Research and Analysis of VAE in Image and Data Analysis

Xuan Sheng

University of Minnesota, Minneapolis MN 55414, the United States
jodie7121212@gmail.com

Abstract. Deep generative models find popular applications in image and data analysis to learn more intricate patterns as well as to generate novel samples. Variational autoencoders are appreciated for ensuring a clear format and stable training and are used in recovery, learning features, and data augmentation. One of the frequent drawbacks is that pixel-based loss will result in output with missing details and blurred images. To better this, researchers have experimented with improvements to the latent space, conditional generation mechanisms, and feature-based perceptual losses, which assist in maintaining coherence of structure as well as maintaining Variational Autoencoder (VAE) stability. This paper examines the history of VAEs, their main technical advance, and their use in reconstruction, representation learning, and data generation. Meanwhile, this paper also points out the main challenges currently faced, such as insufficient retention of details, high computational costs, and limited cross-domain adaptability, and looks forward to possible future research directions.

Keywords: Variational Autoencoders, Perceptual Loss, Deep Generative Models, Image Reconstruction, Representation Learning.

1 Introduction

Deep generative models, with their powerful data distribution modeling capabilities, have become one of the core directions in current machine learning research. In image and data analysis tasks, such models can learn complex structural features from high-dimensional data and generate new data that is highly similar to real samples. Variational Autoencoder (VAE), generative adversarial network (GAN), and Diffusion Models are among the most rapidly developing and widely applied deep generative frameworks in recent years. Among them, VAE is widely applied in tasks such as image generation, reconstruction, compression, and feature learning due to its stable training process, clear probabilistic interpretation, and structured latent space. Since Kingma and Welling proposed the VAE structure [1], this model has become the basis of many generative methods. However, VAE has a prominent drawback in actual image generation and reconstruction. The generated results are often blurry, lacking texture details and edge structure. This issue limits its application value in high-quality visual tasks. Therefore, improving the reconstruction quality of VAE has become a direction worthy of in-depth research.

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_96

The loss design of VAE reconstructions is the major cause of its blurriness. Single pixel-based VAEs are based on a pixel-based loss, which might be mean square error (MSE), to quantify the quality of reconstruction. This loss is simple to optimize, but it considers an image to be made up of independent pixels and disregards other higher-level features such as edges, textures and semantic patterns. In uncertain regions, pixel loss tends to average variations, resulting in excessive smoothing and detail loss. Consequently, the produced images appear blurred and do not align well with human visual perception. This disadvantage has encouraged scientists to discover loss functions that are more acquainted with visual properties.

Another solution that is relevant to this problem is perceptual loss. Johnson et al. demonstrated that pixel-level loss does not preserve the structure information and is also characterised by the lack of performance in the area of style transfer and super-resolution [2]. They use feature maps of a trained VGG network, which is why their approach allows preserving textures and fine details more efficiently. Zhang et al. additionally put forward the LPIPS measure, which reveals that human ratings of similarity between images are highly presumed to be linked to the distance of deep features [3]. Their results offer the theoretical ground on the use of perceptual loss in the reconstruction exercises. Hou et al. introduced the concept of feature-based perceptual loss to autoencoders and VAEs and observed that it enhanced clarity and structural consistency in reconstructed results together with a considerable extent [4]. On the same note, Pihlgren et al. found that perceptual loss does not just increase the quality of reconstruction, but also prevents the strengthening of latent representations, making the learned features more useful for downstream tasks [5]. These two studies combined indicate that the direction of perceptual loss is a practical and effective one to minimise VAE blurriness.

As such developments occur, scholars have paid more attention to the incorporation of deep feature supervision into the probabilistic VAE system. By introducing perceptual constraints to the reconstruction goal, the model is able to support structured latent spaces whilst being able to maintain the high-level semantics and finer-grained details. This changes the objective of focusing only on pixel accuracy to perceptual realism and enables VAEs to create images which are more aligned with human visual expectations.

This review examines the development of deep generative models in image and data analysis, with emphasis on VAE-based approaches and improvements derived from perceptual loss. It summarizes the core principles of major generative techniques, the evolution of VAE variants, and methods designed to address reconstruction limitations. The review also discusses remaining challenges in achieving high-quality generation and highlights potential directions for future research.

