On the Effectiveness of Generative Adversarial Networks as HEp-2 Image Augmentation Tool

Majtner, Tomas; Bajic, Buda; Lindblad, J; Sladoje, N; Blanes-Vidal, Victoria; S. Nadimi, Esmaeil

Published in:
Image Analysis

DOI:
10.1007/978-3-030-20205-7_36

Publication date:
2019

Document version:
Accepted manuscript

Citation for pulished version (APA):
Majtner, T., Bajic, B., Lindblad, J., Sladoje, N., Blanes-Vidal, V., & S. Nadimi, E. (2019). On the Effectiveness of Generative Adversarial Networks as HEp-2 Image Augmentation Tool. In M. Felsberg, P.-E. Forssén, I.-M. Sintorn, & J. Unger (Eds.), *Image Analysis: Proceedings of the 21st Scandinavian Conference, SCIA 2019* (pp. 439-451). Springer. https://doi.org/10.1007/978-3-030-20205-7_36

On the Effectiveness of Generative Adversarial Networks as HEp-2 Image Augmentation Tool

Tomáš Majtner[1], Buda Bajić[2], Joakim Lindblad[3,4], Nataša Sladoje[3,4], Victoria Blanes-Vidal[1], and Esmaeil S. Nadimi[1]

[1] Group of Machine Learning and AI, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
[2] Faculty of Technical Sciences, University of Novi Sad, Serbia
[3] Centre for Image Analysis, Department of Information Technology, Uppsala University, Sweden
[4] Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Serbia
Contact email: tomaj@mmmi.sdu.dk

Abstract. One of the big challenges in the recognition of biomedical samples is the lack of large annotated datasets. Their relatively small size, when compared to datasets like ImageNet, typically leads to problems with efficient training of current machine learning algorithms. However, the recent development of generative adversarial networks (GANs) appears to be a step towards addressing this issue. In this study, we focus on one instance of GANs, which is known as deep convolutional generative adversarial network (DCGAN). It gained a lot of attention recently because of its stability in generating realistic artificial images. Our article explores the possibilities of using DCGANs for generating HEp-2 images. We trained multiple DCGANs and generated several datasets of HEp-2 images. Subsequently, we combined them with traditional augmentation and evaluated over three different deep learning configurations. Our article demonstrates high visual quality of generated images, which is also supported by state-of-the-art classification results.

Keywords: Deep learning · Image recognition · HEp-2 image classification · GAN · CNN · GoogLeNet · VGG-16 · Inception-v3 · Transfer learning.

1 Introduction

Human Epithelial (HEp-2) cells are commonly used in the Indirect Immunofluorescence (IIF) tests to detect autoimmune diseases. Nowadays, the evaluation of IIF test is done mostly by humans and therefore it is a subjective method too dependent on the experience of the physician. Usually, two or three specialists need to analyze patients' specimen images via fluorescence microscopes and vote to decide the staining patterns. Thus, computer-aided systems aim to assist doctors with the diagnosis by automatic classification of HEp-2 images.

A number of automated methods addressing the problem of cell staining pattern recognition have been proposed in the literature. Many of them are the result of the HEp-2 cell classification contests [7, 12, 13], where datasets of samples were made publicly available for method evaluation. While most of the research groups at the time of these competitions still approached the problem by using methods based on extracting so-called *hand-crafted features* for pattern discrimination, nowadays the deep convolutional neural networks (also known as CNNs) are used almost exclusively [2, 9, 16, 21].

To train a successful deep neural network, a large amount of training images is required. It is typically very difficult to collect and label biomedical images due to the lack of experts' time and the cost of imaging devices. It is therefore common to increase the number of training samples by various methods of image augmentation. For HEp-2 images, the flipping operation and the rotation around the central image point are the most common approaches [2, 9, 16, 21].

