

Article

A Comparative Study of Engraved-Digit Data Augmentation by Generative Adversarial Networks

Abdulkabir Abdulraheem and Im Y. Jung * 

School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, Korea

* Correspondence: iyjung@ee.knu.ac.kr; Tel.: +82-53-950-7237

Abstract: In cases where an efficient information retrieval (IR) system retrieves information from images with engraved digits, as found on medicines, creams, ointments, and gels in squeeze tubes, the system needs to be trained on a large dataset. One of the system applications is to automatically retrieve the expiry date to ascertain the efficacy of the medicine. For expiry dates expressed in engraved digits, it is difficult to collect the digit images. In our study, we evaluated the augmentation performance for a limited, engraved-digit dataset using various generative adversarial networks (GANs). Our study contributes to the choice of an effective GAN for engraved-digit image data augmentation. We conclude that Wasserstein GAN with a gradient norm penalty (WGAN-GP) is a suitable data augmentation technique to address the challenge of producing a large, realistic, but synthetic dataset. Our results show that the stability of WGAN-GP aids in the production of high-quality data with an average Fréchet inception distance (FID) value of 1.5298 across images of 10 digits (0–9) that are nearly indistinguishable from our original dataset.

Keywords: data augmentation; generative adversarial networks; engraved digit image; Fréchet inception distance



Citation: Abdulraheem, A.; Jung, I.Y.

A Comparative Study of Engraved-Digit Data Augmentation by Generative Adversarial Networks.

Sustainability **2022**, *14*, 12479.

<https://doi.org/10.3390/su141912479>

Academic Editors: Taqwa Hariguna and Chung-Wen Hung

Received: 31 August 2022

Accepted: 27 September 2022

Published: 30 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As machine learning and big data engineering have grown in popularity, information retrieval (IR) from various sources has become a subject of discussion. In general, an efficient IR system based on machine learning techniques requires a large collection of data sources because it learns by training with a large amount of data. To date, several issues have been raised in the early stages of applying recommender systems in the healthcare industry to assist health practitioners and users in making efficient and accurate health-related decisions [1]. We are particularly concerned about the implications of insufficient datasets used to train some of these models. One of the system applications involves the retrieval of expiry dates found on daily-life products administered to the human body. It is important to help people notice the expiry date, as it can help convey additional information about the product and regulate their health. We introduced generative adversarial network (GAN) as a way to use IT technology to support the continuous healthy life of people. The disabled, the elderly, and patients with vision loss, in particular, have a difficulty in checking dates because the digits are small and sometimes blurry. Expiry dates are often expressed as engraved digits, as shown in Figure 1. To automatically retrieve the engraved digits in expiry dates, we need a large dataset of engraved digits to train the IR system. We collected image (photo) data from medicines, consumables, cosmetic products, and tube-type ointments to recognize the expiration dates expressed in engraved digits. However, the performance of the classification results was not good because of the small amount of data. Our dataset had images that showed both the properties of digits and unusual shadows created by the engraved shape in the images, and it was differentiated from the MNIST dataset [2] by these properties.



Figure 1. Expiry dates on medicines, creams, ointments, and gels in squeeze tubes.

The popularity of generative adversarial networks (GANs) is increasing nowadays, and these are used to generate synthetic datasets that are close and almost indistinguishable from the real data [3]. GAN is a type of machine learning technique that involves two models being trained simultaneously: the generator, G, and the discriminator, D. G is trained to generate fake data samples, and D is trained to distinguish between fake and real data [3,4]. These two models are trained together in an adversarial zero-sum game until the discriminator model is tricked half of the time, implying that the generator model generates satisfactory outputs [3,4]. This means that D effectively distinguishes between actual and false samples. It is either rewarded or no changes to the model parameters are required, whereas G is penalized with significant model parameter changes. D determines whether a batch of samples produced by G is authentic or fake by comparing it with actual samples from the original datasets. To improve its ability to distinguish between genuine and fake samples in the following round, D is updated. More crucially, G is also updated based on whether D is successfully tricked by the generated samples [3,4]. Therefore, the fundamental building blocks of GANs include the generator G, discriminator, D or critic, C, and the associated loss functions. The G model is used to generate new plausible examples from the problem domain, and the D model is used to classify examples from the domain as real or fake. The loss function helps to evaluate the model performance [3,5]. It measures the accuracy of the model in terms of predicting the expected outcome. Basically, it improves the stability of the trained GAN model.

The GAN architecture is shown in Figure 2. First, we sample the noise, z , using a normal or uniform distribution. With z as an input, we use a generator, G, to create an image x ($x = G(z)$) after performing multiple transposed convolutions to upsample z . The discriminator processes the real images (training samples) and generated images separately. It distinguishes whether the input image is real or generated. The output $D(x)$ is the probability that the input x is real. If the input is real, $D(x)$ should equal 1. If it is generated, it should be zero. Through this process, the discriminator identifies the features that distinguish real images from generated images. We train the generator to create images that may be interpreted as real by the discriminator. We train both networks in alternating steps and lock them into a fierce competition to improve themselves. Eventually, the discriminator detects a small difference between the real and generated images, and the generator generates images that cannot be distinguished by the discriminator. The GAN model repeats this process until the loss converges to zero, and the distributions are approximately equal.

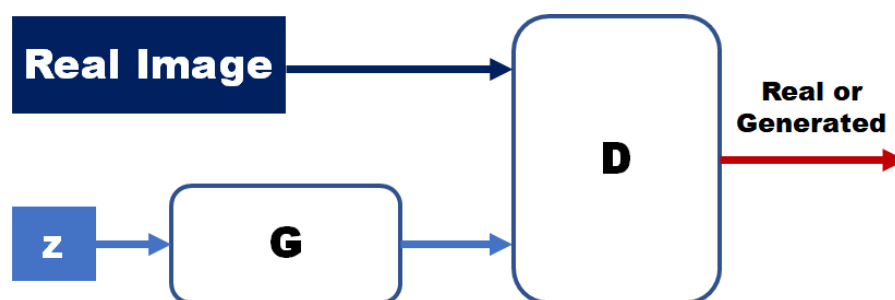


Figure 2. GAN architecture. Adapted from ref [6].