2 Overview of Main Techniques

2.1 Generative Model Foundations

Deep generative models play a crucial role in image and data analysis by learning potential data distributions and generating new samples with similar statistical

characteristics. Early models, such as autoencoders (AE), relied on encoder-decoder structures to learn the latent representations of compression, but they lacked explicit probabilistic modeling capabilities and thus were unable to generate realistic samples. The variational autoencoder (VAE) overcomes this limitation by introducing variational inference to learn continuous and structured latent Spaces, marking an important milestone in the field of generative modeling [1].

Unlike variational autoencoders (VAEs), Generative adversarial networks (GANs) employ an adversarial learning mechanism, where generators and discriminators compete with each other to implicitly model the data distribution. GAN can achieve highly realistic output, but it has some well-known problems, such as unstable training and mode crashes. In recent years, diffusion models have emerged as a powerful generation framework. They generate samples through an iterative denoising process and gradually refine the noise input. Although diffusion models can achieve the most advanced quality, their computational costs are still quite high. Autoencoders (AE), VAE, GAN, and diffusion models together form the current infrastructure of the field of deep generative learning.

2.2 VAE Architecture and ELBO Formulation

VAE adopts an encoder-decoder structure, but its main feature is that it has a potential space in the form of probability. The function of the encoder is to convert the input data into parameters of the latent distribution, namely the mean vector and the variance vector. To be able to perform backpropagation, the model uses reparameterization techniques to sample from this distribution. In this way, the entire model can be trained end-to-end. Afterwards, the decoder uses the sampled latent variables to reconstruct the data, with the aim of learning the complex patterns in the data.

The training objective of VAE is to maximize the lower bound of evidence (ELBO). This goal requires balancing two aspects, one is to rebuild the quality of data, and the other is to standardize the potential space. ELBO consists of two items. The first one is the reconstruction item, which requires the decoder to output data as similar as possible to the input; The second term is the KL divergence term, which, as a regularization term, makes the distribution generated by the encoder approach the preset prior distribution (usually the standard normal distribution). The reconstruction term is to preserve the input information, while the KL divergence term is to keep the potential space in order.

However, a common problem with VAE is that the reconstruction results are rather ambiguous. This is mainly because the reconstruction term usually employs a pixel-by-pixel loss function like the mean square error (MSE). This loss function compares images pixel by pixel, resulting in the model's inability to effectively learn the overall texture and structural information. Therefore, many subsequent studies have been dedicated to improving VAE, such as enhancing the quality of reconstructed images by introducing a perceptual loss function.

2.3 Extensions and Variants of VAEs

Since the debut of the VAE, numerous variants have been suggested in order to overcome its weaknesses and widen its modeling capacity. The standard representative extension is β -VAE, that puts a weighting term in front of the KL divergence term. An

increase in β will promote more disentangling of the latent dimensions, leading to more interpretable features at the expense of reduced reconstruction validity. It represents the inherent trade-off between visual fidelity and semantic clarity in the latent space.

Another way of development is the conditional variational autoencoder (CVAE). Integrating labels or attribute information at both encoding and decoding steps, CVAE allows generation to be controlled and is effective on attribute editing and conditional generation problems. Conditional constraints assist the model in learning more conditional and structured information distributions.

Other work is aimed to increase the expressive power of VAE as such. Such models as NVAE demonstrate that hierarchical latent representations and more powerful decoders can be of significant benefit to the quality of images, which is why the design of latent variables and decoding processes has a direct impact on the quality of generation.

Another line of research integrates perceptual supervision into the VAE framework. Hou et al. [4] introduced feature-level constraints derived from deep networks to help the model preserve semantic and structural information, which effectively reduces the excessive smoothness typical of standard VAE reconstructions. Similarly, Yang and Zhang proposed a perceptual-loss-based improvement that strengthens semantic consistency during decoding, producing clearer results while maintaining training stability [6].

Other studies extend this concept and place perceptual information directly into the architecture in addition to modifying losses. Indicatively, Zhang et al. came up with the Perceptual Generative Autoencoder (PGA) [7], which places perceptual constraints on the encoder and decoder. Their findings demonstrate that visuality is not the only benefit as visual guidance enhances more informative latent representations.