Our paper investigates an alternative method for data augmentation by utilizing Generative Adversarial Network (GAN). This method has been demonstrated to be a powerful technique to perform an unsupervised generation of new synthetic images with the visual appearance of the real ones. We have employed deep convolutional GAN (DCGAN) [19] for this particular purpose. Our motivation is supported with the fact that the original DCGAN architecture was demonstrated to be stable for images of size 64×64, which is very close to the average size of HEp-2 cells images. The comparison of different augmentation techniques is done using the transfer learning framework. We compare the performances of fine-tuned GoogLeNet, VGG-16, and Inception-v3 with augmented data obtained by traditional methods and by utilization of DCGAN.

The next section of the article presents the current state-of-the-art in HEp-2 image recognition and development of GANs. Subsequently, we describe the dataset used in this article and our methods of preprocessing and augmentation of the images. The last sections are dedicated to evaluation together with presentation and discussion of experiments and results, where we demonstrate the effectiveness of our solution.

2 Related Work

The recent progress of pattern recognition techniques for IIF image analysis has been covered by a special issue of Pattern Recognition Letters [11]. Novel techniques, including those examining the role of Gaussian Scale Space theory as a pre-processing approach [18], a superpixel based classification method calculating the sparse codes of image patches [6], a multi-process system based on ensemble of 15 support vector machines [4], and many others, were introduced.

Even more recently, Gao et al. [9] analyzed the impact of hyper-parameter settings of proposed fully-connected CNN on the classification accuracy. The influence of several preprocessing techniques on HEp-2 image classification has been studied by Bayramoglu et al. [2]. Shen et al. were focusing on a very deep residual network for HEp-2 pattern classification [21] and some authors tried

simultaneous cell segmentation and classification by utilizing proposed residual network [16]. All of these papers are focusing on very specific problems but none of them deals with comparison of various augmentation methods, which is the main focus of our paper.

GANs, a class of neural networks, were introduced in 2014 [10]. They typically consist of two CNNs - the generator and the discriminator, which compete with each other in a zero-sum game. The role of the generator is to produce random samples that look like real images, while the role of the discriminator is to correctly classify and recognize these generated images. GANs have been successfully used for biomedical imaging tasks including the image synthesis and classification [25], and also for medical segmentation [3].

In the context of automated analysis of HEp-2 images, GANs were used only for segmentation task [15], while for the HEp-2 images classification there are no peer-reviewed publications focusing on exploring the possibilities. Our article aims at filling this gap with an extensive comparison over three different network configurations.

3 Dataset

In this article, we are using publicly available dataset of HEp-2 images, which was also previously used for benchmarking [13]. The entire dataset contains 13,596 pre-segmented and annotated cell images with their ground truth classes. It utilizes 419 unique positive sera extracted from 419 randomly selected patients. The specimens, one for each patient serum, were automatically photographed using a monochrome high dynamic range cooled microscopy camera. The image dataset is divided into six categories: Centromere (Ce), Golgi (Go), Homogeneous (Ho), Nucleolar (Nu), Nuclear Membrane (Nm), and Speckled (Sp). See the top most part of Fig. 1 for illustration.

Since there are no official independent publicly available test samples, some researchers opt for N-fold cross-validation over the all available images to evaluate the performance of their algorithms. However, this approach is criticized from statistical point of view [1] and it leads to biased results, where the performance tends to drop significantly when the algorithm is applied on new, previously unseen data. Therefore, we use a holdout validation approach on the available part of the dataset. We randomly partitioned the dataset into 70 % for training, 10 % for validation, and 20 % for testing. The validation part is used to evaluate the performance during the training of deep learning, whereas independent testing part is used at the very end to report the final performance. The total number of images in each class, before any form of augmentation, is summarized in Table 1.

4 Proposed Method

When we look at the entire dataset, the average size of an image is 68.75×68.73 pixels with a standard deviation of 6.32 and 6.19 pixels, respectively. For comparison purposes, all images were resized to the same size of 64×64 pixels

Table 1: The division of images before augmentation of the training part of the dataset.

	Ce	Go	Ho	Nu	Nm	Sp	Total
Training	1,918	506	1,745	1,819	1,546	1,981	9,515
Validation	274	72	249	259	220	283	1,357
Testing	549	146	500	520	442	567	2,724
Total	2,741	724	2,494	2,598	2,208	2,831	13,596

using bicubic interpolation. Since the brightness and contrast of the images vary a lot, we employed normalization of image intensities. The intensity adjustment was performed by linear stretching, where 1 % of the pixels are saturated at low and at high end of the intensity range in order to maximize the contrast. The following two subsections describe the two forms of augmentation employed for the training images in this study. The version of training dataset without any form of augmentation is further referred to as *original*.

