Moritz Böhland*, Roman Bruch, Katharina Löffler, and Markus Reischl

Unsupervised GAN epoch selection for biomedical data synthesis

https://doi.org/10.1515/cdbme-2023-1117

Abstract: Supervised Neural Networks are used for segmentation in many biological and biomedical applications. To omit the time-consuming and tiring process of manual labeling, unsupervised Generative Adversarial Networks (GANs) can be used to synthesize labeled data. However, the training of GANs requires extensive computation and is often unstable. Due to the lack of established stopping criteria, GANs are usually trained multiple times for a heuristically fixed number of epochs. Early stopping and epoch selection can lead to better synthetic datasets resulting in higher downstream segmentation quality on biological or medical data. This article examines whether the Fréchet Inception Distance (FID), the Kernel Inception Distance (KID), or the WeightWatcher tool can be used for early stopping or epoch selection of unsupervised GANs. The experiments show that the last trained GAN epoch is not necessarily the best one to synthesize downstream segmentation data. On complex datasets, FID and KID correlate with the downstream segmentation quality, and both can be used for epoch selection.

Keywords: Generative Adversarial Network, Data Synthesis, Segmentation, Computer Vision

1 Introduction

Image segmentation is a crucial task in many biological and biomedical applications, and Neural Networks (NNs) are the state-of-the-art method for cell segmentation [14]. However, supervised training of NNs requires labeled datasets. Manual labeling is time-consuming, tedious, prone to errors, and the inter-observer variability can be high [8]. To overcome these issues, unsupervised (unpaired) Generative Adversarial Networks (GANs) like CycleGAN can be used to synthesize labeled datasets without manual labeling [3, 4]. CycleGAN [18] can be used to transfer images between the unpaired domains of synthetic label images $\mathcal{X}$ and real target microscopy images $\mathcal{Y}$. Doing so enables creation of synthetic paired training data for NNs [2, 3, 12]. CycleGAN consists of two generator networks: One transfers images from domain $\mathcal{X}$ to $\mathcal{Y}$ (G_{XY}), and the other one transfers images from domain $\mathcal{Y}$ to $\mathcal{X}$ (G_{YX}). Two discriminators try to distinguish the real images from the synthetic images in each domain. Training GANs for unsupervised image-to-image translation requires extensive computation and is often unstable. As a result, GANs are trained multiple times for a fixed number of iterations, and the quality of the output is assessed after training by visible inspection or metrics like the FID. Enabling early stopping with metrics usable during GAN training or selection of the best epoch after training can enable scientists to get better results faster.

The Fréchet Inception Distance (FID) [7] and the Kernel Inception Distance (KID) [1] are metrics to quantify the quality of unpaired synthetic images. The FID metric is calculated by extracting features from synthetic images and real target images acquired e.g. by microscopy. Image features are usually extracted with an Inception-v3 network trained on ImageNet [7, 16]. Afterwards, multivariate Gaussians are fitted to the representations of real images and synthetic images in feature space, and the Fréchet distance is used to quantify the distance between both feature space representations. A lower score indicates more similar images. The KID metric measures the maximum mean discrepancy between the feature space representations of real and synthetic images using a polynomial kernel. Also, for KID, a lower score refers to synthetic images being more similar to real images. WeightWatcher (WW) is used to analyze NN training without data [13]. The layers of the network are examined based on the theory of heavy-tailed self-regularization. The analysis is carried out by performing a singular value decomposition of the weight matrix $\boldsymbol{W}$ of a layer and examining the histogram of eigenvalues afterwards. A power law function with the exponent α is fitted to the tail of the histogram. The mean of α over all NN layers can be used to determine whether a network is well-trained. A lower mean α indicates a better NN. It must be evaluated whether the α metric of the CycleGAN generator G_{XY} can potentially be used as an image quality measure.

To the best of our knowledge, the above metrics have not been used for early stopping or epoch selection of unsupervised GANs. Therefore, this article derives a workflow to quantitatively analyze the suitability of metrics available dur-

***Corresponding author: Moritz Böhland,** Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany, e-mail: moritz.boehland@kit.edu

Roman Bruch, Katharina Löffler, Markus Reischl, Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

