

Studies Advanced in Image Generation Techniques based on Deep Learning

Chang Liu

Department of Software Engineering, Software Engineering Institute of Guangzhou, Guangzhou, Guangdong, 510000, China

vergile@mail.imu.edu.cn

Abstract. A popular area of study in computer vision is image generation. How to improve the authenticity and diversity of generated images, and how to reduce the use of hardware resources and the size of model storage to make it more conducive to model landing are the focus and difficulty of image generation algorithms. Moreover, in practical applications, many other tasks are closely related to the image generation task, utilizing the created picture can lower the cost of data collecting, and the produced picture's quality directly influences how well some activities work. This article first introduces the definition of the generated model in supervised learning and unsupervised learning, and analyzes several reasons why learning should be generated into a model. It also introduces the development of the generated models in China and abroad, and makes a simple classification of the generated models with the previous depth degree.

Keywords: Image Generation Techniques; Generated Models; Deep Learning.

1. Introduction

Nowadays, with the rapid development of network [1,2] information technology, in the context of such a big data era, humans are receiving and transmitting information from all over the world at all times. As the most important information acquisition system for human beings, vision has the characteristics of real-time and high efficiency, which can help people to perceive the surrounding environment more objectively and comprehensively. As a medium for the transmission of visual information, images have the advantages of being vivid, intuitive and easy for people to understand, playing an important role in the transmission and utilization of information. On this basis, the development of image acquisition technology and processing [3] technology is particularly critical in today's digital age. With the development and popularization of computer hardware and Internet of Things technology, image processing technology in computer vision can be applied in many fields, such as face image recognition [4], industrial image detection [5], astronomical image processing, resource exploration, video image style transfer and image super-resolution, etc.

In this research background, during the training of large-scale image processing models, network parameter updates are usually performed in a data-driven manner, it indicates that the algorithms need a lot more image information. However, in some cases, such as the training data of medical images, the acquisition method is often difficult or expensive. At the same time, if the algorithm [6] can make the model complete the image generation task well, it means that the model has a good understanding of the characteristics of the original image. With more comprehensive learning. Therefore, how to learn from the existing real images and generate enough real images to serve as datasets is a very valuable research direction.

The basic idea of image generation technology can be traced back to the most primitive copying and re-creation. For most painters, the learning process is usually to first learn how to describe the real world objectively and correctly, and learn the objective characteristics of various things through sketching of still life's and geometric objects. After a long period of learning to copy, learners can combine the learned drawing methods with their personal thoughts to create the desired artwork. With the continuous development of computer vision technology, the storage medium of image information has also changed from the original paper object to a digital matrix in the computer, which is finally displayed on the display through pixel points. Under the trend of combining digital information

technology with visual images, the preservation and display of images has become easier and easier, but for the learning and regeneration of images, traditional feature engineering methods have not yet achieved effective progress. The regeneration of image features in the field often requires more efficient learning and generalization of the overall style of the image. The images generated by the combination of simple feature arrangement are lacking in resolution clarity and content reliability.

Convolutional neural networks' quick growth has allowed for advancements in the speed of based on deep learning picture production., accuracy and visualization. By learning the minimum and maximum properties of real visual information, our deep learning-based image generation model seeks to recreate pictures that are close to the genuine image style. More than simply the picture's fundamental elements are learned by the deep generative algorithm, such as shape and texture, but also learn the high-level semantic features of the image, such as the overall feature space distribution of the image. Therefore, the generative model can complete the generation of high-resolution images and have high semantic content. reliability. At present, the mainstream image generation models based on deep learning can be roughly divided into two categories: image generation models based on variational autoencoders (Variational Auto Encoder, VAE) [7] and image generation models based on adversarial learning (Generative Adversarial Network, GAN) [8].

Focusing on the above two frameworks, in this paper, we introduce the latest research progress of image generation models based on VAE and GAN. In addition, in Section 3, in order to analyse the application effect of different methods, we also quantitatively compare the generation results of different methods on common datasets. Finally, we summarize the remaining research issues which the area of picture production, as well as the future directions this area of study will take.

2. Image Generation Models based on GAN

The image generation model based on adversarial learning is the generative adversarial network GAN, which was developed from the original SGAN (Standard GAN) proposed by Goodfellow and others in 2014. As an unsupervised [9] generative model, it can generate images that are realistic enough to be trained with knowledge of the zero-sum game strategy. Since the first proposal of SGAN, the theory and application of generative adversarial networks have developed rapidly. In recent years, based on image generation, they have been applied to tasks such as image super-resolution reconstruction [10], image style transfer, and image inpainting, and have achieved remarkable results.