Altogether, VAE extensions primarily enhance four features, namely, interpretability of latent space, controllable generation, ability to model complex data, and reconstruction quality of feature supervision. All of these developments increase the scope of VAE in image generation and representation learning.

2.4 Perceptual Loss Integrated into VAEs

A common problem with standard VAEs is that their reconstruction results are often overly smooth and lack clear textures. This is mainly due to the use of a pixel-based loss function, which compares two images point by point. Although such loss functions are easy to optimize, they have difficulty capturing elements such as edges, shapes or overall structures, which are crucial for generating visually convincing results. To improve this, researchers began to explore perceptual loss, a method for measuring image similarity through the feature space of deep neural networks rather than the original pixels. Johnson et al. revealed that pre-trained VGG network feature maps have more informative structural and textural features as compared to pixel-level comparisons [2], which suggests that deep feature–driven reconstruction produces clearer and more coherent outputs. Building on this idea, Zhang et al. introduced the LPIPS metric, and it was proven that the distances calculated using deep feature space are more aligned with human perceptual judgments [3]. These results gave a solid basis for the use of perceptual loss in reconstruction tasks. Hou et al. subsequently directly

incorporated perceptual supervision into the VAE objective which enhanced structural retention and minimizing blurring in standard VAE outputs [4].

Though the use of VGG based perceptual loss is a popular approach, it was later examined to find alternatives to decrease dependency on ImageNet pre-trained networks. Czolbe et al. have suggested a perceptual loss based on Watson model, whereby perceptual similarity is obtained using human vision based principles not necessarily involving deep networks [8]. Equally, Liu et al. proposed a perceptual loss with random initialised CNNs [9], which evidences that untrained networks can offer sufficiently rich feature spaces with lower computational intensity and with lower domain mismatch than VGG.

Altogether, perceptual loss is a good answer to pixel-level goals. VAEs can be used to reproduce images with a higher degree of clarity and structural detail and exhibit constant training behaviour by using feature representations that are more aligned more closely with visual perception.

2.5 Comparative Analysis of Strengths and Weaknesses

In practice, different generative models show distinct strengths and limitations. VAEs benefit from stable training and well-structured latent spaces, which makes them suitable for feature disentanglement and representation learning. However, their reliance on pixel-level reconstruction loss causes outputs to lose detail and appear overly smooth. As a result, VAEs often struggle in tasks that demand high visual fidelity. Therefore, it often falls short in scenarios where high visual fidelity is pursued.

Unlike VAE, GAN promote outstanding performance of generated samples in terms of detail richness and visual fidelity through adversarial learning between the generator and the discriminator. However, GAN also has obvious shortcomings. The training process is unstable, the pattern coverage is incomplete, and it is prone to pattern collapse problems, which affects the diversity of the generated samples.

The diffusion model adopts a progressive denoising generation strategy, which can achieve a very high level of image quality, especially performing well in high-resolution image generation tasks. However, such models often have a large number of sampling steps and a slow generation speed, which limits their application in real-time scenarios.

From the perspective of loss functions, pixel-based losses (such as MSE) are simple and stable, but they cannot effectively depict the structure and semantic information of images. Perceptual loss takes advantage of pretrained network deep features to form similarity, giving more accurate results that can better align with human visual perception. This enhances consistency in structures and the details at the expense of higher computational cost and reliance on huge pretrained models.

Pihlgren et al. provide a systematic comparison of perceptual loss between various architectures [10], which demonstrates that its performance is dependent on the selection of the feature extractor, depth of the network, and how well the pretraining data matches the target domain. Their results show that the effectiveness of perceptual-loss improvements is not reliable across all feature spaces, and consideration of feature space must be considered in terms of opposing reconstruction quality, computation and stability.

In general, there are a number of trade-offs between the visual quality, stability, efficiency, and capacity to represent information between different models of generative and loss functions. In practical applications, a comprehensive selection should be made based on task requirements and resource constraints.