4.1 Augmentation by Rotation and Flipping

There are multiple different forms of augmentation, where their usability is typically subject to the nature of the data. Since we are working with pre-segmented cell images that were acquired using the same microscope settings, the samples are centered and have the same resolution. Therefore, augmentation by shifting or zooming is not appropriate here. On the other hand, the most common and natural technique to augment these biomedical datasets is to use image rotation around the image center. We rotated each image by 90°, 180°, and 270°, which, together with the flipping operation, results in seven unique images generated out of each original input.

The original dataset is unbalanced, with one class (Golgi) having $3-4\times$ lower number of images than the remaining five classes (see Table 1). We therefore additionally rotated each Golgi image by angles of size $23° \times i$, where $i \in \{1, 2, 3\}$. After adding three more rotations, Golgi class reached similar number of images (4×506) as the remaining classes. In this augmentation step, rotated images are first cropped to the size of the largest rectangle within the input image and later resized back to the size of 64×64. The bicubic interpolation is used in both cases. The training part of the dataset derived by this sequence of steps is further referred to as *rotated*. The problem of unbalanced classes is addressed in literature by different approaches, e.g., by using RUSBoost [20] approach to alleviating class imbalance. These methods, however, usually follow the strategy of under-sampling the majority class or classes, which is not optimal in this study, where we have one minority class.

In addition, we also wanted to examine the effect of even stronger augmentation by adding more image rotations. Therefore, we created another version of training dataset, where each image from the *rotated* dataset is further rotated by 45°. This leads to doubling the number of training samples. Also here, the

images are cropped and resized in the same fashion as previously described for Golgi class. This version of training dataset is further referred to as $rotated_{+45^\circ}$. The exact sizes of both *rotated* and $rotated_{+45^\circ}$ datasets are specified in Table 3.

4.2 Augmentation by Generative Adversarial Networks

As aforementioned, we use the DCGAN [19] to generate more HEp-2 samples for increasing the size of the training dataset. The authors of DCGAN introduced several techniques for successful learning: converting the max-pooling layers to convolution layers, converting the fully connected layers to global average pooling layers in the discriminator, using batch normalization layers in the generator and the discriminator, and using leaky ReLU activation functions in the discriminator. In their configuration, a 100 dimensional uniform distribution is projected to a small spatial extent convolutional representation. Subsequently, the series of four fractionally-strided convolutions convert the representation into a 64×64 pixel image. For more details about the network configuration, we refer the reader to the original paper introducing DCGAN [19].

For application of this approach to the HEp-2 images, we train individual DCGAN for each of the six classes. In total, two different training scenarios are followed. In the first one, we use the *original* dataset to train the DCGANs, while in the second one, we use the *rotated* dataset. To distinguish between images generated from GANs trained on *original* dataset and those generated from GANs trained on *rotated* dataset, we use the subscript *rot* for the latter version, i.e., we refer to these datasets as *generated* and $generated_{rot}$, respectively. The motivation is to test the influence of larger and already pre-augmented dataset by rotation and flipping on the quality of generated images via DCGANs. All our models are trained with mini-batch stochastic gradient descent with a mini-batch size of 128. All weights are initialized from a zero-centered normal distribution with standard deviation 0.02. The learning rate is set to 0.0002 and we train all models for 300 epochs. Fig. 1 illustrates both versions of generated datasets.

Since there is no limit in the number of derived images using DCGANs, we use this fact to create also the perfectly balanced classes. In this scenario, we start from the *rotated* set, however, we did not use the additional rotation of Golgi class, where bicubic interpolation and resizing was needed. Therefore, each image from *original* set is only rotated by 90°, 180°, and 270° and flipped, which leads to higher unbalance between classes than in previous scenarios. We subsequently use generated images to fill up those classes having lower number of samples than the most populated class, the Speckled class. The new datasets created using this approach are further referred to as *balanced* and $balanced_{rot}$.

