Diffusion Models in Vision: A critical review

Ruthwik Ganesh (230702930)

***Abstract*—Diffusion models represent an emerging topic in computer vision, demonstrating remarkable results in generative modeling. The objective of this survey is to furnish a comprehensive overview of Diffusion models within the realm of computer vision, elucidating their impact and potential applications in this discipline.**

I. Introduction

Generative modeling, a cornerstone of deep learning over the past decade, has wielded significant influence across diverse domains including images, audio, text, and point clouds. These models aim to produce samples that closely match the distribution of the training data. While energy-based models (EBMs) achieve this through unnormalized probability densities, they rely on Markov Chain Monte Carlo (MCMC) sampling, known for its computational inefficiency. Recent advancements in deep learning architectures, however, have revitalized interest in generative models, particularly with the emergence of generative adversarial networks (GANs), variational autoencoders (VAEs), and normalizing flows. Diffusion-based generative models offer an alternative that circumvents challenges such as posterior distribution alignment, intractable partition function estimation, additional discriminator networks, or network constraints. These models find utility in various tasks including image generation, super-resolution, inpainting, segmentation, classification, and anomaly detection.

Diffusion models represent a class of probabilistic generative models that learn to reverse a process that progressively degrades the structure of training data. This entails two fundamental phases: the forward diffusion process and the backward denoising process. In the forward diffusion process, low-level noise is iteratively added to each input image, with the scale of noise varying at each step. The training data is systematically corrupted until it ultimately transforms into pure Gaussian noise. The backward denoising process involves reversing the forward diffusion process through an analogous iterative procedure, whereby noise is sequentially removed, thereby reconstructing the original image. Consequently, during inference, images are generated by gradually reconstructing them starting from random white noise, with the noise subtracted at each time step being estimated via a neural network, typically based on a U-Net architecture to preserve spatial dimensions.

II. Diffusion Architectures

This section dives into three well adopted formulations of diffusion models: Denoising diffusion probabilistic models (DDPMs), Latent diffusion models (LDMs), and an approach grounded in stochastic differential equations (SDEs) that generalizes the first two methods. For each formulation, the process of corrupting data with noise, the method for learning to reverse this process, and the procedure for generating new samples during inference are elucidated.

Denoising Diffusion Probabilistic Models (DDPMs): DDPMs [1] are generative models that reverse a diffusion process, adding Gaussian noise to data progressively. They consist of a forward process, degrading data into noise, and a backward process, reconstructing original data. Each step involves generating noised versions of the original data, modeled by a Markovian process. DDPMs efficiently sample any noisy version in a single step using a fixed variance schedule. The backward process employs a neural network to predict and remove noise, transforming pure Gaussian noise into structured data. DDPMs excel in data reconstruction but face challenges in sampling efficiency.

Latent Diffusion Model (LDM): LDMs [2] shifts the diffusion process from pixel space to latent space, reducing computational costs and enhancing speed. Despite significant contributions of perceptual details to image information, semantic and conceptual compositions persist even after compression. Hence, LDM incorporates autoencoders to reduce pixel-level information, facilitating a separation of perceptual and semantic compression. Subsequently, a diffusion process operates on latent learning. The denoising phase employs a time-conditioned U-Net with cross-attention mechanisms to handle flexible condition information. This model has spurred research into enhancing model efficiency.

Stochastic Differential Equations (SDEs): SDEs models diffusion as a continuous-time stochastic process. The forward diffusion transforms data into noise via drift and diffusion coefficients. Reverse diffusion employs reverse-time SDEs, requiring score function estimation at each time step. SDEs use neural networks to estimate these score functions, generating samples through numerical SDE solvers. This framework offers a more general and theoretically grounded approach, enabling efficient generation strategies and stronger theoretical results. SDEs represent a significant advancement, unifying and extending capabilities of DDPMs, albeit with computational complexities.

III. Selected Applications and Trends

Text-to-Image Synthesis: The achievements of diffusion models in text-to-image synthesis are particularly noteworthy, showcasing their capacity to combining disparate concepts, such as objects, shapes, and textures, to generate unconventional examples.