There are several application domains of GANs, and significant progress has been made in areas, such as image super-resolution, art creation, speech synthesis, image-to-image translation, video prediction, and 3D object generation [3,4,7–10]. GANs often evaluate input images and leverage convolutional neural networks (CNNs) as generator and discriminator models, and the performance of these neural networks often improves with the amount of available data. Data augmentation is a technique to artificially generate data. Data augmentation using GANs helps to generate more plausible training datasets from existing data [7–10]. This helps to develop better models by improving the model characteristics and providing a regularizing effect, which reduces the generalization error.

In addition to GANs, restricted Boltzmann machines (RBM), generative stochastic networks (GSN), deep belief networks (DBN), deep Boltzmann machines (DBM), and variational autoencoders are used in deep learning [4,11]. Unlike classic augmentation algorithms, GANs can handle invariances represented by simple transformations and strengthen weak points in the learned decision boundaries [7,12]. In addition, GANs have exhibited tremendous computation speed and quality of results compared with other methods [7]. GANs generate new training data, which lead to better classification performance [4,7]. To select the best performing model for data augmentation, the output from each model is evaluated by human visual inspection [13–16] and by calculating the Fréchet inception distance (FID). The FID is a metric used in GAN to compute the Wasserstein distance between feature vectors calculated between a fake and real image [14,17–19]. With the FID, it is possible to identify the output images that show close diversity to the original images. A very low score indicates that the output and original images are identical and have comparable statistics [18]. The FID is expressed as given by Equation (1):

$$\text{FID} = \|\mu_x - \mu_y\|^2 + \text{Tr}(\Sigma_x + \Sigma_y - 2\sqrt{(\Sigma_x \Sigma_y)}) \quad (1)$$

where x and y are the real and generated samples, respectively, i.e., the activations from the pretrained Inception-v3 model [20], and μ_x and μ_y are the feature-wise means of the real and generated images, respectively. Tr is the trace of the matrix, and Σ_x and Σ_y are the covariance matrices of the vectors [18–21]. In [19], the FID evaluation measure was empirically demonstrated to be a viable metric because of its resistance to mode dropping and encoding networks.

The inception score (IS) is another metric that is used to evaluate the image quality and diversity. The IS can be used to evaluate clear images generated by GANs, and the model used to train a GAN should output a high diversity of images from all different classes in ImageNet [13,22]. The most common use of IS is in non-ImageNet datasets for generative models trained on CIFAR-10 [22]. This is because CIFAR-10 is significantly smaller and is more manageable for training than ImageNet. Hence, applying the IS to generative models trained on datasets other than ImageNet yields misleading results [13,23]. The IS score is high for clear images that are well classified into specified object types [23,24]. Owing to the generated images not being present in the classifier's training data, the IS is not ideal for our dataset.

In this study, we leveraged GAN models to produce more datasets to provide accurate results for automatic recognition of expiry-dates in photos, which requires a large amount

of learning data. We implemented and evaluated state-of-the-art GAN models, such as WGAN [5], WGAN-GP [21], WGAN-DIV [25], MMGAN [26], NSGAN [26], LSGAN [27], DRAGAN [28], ACGAN [29], DCGAN [30], EBGAN [31], VAE [32], and BEGAN [33], to augment the data for our dataset. We concluded that Wasserstein GAN with a gradient norm penalty (WGAN-GP) is a suitable data augmentation technique for our dataset. Our results show that the stability of WGAN-GP aids in the production of not only high-quality data but also images of a variety of styles. The images have an average Fréchet inception distance (FID) value of 1.5298 across 10 digits (0–9) that are nearly indistinguishable from our original dataset.

The remainder of this paper is structured as follows. The background section introduces a GAN for data augmentation and describes two metrics to evaluate the GAN performance: FID and inception score (IS). In the section on relevant literature, we introduce recent studies on data augmentation using GAN and state-of-the-art GAN models. In the section, Data Augmentation for Engraved-Digit Images, we establish the method of analysis and discuss the evaluation metrics that are used. In the Evaluation and Analysis section, we provide brief details of the WGAN-GP model, describe its hyperparameters, and present loss plots to demonstrate optimal results. Finally, we summarize our study in the Conclusions section.

2. Relevant Literature

2.1. Data Augmentation by GAN

Researchers have proposed and applied several GAN models to supplement datasets that are insufficient for classification tasks. These datasets vary across different domain areas of application, and the GAN models that are the most suitable for the augmentation process are often adopted. The augmented datasets are applied to a classifier for validation, and a set of benchmark results for the datasets is presented, which may then be used to further characterize and validate the datasets. Our work targets a dataset of few-digit images that are engraved on medicines, ointments, and other forms of squeeze tubes, and the images are usually distorted and blurry. We evaluated the state-of-the-art GAN models to determine the most preferred model for the data augmentation task.

Wang et al. [34] integrated a GAN into an all-encompassing information-retrieval framework. Their approach demonstrated that GAN-based information retrieval systems are promising, but more work is required. The idea of data augmentation across several domains and simple models aided by large datasets can be extremely helpful in improving the performance of object detection applications [5,10,35]. In [36], the authors evaluated the effectiveness of two augmentation methods for CNN-based MODI script handwritten character recognition using the same CNN architecture. They confirmed that the on-the-fly data augmentation technique was more accurate than an offline approach. This strategy enabled the network to view a fresh collection of data each time, boosting the effectiveness of the system. Using the standard MNIST handwritten digit dataset in [37], the authors conducted an experimental evaluation of the advantages of data augmentation for convolutional backpropagation-trained neural networks, convolutional support vector machines, and convolutional extreme learning machine classifiers. Their work showed that in cases where reasonable data transforms are available, augmentation in the data space offers an advantage for enhancing performance and reducing overfitting. The authors investigated the advantages of training a machine learning classifier with examples that were artificially manufactured to supplement the data. In addition, in [38], the authors demonstrated the method of training a deep neural network (DNN) for optical character recognition (OCR) on historical manuscripts with a small amount of data. They examined various methods for data augmentation of palimpsests and evaluated the impact of various techniques on the performance of the DNN. In [8], the authors demonstrated the effective enhancement of conventional vanilla classifiers by a data augmentation generative adversarial network (DAGAN). Additionally, they demonstrated the usage of a DAGAN to improve

few-shot learning systems, such as matching networks. Their approach was evaluated on the Omniglot [39], EMNIST [40], and VGG-face datasets [41].