Lastly, we create two more datasets that match the number of images in $rotated_{+45^\circ}$. We start with *rotated* dataset here and instead of employing additional rotations that were used to create $rotated_{+45^\circ}$ set, we utilize images generated from GANs to match the number of samples in $rotated_{+45^\circ}$. These new datasets are referred as *rotated&generated* and $rotated\&generated_{rot}$, depending on the type of images used to train GANs. The overview of all created training datasets is in Table 2 and the summary of their exact size is in Table 3.

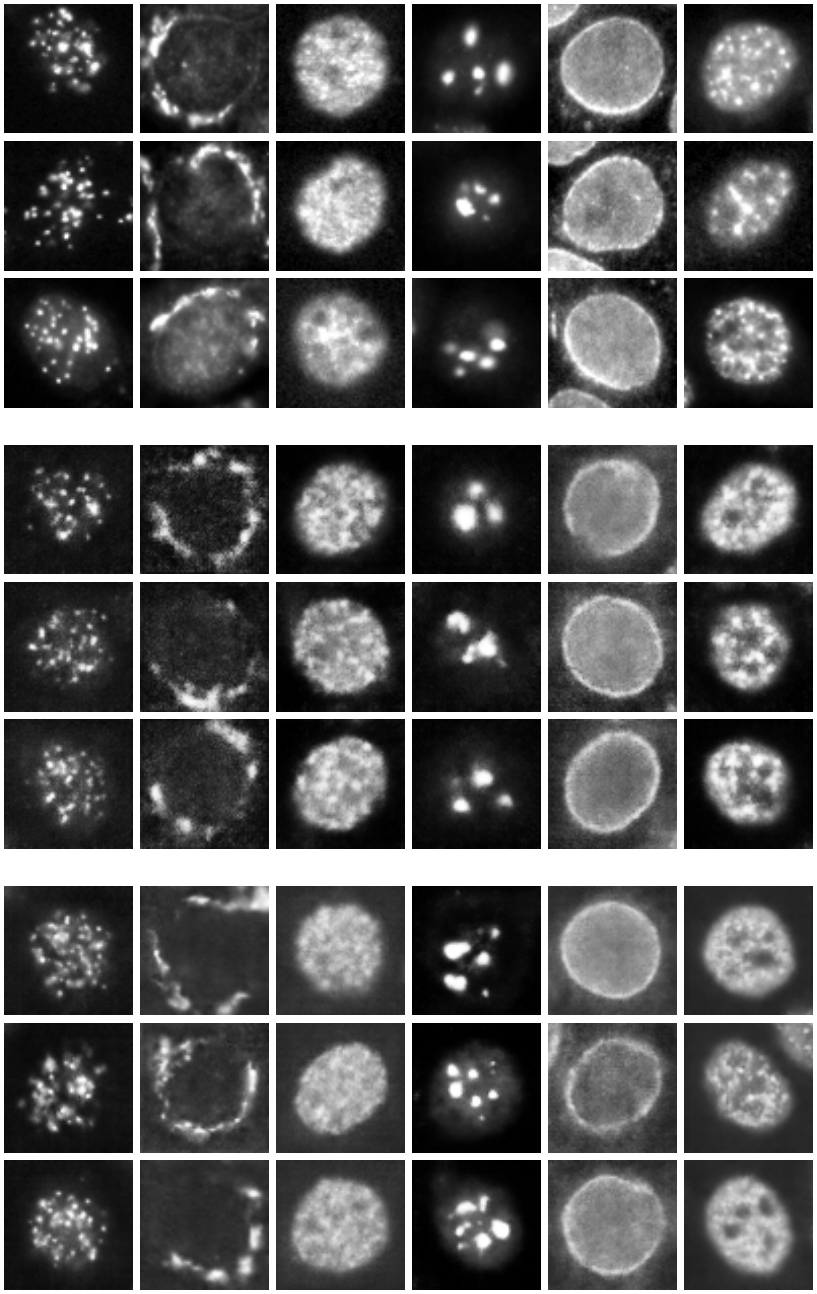

Fig. 1: Examples of *original* HEp-2 images (first three rows), images **generated by DCGAN** from *original* dataset (three middle rows), and images **generated by DCGAN** from *rotated* dataset (last three rows). Each column represents a different image class, in order: Ce, Go, Ho, Nu, Nm, Sp.

