ISING-GAN: ANNOTATED DATA AUGMENTATION WITH A SPATIALLY CONSTRAINED GENERATIVE ADVERSARIAL NETWORK

P. Dimitrakopoulos[1], *G. Sfikas*[1,2], *C. Nikou*[1]

[1]Dept. of Computer Science & Engineering, University of Ioannina, Greece
[2] CERTH/ITI, Thessaloniki, Greece

ABSTRACT

Data augmentation is a popular technique with which new dataset samples are artificially synthesized to the end of assisting training of learning-based algorithms and avoiding overfitting. Methods based on generative adversarial networks (GANs) have recently rekindled interest in research on new techniques for data augmentation. With the current paper we propose a new GAN-based model for data augmentation, comprising a suitable Markov random field-based spatial constraint that encourages synthesis of spatially smooth outputs. Oriented towards use with medical imaging sets where a localization/segmentation annotation is available, our model can simultaneously also produce artificial annotations. We gauge performance numerically by measuring performance through U-Net trained to detect cells on microscopy images, by taking into account the produced augmented dataset. Numerical trials, as well as qualitative results validate the contribution of our model.

***Index Terms*—** Data augmentation, Ising model, Generative Adversarial Networks, Markov Random Field, Microscopy imaging

1. INTRODUCTION AND RELATED WORK

One of the biggest issues facing the use of machine learning in medical imaging is the lack of availability of large datasets. The annotation of medical images is expensive and time consuming, and typically requires expert medical knowledge to be performed accurately. The limited amount of training data can dramatically affect the performance of deep neural networks which often need a very large amount of data on which to train in order to avoid overfitting. "Traditional" data augmentation methods work by applying a random parameterized transform on the available data, such as simple affine transforms (translation, rotation, scaling, horizontal shearing), or non-rigid transforms such as elastic deformations [1]. Other schemes include patch extraction and channel intensity permutation [2]. An important point is that the type of transforms that may be suitable always depends on the dataset and/or related inference problem; for example, flipping inputs may be a suitable strategy for a natural image dataset, but not suitable for a word image dataset.

GANs are a powerful class of generative models, the training of which can be viewed as a two-player game between two neural networks, named the generator and the discriminator [3]. Models that use image-to-image translations GANs [4, 5] or models closest to the originally proposed (noise to image) GANs [6, 7] have started been used recently for data augmentation with success. Data augmentation with GANs has been used in medical imaging applications in a number of recent works. Methods for generating synthetic computed tomography images that include liver lesions are presented in [7]. Using Deep Convolutional GANs (DCGANs) and conditional GANs they manage to subsequently improve the performance of medical imaging classification model training on the augmented data. Synthetic CT and FLAIR images were also generated along with an annotation comprising Cerebrospinal Fluid and White Matter Hyperintensity masks respectively, for each image instance [6]. A Progressive Growing of GANs (PGGAN) network has been proposed in [8] to generate synthetic data in two brain segmentation tasks, with which improvements of 1 up to 5 Dice Similarity Coefficient (DSC) units are achieved. A GAN architecture specific to data augmentation has been proposed in [9]. The model's generator network is composed of an encoder taking an input image and projecting it down to a lower dimensional feature vector. A random vector is then transformed and concatenated with the encoder output. The result is passed to a decoder network which generates an augmentation image.

In this work, a method for data augmentation is presented and tested succesfully in the context of augmenting sets of microscopy images. The model is able to produce augmentations of annotated sets, in the sense of simultaneously generating tuples of synthetic images and corresponding segmentation masks that localize an object, tissue or organ of interest [6, 7]. While GAN-based models that are capable of producing annotated images have previously been proposed, in this work we extend the adversarial training loss of GANs with a Markov random field (MRF)-based loss. By using the proposed loss function, we are capable of producing more robust segmentation masks by explicitly requiring them to be locally homogeneous. Finally, in order to evaluate the proposed model and the generated cell images, we have used the performance of a U-Net segmentation model trained with the augmented set as an evaluation metric, akin to [6], as well as the recently proposed Fréchet Inception Distance metric [10]. Numerical, as well as qualitative results show that the proposed MRF-based model indeed produces results superior to GAN architectures not comprising the proposed loss.