In recent times, GANs have been extensively used in medical applications. In [10], the authors presented a literature review on the application of GANs in ophthalmology image domains to discuss key contributions and suggested probable future research trajectories. Whereas medical records can contain sensitive and personal data of patients, training of GANs with these original datasets may be regulated and a need for synthetic datasets for research in this domain may arise [42]. Other applications of GANs in medicine are discussed in [43,44].

2.2. GAN Models for IR

The Wasserstein GAN with gradient penalty, WGAN-GP [21], was proposed as an alternative for clipping weights, i.e., to penalize the norm of the gradient of the critic with respect to its input. This model has made progress towards the stable training of GANs with almost no hyperparameter tuning, unlike WGAN [5], which suffers from training instability and sometimes generates poor samples or suffers convergence failure. WGAN-GP proposes a gradient penalty to be added to the WGAN discriminator loss as an alternative method for enforcing the Lipschitz constraint, which was previously performed by weight clipping. This penalty does not suffer from bias of the discriminator toward simple functions owing to weight clipping. Additionally, the reformulation of the discriminator by adding a gradient penalty term makes batch normalization unnecessary. Wasserstein divergence, WGANDIV [25], a symmetric divergence, has been proved to faithfully approximate the corresponding Wasserstein divergence through optimization. It has been demonstrated to be stable under various settings, including progressive growing training and has exhibited superior results when compared with state-of-the-art methods, both quantitatively and qualitatively.

In NSGAN [26], two models are simultaneously trained: G, which captures the data distribution, and D, which estimates the probability that a sample is from the training data rather than G. The objective of the training process for G is to maximize the probability that D makes a mistake. The generator loss is the only difference when compared with the MMGAN [26]. In both the NSGAN and MMGAN, the output of G can be interpreted as a probability. However, the output of D in LSGAN [27] is unbounded unless it passes through an activation function. A sigmoid activation function has also been implemented. LSGAN tackles the vanishing gradient problem associated with GANs by swapping out the cross-entropy loss function with the least-squares (L2) loss function. In [27], the authors claimed that the L2 loss function penalizes samples that are identified by the discriminator as real but located outside the decision boundary. Hence, the generated visuals are meant to closely resemble real data. It also stabilizes the training process. The D in DRAGAN [28] can be interpreted as a probability, similar to the D models in MMGAN and NSGAN. DRAGAN is similar to WGAN-GP but seems to be less stable. This model is very similar to WGAN-GP as it applies a gradient penalty to try and obtain an improved training objective based on the optimal performance of D and G. The gradient penalty is only applied close to the real data manifold, whereas WGAN-GP selects the gradient location on a random line between a real and randomly generated fake sample.

The auxiliary classifier GAN (AC-GAN) [29] is an extension of the conditional GAN that changes the discriminator to predict the class label of a given image rather than receive it as an input. It stabilizes the training process and allows the generation of large high-quality images while learning a representation in the latent space that is independent of the class label. A deep convolutional GAN (DCGAN) is a generative adversarial network with convolutional neural networks as the generator and discriminator [30]. The DCGAN was proposed after evaluating a set of constraints on the architectural topology of convolutional GANs that allowed them to be trained in a stable manner in most settings. The trained discriminators were used for image classification tasks, showing competitive performance with other unsupervised algorithms.

To model the discriminator, $D(x)$, as an energy function that assigns low energies to regions close to the data manifold and greater energies to other regions, a family of GANs called EBGANs has been proposed [31]. The discriminator uses an autoencoder. The encoder extracts latent features from the input image, and subsequently, the decoder performs image reconstruction. Instead of a probability value similar to the original GAN, the discriminator outputs the reconstruction mean square error, MSE, between the input image and reconstructed image VAEs [32]. Thus, it encodes an input into a given dimension, z , reparametrizes z using its mean and standard deviation and then reconstructs the image from the reparametrized z . BEGAN [33] optimizes a lower bound of the Wasserstein distance between autoencoder loss distributions on the original and generated data, using an autoencoder as a discriminator. To maintain the discriminator and generator in equilibrium, the authors added an additional hyperparameter $\gamma \in [0, 1]$.

3. Data Augmentation for Engraved-Digit Images

The objective of this study is to compare the state-of-art GAN models and identify the models that are suitable for data augmentation for our specific types of datasets after evaluating the generated data images, variety of the images produced, and performance of the GAN model from the FID score after training. The proposed GAN models were trained using optimal tuning with the same epoch time.

3.1. Engraved-Digit Image Dataset

As shown in Figure 3, our dataset has the properties of both digits and unusual shadows by engraved shapes in the images. They were collected from medicines, gels, and tube-type ointments. Classification models perform poorly when insufficient or blurry data are used for model training. We trained these limited datasets on state-of-the-art GAN models to produce high-quality grayscale fake images. The images were separated into classes of 0–9. We collected approximately a hundred images per digit, which were selectively trained one class at a time to evaluate the diversity of the images produced by each GAN. The data were pre-processed for standardization. We reshaped and converted all input images to a grayscale of $128 \times 128 \times 1$ in the range of $[-1, 1]$. This was done to maintain the same color in the images.

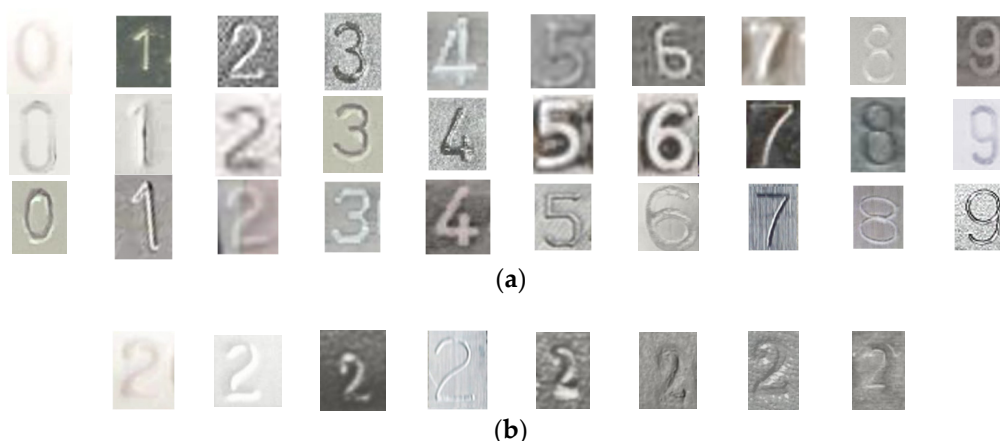


Figure 3. (a) The best selection of engraved digits cropped from the expiry dates in our source datasets (as seen in Figure 1), and (b) more samples of digit-2 images.