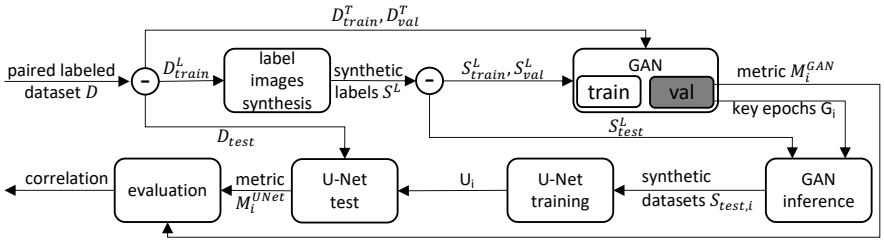


Fig. 1: Workflow to assess the correlation between GAN metrics acquired during training and U-Net segmentation quality metrics. The metrics during GAN training are compared to the downstream U-Net segmentation metrics. The minus sign indicates a split of data.

ing GAN training as stopping criteria and for epoch selection for downstream segmentation tasks. The workflow is applied to a biological and a biomedical dataset, and FID, KID, and α are compared. FID and KID are widely used to assess the quality of synthetic images, while α could potentially be highly useful because it does only rely on network weights.

2 Method

To assess whether the selected metrics are a suitable indicator for epoch selection or early stopping of GANs, the scores of GAN metrics have to be compared to the downstream task metrics. In addition to standard GAN training, metrics are calculated in a validation step after each epoch.

Therefore, we propose the following workflow shown in Figure 1: A paired labeled dataset $D = (D^T, D^L)$ with target images T and labels L is split into training D_{train}, validation D_{val}, and test data D_{test}. The train labels D^L_{train} are used to synthesize unpaired label images S^L. Synthetic label images for cell nuclei can, for example, be generated by randomly placing ellipsoids in an image. The synthetic label images are also split into training, validation, and test images. The GAN is trained with the target images D^T_{train}, D^T_{val} and the unpaired synthetic labels S^L_{train} and S^L_{val}. After each epoch during validation, the GAN metrics M_i^{GAN} for FID, KID, and α are calculated. FID and KID are calculated with D^T_{val} and synthetic target images S^T_{val} synthesized by the GAN from S^L_{val}. The α metric is calculated from the G_{XY} layers without the need for data. For key epochs G_i, model checkpoints are saved and used to infer paired synthetic images $S^T_{\text{test},i}$ from $S^L_{\text{test},i}$. Each paired synthetic dataset $S_{\text{test},i}$ is used to train a U-Net for instance segmentation. Afterwards, each U-Net U_i is evaluated with the advanced Aggregated Jaccard Index (AJI+, $\in [0, 1]$ where 1 is a perfect segmentation) on D_{test} [10]. The U-Net metrics M_i^{UNet} are finally compared to the GAN metrics M_i^{GAN}. A good GAN metric should have a strong correlation with the U-Net segmentation metric. To evaluate whether the GAN epoch with the best GAN metric yields a better downstream performance than the last GAN epoch, the downstream metrics M_i^{UNet} of both can be compared.

3 Results

Datasets

The experiments are carried out on the BBBC039v1 [11] and the Lizard dataset [6]. The BBBC039v1 dataset contains 200 images of U2OS cells. The background is monotonous, and the variation of the cells is limited. Synthetic label images are created by sampling from Elliptic Fourier Descriptors (EFDs) and placing the objects randomly in the images [9]. The train/val/test split is 120/40/40. An example target image with the corresponding label image and an example of a synthetic label image with the preprocessed synthetic label image is shown in Figure 2. We applied preprocessing to label images used by the GAN to ease the learning task. For GAN training, we synthesize 120 training label images and 40 validation label images, whereas, for the U-Net, we synthesize 480 images with a train/val split of 360/120 for each GAN key epoch.

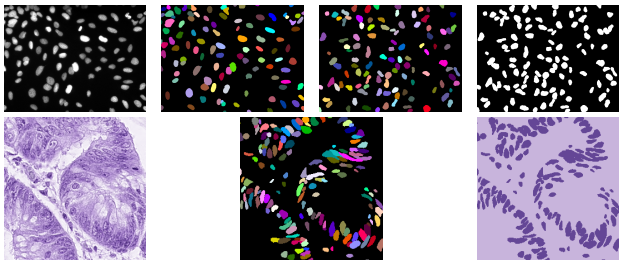
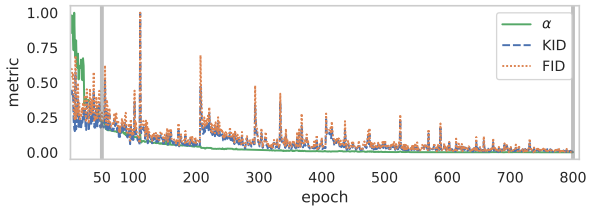
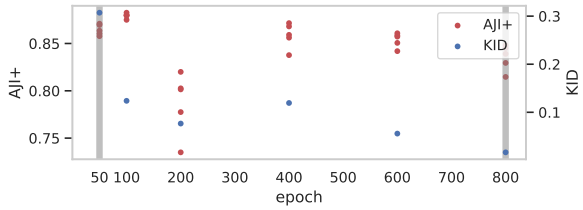


Fig. 2: Top row: Images from the BBBC039v1 dataset with a target image, the corresponding label image, and a synthetic label image with the corresponding preprocessed synthetic label image. Bottom row: Images from the Lizard dataset, with target image, label image, and the corresponding preprocessed label image.

