Journal of Engineering Science and Technology Review 16 (5) (2023) 186 - 195

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

A Novel Variant Autoencoder for Class Imbalance

Chen Xinyuan[1], Zhao Yihang[2], Feng Ao[3,*], Shi Xiaoshi[3, 4] and Tang Zuoliang[3, 4]

[1]*College of engineering, Chinese University of Hong Kong, HorseMaterial Water, Shatin, New Territories, 999077, Hong Kong SAR, China*
[2] *College of Information Engineering, Sichuan Agricultural University, 46 Xinkang Road, Yucheng District, Ya'an 625014, China*
[3] *College of mechanical and electrical engineering, Sichuan Agriculture University, Ya'an 625014, China*
[4] *College of Resources, Sichuan Agriculture University, Chendu 610000, China*

Received 16 August 2023; Accepted 30 October 2023

Abstract

Deep learning requires a large amount of data, and enhanced processing of the data is particularly important. This study aims to investigate the enhancement of the dataset from the perspective of the underlying data distribution and solve the problem of unbalanced data samples. In this study, a novel approach called the single-sample sampling variant autoencoder (S3VAE) was proposed to generate data which were then compared. Experimental results demonstrate that, under the same data discard rate, the data generated by the S3VAE architecture exhibit a test accuracy closer to that of the original data, which proves the ability of the S3VAE architecture to generate results closer to the original data. Furthermore, the reconstruction abilities of C-VAE and S3VAE were compared using two public datasets and conducted five different discard rate experiments. As observed, the test accuracy of S3VAE is higher than that of C-VAE in all cases. With an increase in the data discard rate, the advantage of S3VAE becomes more pronounced. When the data discard rate is 97.5%, the test accuracy of S3VAE is 2.7% higher than that of C-VAE. These results confirm that the method has a significant positive effect on data enhancement and can be effectively used in practical scenarios. Moreover, this method can be extended to most advanced variant autoencoders.

Keywords: Data Augmentation; Computer Vision; Improve VAE

1. Introduction

With the introduction of efficient neural networks such as LeNet[1], AlexNet [2], and ResNet [3], research scholars have progressively focused on network structure in the previous decade. They proposed novel strategies to improve the network structure and hence increase model accuracy. However, in addition to network topology, the dataset used for model training can have a considerable impact on prediction accuracy [4,5]. When using the CIFAR-10 dataset for model evaluation after training, depending solely on its test set may result in inferior real-world performance of the final model, and the actual performance of the model is frequently lower than its performance in the experimental setting [6].

Three factors contribute to degradation in recognition accuracy: the generalization, adaptation gap, and distribution gaps. Theoretical analysis, supported by experiments in the study of Recht et al., suggests that the generalization and adaptation gaps have a relatively minor effect on the results. Consequently, the main factor responsible for decreased model accuracy during the testing phase is the distribution gap. In practical applications of the model, the model can continuously fit only the training dataset. This limitation causes difficulty in addressing the distribution gap between the test datasets, regardless of how well the model is designed. Therefore, data augmentation emerges as the key solution to this problem. Historical studies have traditionally assessed the effect of data augmentation through a series of comparative experiments, which consistently indicate that the classification model obtained by training with data augmentation is more accurate than the baseline model. In terms of experimental results, data augmentation techniques indeed improve the experimental results. Nevertheless, some methods of augmenting data may cause particular dataset distributions to be overfit, leading to models that underperform in real-world scenarios. This problem can be solved by controlling the underlying pixel distribution of the image data. If this distribution can be effectively controlled, then the distribution gap between different test datasets can be managed. However, modeling the underlying pixel distribution of image data is highly intricate. From the perspective of dataset production, manually manipulating the data to fit the underlying pixel distribution of the original test set is difficult. Early data enhancement techniques encompassed numerous manual image processing techniques [7-10]. These methods included geometric transformations such as flipping, rotating, cropping, and random erasing, as well as pixel operations like noise injection, color space transformations, and image mixing. These data enhancement techniques considerably augment the size of the deep learning training dataset. This increase in diversity and capacity of the original training dataset enhances the generalization ability and robustness of the deep learning model, which prevents overfitting in practical applications. However, these data enhancement techniques, which are influenced by human factors, may occasionally produce adversarial examples [11]. Adversarial samples are formed by introducing perturbations to the original dataset. These

*E-mail address: fengao@stu.sicau.edu.cn

doi:10.25103/jestr.165.23

perturbations are often imperceptible for the human eye or have a minimal effect on human recognition. However, they can easily interfere with the model and lead to incorrect judgments by the machine. Therefore, simple manual image processing techniques not only fail to improve the generalization ability of the recognition network but also may affect the robustness of the network in some cases [12].

This study broadens this view by rebuilding the original dataset. The substantial body of literature on deep generative models, such as the variation autoencoder (VAE) [13-26] and generative adversarial network [27-32], is cited numerous times. These models use hidden variables in various ways to regulate the production of images and their underlying pixel distribution. In this approach, they address the distribution gap between datasets indirectly. This paper, in particular, proposes a novel technique to VAEs known as the variable self-encoder. It uses a single-sample sampling training method, which means that each VAE training uses only one sample. The generated model can, to some extent, adjust the distribution gap between datasets and performs well in subsequent data augmentation tasks.

2. State of the art

Neural networks are the foundation of some contemporary data augmentation approaches. Wang et al. [33] employed a small neural network named "Small Net" to select the most successful data augmentation strategies for a given dataset, and their method was dubbed "Neural Augmentation". Cubuk et al. [34] suggested a reinforcement learning search strategy and data "Auto Augmentation". There are also approaches based on generative models. To accomplish effective data enhancement, Antoniou et al. [35] presented "DAGAN" (Data Augmentation GAN), which combined principles from CGAN [36] and used an UResNet generator structure. Recht et al. [37] constructed a new test set in the same way they created the original dataset and tested it with a variety of models. On this new test set, their experimental results revealed a drop in recognition accuracy for image classification networks. Despite their success in data improvement tasks, these models nevertheless face several challenges: 1) Model complexity: The intricacy of these approaches makes them challenging to employ in real-world circumstances. In some circumstances, the time necessary for data improvement may surpass the time required for the recognition network to train. 2) Lack of explanation: These models frequently fail to provide a thorough explanation of why their strategies produce effective data improvement results.