3.2. Data Augmentation for Our Dataset by GANs

We trained our dataset using the following state-of-the-art GAN models: MMGAN, NSGAN, LSGAN, ACGAN, DCGAN, WGAN, WGAN-GP, WGAN-DIV, DRAGAN, BEGAN, EBGAN, and VAE. We implemented these models and examined the images produced by each GAN. In addition, we closely analyzed the graph showing the relationship between the generator and discriminator loss plots of each GAN.

3.2.1. A Classification of GANs

As shown in Table 1, we classified the GAN models into four groups. Group 1 is based on the way D, also known as the critic C, wants to maximize the expression to differentiate between the real and fake images. Group 2 is composed of models based on the way D estimates the probability that a sample originates from the training data rather than G. The training procedure for G maximizes the probability of D making a mistake. Group 3 is based on D's main task, i.e., to perform image classification. The basic architecture of the GANs in Groups 1, 2, and 3 is shown in Figure 2. In contrast, the GANs in Group 4 have an autoencoder in their architecture. The VAE architecture comprises an encoder and a decoder trained to minimize the reconstruction loss, with the input being encoded as a distribution over the latent space. Figure 4 shows the embedding of autoencoders in the EBGAN and BEGAN architectures. Both GANs can be represented using the same architecture; however, in each GAN, the reconstruction loss is calculated differently. Whereas EBGAN uses the mean square error (MSE) to calculate the reconstruction loss, BEGAN uses the Wasserstein distance.

Table 1. GAN classification.

Group	1	2	3	4
GANs	WGAN WGAN-DIV WGAN-GP	LSGAN MMGAN NSGAN DRAGAN	ACN DCGAN	BEGAN EBGAN VAE

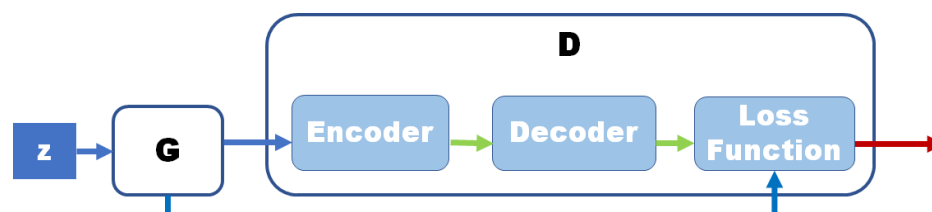


Figure 4. BEGAN/EBGAN architecture.

3.2.2. Digit Image Generation by GANs

The 0–9-digit images generated with our dataset by the GANs are shown in Table 2, along with some selected outputs. The number under each image is the epoch in which the image was generated. We selected the best and most stable images produced during 50,000 epochs of execution. We trained each class of our datasets (digits 0–9) for 50,000 epochs with sample images saved every 1000 epochs.

3.2.3. FIDs and ISs

We calculated the FID [10,17], which measures the statistical difference between the features of our original data and generated images, as shown in Table 3. FID addresses the flaw of IS, in which the statistics of the original and generated samples are not compared [20]. A perfect score of 0.0 indicates that the two sets of images are identical, whereas lower scores indicate that the two groups of images are more similar or have more in common statistically.

Table 4 lists the IS values calculated for each GAN output image. The average IS calculated across all GANs in Table 4 is approximately 1.0. IS measures the realism of a GAN output by using a pre-trained deep learning neural network model to predict the class probabilities for each class of the generated image, as opposed to FID, which is calculated by comparing the statistics of the generated samples with the original samples. A high IS score is an excellent score and usually ranges from 1 to the number of classes in a pretrained network [23]. In our case, the IS is very low because the generated images were not present in the training data of the classifier. Hence, the IS is not ideal for our dataset.

Table 2. The generated images by GANs. The epoch when each image was generated is specified under the image.

GANs	0	1	2	3	4	5	6	7	8	9
WGAN	 48,000	 38,000	 42,000	 49,000	 49,000	 47,000	 49,000	 49,000	 49,000	 49,000
WGAN_DIV	 49,000	 42,000	 43,000	 48,000	 41,000	 50,000	 26,000	 43,000	 46,000	 28,000
WGAN_GP	 46,000	 48,000	 34,000	 38,000	 48,000	 45,000	 47,000	 44,000	 46,000	 42,000
LSGAN	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000
MMGAN	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000
NSGAN	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000
DRAGAN	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000
ACGAN	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000
DCGAN	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000
BEGAN	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000
EBGAN	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000	 50,000
VAE	 50,000	 22,000	 32,000	 49,000	 49,000	 49,000	 50,000	 49,000	 49,000	 49,000

Table 3. FID values for the evaluated GANs.

GANs	0	1	2	3	4	5	6	7	8	9	Average
WGAN	3.985	5.415	8.268	5.617	6.083	8.292	3.231	2.890	5.590	6.202	5.5573
WGAN_DIV	0.874	1.445	1.008	1.550	1.869	0.952	1.072	2.014	1.742	2.407	1.4933
WGAN_GP	1.288	1.442	0.761	1.451	2.210	1.089	1.001	2.492	1.704	1.860	1.5298
LSGAN	5.029	4.407	4.253	4.389	4.812	5.535	4.087	4.734	4.821	5.224	4.7291
MMGAN	30.134	4.743	5.389	13.582	6.267	6.769	6.683	4.630	6.440	6.769	9.1406
NSGAN	7.374	7.655	7.292	4.794	8.671	6.769	6.713	8.452	6.713	6.918	7.1351
DRAGAN	5.012	3.355	3.982	3.161	4.583	4.133	4.823	5.256	3.588	8.404	4.6297
ACGAN	7.169	6.073	6.086	6.148	6.631	7.145	6.297	7.457	8.055	8.448	6.9509
DCGAN	3.392	3.849	3.126	3.511	4.525	3.386	3.106	3.751	4.013	4.044	3.670
BEGAN	1.267	0.712	0.296	0.340	0.491	2.143	0.957	1.478	1.016	2.245	1.0945
EBGAN	2.718	2.681	1.402	3.476	0.545	3.597	2.940	2.608	3.532	4.932	2.843
VAE	0.796	0.309	0.114	0.861	0.353	0.437	0.432	0.770	0.722	0.81	0.484

Table 4. IS values for the evaluated GANs.