The remainder of the paper is structured as follows. In section 2, we present the proposed model. In section 3 we show samples produced with our model, as well as validate our model with extensive numerical trials on a microscopy imaging dataset. Finally, we discuss conclusions and future work in section 4.

2. PROPOSED MODEL

The GAN objective function in its originally proposed form [3] is:

$$L_{\text{GAN}} = E_x \log D(x) + E_z \log(1 - D(G(z))), \qquad (1)$$

where the discriminator network aims to maximize it, while the generator network aims to minimize it. These terms and their optimization objective constitute a two-player game, and training the GAN amounts to finding a Nash equilibrium for the game. The generator

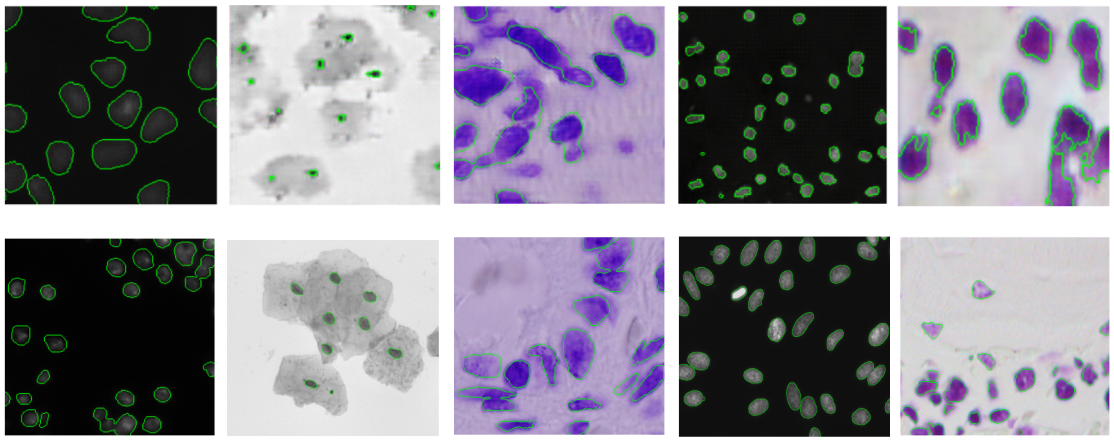

Fig. 1. Indicative images and annotations generated with the proposed Ising-ResGAN model (top row) juxtaposed to real samples (bottom row). Cell segmentation annotation is overlayed as a green border on all images.

takes as input a random vector $z \in \Re^{100}$ sampled from a uniform distribution in $[-1, 1]^{100}$, and outputs a colour cell image and a binary mask with a resolution of 256×256 pixels for both. Variables x stand for data sampled from the true distribution of training images. Both the discriminator and the generator in our model are defined as deep convolutional neural networks. Specifically, the first part of the network consists of a fully connected layer comprising $4 \times 4 \times 512$ neurons, reshaped as a stack of 4×4 feature maps. These are followed by a series of 3 transpose convolutional layers, which gradually upsample (double each axis) the feature map until size 32×32. This map is fed into two sibling branches. The first branch is responsible for generating the synthetic colour image and consists of 3 transpose convolutional layers.The second branch generates the corresponding binary segmentation mask, it has also 3 transpose convolutional layers. We have experimented with two versions of the architecture, concerning including residual connections or not. In the first case, each transpose convolutional layer in both branches is topped by a stack of 2 convolutional layers with stride 1. The output of those stacks is fed to a shortcut connection, which introduces neither extra parameter nor computation complexity. Each upsampling convolutional layer has stride 2 and a 4×4 kernel size. Batch normalization is applied to all layers except for the output.

The discriminator network accepts a 4-channel stack as input, comprising generated RGB channels plus a binary mask. It consists of 7 convolutional layers with a kernel size of 5×5, topped by average pooling, and then followed by a fully connected layer with a sigmoid activation. The sigmoid-activated output corresponds to the probability of whether the input tuple of cell image and its generated mask is perceived by the network as fake or not. Batch-normalization is applied to all layers except for the input and output layers, and all all layers are activated by leaky ReLU units [11] (with leak value set at 0.2) everywhere except the output.