2.1 Principle of GAN

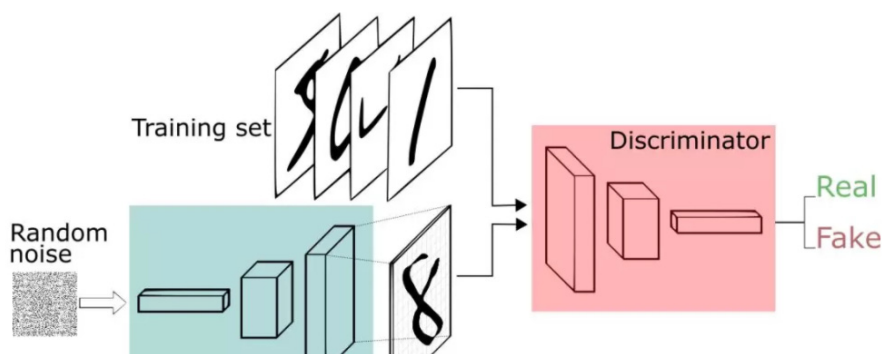


Figure 1. The schematic diagram of classical GAN

The generating network (Generator) and the discriminator network make up the generative adversarial network as a whole (Discriminator). Figure 1 displays the model's schematic diagram. The training goal of the discriminator D is to determine as much as possible whether the sample data input to the network is real or a fake sample generated by the generator G . The output value of the discriminator is between 0 and 1 due to the function of the activation function. When the input data is more real, the output value of the discriminator D is closer to 1, and the input data is falsier, and

the output value of the discriminator D is the value is closer to 0. So, for any discriminator input sample, the discriminator output value $D(x)$ reflects the likelihood that the input specimen was drawn from a genuine sample as opposed to a synthetic one. The training goal of the generator G is to generate fake samples of high enough quality as much as possible so that the discriminator D judges them as real data.

In the training process of the generative adversarial network, it is first necessary to sample from obtaining the noise tensor z from a random noise that complies with the normal distribution, and input the noise into the generator network to discover the created sample $G(z)$, whose dimension is as same as the real. The sample x is the same. Input the real sample x and the generated sample $G(z)$ into the discriminant network D to obtain the corresponding outputs $D(x)$ and $D(G(z))$, and the classifier may be built based on the discriminator's response to the specimen. The loss function L_D , the goal of this loss function is to make $D(x)$ as close to 1 as possible and $D(G(z))$ as close to 0 as possible. After the discriminator loss function L_D completes the gradient back-propagation to update the discriminator network parameters, the model will resample to obtain a new generated sample $G(z)$ and the corresponding discriminator output $D(G(z))$, and build a generator around it Loss function L_G , the goal of this loss function is to be as close to 1 as possible, even if the discriminator D misclassifies the generated samples, after the generator loss function L_G completes the gradient backpropagation to update the generator network parameters, the model will be trained again to discriminate The generator network is trained in this way, and finally high-quality generated samples are obtained.

2.2 Representative GAN-based Image Generation Methods

2.2.1 Text-image Generation Algorithm

Hu Bei-Chen et al. proposed a text generation algorithm. Due to the abstract nature of natural language, the information it expresses is limited, and a single text often lacks detailed information. Therefore, in a bid to create a top-notch picture, we need a generator with constraints to complement the text details. In this paper, a novel text-generating image algorithm is proposed to solve the problem of limited information and improve the quality of the synthesized image by using the prior knowledge formed from the training data set to enrich the given title [11].

As shown in Figure 2, given a title, the text information it contains is first fleshed out, rather than the image being synthesized directly. In the process of title matching, since an image may correspond to multiple titles in the general data set, it is necessary to take the title of each image and its training part as an item in the knowledge base (memory). Then, the compatible ones are retrieved from the repository. Project and refine the title of the project to get the best complement. Finally, the corresponding image has been created using GAN equipped with a personality mechanism.

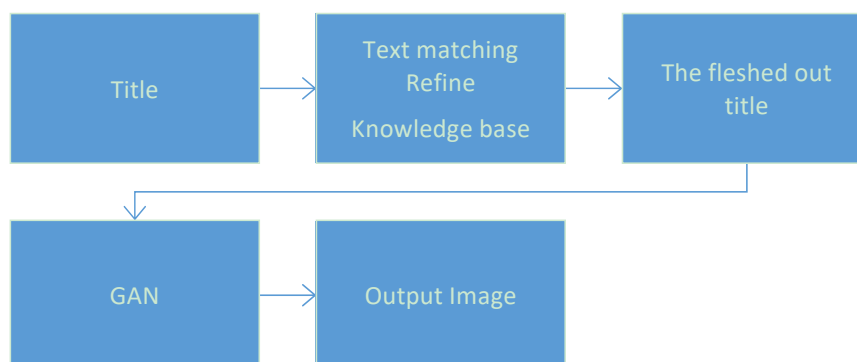


Figure 2. The basic framewok of text-image generation

The researchers present a text image generation algorithm based on GAN. Experiments on the Caltech-UCSDBirds dataset show that the proposed algorithm can synthesize more realistic images than the existing algorithms. However, it is not easy to improve the generation of high-quality images

from multiple headings, and the process needs to be combined with methods such as natural language processing to further improve performance.

