backbone, trained using fixed-size bounding boxes of size 50 × 50 pixels, centered on the mitosis coordinate. We then gradually increased the complexity of our strategy; first by introducing geometric augmentations e.g. rotations, flips, skewing. Then using the exact bounding box obtained from borders of the mask annotations and finally, using a Mask-RCNN with the actual mitosis masks and an offline GAN-based data augmentation method where we transformed the data from each scanner to a different scanner.

3 Results

In Table 1 we show the results of our ablation study to find what worked best for our mitosis detection algorithm. Note that ”F1 val” indicates F1 score on Scanner 1,2 and 3 images, whereas ”F1 val S4” is the F1 score on the same validation images but GAN-transformed to look like the scanner 4 domain.

In Table 2 we finally show a summary of our model's scanner-wise performance statistics on the MIDOG challenge test set and our validation set after training. Note that NA means ”Not Available” as these scanners were not available in either the test or training set. We discuss our results in the next section.

For reference, our model's aggregate validation PR-AUC was 0.8823 and F1 was 0.8287. On the *preliminary* test set our approach resulted in the second-highest aggregate F1-score of 0.7577, resulting from a 0.7820 precision and a 0.7349 recall.

4 Discussion and Conclusion

From the MIDOG Challenge results the pattern emerged that the first, second, and third place winners (us) all enhanced the mitosis annotations before using

Table 1. Ablation study to find optimal mitosis detection strategy. From top to bottom, the algorithm becomes increasingly complex. Describing columns from left to right, first, there is experiment ID, which model we use (Fast-RCNN or Mask-RCNN), whether we used fixed-size or adaptive bounding boxes (based on masks), whether we use the mask itself (for Mask-RCNN), use of skew augmentation and use of domain-specific GAN augmentation. Finally, we report F1 score statistics on the validation set (F1 Val) and the F1 score on the GAN-transformed validation set to scanner 4 (F1 val S4).

exp #	model	adapt. bboxes	mask	skew aug	GAN aug	F1 val	F1 val S4
(1)	Faster-RCNN	×	×	×	×	0.824	0.536
(2)	Faster-RCNN	×	×	✓	×	0.835	0.425
(3)	Faster-RCNN	✓	×	✓	×	0.818	0.493†
(4)	Faster-RCNN	✓	×	✓	✓	0.823	0.815*
(5)	Mask-RCNN	✓	✓	✓	×	0.812	0.705†
(6)	Mask-RCNN	✓	✓	✓	✓	0.813	0.812*

Table 2. Precision, recall and F1 scores for all scanners available in the MIDOG train and test set. NA indicates ”Not Available”, as these scanners were either not available in the test or train set. Note that the validation scores for the Leica GT450 scanner have an asterisk, as this scanner was not annotated in the training data, but we used our GAN approach to evaluate the annotated validation set transformed to the Leica GT450 domain.

	Test			Validation		
Scanner	Precision	Recall	F1	Precision	Recall	F1
Hamamatsu XR	0.669	0.572	0.617	0.618	0.871	0.723
Leica GT450	0.693	0.690	0.692	0.798*	0.825*	0.812*
3DHistech P1000	0.851	0.696	0.766	NA	NA	NA
Hamamatsu RS	0.669	0.572	0.617	NA	NA	NA
Hamamatsu S360	NA	NA	NA	0.775	0.968	0.861
Aperio CS2	NA	NA	NA	0.860	0.804	0.831

some detection algorithm. For this reason, we studied the effect on domain generalizability of adding either instance segmentation for the annotated mitoses or domain-specific GAN augmentation, shown in Table 1. As the MIDOG data does not have a separate test set available to evaluate generalizability for different algorithm variants, we used our GAN domain augmentation to transform our validation set to resemble the non-annotated scanner 4 (Leica GT450). We observe that the ”F1 val” score is similar for all experiments regardless of model or augmentation strategy, indicating that for in-training domains there is no significant effect of adding instance masks or domain-augmentation. However, for ”F1 val S4” we found that just adding the instance masks for scanners 1-3 already improved generalizability to simulated scanner 4, going from F1 score 0.493 (Exp 3) to 0.705 (Exp 5). Moreover, we see that adding GAN augmentation improves F1 score for F-RCNN, going from F1 score 0.493 (Exp 3) to 0.815 (Exp 4). The same observation is true for Mask-RCNN, going from F1 score

0.705 (Exp 5) to 0.813 (Exp 6). We note that adding the GAN augmentation seems to obviate the benefit of adding masks (F-RCNN versus mask-RCNN), but we chose to submit the Mask-RCNN approach nonetheless. We note, however, that we don't have access to the test set for this ablation study so don't know if our findings on the simulated validation set generalize to the test set.