GANs	0	1	2	3	4	5	6	7	8	9
WGAN	1.0000001	1.0000002	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0000002
WGAN_DIV	1.0	1.0	1.0	1.0	1.0000001	1.0	1.0000002	1.0	1.0	1.0
WGAN_GP	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
LSGAN	1.0	1.0000002	1.0	1.0	1.0000001	1.0000001	1.0000001	1.0000002	1.0000002	1.0
MMGAN	1.0000002	1.0000007	1.0	1.0000002	1.0000002	1.0000002	1.0000004	1.0	1.0000002	1.0000002
NSGAN	1.0000002	1.0000002	1.0000001	1.0	1.0000002	1.0000002	1.0000002	1.0	1.0000001	1.0000005
DRAGAN	1.0	1.0	1.0	1.0000002	1.0	1.0000001	1.0000002	1.0	1.0000001	1.0000006
ACGAN	1.0	1.0	1.0	1.0000001	1.0000002	1.0000002	1.0000001	1.0000001	1.0000002	1.0000002
DCGAN	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0000001	1.0
BEGAN	1.0	1.0000002	1.0	1.0000001	1.0000002	1.0000001	1.0000002	1.0000002	1.0000001	1.0000002
EBGAN	1.0	1.0	1.0	1.0000001	1.0000001	1.0000002	1.0000001	1.0	1.0	1.0
VAE	1.0	1.0000002	1.0000002	1.0000002	1.0000002	1.0	1.0	1.0	1.0	1.0

4. Evaluation and Analysis

We evaluated the images by visually inspecting the quality of the images in Table 2 and the average FID score of the GAN model, as shown in the last column of Table 3.

In Table 2, the BEGAN and VAE outputs are not considered suitable for data augmentation, even though their FID values are the lowest at 1.0945 and 0.484, respectively. This is because the images are of poor quality and have little or no diversity. The DRAGAN model is similar to WGAN-GP, but the generated images are not of good visual quality compared with the output images from WGAN-GP and WGAN-DIV. Considering our evaluation metrics, both WGAN-GP and WGAN-DIV outputs are the most preferred. However, WGAN-DIV slightly outperforms WGAN-GP considering the FID score. The corresponding average FID scores of WGAN-GP and WGAN-DIV are 1.5298 and 1.4933, respectively, which corroborates their image quality. The low FID values translate to smaller distances between the fake and original data distributions. Figures 5–7 are the plots for WGAN-GP, WGAN-DIV and BEGAN for digits 3 images, respectively while Figures 8–10 shows the plot for WGAN-GP, WGAN-DIV and BEGAN for digits 0 images, respectively. However, we prefer WGAN-GP simply because of the noticeable quality of images across the digits 0–9. While each GAN displays similarities in its digit-plots from 0–9, Figures 5–7 show the loss plots for both WGAN-GP, WGAN-DIV and BEGAN for the digit-3 images, and Figures 8–10 do for the digit-0 images, respectively. We have only selected the digit 3 and 0, but the WGAN-GP plots have shown consistency and stability for the entire digit images in the range 0–9.

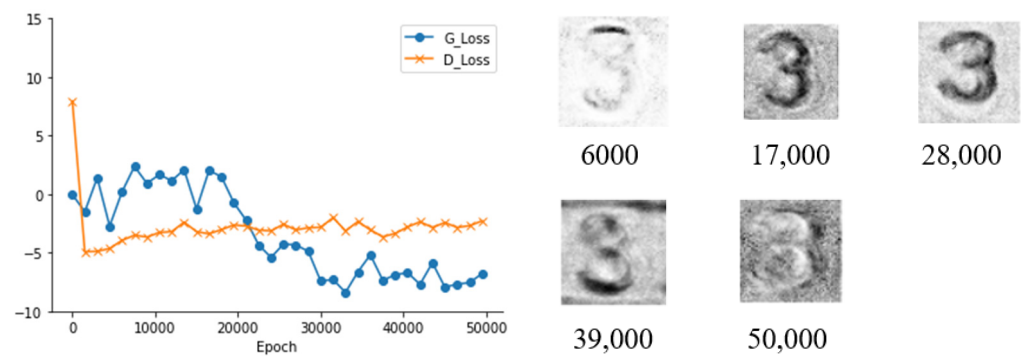


Figure 5. WGAN-GP (digit 3) G_Loss/D_Loss, and the image samples generated at the epochs under the images.

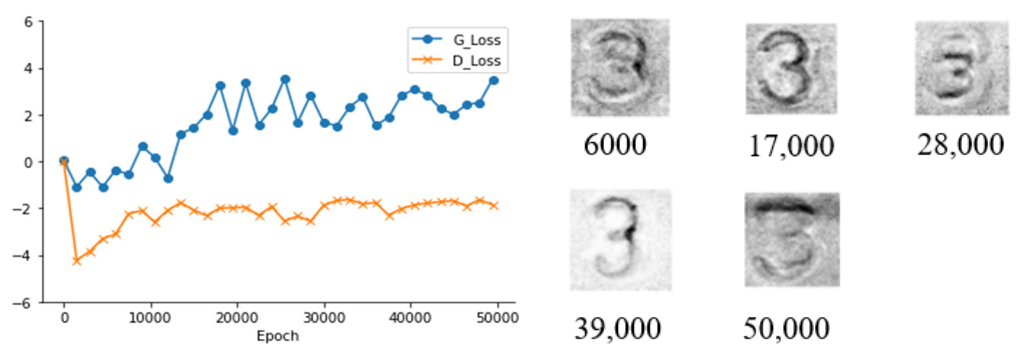


Figure 6. WGANDIV (digit 3) G_Loss/D_Loss, and the image samples generated at the epochs under the images.