3 Applications of Deep Generative Models

3.1 Image Reconstruction and Super-Resolution

Deep generative models can restore image structures based on limited data, and thus are often used in reconstruction and super-resolution tasks. Among them, variational autoencoders are widely adopted because they are stable in training and can learn compact representations. However, its reconstruction relies on pixel-level loss, and the generated images are prone to smoothing, with details and textures often appearing blurry.

Generative adversarial networks employ various mechanisms to generate visually realistic images through adversarial training, with outstanding detail presentation. However, its training process is unstable and prone to mode collapse, which limits its practical application. The diffusion model adopts a progressive denoising strategy, resulting in high generation quality and rich details, but it has a relatively high computational cost and a slow generation speed.

Perceived loss has become an important auxiliary means in recent years. It measures image similarity in the deep feature space and can better preserve edge and structural information. When combined with VAE or GAN, perceptual loss helps reduce smoothing and improves visual consistency. Each model has different strengths. VAE offers stable training, GAN produces detailed images, and diffusion models achieve high quality but require heavy computation. Selecting a method therefore involves balancing image quality, stability, and efficiency.

3.2 Representation Learning and Feature Embedding

Deep generative models have an important use in representation learning. Variational autoencoders are extensively utilized due to the fact that the encoder can generate a continuous latent space which not only reduces the amount of data but also retains important semantic information. Learned representations are capable of capturing patterns in the form of shapes, texture, and category, which can be used in clustering, retrieval, and visualization. The flow of the latent space also allows a smooth transition between samples.

To further enhance the quality of representation, a number of different VAE variants have been suggested. β -VAE enhances regularization to promote disentangled semantic factors and Conditional VAE includes label information to permit controllable generation.

Compared with generative adversarial networks, VAE has more advantages in representation learning. Although GAN can generate more realistic images, its potential space usually lacks a clear structure. Although the diffusion model has a powerful modeling ability, its representation form is not as intuitive and easy to use as VAE.

Overall, VAE has achieved a good balance among the stability, interpretability and practicality of representation learning. This makes it the fundamental framework for many feature learning tasks and also provides an important reference for subsequent representation learning methods.

3.3 Data Generation and Augmentation

Deep generative models can be applied in data generation and other tasks like augmentation where data is small and unbalanced. Variational autoencoders (VAEs) are trained to find smooth latent space that enables learning how to sample diverse but structurally consistent data, which can be used to extend small sets of data.

Conditional VAEs (CVAEs) also add information about the label to allow the controlled production of samples, which can help overcome class imbalance through generating extra useful data about underrepresented classes. Although generative adversarial networks (GANs) are renowned for generating clearer and more realistic images, the instability of their training makes their performance unpredictable when a large amount of diverse generative data is required. In contrast, although the images generated by VAE may be slightly smoother, the results are usually more stable and controllable.

Diffusion models have also been used for data augmentation in recent years. They can generate high-quality synthetic samples, but the computational cost required to generate each sample is significantly higher than that of VAE. This feature may limit its application in large-scale data augmentation tasks.

In practical applications, the synthetic data generated by these models have been proven to effectively enhance the performance of downstream tasks. VAE has achieved a good balance among sample diversity, generation controllability and training reliability, making it a practical and stable data generation and augmentation tool in a wider range of machine learning processes.

4 Challenges and Future Directions

4.1 Main Challenges Currently faced

Although the application of generative models is becoming increasingly widespread, their actual performance is still constrained by several key issues. As for VAE, reconstructed images often have problems such as blurred details and a soft visual effect. Although methods such as perceptual loss can improve this problem to a certain extent, VAE still has a significant gap in visual clarity compared with GAN or diffusion models. Meanwhile, perceptual loss usually relies on a network pre-trained on a specific dataset (such as ImageNet) to extract features. If the distribution of the target data differs significantly from that of the pre-trained data, the effectiveness of feature extraction may decline, thereby affecting the quality of reconstruction.

The training cost brought about by model complexity cannot be ignored either. As the structure becomes increasingly complex, the requirements for computing resources and training time of the model have significantly increased, which poses an obstacle to actual deployment. In addition, the generation of high-resolution images remains a challenge because higher resolutions require potential representations to have stronger

information carrying capacity, and the decoding process also needs to handle more refined structural and texture information.