Table 2: The brief overview of all created training datasets. In *balanced* and $balanced_{rot}$ datasets, we eliminated the additional rotation of Golgi class.

	Description
original	original data, no augmentation
rotated	each image flipped and rotated, additional rotation for Golgi class
generated	GANs trained on *original*, equal output size as *rotated*
$generated_{rot}$	GANs trained on *rotated*, equal output size as *rotated*
balanced	*rotated* dataset perfectly balanced using GANs trained on *original*
$balanced_{rot}$	*rotated* dataset perfectly balanced using GANs trained on *rotated*
$rotated_{+45°}$	*rotated* dataset plus additional rotation by 45°
rotated&generated	*rotated* dataset plus *generated* dataset
$rotated\&generated_{rot}$	*rotated* dataset plus $generated_{rot}$ dataset

Table 3: The total number of images in different versions of the training dataset after various forms of augmentation. In *balanced* and $balanced_{rot}$ datasets, we eliminated the additional rotation of Golgi class. Therefore, balanced classes have lower number of samples than the maximum of *rotated* dataset.

	Ce	Go	Ho	Nu	Nm	Sp
original	1,918	506	1,745	1,819	1,546	1,981
rotated, *generated*, $generated_{rot}$	15,344	16,192	13,960	14,552	12,368	15,848
balanced, $balanced_{rot}$	15,848	15,848	15,848	15,848	15,848	15,848
$rotated_{+45°}$, *rotated&generated*, $rotated\&generated_{rot}$	30,688	32,384	27,920	29,104	24,726	31,696

5 Evaluation

In the experimental part, we are using three different pretrained convolutional neural networks, namely GoogLeNet [23], VGG-16 [22], and Inception-v3 [24]. All three networks were pretrained on ImageNet [5]; we perform fine-tuning, also known as the transfer learning [26], to adjust them for HEp-2 image recognition. This implies that, for all three networks, we replace their last three layers with a fully-connected layer, a softmax layer, and a classification layer, which classifies images directly to the six categories of HEp-2 images.

For this fine-tuning, we utilize stochastic gradient descent with momentum optimizer, initial learning rate of 0.001, and a mini-batch size of 32 images. All the networks are trained for 50 epochs, to be sure that the training is stabilized (see the stable curves in Fig. 2 with almost no fluctuations at the end). Images are resized to appropriate input size for each network separately. All tests are performed using MATLAB R2018b. During the training, we validate the performance using an independent validation dataset and at the end, the final version of each model is evaluated using the test dataset. For illustration of the development of training process, Fig. 2 depicts accuracy and loss for VGG-16 network trained on *generated* dataset.

Evaluation of classification performance is performed using two different metrics. The first one is the overall accuracy (OA), defined as the overall correct classification rate of all images. In some previous works on HEp-2 image recog-

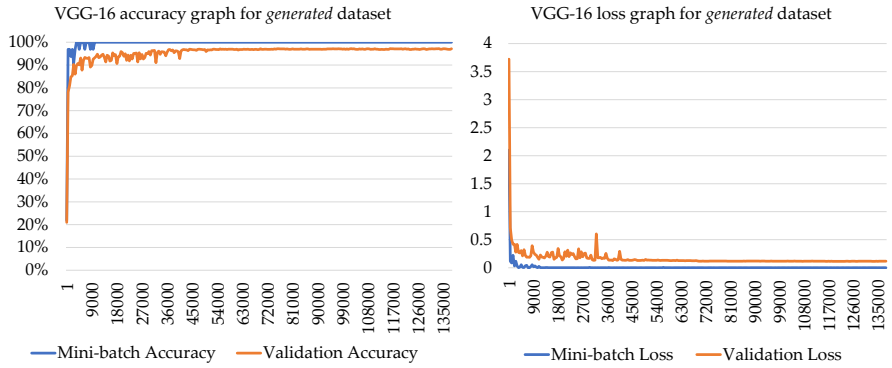

Fig. 2: Accuracy (left) and loss (right) for VGG-16 network during its training on *generated* dataset. The number of iterations is displayed on the x-axis.