Our model improves on the classical GANs architecture by extending the classical adversarial loss with a Markov Random Field-based loss, applied on the generator output. We define an additional loss term, depending only on the generator output:

$$L_{\text{smooth}}(G(z)) = \sum_{p \in \Omega} \sum_{q \in A(p)} E_{\text{smooth}}(f_p, f_q), \tag{2}$$

where function E_{smooth} assigns non-negative penalties by comparing values f_p and f_q at adjacent pixel positions p and q. Ω represents all values and $A(p)$ is an adjacency function, representing the set of neighbours for position p. We have assumed 4-adjacency for the adjacency function in this work.

We have found that high-pass smoothness terms defined over the cell image easily resulted in over-smoothed outputs of inferior quality. On the other hand, enforcing smoothness almost consistently improved the quality of the produced output. We believe that this happens because the GAN loss can encode the optimal level of object smoothness required on the synthetic image, as the discriminator provides an implicit way to judge how realistic the synthetic image is as a whole. In the case of the mask however, while again the GAN loss by itself can create largely coherent object masks, it does not avoid creating abnormalities such as object holes or producing over-fragmented masks. A smoothness term over the mask explicitly penalizes this case. Therefore, we define the smoothness terms according to an Ising model [12]:

$$E_{\text{smooth}}(f_p, f_q) = \begin{cases} 0 & \text{if } f_p = f_q \\ 1 & \text{if } f_p \neq f_q \end{cases} \tag{3}$$

The proposed objective is hence written as:

$$L_{\text{Ising}} = L_{\text{GAN}} + \lambda L_{\text{smooth}}(G(z)), \tag{4}$$

extending eq. (1), and where the standard GAN loss plays the role of the data term, while λ is a hyperparameter that controls the trade-off between the data and the smoothing term [13]. Implementation-wise, the Ising loss can be approached via a composition of two differentiable operations, namely a convolution and a norm. In particular, we define it as a convolution with a Laplacian kernel followed by a l_1 norm over the result.

3. EXPERIMENTAL RESULTS

We have run experiments on the microscopy imaging dataset BBBC038v1, available from the Broad Bioimage Benchmark Collection [14]. These data were used at the annual 2018 Data Science Bowl competition, hosted on the data science website *kaggle*. The data consist of 729 microscopy images, all annotated with segmentation masks by experts, localizing nuclei in each image. Nuclei

have been treated and imaged in a variety of conditions, including fluorescent and histology stains, several magnifications, and a varying quality of illumination.

As the dataset contains many different modalities, we have run a k-means algorithm on it and subsequently treated each modality separately. We used the following features [15] to cluster the dataset: average intensity, average contrast, texture smoothness, texture uniformity, third moment and entropy in all three colour channels. For the remainder of the paper we refer to the computed clusters as modalities numbered 1 to 5; these comprise $459, 37, 66, 16$ and 85 cell images respectively. Example images from each modality can be viewed in Fig. 1. Treating the whole set at once would have been possible with a conditional GAN; however, as shown in [7], training k separate GANs for k modes gives better results than a single conditional GAN, and we have chosen to follow this approach as well. The training process was done iteratively for the generator and the discriminator, with a mini-batch size of 64. We applied the ADAM optimizer [11], with parameters $\beta_1 = 0.9, \beta_2 = 0.999$ and learning rate 2×10^{-4} for both the generator and the discriminator, and initialized weights with Xavier initialization. The slope of the leak for leaky ReLU units was set to 0.2. We cross-validated the most suitable value for the parameter λ controlling the trade-off between the standard GAN loss and the MRF loss; this was found to be $\lambda = 1.5/(D * N)$, where D stands for total number of output pixels (256×256) and N is the batch size (64). We have trained all GANs with the available training sets augmented with simple rotations (rotations at 90,180,270 degrees and left-right / up-down flips). We have replaced $E_z \log D(1 - G(z))$ instead of $-E_z \log D(G(z))$ for optimizing the generator [3]. Concerning image preprocessing, the only operations involved were scaling to a fixed spatial size (256×256) and transforming to the range $(-1, 1)$.