2.2.2 GAN in image-to-image Applications

A conditional GAN-based globally benchmarked for image interpretation is called Pix2Pix. It has produced excellent achievements in several image translation sources of data and realized the extension of conceptual framework and stochastic gradient descent. The generator structure of Pix2Pix is based on UNet, which belongs to the encoder-decoder framework, but adds skip connections from encoder to decoder in order to bypass the bottleneck of sharing low-level information such as object edges [12].

One problem with GAN is that it cannot control the data generated by the generated model. To solve this problem, Conditional GAN proposes to add conditional information to both the generated model and the discriminant model to guide the model training, so as to realize the controllability of the generated content. Conditional GAN losses are defined as:

$$\mathcal{L}_{CGAN}(G, D) = \mathbb{E}_{x, y \sim p_{\text{dat}}(x, y)} [\log D(x, y)] + \mathbb{E}_{x \sim p_{\text{data}} \sim p_z(z)} [\log (1 - D(x, G(x, z)))] \quad (1)$$

The total loss is:

$$G^*, D^* = \arg \min_G \max_D \mathcal{L}_{CGAN} + \lambda \mathcal{L}_{L1}(G) \quad (2)$$

Where λ is a hyperparameter to balance the two loss terms,

2.2.3 Least Squares Generative Adversarial Networks

The traditional GAN Discriminator network adopts sigmoid cross entropy loss,[13] which is prone to gradient dispersion during training. Therefore, Mao et al. chooses another loss function: Least Squares loss function, which has the following two advantages: (1) The image quality is higher than that of the traditional gan network; (2) The training process is more stable.

Replace the loss function with least squares loss, as:

$$\begin{aligned} \min_Q V_{LSGAN}(Q) &= \frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}(x)} [(Q(x) - b)^2] + \frac{1}{2} \mathbb{E}_{z \sim p_Y(y)} [(Q(T(z)) - a)^2] \\ \min_T V_{LSGAN}(T) &= \frac{1}{2} \mathbb{E}_{z \sim p_Y(y)} [(Q(T(z)) - c)^2] \end{aligned} \quad (3)$$

For some samples generated by G, if discriminator D decides that they are real, In the case that the traditional loss function sigmoid cross entropy loss is used: then the samples will no longer be optimized. But with least squares loss, these images continue to be optimized, so the resulting images are of higher quality. The researchers aimed to improve the quality of the images generated by standard GAN and the instability of the training process. The improvement method is to change the objective function of GAN from cross entropy loss to least squares loss, and this change solves two defects at the same time.

They take the decision boundary as the intermediary and believe that the separation among the output picture or the decision border can represent the separation between the created image and the actual data. This is because the learned decision boundary must pass through real data points, otherwise the learning process is saturated. In the future work, it is also mentioned that the generated image can be directly pulled towards the real data, rather than the decision boundary.

2.2.4 Boundary Equilibrium GANs (BEGAN)

BEGAN [14] was a paper published by Google in the first half of '17. This paper made further improvements to GAN, and proposed a new way to evaluate the quality of the generators. GAN could use a very simple network without some training, such as BN, minibatch, SELU activation function, and so on. You'll also be able to achieve great training without having to worry about model collapses and training imbalances.

When creating the information spread is equal to the actual data distribution, Past GAN and its variations are designed to produce data distributions that are as near to the genuine distribution of data as feasible, we can determine the generator G trained the same sample can be generated and real

data distribution, which gains the ability to generate enough genuine data, As a result, the scientists developed a number of degradation algorithms to bring the produced information population of G as near to the actual information dispersion as feasible.

Intuitively, the closer two distributions are, the more similar we can assume they are. The separation among the output picture or the decision border can represent the separation between the created image and the actual data. Among them, the better improvement results mainly include DCGAN, WGAN, WGAN-GP and so on. This technique for calculating risk distributions was superseded by BEGAN. Rather than calculating the gap between the produced probability P_g and the actual distribution, P_x , it estimated the difference between the distribution error distributions. The author believes that these distributions can be considered similar as long as the error distributions among the distributions are similar.

2.2.5 Deep Convolution Generative Adversarial Networks.

In 2014, Alec Radford et al. proposed DCGAN [15], which can stabilize training and thus be used for unsupervised learning. The researchers hope that CNN can achieve the same success as supervised learning in unsupervised learning. Through the architectural constraints, the Deep convolutional Generative Adversarial Network (DCGAN) is constructed. It is proved that DCGAN is an advanced unsupervised learning network. The generator and discriminator of DCGAN have learned the multi-level representation from the object details to the whole scene, and it is proved that the image features extracted by DCGAN discriminator have good generalization.