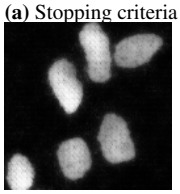
We used the CoNiC Challenge data preprocessing of the Lizard dataset, which contains 4981 images, and applied



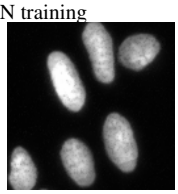
(a) Stopping criteria during GAN training



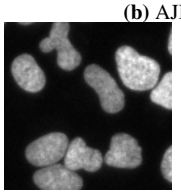
(b) AJI+ and KID for selected epochs



(c) synthetic target image (epoch 50)



(d) synthetic target image (epoch 800)



(e) real unpaired target image

(f) segmentation (epoch 50)

(g) segmentation (epoch 800)

Fig. 3: Results on the BBBC039v1 dataset. (a) metrics during GAN training normalized to range [0,1], (b) AJI+ on test data, and KID during GAN training for manually selected key epochs 50, 100, 200, 400, 600 and 800, (c-d) synthetic target images and (e) target image with segmentations (f-g). The quality of the synthetic target image is reduced for early epochs, but it is not relevant for segmentation.

color normalization according to a training reference image [5, 17]. The spatial distribution and background structures in histopathological images are complex, and the data is diverse because of the variety of patients. Thus, instead of synthesizing label images, the training data D_{train} and the validation data D_{val} are each split in half. The label images of the first half mimic synthetic labels S^L, and the corresponding target images are discarded. The target images of the second half are used for GAN training, and the corresponding label images are discarded. 70% of the dataset is used for GAN training, and the remaining 30% are used for the U-Net downstream task. An example target image with the corresponding label image and a preprocessed label image are shown in Figure 2.

Architecture, Training, and Implementation

We trained for 800 epochs (24000 steps) on the BBBC039v1 dataset and for 400 epochs (135000 steps) on the Lizard dataset. For the U-Net, we used the implementation from [15]. Each U-Net was trained for a maximum of 200 epochs, while early stopping was applied when the validation loss did not improve for 50 epochs. To reduce the variation introduced by the U-Net, we trained five U-Nets for each synthetic dataset. The code is available at https://github.com/MoritzBoe/BMT_GAN_stopping.

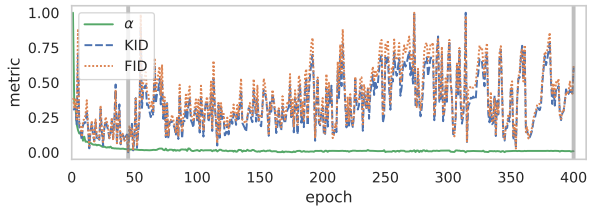
Experiments

All GAN metrics for the BBBC039v1 dataset (Figure 3) improve with increasing number of epochs, while FID and KID yield a Pearson correlation coefficient of r = 0.99. Therefore, we only consider the KID. Early epochs (c) show checkerboard artifacts within cells, while late epochs (d) do not. However, the performance of the AJI+ metric for the downstream task shows comparable performance over all key epochs, with the lowest AJI+ scores for epoch 200.

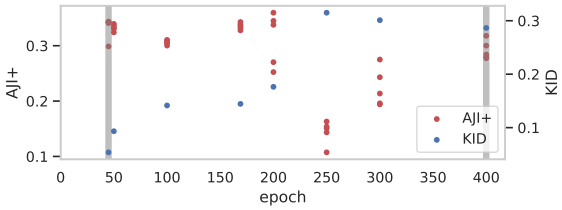
For the Lizard dataset (Figure 4), the α metric saturates during training, while KID and FID are strongly correlated (r = 0.99) but do not saturate. Because of the correlation, we again focus on the evaluation of the KID. AJI+ varies substantially between epochs. The mean of the AJI+ for each epoch compared to KID results in r = -0.82. Epoch 45 yields the best KID metric, which is compared to the last epoch. Furthermore, the results show that the last epoch is not the best epoch, and segmentation quality can change quickly during training. In (g) and (h), example segmentations for the best networks for epoch 45 (AJI+ = 0.343) and 400 (AJI+ = 0.318) are shown. Using epoch 45 instead of 400 increases the performance by 8%. The correlation between α and the AJI+ is low (r = -0.24).