We have evaluated the proposed model numerically with two different evaluation schemes. Furthermore, we have compared two different network architectures in order to gauge the robustness of the proposed loss over the underlying network, and also have compared the proposed loss over a standard adversarial loss. We also compare our model against classical, non GAN-based data augmentation methods. In the first set of evaluations, we measure the usefulness of the produced images by comparing how effective they are in each case in helping train a separate model for a segmentation task. To this end, we have employed a standard U-Net model [1]. Each U-Net instance was trained for 30 epochs, with an ADAM optimizer and learning rate set to 10^{-4}. Intersection over Union (IoU) segmentation results for trained U-Net over different experimental data augmentation setups are presented in Table 1 (IoU threshold set to 0.5). In all cases, using augmentations provided by the proposed Ising-based GAN models consistently leads to the best U-Net segmentation performance. Following the work in [6], we also tested performance when only a limited subset (20% of initial images picked at random) of the training data is available. In this case too, we can see from the results on Table 1 that again the proposed models fare consistently better. Results are also reported across different estimated modalities in Fig. 2, where again one of the proposed model versions leads to the best scores in all cases.

We have furthermore used the Fréchet Inception distance (FID) to gauge numerically the effectiveness of the proposed model. FID has been shown to be consistent with human evaluation in assessing the realism and variation of GAN-generated samples [10]. FID first uses the Inception-$v3$ network to describe each image as multi-dimensional vector, then compares the statistics of the training images against those of the synthetic images. A lower FID value corresponds to more realistic synthetic images, hence better GAN performance. A Fréchet distance metric is used to compare the two distributions. Assuming that (μ_x, Σ_x) and (μ_g, Σ_g) are the mean and covariance of the true and generated samples respectively, the FID is defined as:

$$FID(x, g) = ||\mu_x - \mu_g||^2 + Tr[\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{\frac{1}{2}}] \quad (5)$$

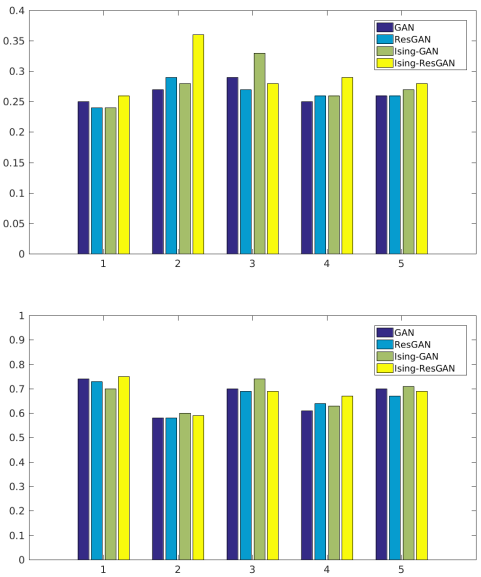

Fig. 2. Comparison of data augmentations using the improvement in performance of a segmentation network (U-Net) as an evaluation measure. IoU segmentation accuracy figures are shown (vertical axis, higher results are better) against estimated modality number (horizontal axis). Tests were run using all original data (bottom plot) or a fraction of it (top plot). The proposed Ising model-based GANs consistently gives the best results.

All models have been trained for 5000 epochs, and the best FID values are reported. Numerical values are shown in Table 2. Results are reported with respect to each estimated modality. Note that in the majority of the modalities, the proposed models (Ising-GAN, Ising-ResGAN) outperform their vanilla GAN versions. Also, in absolute terms either Ising-GAN or Ising-ResGAN scores the best FID in the first three modalities, which combined number approximately 84.7% cell images of the full dataset.

Qualitative results can be seen in figures 1 and 3, where the images produced with the proposed models can be observed to produce visually coherent and convincing synthetic segmentations and cell images.

4. CONCLUSION AND FUTURE WORK

We have presented a new model for data augmentation based on the GAN paradigm, that is suitable to produce synthetic cell images simultaneously with their segmentation maps. The model incorporates an Ising-based smoothing term that forces the synthesized annotations and implicitly their synthesized image counterpart to be visually coherent. Numerical, as well as qualitative results, validate the usefulness of our model. Regarding future work, we envisage exploring other forms of smoothing penalties, including using an edge-

Table 1. Comparison of data augmentations using the improvement in performance of a segmentation network (U-Net) as an evaluation measure. IoU segmentation accuracy figures are shown (higher results are better). Tests were run using all original data (column marked 100%) or a fraction of it (column marked 20%). The proposed Ising model-based GANs have the best performance, in absolute terms (Ising-ResGAN) as well as with respect to their respective vanilla GANs (Ising-GAN vs GAN, Ising-ResGAN vs ResGAN).