The researchers propose a more stable framework for training the generation of adversarial networks, showing that adversarial networks can learn the representation of images well and use it in supervised learning and generative modeling. But the mode crash problem still exists and can be extended to other areas, such as video (for frame-level prediction) and audio (for pre-training features of speech synthesis), so more research can be done on latent space.

3. Image Generation Models based on VAE

Similar to adversarial generative networks, variational autoencoders [16] are designed to solve the problem of data generation. In an autoencoder structure, an input data is usually required, and the generated data is identical to the input data. However, the data generated are usually expected to be different to a certain extent, it recognizes the compositional properties of created pictures and requires the input of random vectors. Therefore, the adversarial generative network structure with random vectors as the input to generate specific samples is generated in subsequent studies. Variational autoencoders also take random samples with specific distribution as input and can generate corresponding images, which is similar to adversarial generative network in this respect. But variational autoencoders do not require discriminators, and instead use encoders to estimate specific distributions. The overall structure is similar to the autoencoder structure, but the intermediate transfer vectors are random vectors of a specific distribution, where special distinctions need to be made: encoder, decoder, generator, and discriminator.

3.1 Principle of VAE

The variational autoencoder or VAE was proposed by Diederik et al. in 2013. Its theoretical basis revolves around variational and Bayesian[17] theory. The learning and generation of the co-training of the generator encoder as well as the detector decoder, whose fundamental architecture is displayed through Figure 3, completes the distribution of picture data. For a series of real samples x , it is theoretically believed that the real sample data should obey a certain distribution p_θ . If the specific form of this distribution can be obtained, then sampling directly from this distribution, the obtained samples can be regarded as real samples. It is the task goal of the generative model, and the distribution of the real samples is changed in the variational autoencoder.

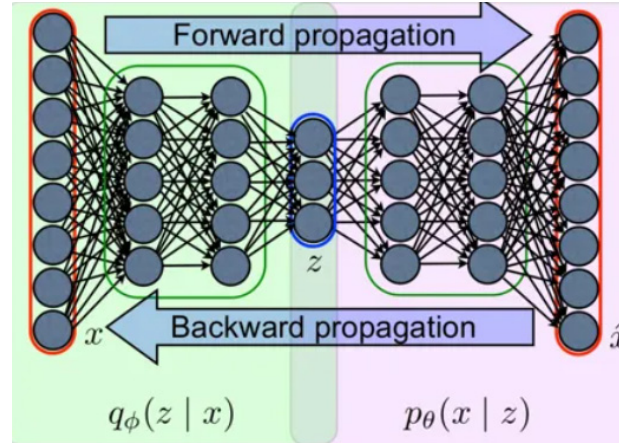


Figure 3. The basic framework of variational autoencoders

In the formula, z is a known distribution set. As long as $p_\theta(x|z)$ can be obtained, the real sample can be obtained directly from the known distribution z by sampling and then using $p_\theta(x|z)$, where z is set as a standard positive A collection of state distributions. Since $P(z)$ is known, and $p_\theta(x|z)$ is unknown, our initial goal is to solve for $p(x)$. We want $p(x)$ to be as large as possible. This is equivalent to solving for the maximum log-likelihood of x :

$$\hat{L} = \sum_x \log P(x) \quad (4)$$

And $\log P(x)$ can be transformed into:

$$\begin{aligned} \log P(x) &= \int_z i(z|x) \log P(x) dz \\ &= \int_z i(z|x) \log \left(\frac{P(z,x)}{P(z|x)} \right) dz \\ &= \int_z i(z|x) \log \left(\frac{P(z,x)}{i(z|x) P(z|x)} \right) dz \\ &= \int_z i(z|x) \log \left(\frac{P(z,x)}{i(z|x)} \right) dz + \int_z i(z|x) \log \left(\frac{i(z|x)}{P(z|x)} \right) dz \\ &= \int_z i(z|x) \log \left(\frac{P(x|z)P(z)}{i(z|x)} \right) dz + \int_z i(z|x) \log \left(\frac{i(z|x)}{P(z|x)} \right) dz \end{aligned} \quad (5)$$

The original formula can also be expressed as:

$$\log P(x) = L_b + \text{KL}(q(z|x) \parallel P(z|x)) \quad (6)$$

3.2 Representative VAE-based Image Generation Methods

Yunchen Pu et al. developed a new VAE setting to evaluate the visuals. The encoder for the underlying parameter distribution of DGDN[18] is based on CNN, while DGDN is utilized as that of the decoder. The pace is way more efficient when tested now then when implemented with in middle since CNN is employed in the recognition model.