In practical applications, users have put forward higher requirements for the reliability, interpretability and output consistency of generative models, while existing methods still have obvious room for improvement in these aspects.

4.2 Possible future research directions

To address the above challenges, future research can be carried out from multiple directions. One is to design perceptual losses that are more adapted to the characteristics of the target data, reduce reliance on fixed pre-trained networks, and thereby enhance the quality of cross-domain generation. Second, it combines VAE with emerging architectures such as visual Transformer and diffusion model components to enhance the ability to generate details while maintaining training stability.

High-resolution generation will continue to be a research focus. By designing more flexible latent structures or multi-scale decoding strategies, it is expected to better capture fine-grained visual patterns. In addition, developing a more lightweight and efficient architecture will help lower the threshold for training and deployment, and promote the application of generative models in a wider range of scenarios.

As generative models are increasingly applied in practical situations, enhancing their robustness and interpretability, and ensuring that their generative behaviors comply with ethical norms, will also become important directions worthy of attention in the future.

5 Conclusion

Variational autoencoders and other deep generative models have gradually become practical tools in image and data analysis tasks. These models can learn potential patterns from high-dimensional data and achieve functions such as new sample generation or feature extraction. Among them, the variational autoencoder (VAE) has received extensive attention due to its relatively stable training process and the ability to construct a clearly structured latent space. These features enable VAE not only to be used in image reconstruction and generation, but also to play a role in tasks such as representation learning and data augmentation.

However, the limitations of VAE are also quite obvious. Compared with generative adversarial networks or diffusion models, the images they generate often appear to have blurred details and a softer visual effect. Although many improvement schemes have been proposed, such as introducing perceptual loss, optimizing potential structures or designing more flexible decoders, there is still a significant gap in visual quality. In addition, the reliance of perceptual loss on pre-trained networks also brings practical problems, especially when there are differences between the target data and the conventional training dataset, the effect of feature extraction may decline. At both the computational and model-design levels, producing high-resolution images remains a challenge.

Nevertheless, VAEs retain distinct advantages. They offer stable training and an interpretable latent space, qualities that many other generative models do not easily

replicate. Future developments may involve integrating modern architectures such as vision transformers, creating more adaptive loss functions, or designing lightweight models that can handle high-resolution data more efficiently.

Overall, VAEs continue to play an important role in generative modeling. As research progresses, they are expected to become more efficient, flexible, and practical while maintaining their core strengths.

References

1. Kingma, D. P., & Welling, M.: Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114. (2014)
2. Johnson, J., Alahi, A., & Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In Computer Vision – ECCV 2016 (pp. 694–711). Springer. (2016)
3. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 586–595). (2018)
4. Hou, X., Sun, K., Shen, L., & Qiu, G.: Deep feature consistent variational autoencoder. arXiv preprint arXiv:1610.00291. (2016)
5. Pihlgren, G. G., Sandin, F., & Liwicki, M.: Improving image autoencoder embeddings with perceptual loss. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1–7). IEEE. (2020)
6. Yang, J., & Zhang, J.: Research on improved algorithm based on perceptual loss for variational autoencoder. In Proceedings of the 2023 International Conference on Big Data, Artificial Intelligence and Robotics (ICBAR 2023) (pp. 441–445). ACM. (2023)
7. Zhang, Z., Zhang, R., Li, Z., Bengio, Y., & Paull, L.: Perceptual generative autoencoders. In Proceedings of the 37th International Conference on Machine Learning (ICML 2020) (pp. 11298–11308). PMLR. (2020)
8. Czolbe, S., Krause, O., Cox, I., & Igel, C.: A loss function for generative neural networks based on Watson's perceptual model. In Advances in Neural Information Processing Systems (NeurIPS 2020). (2020)
9. Liu, Y., Chen, H., Chen, Y., Yin, W., & Shen, C.: Generic perceptual loss for modeling structured output dependencies. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021) (pp. 5424–5432). (2021)
10. Pihlgren, G. G., Nikolaidou, K., Chhipa, P. C., Abid, N., Saini, R., Sandin, F., & Liwicki, M.: A systematic performance analysis of deep perceptual loss networks: Breaking transfer learning conventions. arXiv preprint arXiv:2302.04032. (2023)