The primary goal of model development is clearly not to simply replicate the dataset. The model's purpose is to learn similar distributions from the original dataset. According to Hou et al. [25], images created by the ordinary variant self-encoder (plain VAE, P-VAE) are greatly fuzzy due to the intrinsic constraints of pixel-to-pixel reconstruction errors. The precise formula is as follows:

$$L_{rec} = -E_{q(Z|X)}(\log p(X|Z)) \tag{1}$$

Among them, L_{rec} represents reconstruction loss, $E_{q(Z|X)}$ represents the posterior distribution of the hidden variable Z under the given input data X, $P(X|Z)$ represents the probability distribution of input data X given the hidden variable Z.

The equation cannot account for the data's spatial connectedness and perceptibility. To capture information that cannot be conveyed by the reconstruction error, Hou et al. substituted the P-VAE reconstruction error loss function with a feature perceptual loss. This adjustment resulted in superior results and a significant improvement in the model's generating ability.

Hou et al. concentrated on the VAE approach to reproducing datasets. P-VAE just reduces the reconstruction error, according to their findings. Such a narrow focus, however, would result in a data improvement network that effectively functions as a "copy and paste" method. From this vantage point, we can see that the issue arises during the optimization of the loss function. The following is the P-VAE loss function:

$$L_{kl} = D_{kl}(q(Z|X)||p(Z)) \tag{2}$$

$$L_{vae} = L_{rec} + L_{kl} \tag{3}$$

Among them, L_{kl} represents KL loss, D_{kl} represents Kullback-Leibler divergence, $q(Z|X)$ represents the posterior distribution of the hidden variable z given the input data X, $p(Z)$ represents the prior distribution of hidden variable Z, L_{vae} represents the loss function of the variational autoencoder, L_{rec} represents reconstruction loss.

If we over-optimize the L_{rec}, P-VAE will continue to generate data that is very similar to the input photos. In this case, the model learns the original dataset's surface distribution rather than investigating the underlying distribution. Given these factors, we argue that improving the model's ability to interpolate the potential space is essential for improving its generative performance. The model's capacity to interpolate the potential space results in a more thorough study of the potential distribution based on the original dataset, which leads to increased generative ability.

To overcome the challenges raised by the probalitity generative model outlined above, several broad kinds of techniques can be applied.

1) KL annealing at a low cost: The KL cost annealing [38] method is simple to utilize. At the start of training, the KL term is multiplied by a weighting factor of 0, giving $q(Z|X)$ additional time to learn for encoding information from X into Z. As training proceeds, the weighing factor is eventually increased to 1. This strategy is generally paired with word dropout, which is a standard method for weakening the decoder.

2）Free Bits: The idea of Free Bits is also simple: each dimension of the KL term is allowed to "reserve a little space" for allowing more information to be encoded into the latent variable. Specifically, if the KL value in a dimension is too small, then it remains untouched until it increases beyond a threshold before optimization. This process leads to the loss function:

$$\max_{\theta,\phi} E_{q(x)}E_{q\phi}(Z \mid X)\log p_\theta(X \mid Z) - \sum_{i=1}^{D}\max(KL(q_\varphi(Z_i \mid X) \| p_\theta(Z_i)),\in) \quad (4)$$

Among them, $\max_{\theta,\phi}$ Represented in parameter θ and ϕ Find maximum value on. $E_{q\phi}(Z \mid X)$ represents the expectation on the data distribution $q(X)$. $E_{q\phi}(Z \mid X)$ represents the expectation on the posterior distribution $q(Z \mid X)$ of the hidden variable Z given the input data X. $p_\theta(X \mid Z)$ represents the conditional probability distribution p of input data X, given the hidden variable $q_\phi(Z_i \mid X)$. $q_\phi(Z_i \mid X) \| p_\theta(Z_i)$ represents the hidden variable Z under the given input data X. $q_\phi(Z_i \mid X)$ represents a posterior distribution and p_θ represents a prior distribution. $\in$ represents a non-negative threshold.

The entire KL can be controlled without breaking it down into each dimension. However, this process may result in very few dimensions being actively involved, with the vast majority of the dimensions of Z not containing information about X. An advantage of the Free Bits method is its simplicity. One disadvantage is that the threshold ε needs continuous adjustment.

3) Normalizing flow: Various variants of the normalizing flow concept exist [13], including autoregressive Flow and inverse autoregressive flow. The core idea involves starting with a latent variable sampled from a simple distribution and then increasing its flexibility by iterating a sequence of reversible transformations. Most of these methods aim to improve the posterior distribution given that direct Gaussian modeling often lacks accuracy in addressing realistic problems.

4) CNN decoder: CNNs are worth considering. If only traditional CNNs are used, then the contextual capacity may be limited. Thus, a dilated CNN decoder can be used [39]. The width of the dilated CNN decoder can be adjusted, which allows it to approximate various models, from the simple bag-of-words model to the most complex LSTM. The most suitable configuration can be determined by experimenting, and this method tends to perform well.

5) Additional loss: The introduction of additional losses [40], such as bag-of-words loss, can be valuable. This approach effectively increases the weight of reconstruction, which encourages the model to focus more on optimizing reconstruction terms rather than the KL divergence. It is also highly effective in preventing KL divergence from vanishing.

Our work is designed with the principle of simplicity in mind and aims to explore the fundamental reasons for the effectiveness of data enhancement to shape our data enhancement model. We use a probability generative model and propose a novel training method for variant self-encoders. This method involves improving the number of samples and direct manipulation within the hidden space. As a result, 1) it helps mitigate the distortion of generated images to a certain extent and effectively prevents KL-vanishing; 2) it requires minimal changes to the code while allowing us to target the pixel distribution of images; and 3) this method has yielded positive results in pest identification projects we have led. However, this data enhancement task can be time consuming. The specific details of our work are discussed in Section 3.

3. Methodology

3.1 Improvement method

The three disadvantages of S3 are as follows: a. Assumption correctness: The main difficulty in S3 is ensuring that the single photo extracted from the training set x is suitably representative. b. Overfitting risk: Given that the network is trained with only one image, the risk of overfitting is substantial in the case of S3. c. Generalization ability collapse: Because of the significant loss in training data, the model trained with S3 lacks generalization capability. The essence of problem 1 is to select a representative training sample from a huge quantity of image data.