To clarify the connection between the proposed CNN-based encoder and the decoder described above, we also preliminarily illustrate the $L = 2$ layer model of the encoder. While the two-layer, image $X(n)$:

$$\begin{aligned} L1: \widetilde{A}^{(a,h_1,1)} &= X^{(n)} *_s F^{(H_1,1)}, h_1 = 1, \dots, H_1 \\ \text{Threadpool: } A^{(a,1)} &\sim \text{Threadpool}(\widetilde{U}^{(n,1)}) \\ L2: \widetilde{A}^{(a,h_2,2)} &= A^{(n,1)} *_s F^{(H_2,2)}, h_2 = 1, \dots, H_2 \\ \text{Generation of code: } e_n &\sim \mathcal{N} \left(\mu_\phi(\widetilde{U}^{(n,2)}), \text{diagonal matrix} \left(\sigma_\phi^2(\widetilde{U}^{(n,2)}) \right) \right) \end{aligned} \quad (7)$$

The Encoder based on CNN encoder's output: The construction enables modeling of pictures with or without related tags and titles., and the new semi-supervised setup is embodied in CNN image learning. Even unsupervised CNN learning is permitted under the construction (solely using images).

There is a serious problem when VAE model is implemented. Hidden variables of sampling from the encoder output $q_\phi(z|x)$ as shown in figure on the left. When $q_\phi(z|x)$ and $p(z)$ are assumed as

normal distribution, the encoder output μ mean value and variance of normal distribution σ , the decoder input sampling since $N(\mu, \sigma)$. Due to the existence of sampling operation, the ladder propagation is not continuous and the gradient descent algorithm cannot be used to train the VAE network end-to-end. To solve this issue, Kingma et al. proposed a continuously differentiable solution called the Reparameterization Trick. The ReLU activation function is used for the layer and returns logits or log-softmax from the forward propagation.

β -VAE (Higgins et al., 2017) [19] is a modification of mutated autoencoders, with special emphasis on detecting potential factors for decomposition. By using the same VAE reward system, they aim to increase the likelihood of producing genuine data while maintaining δ a small enough gap between the true distribution as well as the calculated prior distribution:

$$\begin{aligned} \max_{\phi, \theta} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z} | \mathbf{x})} \log p_{\theta}(\mathbf{x} | \mathbf{z}) \right] \\ \text{subject to } D_{\text{KL}}(q_{\phi}(\mathbf{z} | \mathbf{x}) \parallel p_{\theta}(\mathbf{z})) < \delta \end{aligned} \quad (8)$$

A potential representation is said to be decomposed or factored when every parameter inside the detected potential depiction is also only highly sensitive to one generating component and generally insensitive to the rest. Good interpretability and simplicity in generalization to a variety of activities are two advantages that decomposition representations frequently deliver.

3.3 Applications in Cells Image Generation

Despite the fact that we have reached the period of big data, for a number of reasons, including lack of availability, there are typically few image statistical metrics with appreciable 's vital in the context of healthcare. Currently, VAE and GAN are the 2 main models with the ability to produce pictures. White blood cell categorization is frequently hampered by incomplete data collected and uneven data[20]. The researcher converts the real picture to a potential matrix $\mathbf{z}$ using an encoding system, where $\mathbf{z}$ follows a Distribution function and the characteristics of the real picture coincide with the potential matrix. The picture is encoded, entered, and afterwards reflected over onto intermediate space $\mathbf{w}$ using the translation network rather than providing random matrices directly to the generator. While $\mathbf{z}$ is uninterrupted, real data may be discontinuous (regular gan could be learned to optimize). $\mathbf{Z}$ can be adjusted to provide a more accurate sample. In other terms, the latest design does not have the drawback of GAN being unable to produce discontinuous data.

To properly increase the training data response rate and address the training data mismatch Information has been a controversial subject in the studies. To resolve this concern, Kaikai Liu and colleagues suggested a picture generating approach based on VAE and GAN merging nets, which overcomes the current drawbacks and improves the generation efficiency of power pictures when the two generation methods are used separately. The images obtained with VAE are fuzzy, but those obtained with GAN are closer to the original ones. The network uses the hidden codes VAE to generate images as input to StyleGAN, this is able to provide actual cell pictures. Results from a statistical and multimodal experimental investigation demonstrate the method's viability. This technique can provide superior cell pictures to address the issue from insufficient information. When there is a sufficient amount of excellent patient records, AI can use supervised learning to extract rules from huge data. Nevertheless, due to the limitations of human ability, conventional techniques of therapy and diagnostics are not always the most effective. Ai (VAE paired with StyleGAN) offers useful guidance for aid in diagnosis and lessen the incidence of misinterpretation.

