Image Synthesis as a Pretext for Unsupervised Histopathological Diagnosis

Dejan Štepec[1,2] and Danijel Skočaj[1]

[1] University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana, Slovenia
[2] XLAB d.o.o.
Pot za Brdom 100, 1000, Ljubljana, Slovenia
dejan.stepec@xlab.si

Abstract. Anomaly detection in visual data refers to the problem of differentiating abnormal appearances from normal cases. Supervised approaches have been successfully applied to different domains, but require abundance of labeled data. Due to the nature of how anomalies occur and their underlying generating processes, it is hard to characterize and label them. Recent advances in deep generative based models have sparked interest towards applying such methods for unsupervised anomaly detection and have shown promising results in medical and industrial inspection domains. In this work we evaluate a crucial part of the unsupervised visual anomaly detection pipeline, that is needed for normal appearance modelling, as well as the ability to reconstruct closest looking normal and tumor samples. We adapt and evaluate different high-resolution state-of-the-art generative models from the face synthesis domain and demonstrate their superiority over currently used approaches on a challenging domain of digital pathology. Multifold improvement in image synthesis is demonstrated in terms of the quality and resolution of the generated images, validated also against the supervised model.

Keywords: Anomaly detection · Unsupervised · Deep-learning · Generative adversarial networks · Image synthesis · Digital pathology

1 Introduction

Anomaly detection represents an important process of determining instances that stand out from the rest of the data. Detecting such occurrences in different data modalities is widely applicable in different domains such as fraud detection, cyber-intrusion, industrial inspection and medical imaging [1]. Detecting anomalies in high-dimensional data (e.g. images) is a particularly challenging problem, that has recently seen a particular rise of interest, due to prevalence of deep-learning based methods, but their success has mostly relied on abundance of available labeled data.

Anomalies generally occur rarely, in different shapes and forms and are thus extremely hard or even impossible to label. Supervised deep-learning approaches have seen great success, especially evident in the domains with well known characterization of the anomalies and abundance of labeled data. Obtaining such detailed labels to learn supervised models is a costly and in many cases also an impossible process, due to unknown set of all the disease biomarkers or product defects. In an unsupervised setting, only normal samples are available (e.g. healthy, defect-free), without any labels. Deep generative methods have been recently applied to the problem of unsupervised anomaly detection (UAD), by utilizing the abundance of unlabeled data and demonstrating promising results in medical and industrial inspection domains [2, 3, 4]. Deep generative methods, in a form of autoencoders [5] or GANs [6] are in a UAD setting used to capture normal appearance, in order to detect and segment deviations from that normal appearance, without the need for labeled data.

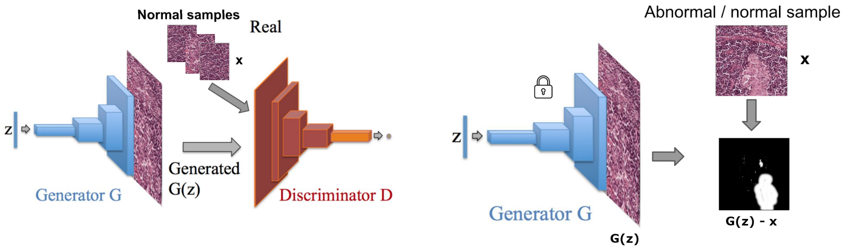
AnoGAN [7] represents the first method, where GANs are used for anomaly detection in medical domain. A rich generative model is constructed on healthy examples of optical coherence tomography (OCT) images of the retina and a methodology is presented for image mapping into the latent space. The induced latent vector is used to generate the closest example to the presented query image, in order to detect and segment the anomalies in an unsupervised fashion. AnoGAN [7] and the recently presented f-AnoGAN [2] improvement, utilize low resolution vanilla DCGAN [8] and Wasserstein GAN [9] architectures, for normal appearance modelling, with significantly lower anomaly detection performance in comparison with autoencoder-based approaches [10, 11]. This does not coincide with superior image synthesis performance of the recent GAN-based methods. We argue, that adapting recent advancements in GAN-based unconditional image generation [12, 13, 14], currently utilized mostly for human face synthesis, should also greatly improve the performance of image synthesis in different medical imaging domains, as well GAN-based UAD methods.

In this work we focus on normal appearance modelling part of the UAD pipeline and evaluate the ability to synthesize realistically looking histology imagery. This presents a pretext for an important problem of metastases detection from digitized gigapixel histology imagery. This particular problem has been already addressed in a supervised setting [15], by relying on the limited amount of expertly lesion-level labeled data, as well as in a weakly-supervised setting [16], where only image-level labels were used. Extremely large histology imagery and highly variable appearance of the anomalies (i.e. cancerous regions) represent a unique challenge for existing UAD approaches. We investigate the use of the recently presented high resolution generative models from the human face synthesis domain [12, 13, 14], for normal appearance modelling in a UAD pipeline, which could consequently improve the performance and stability of the current state-of-the-art approaches [2, 11].