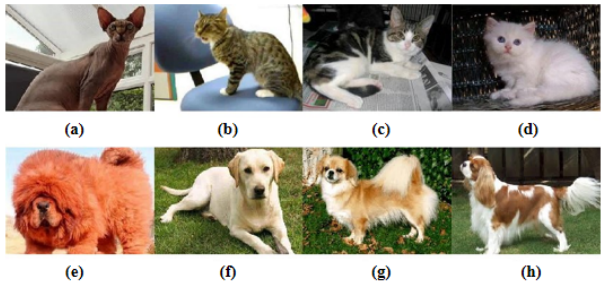

Fig. 1. Representative and nonrepresentative samples.
Note: Image (a) features a hairless cat, and image (e) showcases a pig mastiff, each belonging to a particular breed of cat and dog. The remaining six images depict more common cats and dogs. When using S3 to sample from a training set comprising these images, we should obviously refrain from selecting training samples (a) and (e) to satisfy the assumptions of S3.

Fig. 2. Confusion caused by a non-representative sample

The hairless cat is on the left in Fig.2, and the deerhound is on the right. Given its significant resemblance to the deerhound, the hairless cat is not typical of the cat group. This resemblance could easily lead to classification errors in an image recognition network. We believe that choosing a representative image is as simple as selecting images from the same class that have similar qualities to the majority of images in that class. Fig.1 depicts representative and non-representative samples. Numerous picture samples satisfy such constraints for a specific type of image dataset, making S3 applicable to practically all images in that dataset. Because the picture collection belongs to the same class, the image samples in it automatically share common characteristics.

In S3 of problem 2, we use Gaussian noise to disturb the propagation stream before we output the mean and variance of the VAE to reduce the danger of overfitting. Fig.3 depicts the network structure. This method successfully reduces model overfitting. Notably, the Gaussian noise layer is important not just during backpropagation but also during

testing. With this dual feature, the VAE can learn how to introduce noise for better generating results.

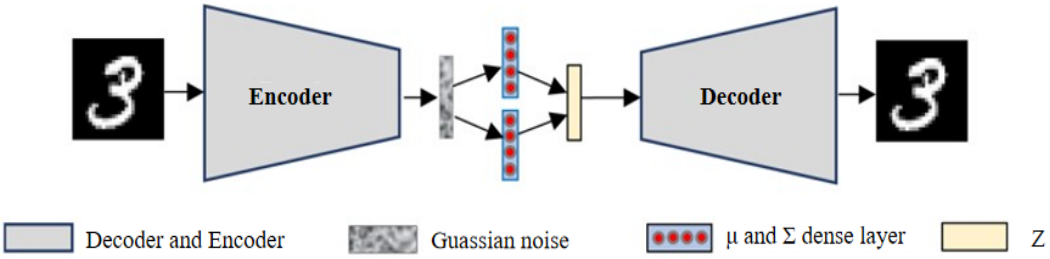

Fig. 3. Gaussian noise-based S3VAE

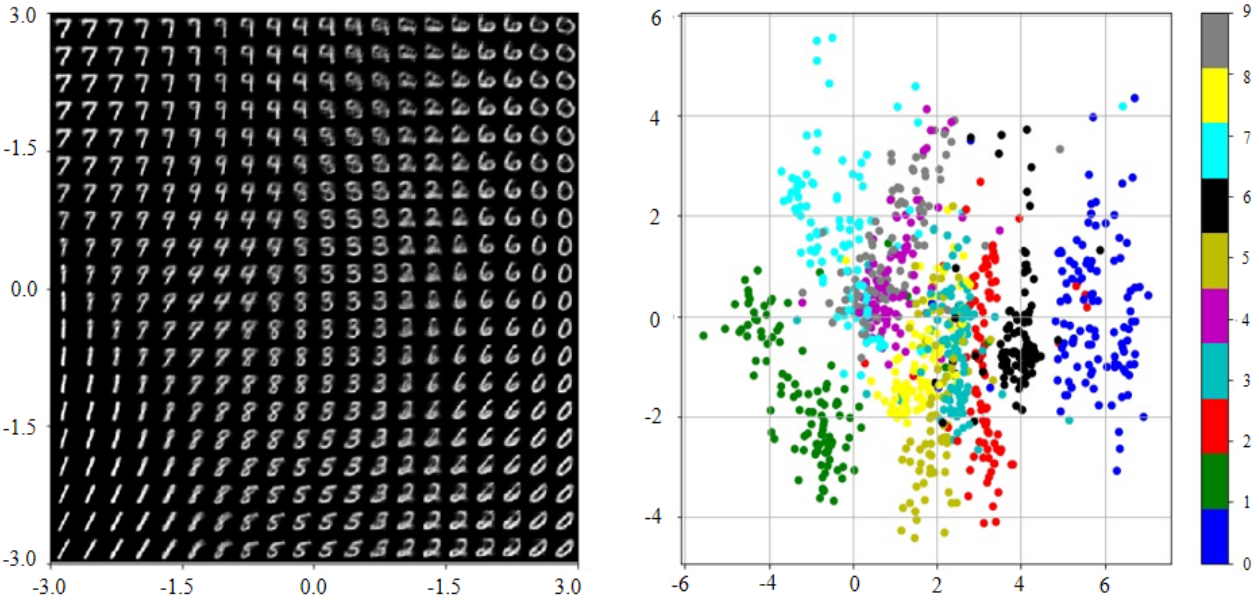

Fig. 4. Gaps in the distribution of samples in the hidden space

In the hidden space, the sample set consists of numerous Gaussian distributions that must all converge into a single Gaussian distribution during coding. The mean and variance of a Gaussian distribution define it. The variance dictates the size of the distribution, whereas the mean specifies where the center lies in the hidden space. When many Gaussian distributions converge into one, a gap in the hidden space appears, as illustrated in Fig. 4 All distributions cannot be spread out in this gap. Because the hidden space lacks a comparable dataset, sampling in this place may produce unfavorable results. Furthermore, if the sampled distribution substantially fuses with the encoded distribution, the previously noted problem of inferential collapse may occur. We examine the experimental advantages of single-sample sampling, which overcomes this difficulty. Single-sample sampling VAE (S3VAE) works reasonably well despite its fundamentally simplistic approach. It has various advantages, including reducing the VAE's posterior crash problem and enabling excellent data improvement. The Experiments section contains detailed information about the experiments, network structure, and training parameters.