nition, this metric is also known as the average classification accuracy (ACA). The second one, the mean class accuracy (MCA), is defined as

$$MCA = \frac{1}{K}\sum_{k=1}^{K} CCR_k \tag{1}$$

where CCR_k is the classification accuracy of a particular cell class k and K is the number of cell classes.

6 Results and Discussion

The comparison of all tested variants is summarized in Table 4 and the overall accuracy is also plotted in Fig. 3. From the results we can see that already the performance on the *original* dataset is relatively high, which confirms the quality of our preprocessing and the appropriate choices of deep learning techniques. The performance on the *generated* and $generated_{rot}$ datasets is lower, when compared to corresponding *rotated* dataset (see also Table 5 for confusion matrices of *generated* and *rotated* versions). Despite of the very good visual appearance of generated images (see Fig. 1), their standalone classification performance is not as convincing.

This result confirms the observation made by Perez and Wang [17] for real-world images. They also concluded that GANs do not perform better than traditional augmentations. However, there is still a potential in combining them together, as was shown by Frid-Adar et al. [8] for liver lesion classification, where inclusion of the GAN-based augmentation does help. Our results for VGG-16 and also for Inception-v3 support this conclusion also for HEp-2 images, since we observe a slight increase in accuracy achieved by combining rotated and generated images during the training.

Table 4: The comparison of performances of all three network configurations on all derived training datasets for both tested metrics. In the table, G-net stands for GoogLeNet, V-net stands for VGG-16, and I-net stands for Inception-v3. Presented values are in %.

	OA			MCA		
	G-net	V-net	I-net	G-net	V-net	I-net
original	96.84	96.26	95.99	97.10	96.45	96.04
generated	95.96	96.33	96.48	95.94	96.49	96.39
$generated_{rot}$	96.33	96.62	96.26	96.40	96.68	96.07
balanced	97.98	98.13	98.20	98.17	98.31	98.17
$balanced_{rot}$	98.31	98.24	98.49	98.44	98.22	98.53
rotated&generated	98.27	97.91	98.38	98.41	97.94	98.36
$rotated\&generated_{rot}$	98.35	**98.27**	**98.60**	98.47	**98.34**	98.55
$rotated_{+45°}$	**98.60**	97.72	98.42	98.61	97.76	98.38
rotated	**98.60**	98.13	98.49	**98.71**	98.30	**98.62**

Table 5: GoogLeNet confusion matrices for *generated* and *rotated* versions of training dataset. Presented values are in %.

Generated version

	Ce	Go	Ho	Nu	Nm	Sp
Ce	98.00	0.36	0.00	0.36	0.18	1.10
Go	0.00	95.22	0.68	2.74	0.68	0.68
Ho	0.60	0.60	95.00	0.20	0.60	3.00
Nu	0.96	0.96	0.77	95.58	0.38	1.35
Nm	0.00	0.68	0.68	0.23	98.18	0.23
Sp	2.65	0.18	2.12	1.05	0.35	93.65

Rotated version

	Ce	Go	Ho	Nu	Nm	Sp
Ce	98.89	0.19	0.00	0.73	0.00	0.19
Go	0.00	99.32	0.00	0.68	0.00	0.00
Ho	0.20	0.20	98.80	0.40	0.00	0.40
Nu	0.38	0.58	0.19	97.89	0.38	0.58
Nm	0.00	0.00	0.23	0.45	99.32	0.00
Sp	0.00	0.00	0.53	0.88	0.53	98.06

We also observe that the versions with the subscript *rot* have generally slightly higher performance than their corresponding variants without this subscript. This indicates importance of the amount and variability of training samples for performance of DCGANs, as well as an effect that the quality of training data has on the resulting quality of generated images. Finally, both balanced datasets exhibit slightly lower performance. However we note that the original dataset is relatively balanced, with only one class with lower number of training samples, which is primarily compensated by additional rotations. As a result, the effect of perfect class balancing does not turn out to be important.