We demonstrate this with significant improvements in the quality and increased resolution of the generated imagery in comparison with currently used approaches [8, 9], which also represents a novel application of generative models to the digital pathology domain. We also investigate the effectiveness of current latent space mapping approaches, specifically their ability of closest looking normal histology sample reconstruction. We validate the quality of the synthesized imagery against the supervised model and demonstrate the importance of synthesizing high resolution histology imagery, resulting in an increased amount of contextual information present, crucial for distinguishing tumor samples from the normal ones.

2 Methodology

Unsupervised Visual Anomaly Detection The capability to learn the distribution of the normal appearance represents one of the most important parts in a visual anomaly detection pipeline, presented in Figure 1. This is achieved by learning deep generative models on normal samples only, as presented in Figure 1a. The result of this process is the capability to generate realistically looking artificial normal samples, which cannot be distinguished from the real ones. To detect an anomaly, a query image is presented and the closest possible normal sample appearance is generated, which is used to threshold the difference, in order to detect and segment the anomalous region, as presented in Figure 1b. This is possible due to learned manifold of normal appearance and its inability to reconstruct anomalous samples.



(a) Normal appearance modelling

(b) Anomaly detection

Fig. 1: GAN based visual anomaly detection pipeline, consisting out of a) normal appearance modelling and b) the search for optimal latent representation, that will generate the closest normal appearance sample, used for anomaly detection.

Different approaches have been proposed for normal appearance modelling, as well as anomaly detection. Learning the normal visual appearance is based on autoencoders [3], GANs [7, 2], or combined hybrid models [10, 11]. Most of the approaches learn the space of the normal sample distribution Z, from which latent vectors $z \in Z$ are sampled from, that generate the closest normal appearance, to the presented query image. Different solutions have been proposed for latent vector optimization, that are usually independent from the used normal appearance modelling method (i.e. autoencoders, GANs).

Current state-of-the-art GAN-based visual anomaly detection methods [7, 2] are based on vanilla GAN implementations [8, 9], with limited resolution generated normal samples of low fidelity, as well as with stability problems. Autoencoder based anomaly detection methods have shown improved results over pure GAN based implementations [10, 11], but are similarly limited to a low resolution of 64^2. In comparison, we evaluate the feasibility of generating high fidelity histology imagery up to the resolution of 512^2, with the recently presented GAN architectures [12, 13, 14] from the face generation domain, not yet utilized in anomaly detection pipelines, as well as in the digital pathology domain. Note, that we limit the resolution to a maximum of 512^2, due to hardware resource and time constraints. We also investigate the effectiveness of recently presented latent space mapping approaches and the feasibility to be applicable for UAD in the digital pathology domain.

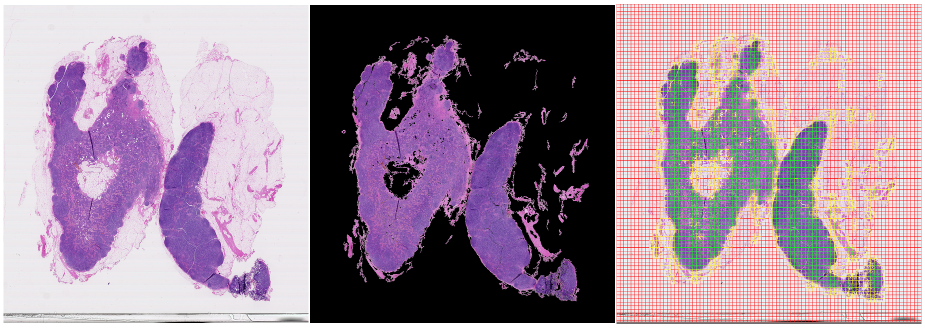
Deep Generative Adversarial Models. Original GAN implementation [6] was based on standard neural networks, which generated images suffered from being noisy and the training process was notoriously unstable. This was improved by implementing the GAN idea using the CNNs - DCGAN [8], by identifying a family of architectures, that result in a stable training process of higher resolution deep generative models. The method was later on also adapted in the AnoGAN anomaly detection framework [7]. Stability problems of GAN methods were first improved by proposing different distance measures for the cost function (e.g. Wasserstein GAN [9]), adapted also by the f-AnoGAN anomaly detection method [2]. The main limitation of those early GAN methods is also the low resolution (up to 64^2) and the limited variability of the generated images.