% of original data used	20%	100%
No GAN-based augmentation	0.20	0.57
GAN	0.22	0.63
ResGAN	0.24	0.67
Ising-GAN	0.23	0.63
Ising-ResGAN	**0.29**	**0.69**

Table 2. Comparison of GAN performance using Fréchet Inception Distace (FID). FID evalutes the quality of GAN synthetic images versus real images (lower value corresponds to better performance). Results are reported with respect to dataset estimated modalities. Proposed models (Ising-GAN, Ising-ResGAN) lead to best values with respect to their vanilla GAN version in the majority of cases.

Model	Mod.1	Mod.2	Mod.3	Mod. 4	Mod. 5
GAN	**106**	129	260	**275**	98
Ising-GAN	110	**125**	**243**	330	**92**
ResGAN	105	215	146	**248**	**55**
Ising-ResGAN	**88**	**201**	**145**	261	58

preserving prior [16] or conditional random fields [13], defined over the synthesized image and annotation.

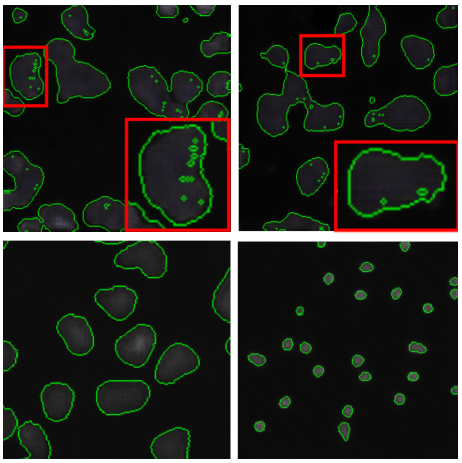

Fig. 3. Comparison of output of vanilla GAN model (top row) versus the proposed model (bottom row). Visible artifacts are prominent on the vanilla GAN output, while the proposed model produces smooth annotations by virtue of the incorporated Ising prior.

5. REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Intl. Conf. on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[2] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NIPS)*, 2012, pp. 1097–1105.

[3] K. Kurach, M. Lucic, X. Zhai, M. Michalski, and S. Gelly, "The GAN landscape: Losses, architectures, regularization, and normalization," *arXiv preprint arXiv:1807.04720*, 2018.

[4] P. Costa, A. Galdran, M. I. Meyer, M. D. Abràmoff, M. Niemeijer, A.M. Mendonça, and A. Campilho, "Towards adversarial retinal image synthesis," *arXiv preprint arXiv:1701.08974*, 2017.

[5] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, "Emotion classification with data augmentation using generative adversarial networks," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 349–360.

[6] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "GAN augmentation: Augmenting training data using generative adversarial networks," *arXiv preprint arXiv:1810.10863*, 2018.

[7] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *arXiv preprint arXiv:1803.01229*, 2018.

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[9] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

[10] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[11] G. Strang, *Linear algebra and learning from data*, Wellesley-Cambridge Press, 2019.

[12] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive GMMRF model," in *IEEE European Conference in Computer Vision (ECCV)*. Springer, 2004, pp. 428–441.

[13] Simon JD Prince, *Computer vision: models, learning, and inference*, Cambridge University Press, 2012.

[14] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, "Annotated high-throughput microscopy image sets for validation," *Nat Methods*, vol. 9, no. 7, pp. 637, 2012.

[15] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, and A. Charchanti, "SIPAKMED: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images," in *IEEE Intl. Conf. on Image Processing (ICIP)*, 2018, pp. 3144–3148.

[16] K. Papadimitriou, G. Sfikas, and C. Nikou, "Tomographic image reconstruction with a spatially varying gamma mixture prior," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 8, pp. 1355–1365, 2018.