To provide a look at the HEp-2 image classification from a broader perspective, we present the comparison of our top performing approach with the methods from the literature in Table 6. To enable a fair comparison, we include only the methods using the same, or almost the same, split technique for training and test datasets as we did. Table 6 suggests that we share the top position together with the Shen et al. [21], depending on the choice of metric used for

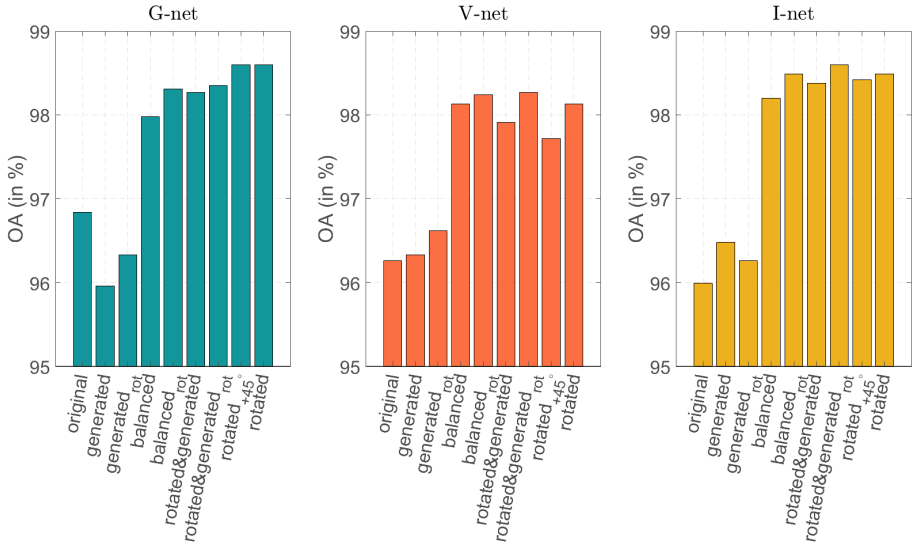

Fig. 3: The overall accuracy (OA) graphs for GoogLeNet (G-net), VGG-16 (V-net), and Inception-v3 (I-net).

Table 6: Comparison with other approaches on the same dataset and with the same division of publicly available part of HEp-2 images. Presented values are in %.

	OA	MCA
Kastaniotis et al. [14]	–	93.6
Gao et al. [9]	97.24	96.76
Shen et al. [21]	**98.82**	98.62
Our top performing method	98.60	**98.71**

evaluation. Shen et al. [21] proposed a deep cross residual network (DCRNet) for HEp-2 cell classification and their method is the winner of the most recent HEp-2 image recognition contest, with achieved accuracy which exceeds all of the top performers in the previous contests. Our solution is based on transfer learning and we used slightly less images for training (70 % vs. 80 %), when compared to their presented solution.

7 Conclusion

In this article, we compare and discuss augmentation techniques for HEp-2 images for their classification. We evaluate the usage of the recently proposed DCGAN and we observe that these type of networks are capable of producing very realistically looking images of HEp-2 cells. However, application of DCGAN for classification purposes does not lead to convincing results, in particular when the

generated images are used independently, without the combination with original ones. This result is not surprising and it supports the conclusions from the similar comparison performed in a different image domain [17]. The potential of combining generated and rotated images is, however, still interesting, as is also demonstrated by our results, especially for the VGG-16 and Inception-v3 network configurations.

For future work, we would like to focus on further improvement of the quality of the generated dataset by an external measure. There is a possible problem of large intra-class variance, which was not discussed and covered in this work and which could lead to the low quality of synthetic images. Despite some of the weak performances presented here, we still see a potential of GANs in biomedical and medical domain for helping to address the problem of small annotated datasets.