Recently, the ideas of progressively growing GANs [12] and style-based generators [13, 14] were presented, allowing a stable training of models for resolutions up to 1024^2, with increased variation and quality of the generated images. In progressively growing GANs [12], layers are added to the generator and discriminator progressively, by linearly fading them in and thus enabling fast and stable training. StyleGAN [13] proposes an alternative generator architecture, based on style transfer literature [17], exposing novel ways to control synthesis process and reducing the entanglement of the latent space. StyleGAN2 [14] addresses some of the main characteristic artifacts resulting from the progressive growing in StyleGAN [13], further boosting generative performance.

In comparison with autoencoders, GANs do not automatically yield the inverse mapping from the image to latent space, which is needed for closest-looking normal sample reconstruction and consequently anomaly detection. In AnoGAN [7] an iterative optimization approach was proposed to optimize the latent vector z via backpropagation, using the residual and discrimination loss. Residual loss is represented with pixel-wise Mean Square Error (MSE) loss, while discrimination loss is guided by the GAN discriminator, by computing feature matching loss between the real and synthesized imagery. In f-AnoGAN method [2], an autoencoder replaces the iterative optimization procedure, using the trainable encoder and the pre-trained generator, as the decoder. For StyleGAN2 [14], authors proposed an iterative inverting procedure, which specifically optimizes an intermediate latent space and noise maps, based on the Learned Perceptual Image Patch Similarity (LPIPS) [18].

3 Experiments and Results

Histology Imagery Dataset. We address aforementioned problems of anomaly detection pipeline on a challenging domain of digital pathology, where whole-slide histology images (WSI) are used for diagnostic assessment of the spread of the cancer. This particular problem was already addressed in a supervised setting [15], as a competition[3], with provided clinical histology imagery and ground truth data. A training dataset with (n=110) and without (n=160) cancerous regions is provided, as well as a test set of 129 images (49 with and 80 without anomalies). Raw histology imagery, presented in Figure 2a, is first preprocessed, in order to extract the tissue region (Figure 2b). We used the approach from IBM[4], which utilizes a combination of morphological and color space filtering operations. Patches of different sizes (64^2 - 512^2) are then extracted from the filtered image (Figure 2c) and labelled according to the amount of tissue in the extracted patch.



(a) Original WSI (b) Filtered WSI (c) Patches from WSI (1024^2)

Fig. 2: Preprocessing of the original WSI presented in a) consists of b) filtering tissue sections and c) extracting patches, based on tissue percentage (green $\geq$ 90%, red $\leq$ 10% and yellow in-between). Best viewed in digital version with zoom.

Image Synthesis. We first evaluate the performance of different GAN based generative approaches on a challenging histology imagery using the Fréchet Inception Distance (FID) [19], by following evaluation procedure from StyleGAN2 [14, Table 1]. The FID score represents a similarity between a set of real

[3] https://camelyon16.grand-challenge.org/

[4] https://github.com/CODAIT/deep-histopath

and generated images, based on the statistics extracted from the pretrained Inception classifier [20]. We train different GANs using 1000 randomly extracted patches, from each of the 160 normal WSIs, with tissue coverage over 90% (Figure 2c). This presents an input to the baseline DCGAN [8] model, used in the AnoGAN [7] anomaly detection framework, Wasserstein GAN (WGAN), used in f-AnoGAN [2], as well as to recently presented GAN architectures, based on progressive growing (PGAN [12]) and style transfer (StyleGAN [13], StyleGAN2 [14]). We evaluate not only the feasibility to generate pathology imagery, but to generate it up to a resolution of 512^2 and present the results in Table 1.

Table 1: FID scores for different methods and different input image sizes DCGAN and WGAN are limited to a maximum resolution of 64^2 and only best performing model is evaluated at 512^2.

Image size	DCGAN	WGAN	PGAN	StyleGAN	StyleGAN2
64^2	88.52	12.65	18.89	7.15	**6.64**
256^2	-	-	17.82	5.57	**5.24**
512^2	-	-	-	-	**2.93**

We can see that generative performance of the recently presented methods significantly outperforms vanilla DCGAN model and should be considered in all the proposed GAN based visual anomaly detection pipelines. They are also capable of generating images of much bigger size and higher resolution, which is particularly important for anomaly detection, due to the increased amount of visual context, available for determining the presence or absence of the anomalies. For WGAN, we used the implementation from f-AnoGAN [2], based on residual neural networks, which also produces high quality, high fidelity imagery up to image size of 64^2. There is no significant difference in terms of the FID score between different style transfer based approaches, but StyleGAN2 represents an incremental improvement over StyleGAN and also offers an additional benefit of generator reversibility, especially interesting for anomaly detection (Figure 1b). Note that in StyleGAN2 a smaller network configuration E was used for an increased image throughput, at small performance expense [14, Table 1]. Visual comparison of generated results is presented in Figure 3, demonstrating high variability and quality of the generated images and also the applicability of such methods for digital pathology.