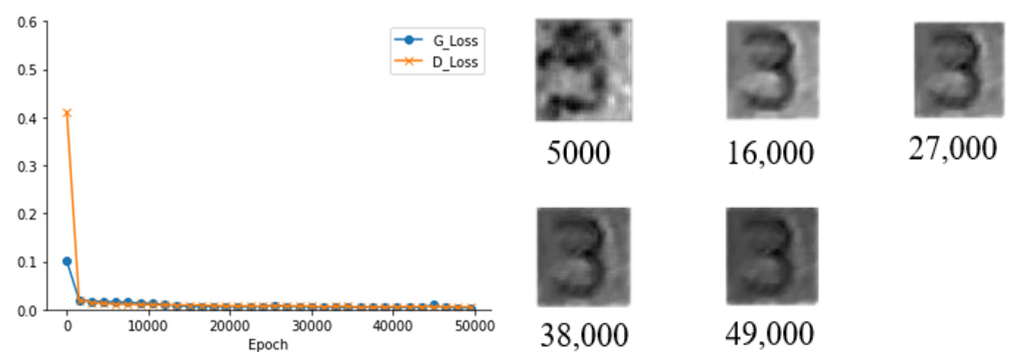


Figure 7. BEGAN (digit 3) G_Loss/D_Loss, and the image samples generated at the epochs under the images.

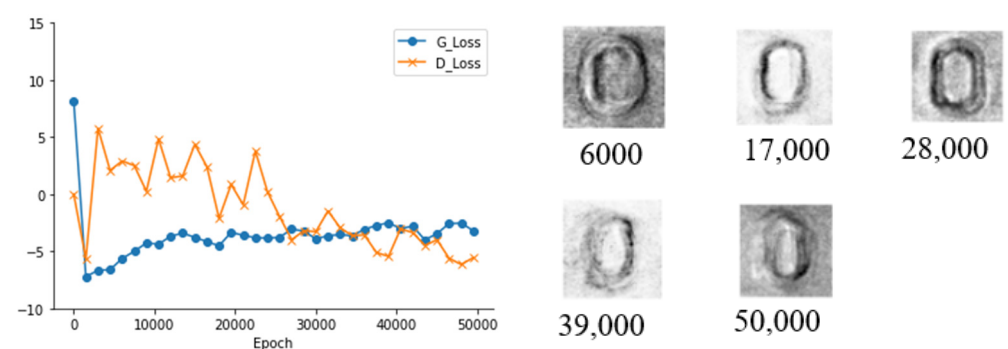


Figure 8. WGAN-GP (digit 0) G_Loss/D_Loss, and the image samples generated at the epochs under the images.

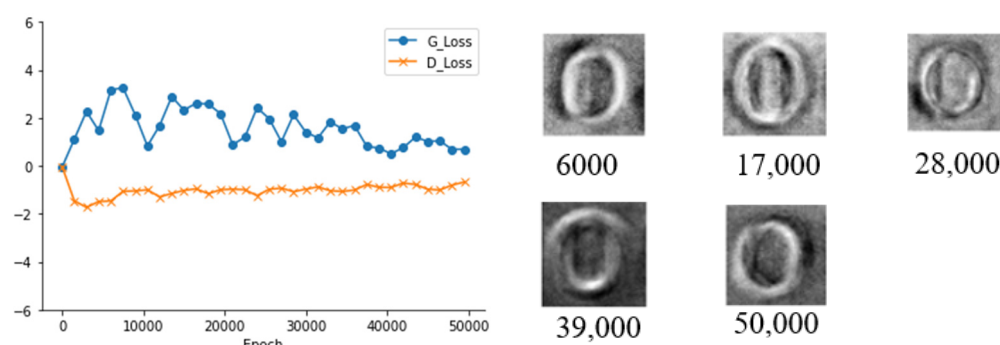


Figure 9. WGANDIV (digit 0) G_Loss/D_Loss, and the image samples generated at the epochs under the images.

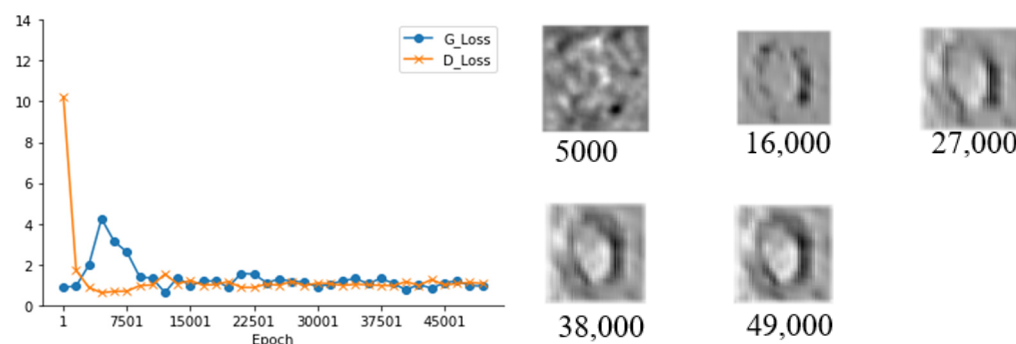


Figure 10. BEGAN (digit 0) G_Loss/D_Loss, and the image samples generated at the epochs under the images.

Unlike BEGAN in Figures 7 and 10, the WGAN_GP plots in Figures 5 and 8 show the increase in discriminator losses (D_Loss), whereas there is a decrease in the generator losses (G_Loss). In all WGAN-GP plots in our experiment, it is evident that whereas the discriminator attempts to maximize its ability to detect real data from fake data, the generator attempts to minimize the distance between the generated and real data.

The performance of a GAN model strongly depends on the dataset, and models are evaluated based on the quality of images generated in the context of a specific targeted domain [19]. After execution, the WGAN-GP on our dataset achieves an FID score of 1.5298. This score and the generated images reflect the high quality of the desired data required for augmentation. Hence, we conclude that WGAN-GP is an appropriate model for replenishing our dataset. In Table 5 [19], WGAN-GP shows good performance in terms of the FID score across several other datasets. We compared this result with the FID score of only one class of our dataset: digit 3. The FID scores obtained in a large-scale hyperparameter search for the FASHION (60,000 training datasets), CIFAR-10 (6000 images/class), and CELEBA datasets (202,599 face images) are displayed [19]. The best scores are the WGAN-GP scores of 24.5, 55.8, 30, and 0.761 for the FASHION, CIFAR-10, CELEBA, and ENGRAVED DIGITS datasets, respectively.