4. Experiments

4.1 Common Data Set

(1) CelebA. Over 200K bikini pics with 40 feature classifications each make up the CelebFaces Characteristics Collection (CelebA), a sizable dataset[21] of facial attributes. This array of photos includes a wide range of postural variations and cluttered surroundings. CelebA has 10,177 identities,

202,599 photos of features, 5 iconic places, and 40 decimal characteristic comments each picture. It is quite diverse and rich in annotations. The main tasks involving computer vision can utilize this dataset as a training and testing set: Facial feature identification, face recognition, and point (or portion of face) positioning. Sample image:

(2) ImageNet. ImageNet is a dataset of images organized according to the WordNet hierarchy. For ImageNet, its aim is to provide an average of 1000 images per synset. Each concept image is quality-controlled and human-annotated. When done, ImageNet is expected to provide tens of millions of cleanly sorted images for most concepts in the WordNet hierarchy. ImageNet is an ongoing research effort aimed at providing easily accessible image databases to researchers around the world. There are now 22000 synsets in ImageNet, that contain a total of 14197122 photos (synsets). Mollusc, mammal, playing instruments, processing facility, reptilian, athletics, framework, instrument, shrub, cutlery, vegetable, car, and individual are some of the major categories. Other main categories include: salamander, animal, device, bird, covering, device, fabric, fish, floral, meals, fruit, fungus, furniture, and geoscience creation.

(3) COCO. For picture identification, segmentation, and sensitively and appropriately, one can utilize the large-scale COCO dataset. It features more than 330K photographs, of which 220K are tagged, 1.5 million items, 80 object types (such as people, automobiles, elephant, etc.), and 91 substance types (such as grassland, brick, air, etc.). Additionally, each image has a 5 explanation, and there are 250,000 pedestrians annotated with keypoints. The first version of the COCO dataset was released in 2014, and it contains 164,000 images divided into a training set (83,000), a validation set (41,000), and a test set (41,000). An additional 81,000-image test set was released in 2015, containing 40,000 fresh photographs in addition to all prior trial images. The allocation of the train and validation sets was altered in 2017 from 80 to 120k, 40k to 05k; the revised divide makes usage of the identical pictures & annotations. The 2017 test set is a portion of the 41,000 picture 2015 test set. Moreover, a new unidentified collection of 123,000 photos is included in the 2017 edition.

(4) AFLW. AFLW (Annotated Facial Landmarks in the Wild) database is a large-scale face database with several poses and perspectives, having 21 facial landmarks assigned to each face. The database is incredibly educational and contains pictures with different stances, emotions, illumination, ethnicities, as well as other things. Approximately 250 million arm facial photos, 59% of whom are female & 41% of them being male, may be found in the AFLW face database.

4.2 Evaluation Indx

The indicators used to evaluate image generation algorithms mainly include:

(1) Peak signal-to-noise ratio (PSNR). PSNR is usually used to measure whether a processing program is satisfactory. Typically, PSNR is used to gauge how well a processing software is working. In relation to the value of $(2^n - 1)^2$ (its greatest signal value squared, where n is the value of the bytes each test), it represents the average square error here between original picture and the image is obtained.

(2) Structural Similarity (SSIM). An indicator called structural similarity is used to assess how structurally related two photographs are to one another. Compared with peak signal-to-noise ratio (PSNR), structural similarity is more consistent with human eyes' judgment of image quality. The characteristic of object - oriented paradigm in a scene is reflected by the structural similarity measure, which defines spatial features as irrespective of light and contrast. Light $l(a,b)$, contrast $c(a,b)$, and shape s are three separate elements that together make up the distortion (a,b) , as:

$$\begin{aligned} \text{SSIM}(a, b) &= [l(a, b)]^\alpha [c(a, b)]^\beta [s(a, b)]^\gamma \\ l(a, b) &= \frac{2\mu_a\mu_b + C_1}{\mu_a^2 + \mu_b^2 + C_1} \\ c(a, b) &= \frac{2\sigma_a\sigma_b + C_2}{\sigma_a^2 + \sigma_b^2 + C_2} \\ s(a, b) &= \frac{\sigma_{ab} + C_3}{\sigma_a\sigma_b + C_3} \end{aligned} \quad (9)$$

(3) Learned Perceptual Image Patch Similarity (LPIPS). To calculate the difference between two photos, one uses LPIPS. The metric causes the producer to acquire that reverse mapping from the fake picture to reconstruct the genuine picture to selectively analyze the proximity of their perceptions by learning the reversal map from the generated image to the underlying data. Learned Perceptual Image Patch Similarity are closer to sensory consciousness than conventional techniques like Peak signal-to-noise proportion & Structural Similarity. As the figure 4 shown, the two inputs were fed into the neural network F (VGG, Alexnet, Squeezenet) for feature extraction. The output of each layer was normalized after activation, and then L2 distance was calculated after weighted multiplies of layers w .