Classification of Real and Synthetic Patches. To additionally asses the quality of the generated images in comparison with the real ones, we evaluated the performance of a discriminative classifier applied on both types of data. We trained supervised DenseNet121 [21] model on extracted normal and tumor histology imagery patches and compared its performance to distinguish the two classes on real and synthesized imagery (Table 2). We extracted 100.000 normal

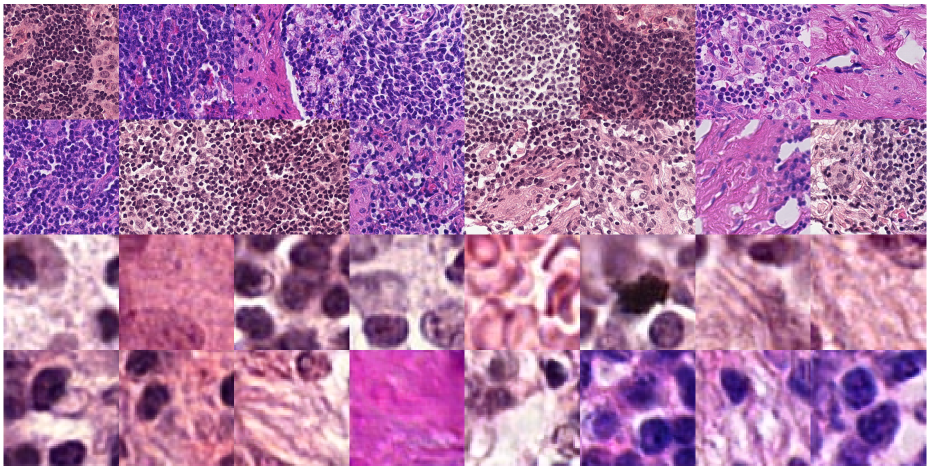


Fig. 3: Examples of real histology imagery at image size of 512^2 (top row), generated images by the best performing StyleGAN2 model at 512^2 (second row), real histology imagery at 64^2 (third row) and images generated by the WGAN model at 64^2 (last row). Best viewed in digital version with zoom.

and tumor patches (with provided annotations) of size 64^2 and 512^2 to train DenseNet121 model and evaluated the performance on a test set of 10.000 (real) normal and tumor patches. Besides on normal histology imagery, we also trained WGAN [9] and StyleGAN2 [14] on tumor patches, in order to be able to generate both classes. We then synthesized a test set of 10.000 normal and tumor patches and evaluated the capability of the DenseNet121 model (trained on real imagery) to distinguish anomalous samples in synthesized imagery.

Table 2: Classification accuracy (CA) for normal vs. tumor patch based classification on real and synthesized imagery (WGAN for 64^2 and StyleGAN2 for 512^2).

Image size	Real (CA)	Synthesized (CA)
64^2	88.23%	87.05%
512^2	98.89%	98.34%

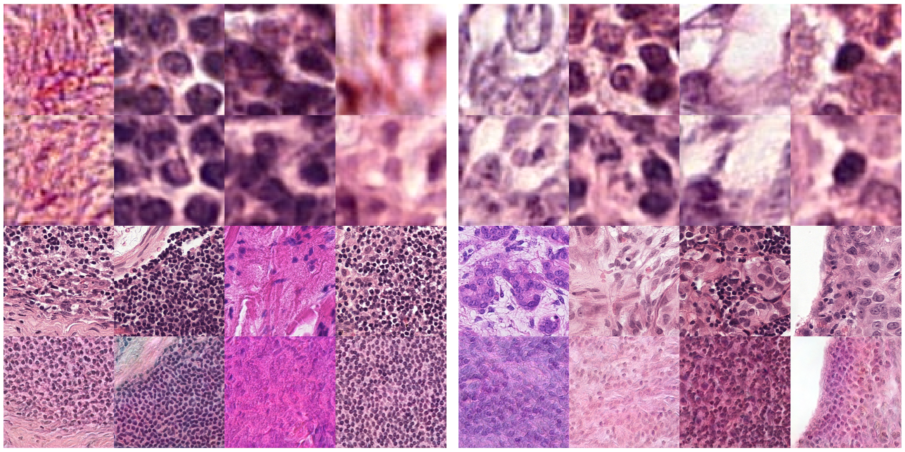
Results in Table 2 demonstrate that supervised model (trained on real imagery) successfully recognizes synthesized imagery, with no drop in performance, compared to real imagery. Supervised model can be seen as a virtual pathologist, confirming to some extent the correctness of image synthesis, on a much larger scale. We can also see significant drop in performance on 64^2 patches, which additionally confirms, that larger patches hold more contextual information, useful for classification.