3.2 Experimental design

S3VAE is a simple optimization that improves all VAE architectures. The first set of tests tries to qualitatively demonstrate the method's usefulness in addressing the sample imbalance problem, while the second set aims to numerically demonstrate how the strategy can improve VAE.

3.2.1 Experiment 1

In Table 1, we validate the proposed method using two publicly available datasets: one is the CIFAR dataset, which is a simple 32 × 32 pixel multiclassification dataset; the other is the PlantVillage dataset, which is a more complex 256 × 256 pixel multiclassification dataset. These datasets are originally balanced. However, our objective is to verify the effectiveness of the proposed method in addressing unbalanced dataset. For this purpose, we intentionally unbalance one of the classes by removing some data from it. The deleted data are referred to as “discarded data” while the data that remain after this removal and the data generated to compensate for the discarded data are termed “generated data”.

Table. 1. Statistical information on the dataset

Dateset	Resolution	Class	Training image per class		
			Min	Mean	Max
CIFAR	32×32	10	5000	5000	500
PlantVillage	256×256	10	500	5000	500

We do a variety of experiments. The first set makes use of the entire balanced dataset before any data is removed. In the second set, we run experiments on the dataset after it has been cleaned up. The generated dataset is used in a third series of tests. In this set, 60%, 80%, and 90% of the data for each dataset category are removed at random. The PlantVillage dataset, for example, may yield only 50 photos

for training while the validation set data remains constant. We discover more optimum discard ratios to apply to the CIFAR dataset based on the experimental results with the PlantVillage dataset, specifically 90%, 95%, and 97.5%. To demonstrate the diversity of the created data, we reproduce the leftover data after elimination in a fourth series of tests. For example, if we eliminate 90% of the data in the PlantVillage dataset, the training set for a category will consist of only 50 photos. Then we rerun the process and add 50 more photographs to the training set. This collection of studies compares duplicated data to created data, which is the topic of the third set of experiments.

The four datasets given before are required for our experimental strategy. Following that, we use an evaluation model to evaluate these datasets based on evaluation criteria. The main contrast is between the third set of experiments and the other groups. If the third set of experiments outperforms the second, the generated data are correct. Similarly, if the third set of experiments outperforms the fourth set, then the collected data are diverse. Section 4 contains a more detailed analysis of the experimental data.

3.2.2 Experiment 2

This series of studies' experimental results and ideas are very similar to those in Experiment 1. However, in these trials, we are more concerned with comparing S3VAE to other VAEs. For this comparison, we used the conditional VAE (C-VAE) as the basic model. Fig.5 depicts the fundamental model architecture. It is called "the basic model architecture" because we use it to demonstrate the universality of the single-sample sampling approach.

S3VAE's capacity to address unbalanced datasets is proved by comparing its accuracy to C-VAE in validation studies done under the same experimental settings. S3VAE's core architecture should be the same as C-VAE's. If the basic architecture differs, then any comparison between the two will not provide a clear indication of whether the VAE benefits from the single-sample sampling strategy. Fig.6 depicts the basic model design of S3VAE.

Notably, our primary focus is on data enhancement for unbalanced datasets, not the expansion of balanced datasets. This task requires the generative model targeting precise generation of individual species, with the ability to control a variable for generating a certain class of images. C-VAE is well suited for this task compared with other VAEs. Thus, our choice of infrastructure is C-VAE. However, the choice of the comparison model (C-VAE) does not significantly affect the conclusions of the experiments. Earlier in our experiments, we explored or constructed various VAEs. For example, we experimented with different VAE network architectures by altering factors such as the weighted value of reconstruction error and KL divergence error. We observed that VAE-like models generate ambiguity when the KL divergence error disappears during the optimization process. Thus, we tried increasing the weight of the KL divergence error in the loss function. We also experimented with changes in the number of encoding and decoding layers and specific parameters. In practice, these infrastructure parameters do not fundamentally affect our analysis of VAE.

These VAEs serve as comparison models in our experiments, and the VAEs that incorporate our methods outperform VAEs with the original structure. The choice of C-VAE as our representative model is justified for several reasons: 1. C-VAE exhibits some structural differences compared with the typical VAE. 2. C-VAE has been employed to deal with some problems, and we have identified common defects (including the theoretical defects mentioned earlier and the generation of unsatisfactory images) that are relevant to other VAEs. 3. C-VAE excels in generating certain types of images given that it is designed for targeted image generation. If our model is superior to C-VAE, then it implies superiority over most VAEs. Our familiarity with C-VAE selection is based on our experience conducting experiments.

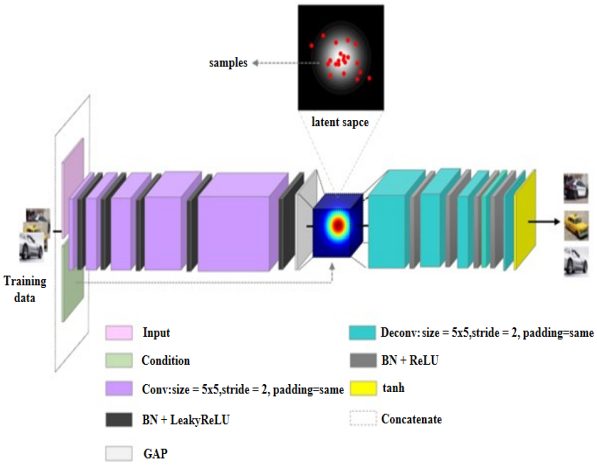

Fig. 5. Basic model architecture of C-VAE

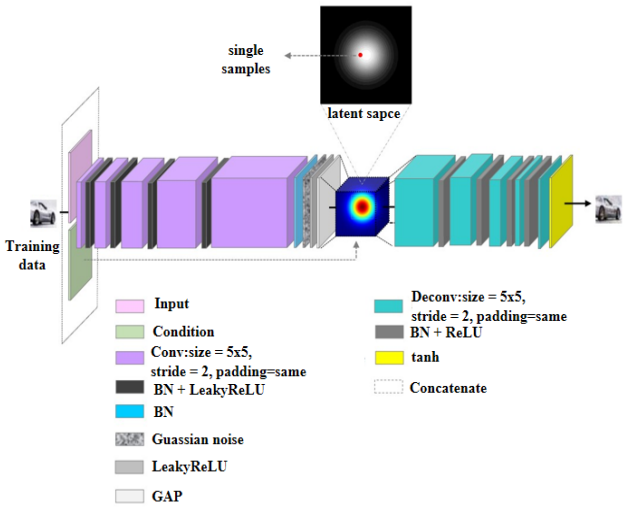

Fig. 6. Basic model architecture of S3VAE.