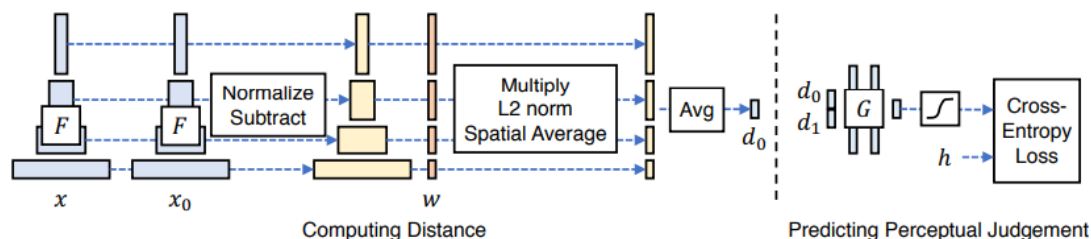


Figure 4. The pipeline of LPIPS calculation

4.3 Performance Comparison

4.3.1 Quantitative Comparison

Figure 5 shows the performance change of different representative algorithms, where examples of labeling caused by different random selections at around 1%, a small percentage of label images (less than 10% of labels) and a larger percentage of label images over 30% when the variance drops to less than 0.2% are taken into account. As can be seen from the figure, 60% of the labeled data of the semi-supervised learning method achieved similar results (61 top-1 reliability) outcomes and the outcome of the complete data set (61.8% top-1 accuracy), proving the effectiveness of the method proposed by the researchers for the semi-supervised method. AlexNet and GoogLeNet had similar accuracy of 80 and 88 when the proportion of tag images was 100. The reason for AlexNet's high accuracy may be that the use of specialized GPU hardware allows us to learn more complex relationships in a reasonable amount of time (deeper and larger network = more variables for prediction), thus improving the accuracy of the overall network. But obviously the GoogLeNet method performs better because the AlexNet structure is based on a huge network's depth (number of hidden layers) to get higher training outcomes. Numerous detrimental consequences, like overfitting, slope vanishing, slope inflation, etc, will result from adding more layers. From another perspective (inception), training results can be improved: additional characteristics may be retrieved with the same amount of work thanks to much optimal use of computational resources. This enhances learning outcomes.

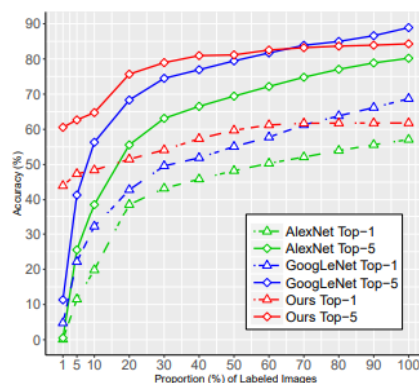


Figure 5. Comparison of different representative methods

4.3.2 Visual Comparison

We also provide a visual comparison between different representative image generation methods, such as β -VAE, VAE and InfoGAN, whose results as depicted in Figure 6. The findings show the beta-vae presents better results in comparison. There are three reasons for this: (1) The training is steady, requires minimal reliance on the data, and is based on tweaking a solitary subset of features, either directly using weakly labeled data for hyper - parameters search or indirectly via intuitive image examination of pure huge datasets. (2) Completely unsupervised automatic discovery of decipherable decomposition potential interpretations from either the original picture data. (3) $\beta < 1$ relaxed the constraint on potential space, while $\beta > 1$ intensified the constraint. The former will get better reconstruction results, while the latter will get better unconditional sampling.

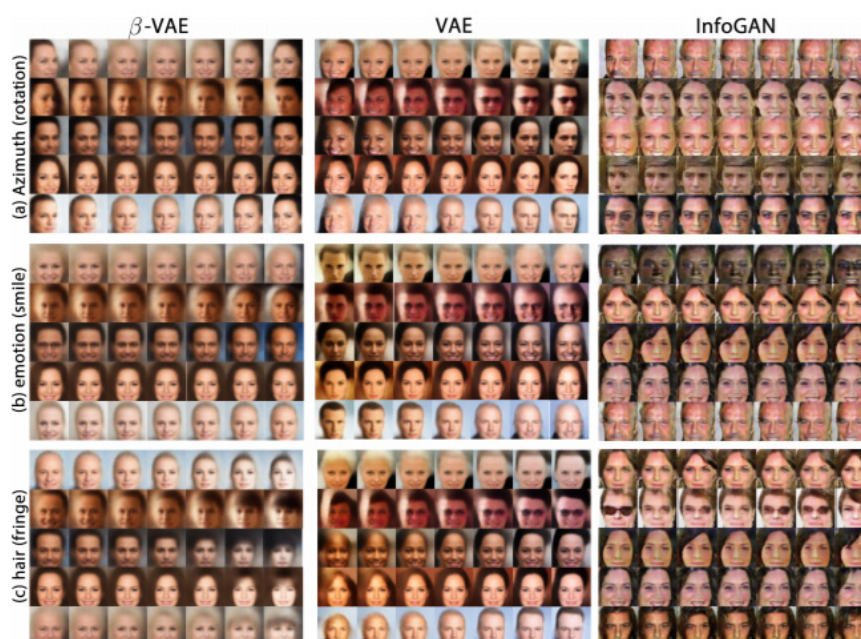


Figure 6. Visualization results of different representative methods

5. Discussion

In the aspect of image generation, it mainly uses the information of the image itself, and does not combine with text, speech and other forms of information. However, with the proposal of generative adversarial network, image generation technology is no longer limited to the original convolutional neural network model, but adopts the adversarial way to train the model, It significantly enhances the impact of picture production and breathes fresh life into the field. Right now, in terms of achieving the further advancement of picture creation technology, it is necessary to adjust the size of the model according to the demand, combine with the model compression method based on "knowledge distillation", optimize the compression of matching requirements, embed it into small software, and obtain more and wider application fields. Further research is needed.