Latent Space Mapping. We qualitatively evaluate the capability of latent space projection using the encoder based approach (izi_f) presented in f-AnoGAN [2] for WGAN [9] based generator and LPIPS-distance [18] based approach for StyleGAN2 [14] based generator, proposed and specifically designed for StyleGAN2 already in the original work [14]. Figure 4b presents the results for both methods, image sizes and histopathological classes (i.e. normal and tumor). The generators are trained on normal samples only and should reconstruct only normal samples, while tumor samples should be poorly reconstructed, thus enabling detection of anomalies. We can see that the encoder based approach, proposed in f-AnoGAN [2] is able to find very similar looking artificial samples in the latent space. The problem is, that it also reconstructs tumor samples, with similar performance. We argue, that this is due to small patch size (64^2), which is in more than 10% ambiguous also for the supervised classifier (Table 2) - such small tumor patch might in fact not contain abnormality, or is represented with insufficient biomarkers. Such samples, even in small percentage [22], cause the generator to learn how to reconstruct anomalous samples.

StyleGAN2 [14] did not yield good reconstructions, capturing only major properties of the query image. This is beneficial for tumor samples, where we noticed consistent failure to capture even the main properties of the query image, with notable exception of the staining color. This can probably be attributed to the larger image sizes and more contextual information present to distinguish anomalous samples, not seen during the generator training. Integration of StyleGAN2 and encoder based mapping (coupling best of the two methods) offers a promising future direction to be investigated, to improve latent space mapping and thus enabling UAD in histopathological analysis.

4 Conclusion

In this work we addressed image synthesis as a pretext for GAN based anomaly detection pipeline in histopathological diagnosis and demonstrated, that histology imagery of high quality and variability can be synthesized, as well as reconstructed. We identified the importance of synthesizing large histology samples, not used in current GAN based anomaly detection pipelines, as well as the drawbacks and future research direction for more effective latent space mapping. The ability to generate realistically looking normal histology imagery of high resolution and size will enable the development of UAD pipeline, in order to apply it to cancer diagnosis, especially important for rare cancer types (e.g. paediatric), where annotated data is scarce, thus preventing the use of supervised approaches. Reducing the performance gap between supervised and unsupervised approaches and increasing the robustness of the UAD approaches will represent a significant contribution to wider adaption of automated visual analysis techniques, well beyond presented medical domain.



(a) normal

(b) tumor

Fig. 4: Projecting real a) normal and b) tumor imagery to the latent space of WGAN at 64^2 (first row) and StyleGAN2 at 512^2 (third row) and resulting closest matches in the latent space for WGAN (second row) and StyleGAN2 (last row). Best viewed in digital version with zoom.

Acknowledgment

This work was partially supported by the European Commission through the Horizon 2020 research and innovation program under grant 826121 (iPC).

References

1. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.
2. T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical image analysis*, vol. 54, pp. 30–44, 2019.
3. C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain mr images," in *MICCAI Brainlesion Workshop*, pp. 161–169, Springer, 2018.
4. P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *CVPR*, pp. 9592–9600, 2019.
5. Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in *ICCV*, pp. 1511–1519, 2015.
6. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

7. T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *IPMI*, pp. 146–157, Springer, 2017.
8. A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR*, 2016.
9. S. Martin Arjovsky and L. Bottou, "Wasserstein generative adversarial networks," in *ICML*, 2017.
10. S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *ACCV*, pp. 622–637, Springer, 2018.
11. S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in *IJNN*, pp. 1–8, IEEE, 2019.
12. T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *ICLR*, 2018.
13. T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, pp. 4401–4410, 2019.
14. T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*, pp. 8110–8119, 2020.
15. B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.
16. G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.
17. X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *CVPR*, pp. 1501–1510, 2017.
18. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, pp. 586–595, 2018.
19. 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*, pp. 6626–6637, 2017.
20. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, pp. 2818–2826, 2016.
21. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, pp. 4700–4708, 2017.
22. A. Berg, J. Ahlberg, and M. Felsberg, "Unsupervised learning of anomaly detection from contaminated image data using simultaneous encoder training," *arXiv preprint arXiv:1905.11034*, 2019.