3.3 Evaluation indicator

The evaluation model employed is ResNet-18. The convergence of this model is used as an evaluation criterion but is not described in detail in this study. The model structure is shown in Fig.7.

For all experiments, the validation set used is from the original balanced dataset, and accuracy on the validation set serves as an evaluation metric, as shown in the Table 2. However, given that these experiments involve unbalanced datasets, recall and precision are also used as evaluation metrics. Therefore, we consider accuracy on the validation set, recall, and precision on the training set as criteria for assessing the performance of the model, all of which pertain to positive samples. We consider the classes that do not undergo the discard operation as positive samples.

Table 2. Evaluation indicator parameters

TP	The true category is positive, and the category predicted by the model is also positive
FP	The predicted category is positive, but the true category is negative
FN	The predicted category is negative, but the true category is positive
TN	The true category is negative, and the predicted category is also negative
Accuracy	Performance accuracy on the validation set
Precision	Proportion of samples with a true positive category among those predicted as positive
Recall	Proportion of samples successfully predicted as true positives by the model
F^{β}	Weighted ratio of precision to recall

To provide a clearer characterization of the experimental results, we assign higher weight to the recall rate, which makes low precision rates more pronounced. Thus, we set the weight as β=1.5. We also normalize the evaluation metrics, excluding precision, to make the comparison results more intuitive.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F^{\beta} = \frac{1 + \beta^2}{\frac{1}{Precision} + \frac{\beta^2}{Recall}} \tag{8}$$

In our model training method, we used a decaying learning rate algorithm and an early stopping strategy. In Experiment 1, the early stopping criterion was based on monitoring the accuracy of the validation set. Meanwhile, in Experiment 2, it was determined by the number of combined samples, which corresponds to the number of combined images from a single sample generated by S3VAE.

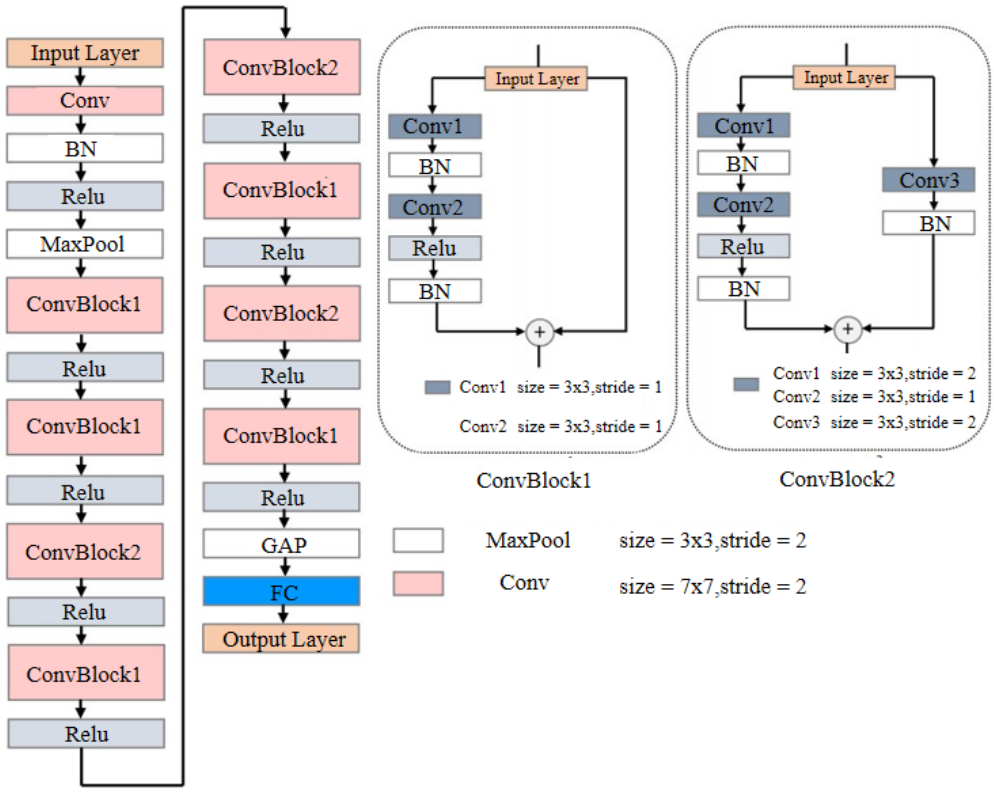

Fig. 7. Network structure of Resnet-18.
Note: BN stands for batch normalization, and GAP stands for global average pooling.

Table 3. Training-related parameters

Parameters	**Learning rate**	**Momentum**	**Decay rate**	**Epoch**	**Early stopping strategy for monitoring targets**	**Early stopping parameters**	**Number of batches OR sample combinations**
Value	0.0001	0.9	0.001	200	Validation set accuracy	0.01	64

Experiment 2 cannot be designed purely on the basis of a specific network architecture or a predetermined set of parameters to demonstrate the superiority of our strategy. As seen in the existing network architecture, this method may introduce an element of chance. To avoid the influence of experimental chance, we systematically altered the network architecture and associated essential parameters.

Given that the Gaussian noise layer is a critical component of our technique, we ran experiments to see how changing the settings of this layer affected the findings. We also changed the weighting of the loss function to see if our proposed method might outperform the techniques described by other researchers in Section 2.Tables 3 and 4 indicate parameter settings.