The research of image generation technology still has the potential of further breakthrough and improvement, including the generation of high-resolution image, the reliability of the text and picture descriptions that are created, and the versatility with the multi-scene application, the generation of image quality evaluation standards and the accuracy of evaluation technology breakthrough, etc. In terms of model architecture, it is suggested to analyze the importance of text embedding and the quality of results, and explore the understanding and application of text-image generation model in different scenarios. In addition, in order to realize the scene application of text in image generation, it is very important to control the process of image generation with fine granularity. Therefore, on the basis of using models to achieve image generation, future research should also focus on the realization of result iteration and intelligent interaction.

6. Conclusion

In this paper, we summarized some of the mainstream GAN and VAE image generation techniques and their applications as well as other derived or improved algorithms of method prototypes (BEGAN, DEGAN; Beta-VAE; The results of each algorithm (including GAN and VAE) are evaluated, displayed and analyzed visually through the general data set. Finally, the existing problems in image generation are summarized and the future development of this subject is discussed. This review is expected to help researchers understand the challenges facing current image generation technologies and provide references for future research developments in this field.

References

- [1] Ummerhofer B, Zhou H, Uhrig J, et al. Demon: Depth and motion network for learning monocular stereo[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5038-5047.
- [2] Qi X, Liao R, Liu Z, et al. Geonet: Geometric neural network for joint depth and surface normal estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 283-291.
- [3] Hinton G E, Sejnowski T J. Learning and relearning in Boltzmann machines[J]. Parallel distributed processing: Explorations in the microstructure of cognition, 1986, 1(282-317): 2.
- [4] Lu X. Image analysis for face recognition[J]. Personal notes, 2003: 36.
- [5] Wang Y, Tan D J, Navab N, et al. Self-supervised Latent Space Optimization with Nebula Variational Coding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
- [6] Vahdat A, Kautz J. NVAE: A deep hierarchical variational autoencoder[J]. Advances in Neural Information Processing Systems, 2020, 33: 19667-19679.
- [7] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview[J]. IEEE signal processing magazine, 2018, 35(1): 53-65.
- [8] Gowdy J N, Subramanya A, Bartels C, et al. DBN based multi-stream models for audio-visual speech recognition[C]//2004 IEEE International conference on acoustics, speech, and signal processing. IEEE, 2004, 1: 1-993. Zou M, Conzen S D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data[J]. Bioinformatics, 2005, 21(1): 71-79.
- [9] S. Benaim and L. Wolf, "One-sided unsupervised domain mapping," arXiv preprint arXiv:1706.00826, 2017.
- [10] An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability[J]. Special Lecture on IE, 2015, 2(1): 1-18.7
- [11] Qiao T, Zhang J, Xu D, et al. Mirrorgan: Learning text-to-image generation by redescription [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1505-1514.
- [12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," arXiv preprint arXiv:1703.10593, 2017.
- [13] Mao X, Li Q, Xie H, et al. Least squares generative adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2794-2802.
- [14] Berthelot D, Schumm T, Metz L. Began: Boundary equilibrium generative adversarial networks[J]. arXiv preprint arXiv:1703.10717, 2017.
- [15] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.
- [16] Dai B, Wipf D. Diagnosing and enhancing VAE models[J]. arXiv preprint arXiv:1903.05789, 2019.
- [17] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.
- [18] Pu Y, Gan Z, Henao R, et al. Variational autoencoder for deep learning of images, labels and captions[J]. Advances in neural information processing systems, 2016, 29.

- [19] Higgins I, Matthey L, Pal A, et al. beta-vae: Learning basic visual concepts with a constrained variational framework[J]. 2016.
- [20] Liu K, Shuai R, Ma L. Cells image generation method based on VAE-SGAN[J]. Procedia Computer Science, 2021, 183: 589-595. Huang H, Yu P S, Wang C. An introduction to image synthesis with generative adversarial nets[J]. arXiv preprint arXiv:1803.04469, 2018.
- [21] Yu F, Seff A, Zhang Y, et al. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop[J]. arXiv preprint arXiv:1506.03365, 2015.
- [22] Pu Y, Gan Z, Henao R, et al. Variational autoencoder for deep learning of images, labels and captions[J]. Advances in neural information processing systems, 2016, 29.
- [23] Higgins I, Matthey L, Pal A, et al. beta-vae: Learning basic visual concepts with a constrained variational framework[J]. 2016.