Acknowledgement This research was supported by a research grant from the University of Southern Denmark, Odense University Hospital, Danish Cancer Society, and Region of Southern Denmark through the Project EFFICACY and by Swedish Research Council through Projects 2015-05878 and 2017-04385, Sweden's Innovation Agency VINNOVA through Project 2017-02447, and Ministry of Education, Science and Technological Development of the Republic of Serbia through Projects ON174008 and III44006.

References

1. Babyak, M.: What you see may not be what you get: a Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. Psychosomatic Medicine **66**(3), 411–421 (2004)
2. Bayramoglu, N., Kannala, J., Heikkilä, J.: Human Epithelial Type 2 Cell Classification with Convolutional Neural Networks. In: 15th Int. Conf. on Bioinformatics and Bioengineering. pp. 1–6. IEEE (2015)
3. Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D., Hernández, M., Wardlaw, J., Rueckert, D.: GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks. arXiv preprint arXiv:1810.10863 (2018)
4. Cascio, D., Taormina, V., Cipolla, M., Bruno, S., Fauci, F., Raso, G.: A Multi-Process System for HEp-2 Cells Classification Based on SVM. Pattern Recognition Letters **82**, 56–63 (2016)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
6. Ensafi, S., Lu, S., Kassim, A.A., Tan, C.: Accurate HEp-2 Cell Classification Based on Sparse Coding of Superpixels. Pattern Recognition Letters **82**, 64–71 (2016)
7. Foggia, P., Percannella, G., Soda, P., Vento, M.: Benchmarking HEp-2 Cells Classification Methods. IEEE Trans. on Medical Imaging **32**(10), 1878–1889 (2013)
8. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification. arXiv preprint arXiv:1803.01229 (2018)

9. Gao, Z., Wang, L., Zhou, L., Zhang, J.: HEp-2 Cell Image Classification with Deep Convolutional Neural Networks. IEEE Journal of Biomedical and Health Informatics **21**(2), 416–428 (2017)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
11. Harandi, M., Lovell, B., Percannella, G., Saggese, A., Vento, M., Wiliem, A.: Executable Thematic Special Issue on Pattern Recognition Techniques for Indirect Immunofluorescence Images Analysis. Pattern Recognition Letters **82**, 1–2 (2016)
12. Hobson, P., Lovell, B., Percannella, G., Saggese, A., Vento, M., Wiliem, A.: HEp-2 Staining Pattern Recognition at Cell and Specimen Levels: Datasets, Algorithms and Results. Pattern Recognition Letters **82**, 12–22 (2016)
13. Hobson, P., Lovell, B., Percannella, G., Vento, M., Wiliem, A.: Benchmarking Human Epithelial Type 2 Interphase Cells Classification Methods on a Very Large Dataset. Artificial Intelligence in Medicine **65**(3), 239–250 (2015)
14. Kastaniotis, D., Fotopoulou, F., Theodorakopoulos, I., Economou, G., Fotopoulos, S.: HEp-2 cell classification with vector of hierarchically aggregated residuals. Pattern Recognition **65**, 47–57 (2017)
15. Li, Y., Shen, L.: cC-GAN: A robust transfer-learning framework for HEp-2 specimen image segmentation. IEEE Access **6**, 14048–14058 (2018)
16. Li, Y., Shen, L., Yu, S.: HEp-2 specimen image segmentation and classification using very deep fully convolutional network. IEEE Transactions on Medical Imaging **36**(7), 1561–1572 (2017)
17. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621 (2017)
18. Qi, X., Zhao, G., Chen, J., Pietikäinen, M.: HEp-2 Cell Classification: The Role of Gaussian Scale Space Theory as a Pre-processing Approach. Pattern Recognition Letters **82**, 36–43 (2016)
19. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
20. Seiffert, C., Khoshgoftaar, T., Van Hulse, J., Napolitano, A.: RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **40**(1), 185–197 (2010)
21. Shen, L., Jia, X., Li, Y.: Deep Cross Residual Network for HEp-2 Cell Staining Pattern Classification. Pattern Recognition **82**, 68–78 (2018)
22. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556 (2014)
23. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)
24. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
25. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: A review. arXiv preprint arXiv:1809.07294 (2018)
26. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems. pp. 3320–3328 (2014)