Table 4. List of related parameters of the network architecture

Parameters	**Symbolic representation**
Number of coding and decoding layers	L
Gaussian noise layer variance	σ
Datasets (resolution)	/
Data discard rate (%)	/
Models	/
Ratio of reconstruction error to KL dispersion error	K
Parameters	Symbolic representation

Given the randomization of data removal, all discard procedures were performed five times to guarantee that the experimental results were consistent despite the randomness of the removal. This experience demonstrated that the randomization of the deleted data had no effect on the experimental results. To acquire the final experimental results, the average of the five experiments was calculated. Given the extremely low probability of chance influencing the optimization process, we feel that repeating the experiment five times effectively neutralized the influence of chance on the outcomes.

4. Result Analysis and Discussion

In this section, we analyze the experimental results of Experiments 1 and 2.

4.1 Experimental Result (I)

To ensure that the photos created by the algorithms under consideration accurately represented the target classes, we classed them using a deep learning model trained on the original dataset. We confirmed that the projected classes corresponded to the target classes. We used the ResNet-18 model in this work to ensure that the accuracy of the third set of tests was equal to or greater than that of the second set; otherwise, the data collected did not represent the intended classes. The experimental results showed that as the percentage of deleted data increased, accuracy on the unbalanced dataset declined dramatically. The model barely learned any characteristics from the dataset as a result of this drop. However, when the second set of data was generated, the third set of experimental results indicated a considerable improvement over the second set. As a result, the approach produced a dataset containing the needed features. These features were learned by the model, which enhanced image accuracy.Fig.8 depicts the results.

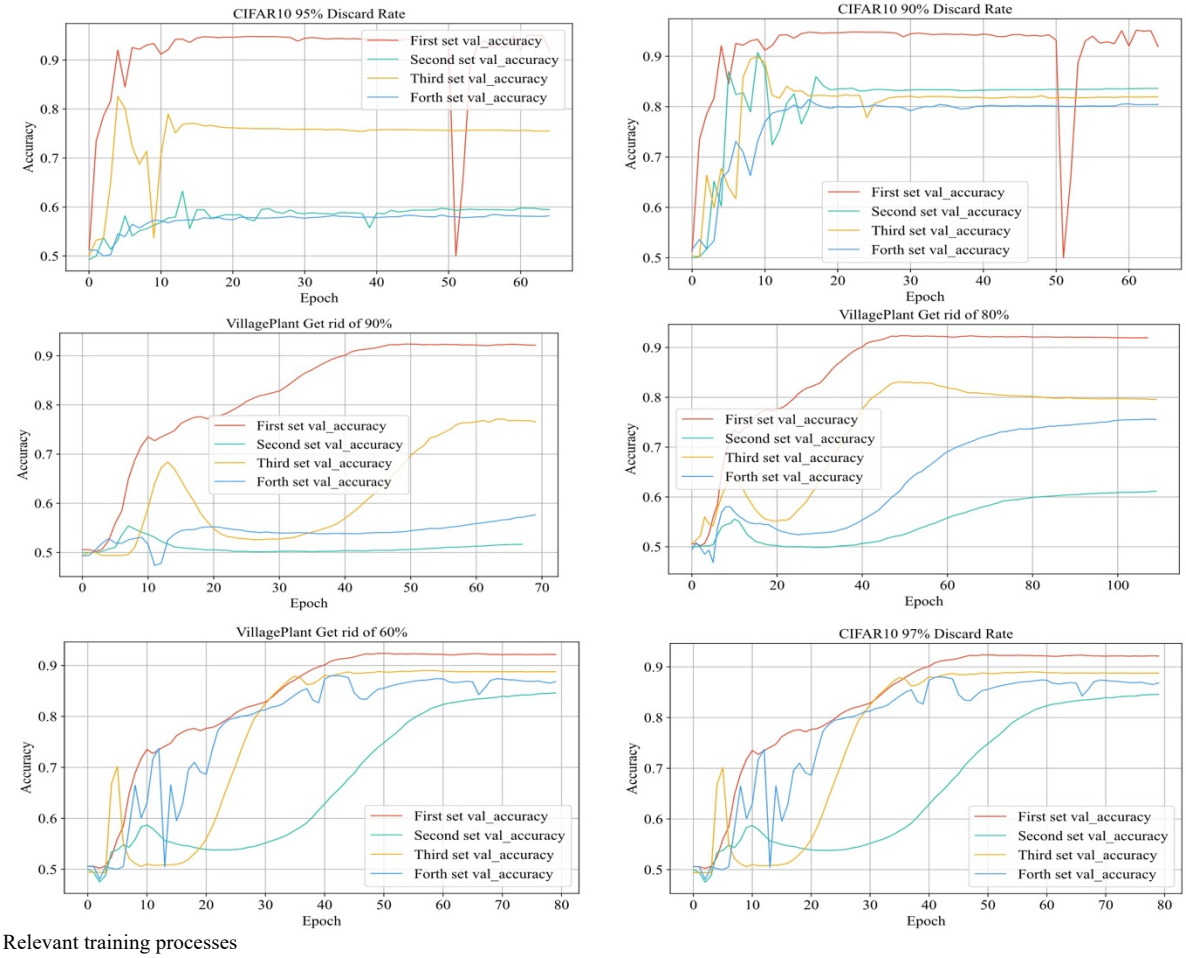

Fig. 8. Relevant training processes

Fig.8 shows the training process for the first experiment. We focus on the yellow training curve, which represents the

accuracy of the generated data. The generated data's diversity is validated by comparing it to the replicated dataset. If the third set of experiments' accuracy is not lower than the fourth set's, the generated data effectively expands the underlying distribution of the training set. Experiment results reveal that when the discard rate increases, the dataset becomes increasingly imbalanced. At this stage, the benefit of creating data becomes more obvious, considerably alleviating the sample imbalance. When comparing the findings of the second and third sets of experiments, this observation becomes clear. The diversity of the generated samples becomes critical in addressing the sample imbalance problem at a discard rate of 90% (for the PlantVillage dataset) and similarly in the CIFAR dataset. The third set of experiments outperforms the fourth, indicating that the generated data broadens the underlying distribution of the training set. The experimental results are shown in Tables 5 and 6.