Finally, we compare the performance of our algorithm between the train, validation, and test set in Table 2. As is expected, our validation scores are always higher than the test scores. Interestingly enough our algorithm generalizes better to an unseen scanner (3DHistech) than a scanner that was actually in the training dataset (Hamamatsu XR), though we note that this scanner also performs worst among the four train scanners in validation. The Leica GT450 scanner, for which we explicitly used our GAN domain augmentation during training, performs second-best in test, suggesting our approach indeed enhanced the model's generalizing properties to this domain.

On the preliminary test set, it was interesting that the MIDOG reference approach [16], which used a RetinaNet with domain adversarial training, was already among the top competitors on the leaderboard. The computational benefit of domain adversarial training over domain-specific GAN augmentation is that it is not necessary to train a cycle-GAN or transform any of the training images. On the other hand, the GAN augmentation can be used for any network architecture without having to choose where to plug in the domain adversarial loss during training - something that the reference approach had to experiment with. It is a subject of future work which of these approaches provides the best domain generalizability.

As for the training of the Residual Cycle-GAN, we note that visually the results illustrated in Figure 1 seem convincing, but the color transformation is not always completely consistent between different frames. As is typical of GANs, it is hard to know exactly when to stop training, and it is hard to assess how these color variations impact the final mitosis detection performance.

In conclusion, while the winning approaches in the MIDOG challenge were different, it seems that injecting more information into the mitosis detection problem improves the final detection performance. It would be interesting to see how using self-supervised contrastive learning as pretraining [5], instead of ImageNet pretraining, could further improve the mitosis detection performance of any approach.

References

[1] Marc Aubreville et al. "MItosis DOmain Generalization Challenge (MIDOG)". In: *24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2021). DOI: `10.5281/zenodo.4573978`.

[2] Marc Aubreville et al. "Quantifying the Scanner-Induced Domain Gap in Mitosis Detection". In: *Medical Imaging with Deep Learning (MIDL)* (2021).

[3] Thomas de Bel et al. “Residual cyclegan for robust domain transformation of histopathological tissue slides”. In: *Medical Image Analysis* 70 (2021), p. 102004.

[4] Christof A Bertram et al. “Are pathologist-defined labels reproducible? Comparison of the TUPAC16 mitotic figure dataset with an alternative set of labels”. In: *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Springer, 2020, pp. 204–213.

[5] Ozan Ciga, Tony Xu, and Anne Louise Martel. “Self supervised contrastive learning for digital histopathology”. In: *Machine Learning with Applications* (2021), p. 100198.

[6] Khrystyna Faryna, Jeroen van der Laak, and Geert Litjens. “Tailoring automated data augmentation to H&E-stained histopathology”. In: *Medical Imaging with Deep Learning*. 2021.

[7] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[8] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE ICCV*. 2017, pp. 2961–2969.

[9] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[10] Tasleem Kausar et al. “SmallMitosis: Small Size Mitotic Cells Detection in Breast Histopathology Images”. In: *IEEE Access* 9 (2020), pp. 905–922.

[11] *MITOS14 Challenge*. 2014. URL: `https://mitos-atypia-14.grand-challenge.org/`.

[12] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: `1506.01497 [cs.CV]`.

[13] Ludovic Roux et al. “Mitosis detection in breast cancer histological images An ICPR 2012 contest”. In: *Journal of pathology informatics* 4 (2013).

[14] Mitko Veta et al. “Assessment of algorithms for mitosis detection in breast cancer histopathology images”. In: *Medical image analysis* 20.1 (2015), pp. 237–248.

[15] Mitko Veta et al. “Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge”. In: *Medical image analysis* 54 (2019), pp. 111–121.

[16] Frauke Wilm, Katharina Breininger, and Marc Aubreville. “Domain Adversarial RetinaNet as a Reference Algorithm for the MItosis DOmain Generalization (MIDOG) Challenge”. In: *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD* (2021).

[17] Frauke Wilm et al. “Influence of Inter-Annotator Variability on Automatic Mitotic Figure Assessment”. In: *Bildverarbeitung für die Medizin 2021*. Springer, 2021, pp. 241–246.