The WGAN-GP we implemented is shown in Figure 11 and Algorithm 1 in detail. During the implementation, the discriminator network (C) was first trained on a real batch of data (x), then it was trained on a batch of data generated from the noise (Z) via the generator (G). This was required to provide a random weighted average between real and generated image samples needed for gradient norm penalty. The discriminator's loss function was arranged such that it estimates the Wasserstein Distance with gradient penalty. The gradient penalty was computed for a batch of average samples to ensure it is 1-Lipschitz-Continuous and was first computed using prediction and weighted real/fake samples. The performance of Algorithm 1 depends on the number of iterations for θ converge, n_θ , the number of critic iterations per generator iteration, η_{critic} , and the batch size, m : $O(n_\theta * \eta_{critic} * m)$.

Table 5. FID of WGAN-GP digit-2 comparison with other datasets.

GANs	MNIST	FASHION	CIFAR	CELEBA	ENGRAVED DIGITS
WGAN	6.7 ± 0.4	21.5 ± 1.6	55.2 ± 2.3	41.3 ± 2.0	8.32 ± 0.1
WGAN_GP	20.3 ± 5.0	24.5 ± 2.1	55.8 ± 0.9	30.0 ± 1.0	0.761 ± 0.5
LSGAN	7.8 ± 0.6 *	30.7 ± 2.2	87.1 ± 47.5	53.9 ± 2.8 *	4.25
MMGAN	9.8 ± 0.9	29.6 ± 1.6	72.7 ± 3.6	65.6 ± 4.2	6.42 ± 1.0
NSGAN	6.8 ± 0.5	26.5 ± 1.6	58.5 ± 1.9	55.0 ± 3.3	6.39 ± 0.9
DRAGAN	7.6 ± 0.4	27.7 ± 1.2	69.8 ± 2.0	42.3 ± 3.0	10.75 ± 0.1
BEGAN	13.1 ± 1.0	22.9 ± 0.9	71.4 ± 1.6	38.9 ± 0.9	0.32
VAE	23.8 ± 0.6	58.7 ± 1.2	155.7 ± 11.6	85.7 ± 3.8	0.13 ± 0.1

The asterisk (*) indicates the presence of significant outlier runs, usually severe mode collapses or training failures.

Algorithm 1. WGAN-GP with our set of parameters [21]

$\lambda=10$, $\eta_{critic} = 5$, $\alpha=0.0002$, $\beta_1=0$, $\beta_2=0.9$.

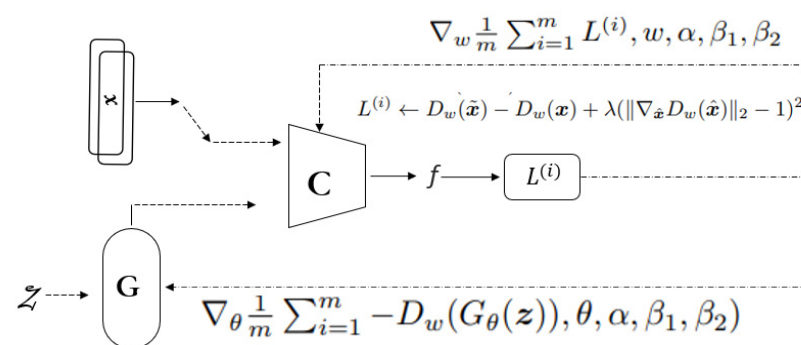
Require: The gradient penalty coefficient λ , the number of critic iterations per generator iteration η_{critic} , Adam hyperparameters α, β_1, β_2 .

Require: initial critic parameters ω_0 , initial generator parameters θ_0 .

```

1:   while  $\theta$  has not converged do
2:     for  $t = 1, \dots, \eta_{critic}$  do
3:       for  $i = 1, \dots, m$  do
4:         Sample real data  $x \sim \mathbb{P}_r$ , latent variable  $z \sim p(z)$ , a random number
            $\epsilon \sim U[0, 1]$ 
5:          $\tilde{x} \leftarrow G_\theta(z)$ 
6:          $\tilde{x} \leftarrow \epsilon x + (1 - \epsilon)\tilde{x}$ 
7:          $L^{(i)} \leftarrow D_w(\tilde{x}) - D_w(x) + \lambda(\|\nabla_{\tilde{x}} D_w(\tilde{x})\|_2 - 1)^2$ 
8:       end for
9:        $w \leftarrow Adam(\nabla_w \frac{1}{m} \sum_{i=1}^m L^{(i)}, w, \alpha, \beta_1, \beta_2)$ 
10:    end for
11:    Sample a batch of latent variables  $\{z^{(i)}\}_{i=1}^m \sim p(z)$ .
12:     $\theta \leftarrow Adam(\nabla_\theta \frac{1}{m} \sum_{i=1}^m -D_w(G_\theta(z)), \theta, \alpha, \beta_1, \beta_2)$ 
13:  end while

```

**Figure 11.** Architecture of a WGAN-GP.

5. Conclusions

In this study, we investigated data augmentation for engraved-digit image datasets on expiry date. Limited, fuzzy, and distorted digit images were required to improve image identification of consumables and cosmetic products. We evaluated the state-of-the-art GAN models, MMGAN, NSGAN, LSGAN, ACGAN, DCGAN, WGAN, WGAN-GP, WGANDIV, DRAGAN, BEGAN, EBGAN, and VAE by visually inspecting the results and calculating the FID values for each GAN. WGAN-GP and WGANDIV show stability and are suitable for the data augmentation task; however, we consider WGAN-GP to be the

preferred GAN owing to its quality output images after visual inspection. The consistency and stability of its G_Loss and D_Loss plots over the digits 0–9 is also satisfactory.

Our future research would focus on automatic recognition of engraved expiry-date digit images. After augmenting the few datasets to an abundance with WGANP. We intend to build a recognition model and train the model with these synthetic datasets to a very high degree of confidence. We conjecture that the image quality and diversity from WGAN-GP would contribute to our model's stability. The new recognition model would not only be tailored to engraved-digit image recognition, but it would also serve as a benchmark for models required to train similar datasets to high performance.