To ensure that the photos created by the algorithms under consideration accurately represented the target classes, we classed them using a deep learning model trained on the original dataset. We confirmed that the projected classes corresponded to the target classes. We used the ResNet-18 model in this work to ensure that the accuracy of the third set of tests was equal to or greater than that of the second set; otherwise, the data collected did not represent the intended classes. The experimental results showed that as the percentage of deleted data increased, accuracy on the unbalanced dataset declined dramatically. The model barely learned any characteristics from the dataset as a result of this drop. However, when the second set of data was generated, the third set of experimental results indicated a considerable improvement over the second set. As a result, the approach produced a dataset containing the needed features. These features were learned by the model, which enhanced image accuracy.

The diversity of the generated data is verified by comparing it with the replicated dataset. If the accuracy of the third set of experiments is not lower than that of the fourth set, then the generated data effectively expand the underlying distribution of the training set. The results of the experiments show that, as the discard rate increases, the dataset becomes more unbalanced. At this point, the advantage of generating data becomes more pronounced, which significantly alleviates the sample imbalance of the sample. This observation is evident when comparing the results of the second and third sets of experiments. At a discard rate of 90% (for the Plant-Village dataset) and similarly in the CIFAR dataset, the diversity of the generated samples becomes crucial in addressing the sample imbalance problem. The third set of experiments outperforms the fourth set, which confirms that the generated data expand the underlying distribution of the training set.

Table 5. Experimental results for the PlantVillage dataset

Dataset	Data discard rate	Experimental group number	Accuracy	$F^{1.5}$
PlantVillage	0%	1	0.917	97.643%
PlantVillage	60%	2	0.861	84.754%
PlantVillage	60%	3	0.889	89.451%
PlantVillage	60%	4	0.870	85.467%
PlantVillage	80%	2	0.613	75.164%
PlantVillage	80%	3	0.832	79.156%
PlantVillage	80%	4	0.772	75.364%
PlantVillage	90%	2	0.553	68.458%
PlantVillage	90%	3	0.765	73.487%
PlantVillage	90%	4	0.618	70.947%

Table 6. Experimental results for the CIFAR dataset

Dataset	Data discard rate	Experimental group number	Accuracy	$F^{1.5}$
CIFAR	0%	1	0.957	96.762%
CIFAR	90%	2	0.836	88.648%
CIFAR	90%	3	0.821	93.154%
CIFAR	90%	4	0.801	86.876%
CIFAR	95%	2	0.590	69.157%
CIFAR	95%	3	0.759	76.364%
CIFAR	95%	4	0.581	72.875%
CIFAR	97.5%	2	0.555	65.645%
CIFAR	97.5%	3	0.720	70.432%
CIFAR	97.5%	4	0.545	65.871%

4.2 Experimental Reulst (II)

Considering the complexity of the experimental results, we have selected two representative tables to present the conclusions. The superiority of S3VAE over C-VAE is evident in each set of comparative experiments, but the details will not be presented. Table 7 shows the experimental results.

Table 7. Experimental results for different data discard rates in Experiment 2

Model	Dataset (resolution)	Data discard rate	L	K	σ	Accuracy	$F^{1.5}$
S3VAE	CIFAR (32×32)	60%	5	0.1	0.1	0.889	89.451%
C-VAE	CIFAR (32×32)	60%	5	0.1	/	0.876	90.367%
S3VAE	CIFAR (32×32)	80%	5	0.1	0.1	0.832	79.156%
C-VAE	CIFAR (32×32)	80%	5	0.1	/	0.823	83.348%
S3VAE	CIFAR (32×32)	90%	5	0.1	0.1	0.765	73.487%
C-VAE	CIFAR (32×32)	90%	5	0.1	/	0.742	76.346%
S3VAE	PlantVillage (256×256)	90%	5	0.1	0.1	0.821	93.154%

C-VAE	PlantVillage (256×256)	90%	5	0.1	/	0.831	82.647%
S3VAE	PlantVillage (256×256)	95%	5	0.1	0.1	0.759	76.364%
C-VAE	PlantVillage (256×256)	95%	5	0.1	/	0.716	69.648%
S3VAE	PlantVillage (256×256)	97.5%	5	0.1	0.1	0.720	70.432%
C-VAE	PlantVillage (256×256)	97.5%	5	0.1	/	0.693	61.248%

4.3 Experiments on real-world networks

Two generation situations are worth exploring: blurring and ghosting of the C-VAE generation in small batches. To highlight the comparison, we have isolated some samples. However, this problem does not occur with the single-sample sampling VAE, as shown in Fig.9.Similarly, the C-VAE does not perform as effective as the single-sample sampling VAE in large batch generation, as shown in Fig.10.We test S3VAE on the CIFAR10 dataset to ensure accuracy. The results are shown in Fig.11.

The generated images are clear and detailed, with some noise points introduced by the Gaussian noise layer. Unlike traditional adversarial samples, these noise points are part of the model training and do not cause the same issues. As a result, overfitting and data enhancement are effectively prevented.

(a) (b) (c)

Fig. 9. Experimental results of single sample sampling.
Note:(a) Small batch sample image generated by C-VAE, which exhibits ghosting. (b) Small batch sample image, which is also generated by C-VAE. (c) S3VAE generation results.

(a) (b) (c)

Fig. 10. Generate experimental results in large quantities.
Note:(a) Partial sample screenshot of the PlantVillage dataset, (b) S3VAE generation results, and (c) C-VAE generation results.

Fig. 11. Effect of S3VAE generation in the vehicle category of the CIFAR10 dataset

5. Conclusions

As the demand for artificial intelligence technology rises in a variety of industries, getting better results with less data has become a critical challenge. The dataset used for model training influences prediction accuracy. Addressing the difficulty of dataset imbalance has become a key area for academics in this setting. The purpose of this study is to investigate the usage of the single-sampling method to improve the traditional VAE. On two available datasets, we assess and compare models with various sampling approaches. The investigation yielded the following conclusions:

(1)C-VAE outperforms S3VAE in creating a huge volume of data in the PlantVillage and CIFAR public datasets. S3VAE eliminates ghosting and blurriness issues encountered in C-VAE small batch generation.