4 Discussion and Conclusion

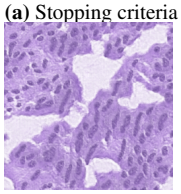
KID cannot be used to predict U-Net performance on the simple BBBC039v1 dataset. We conclude that the U-Net does not need perfect target images because the difference in brightness between the foreground and background is large. On the complex Lizard dataset, KID strongly correlates with AJI+ and can substitute training a U-Net for each GAN epoch. Further, KID can be used for epoch selection, but since KID varies consider-



(a) Stopping criteria during GAN training

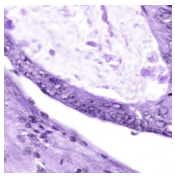


(b) AJI+ and KID for selected epochs

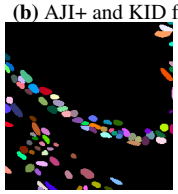


(c) synthetic target image (epoch 45)

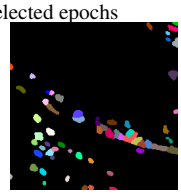
(d) synthetic target image (epoch 400)



(e) real unpaired target image



(f) ground truth segmentation



(g) segmentation (epoch 45)

(h) segmentation (epoch 400)

Fig. 4: Results on the Lizard dataset. (a) metrics during GAN training normalized to range [0,1], (b) AJI+ on test data, and KID during GAN training for manually selected key epochs 50, 100, 200, 250, 300 and 400. Furthermore, the epochs with the best KID metric (45) and the best α metric (169) are selected as key epochs. (c-d) synthetic target images and (e) target image with segmentations (f-h). For epoch 45, four out of five AJI+ scores are close to 0.34. The last epoch yields a worse segmentation, than earlier epochs.

ably during training, it cannot be used as a stop criterion. Our results show that instead of using the last epoch, which is state-of-the-art, it is beneficial to additionally train the downstream segmentation on the GAN epoch with the best KID. This improved the performance by 8% on the Lizard dataset and came with low additional computation costs. We recommend including validation metrics in the GAN training for biomedical segmentation data in the future. While the WW α metric is used to evaluate the quality of NNs, it was not suitable for early stopping or epoch selection. In the future, it needs to be examined whether different WW metrics can provide even better insight into GAN training.

References

[1] Bińkowski M, et al. Demystifying MMD GANs. arxiv:1801.01401 2021.

[2] Böhland M, et al. Influence of Synthetic Label Image Object Properties on GAN Supported Segmentation Pipelines. In: Proceedings - 29. Workshop Computational Intelligence. 2019, 289.

[3] Bruch R, et al. Synthesis of large scale 3D microscopic images of 3D cell cultures for training and benchmarking. *PLOS ONE* 2023;18:e0283828.

[4] Dunn KW, et al. DeepSynth: Three-dimensional nuclear segmentation of biological images using neural networks trained with synthetic data. *Scientific Reports* 2019;9:18295.

[5] Graham S, et al. CoNIC: Colon Nuclei Identification and Counting Challenge 2022. arxiv:2111.14485 2021.

[6] Graham S, et al. Lizard: A Large-Scale Dataset for Colonic Nuclear Instance Segmentation and Classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, 684–693.

[7] Heusel M, et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Advances in Neural Information Processing Systems, volume 30. 2017.

[8] Karimi D, et al. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis* 2020;65:101759.

[9] Kuhl FP, et al. Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing* 1982;18:236–258.

[10] Kumar N, et al. A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology. *IEEE Transactions on Medical Imaging* 2017;36:1550–1560.

[11] Ljosa V, et al. Annotated high-throughput microscopy image sets for validation. *Nature Methods* 2012;9:637–637.

[12] Mahmood F, et al. Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images. *IEEE Transactions on Medical Imaging* 2020;39:3257–3267.

[13] Martin CH, et al. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications* 2021;12:4122.

[14] Maška M, et al. The Cell Tracking Challenge: 10 years of objective benchmarking. *Nature Methods* 2023;:1–11.

[15] Schilling MP, et al. KaIDA: A modular tool for assisting image annotation in deep learning. *Journal of Integrative Bioinformatics* 2022;19.

[16] Szegedy C, et al. Rethinking the Inception Architecture for Computer Vision. arxiv:1512.00567 2015.

[17] Vahadane A, et al. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. *IEEE Transactions on Medical Imaging* 2016;35:1962–1971.

[18] Zhu JY, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In: IEEE International Conference on Computer Vision (ICCV), 2017. 2017.