Author Contributions: A.A. and I.Y.J. conceived and designed the experiments; A.A. performed the experiments; A.A. and I.Y.J. analyzed the data; A.A. wrote the paper and I.Y.J. re-organized and corrected the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2021R1F1A1064345) and by the BK21 FOUR project funded by the Ministry of Education, Korea (No. 4199990113966).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Tran, T.N.; Felfernig, A.; Trattner, C.; Holzinger, A. Recommender systems in the healthcare domain: State-of-the-art and research issues. *J. Intell. Inf. Syst.* **2021**, *57*, 171–201. [\[CrossRef\]](#)
- Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Process. Mag.* **2012**, *29*, 141–142. [\[CrossRef\]](#)
- Chollet, F. *Deep Learning with Python*; Simon and Schuster: New York City, NY, USA, 2021.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2020**, *63*, 139–144.
- Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
- Medium.com. Available online: <https://jonathan-hui.medium.com/gan-energy-based-gan-ebgan-boundary-equilibrium-gan-began-4662cceb7824> (accessed on 27 August 2022).
- Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [\[CrossRef\]](#)
- Antoniou, A.; Storkey, A.; Edwards, H. Data augmentation generative adversarial networks. *arXiv* **2017**, arXiv:1711.04340.
- Iqbal, A.; Sharif, M.; Yasmin, M.; Raza, M.; Aftab, S. Generative adversarial networks and its applications in the biomedical image segmentation: A comprehensive survey. *Int. J. Multimed. Inf. Retr.* **2022**, *11*, 333–368. [\[CrossRef\]](#)
- You, A.; Kim, J.K.; Ryu, I.H.; Yoo, T.K. Application of generative adversarial networks (GAN) for ophthalmology image domains: A survey. *Eye Vis.* **2022**, *9*, 6. [\[CrossRef\]](#)
- Bengio, Y.; Laufer, E.; Alain, G.; Yosinski, J. Deep generative stochastic networks trainable by backprop. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014.
- Wenzel, M. Generative Adversarial Networks and Other Generative Models. *arXiv* **2022**, arXiv:2207.03887.
- Borji, A. Pros and cons of gan evaluation measures. *Comput. Vis. Image Underst.* **2019**, *179*, 41–65. [\[CrossRef\]](#)
- Shmelkov, K.; Schmid, C.; Alahari, K. How good is my GAN? In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
- Esteban, C.; Hyland, S.L.; Rätsch, G. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv* **2017**, arXiv:1706.02633.
- Denton, E.L.; Chintala, S.; Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [\[CrossRef\]](#)
- Zhu, X.; Vondrick, C.; Fowlkes, C.C.; Ramanan, D. Do we need more training data? *Int. J. Comput. Vis.* **2016**, *119*, 76–92. [\[CrossRef\]](#)
- Dowson, D.C.; Landau, B. The Fréchet distance between multivariate normal distributions. *J. Multivar. Anal.* **1982**, *12*, 450–455. [\[CrossRef\]](#)
- Lucic, M.; Kurach, K.; Michalski, M.; Gelly, S.; Bousquet, O. Are gans created equal? A large-scale study. *Adv. Neural Inf. Process. Syst.* **2018**, *31*. [\[CrossRef\]](#)
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [\[CrossRef\]](#)

21. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [[CrossRef](#)]
22. Krizhevsky, A.; Vinod, N.; Geoffrey, H. The CIFAR-10 Dataset 2014. Available online: <http://www.cs.toronto.edu/kriz/cifar> (accessed on 6 August 2022).
23. Barratt, S.; Sharma, R. A note on the inception score. *arXiv* **2018**, arXiv:1801.01973.
24. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
25. Wu, J.; Huang, Z.; Thoma, J.; Acharya, D.; Van Gool, L. Wasserstein divergence for gans. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
26. Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv* **2016**, arXiv:1701.00160.
27. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul, S.S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
28. Kodali, N.; Abernethy, J.; Hays, J.; Kira, Z. On convergence and stability of gans. *arXiv* **2017**, arXiv:1705.07215.
29. Odena, A.; Olah, C.; Shlens, J. Conditional image synthesis with auxiliary classifier gans. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
30. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
31. Zhao, J.; Mathieu, M.; LeCun, Y. Energy-based generative adversarial network. *arXiv* **2016**, arXiv:1609.03126.
32. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
33. Berthelot, D.; Schumm, T.; Metz, L. Began: Boundary equilibrium generative adversarial networks. *arXiv* **2017**, arXiv:1703.10717.
34. Wang, J.; Yu, L.; Zhang, W.; Gong, Y.; Xu, Y.; Wang, B.; Zhang, P.; Zhang, D.I. A minimax game for unifying generative and discriminative information retrieval models. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Tokyo, Japan, 7–11 August 2017.
35. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 84–90. [[CrossRef](#)]
36. Joseph, S.; George, J. Data augmentation for handwritten character recognition of MODI script using deep learning method. In Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems, Singapore, 15 May 2020.
37. Wong, S.C.; Gatt, A.; Stamatescu, V.; McDonnell, M.D. Understanding data augmentation for classification: When to warp? In Proceedings of the 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Gold Coast, Australia, 30 November 2016.
38. Starynska, A.; Easton, R.L., Jr.; Messinger, D. Methods of data augmentation for palimpsest character recognition with deep neural network. In Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, Kyoto, Japan, 10 November 2017.
39. Lake, B.M.; Salakhutdinov, R.; Tenenbaum, J.B. Human-level concept learning through probabilistic program induction. *Science* **2015**, *350*, 1332–1338. [[CrossRef](#)]
40. Cohen, G.; Afshar, S.; Tapson, J.; Van, S.A. EMNIST: Extending MNIST to handwritten letters. In Proceedings of the International Joint Conference on Neural Networks (IJCNN) IEEE, Anchorage, AK, USA, 14 May 2017.
41. Parkhi, O.M.; Andrea, V.; Andrew, Z. Deep face recognition. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–11 September 2015.
42. Tanaka, F.H.; Aranha, C. Data augmentation using GANs. *arXiv* **2019**, arXiv:1904.09135.
43. Wickramaratne, S.D.; Mahmud, M.S. Conditional-GAN based data augmentation for deep learning task classifier improvement using fNIRS data. *Front. Big Data* **2021**, *4*, 659146. [[CrossRef](#)]
44. Wei, K.; Li, T.; Huang, F.; Chen, J.; He, Z. Cancer classification with data augmentation based on generative adversarial networks. *Front. Comput. Sci.* **2022**, *16*, 162601. [[CrossRef](#)]