(2) We notice a considerable decline in the accuracy of imbalanced datasets as the discard rate increases while altering data discard rates to replicate sample imbalances. Surprisingly, S3VAE's accuracy on the verification set is higher than C-VAE's at the same discard rate.

(3) When compared to C-VAE, data generated by S3VAE utilizing the fundamental architecture is more diverse and accurate. It also outperforms the test set in terms of precision in comparative experiments.

The hidden space distribution following picture coding is properly scrambled and sampled in this work. This single-sample sampling method improves data processing and simplifies the time-consuming initialization activities that are common in most generative models. It ensures the algorithm's efficiency in a variety of sectors. We will further improve the S3VAE framework, reduce the number of parameters, and present a data augmentation technique with lower operating costs in future work. We will also undertake comparative verification on additional public datasets in order to provide data support to scholars in other domains.

References

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner,"Gradient-based learning applied to document recognition," *Proc IEEE Inst Electr Electron Eng.*, vol. 86, pp. 2278-2324, Nov. 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, pp. 84-90, May. 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, Las Vegas, NV, USA, Dec. 2016, pp. 770-778.

[4] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and surface variations," 2019. [Online]. Available:https://arXiv.org/abs/1807.01697.

[5] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *Int. Conf. Mach. Learn., ICML*, Long Beach, CA, USA, June. 2019, pp. 3218-3238.

[6] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do CIFAR-10 classifiers generalize to CIFAR-10," 2018. [Online.] Available: https://arXiv.org/abs/1806.00451.

[7] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI - AAAI Conf. Artif. Intell.,* New York, NY, USA, Feb. 2020, pp. 13001-13008.

[8] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, arXiv: 1708.04552.

[9] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: regularization strategy to train strong classifiers with localizable features," 2019. [Online.] Available: https://arXiv.org/abs/ 1905.04899v1.

[10]. D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: a simple data processing method to improve robustness and uncertainty," 2020. [Online.] Available: https://arXiv.org/abs/1912.02781.

[11] C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille, and Q. V. Le, "Adversarial examples improve image recognition," in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, Virtual, Online, USA, June. 2020, pp. 816-825.

[12] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2014. [Online.] Available: https://arXiv.org/abs/1312.6199.

[13] X. Chen *et al.*, "Variational Lossy Autoencoder," 2017. [Online.] Available: https://arXiv.org/abs/1611.02731.

[14] D.P.Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online.] Available: https://arXiv.org/abs/1312.6114.

[15] C. Doersch, "Tutorial on variational autoencoders," 2021. [Online.] Available: https://arXiv.org/abs/1606.05908.

[16] I. Gulrajani *et al*., "PixelVAE: A latent variable model for natural images," 2016. [Online.] Available: https://arXiv.org/abs/ 1611.05013.

[17]J. Tomczak and M. Welling, "VAE with a vampPrior, " in *Int. Conf. Artif. Intell. Stat., AISTATS*, Playa Blanca, Lanzarote, Canary Islands, Spain, Apr. 2018, pp. 1214-1223.

[18] A. Vahdat, W. Macready, Z. Bian, A. Khoshaman, and E. Andriyash, "DVAE++: Discrete variational autoencoders with overlapping transformations," in *Int. Conf. Mach. Learn., ICML*, Stockholm, Sweden, July. 2018, pp. 8008-8023.

[19] J. T. Rolfe, "Discrete variational autoencoders," 2017. [Online.] Available: https://arXiv.org/abs/1609.02200.

[20] A. Vahdat, E. Andriyash, and W. Macready, "DVAE#: Discrete variational autoencoders with relaxed boltzmann priors," 2018. [Online.] Available: https://arXiv.org/pdf/1805.07445.pdf.

[21] C. K. Sø nderby, T. Raiko, L. Maaløe, S. ren K. Sø nderby, and O. Winther, "Ladder variational autoencoders," in *Adv. neural inf. proces. syst.*, Barcelona, Spain., Dec. 2016, pp. 3745-3753.

[22] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," in *Adv. Neural Inf. Proces. Syst.*, Virtual, Online, Dec. 2020, pp. 19667-19679.

[23] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017. [Online.] Available: https://arXiv.org/abs/1701.04862.

[24] A. H. Jha, S. Anand, M. Singh, and V. S. R. Veeravasarapu, "Disentangling factors of variation with cycle-consistent variational auto-encoders," in *Lect. Notes Comput. Sci.*, Munich, Germany, Sept. 2018, pp. 829-845.

[25] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *Proc. - IEEE Winter Conf. Appl. Comput. Vis., WACV*, Santa Rosa, CA, USA, Mar. 2017, pp. 1133-1141.

[26] D. J. Rezende and F. Viola, "Taming VAEs," 2018. [Online.] Available: https://arXiv.org/abs/1810.00597.

[27] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014. [Online.] Available: https://arXiv.org/abs/1411.1784.

[28] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, pp. 139-144, Oct. 2020.

[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein GANs," in *Adv. Neural Inf. Proces. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5768-5778.

[30] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, Long Beach, CA, June. 2019, pp. 4396-4405.

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc IEEE Int Conf Comput Vision*, Venice, Italy, Oct. 2017, pp. 2242-2251.

[32] W. Joo, W. Lee, S. Park, and I.-C. Moon, "Dirichlet variational autoencoder," 2019. [Online.] Available: https://arXiv.org/abs/1901. 02739.

[33] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017. [Online.] Available: https://arXiv.org/abs/1712.04621.

[34] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, Long Beach, CA, USA, June. 2019, pp. 113-123.

[35] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," 2018. [Online.] Available: https:// arXiv.org/abs/1711.04340.

[36] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014. [Online.] Available: https://arXiv.org/abs/1411.1784.

[37] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imageNet," in *Proc. Of The 36th Int. Conf. On Machine Learning, PMLR*, Long Beach, CA, USA, May. 2019, pp. 5389-5400.

[38] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," 2016. [Online.] Available: https://arXiv.org/abs/1511.06349.

[39] Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick, "Improved variational autoencoders for text modeling using dilated convolutions," in *Int. Conf. Mach. Learn., ICML*, Sydney, NSW, Australia , Aug. 2017, pp. 5917-5928.

[40] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," 2017. [Online.] Available: https://arXiv.org/abs/ 1703.10960.