Gu et al. [3] present the VQ-Diffusion model, an approach to text-to-image synthesis devoid of the unidirectional bias present in earlier methods. The proposed method incorporates a masking mechanism to circumvent error accumulation during inference. The model comprises two stages, with the first stage based on a VQ-VAE representing an image through discrete tokens, and the second stage employing a discrete diffusion model operating on the discrete latent space of the VQ-VAE. The VQ-Diffusion model achieved superior outcomes in text-to-image generation when contrasted with traditional autoregressive

(AR) models possessing comparable parameter counts. In comparison to prior text-to-image approaches based on GANs, and exhibits enhanced capabilities in managing intricate scenes, resulting in a substantial improvement in the quality of synthesized images.

Avrahami et al. [4] introduce a text-driven editing of images using diffusion model conditioned on pretrained language-image model (CLIP), the editing process is directed towards a user-supplied text prompt. The synthesis involves blending noised versions of the input image with the locally guided diffusion latent at varying noise levels, ensuring a seamless integration of the edited region with the unchanged portions. Augmentations are introduced to the diffusion process serve to alleviate adversarial results.

Video Generation and Editing: There has been considerable scholarly inquiry into extending diffusion models into the temporal domain to effectively capture the broad spectrum of natural motion inherent in large-scale video datasets.

Ho et al. [5] pioneered the development of the Video Diffusion Model (VDM), which extends the 2D U-Net architecture into the temporal domain. This extension is facilitated through factorized space and time modules, offering enhanced computational efficiency, and enabling joint training on individual images, videos, and textual data.

Building upon this advancement, Imagen Video [6] introduces a scaled-up version of the VDM, featuring cascaded text-to-video (T2V) modelling with an impressive 11 billion parameters. This model comprises a base architecture with low spatio-temporal resolution, followed by multiple cascaded super-resolution models aimed at enhancing both spatial resolution and effective framerate. Imagen Video is trained from scratch using a substantial corpus of high-quality video data alongside corresponding captions, supplemented by diverse text-image datasets.

To leverage learned image priors for video generation, Make-A-Video [7] expands upon a pre-trained text-to-image (T2I) model by incorporating spatio-temporal convolution and attention layers into the existing framework. This approach allows for separate training of individual components: the T2I model is pre-trained on image-text pairs, while the complete T2V system is fine-tuned on a vast unlabelled video corpus, thereby obviating the need for video-caption-paired training data.

These approaches, characterized by factorized or separable spatio-temporal modules and built upon pre-trained text-to-image models, have been further applied to expand image Latent Diffusion Models (LDMs) to videos, as exemplified by Stable Diffusion [2]. Such extensions aim to learn the video distribution within a low-dimensional space, offering efficient training while capitalizing on the rich 2D priors acquired by T2I LDMs. Moreover, the versatility of this approach enables the transfer of learned motion modules to other image models derived from the same foundational T2I model, allowing for personalized video synthesis in specific styles or contexts [2, 8].

IV. Challenges

Research on diffusion models offers substantial potential for advancements in theoretical understanding and practical applications. Key research directions include the development of more efficient sampling methods, the improvement of likelihood, and the exploration of how diffusion models can be integrated with other generative models for various applications. Future research is expected to venture into several new areas:

Addressing Assumptions: Current assumptions, such as the complete erasure of information in data by the forward process of diffusion models, need reevaluation. The ideal point for halting the forward noising process to balance efficiency and sample quality is a topic of significant interest. Recent developments in Schrödinger bridges and optimal transport suggest novel approaches for diffusion models, potentially allowing convergence to specified prior distributions within finite timeframes.

Theoretical Understanding: Understanding why diffusion models are effective over other types, such as GANs, VAEs, or EBMs, is crucial. Identifying their unique characteristics could clarify why they produce high-quality samples and achieve top likelihoods. Additionally, establishing theoretical guidance for systematic hyperparameter selection is necessary.

Latent Representations: Unlike VAEs or GANs, diffusion models struggle with generating meaningful latent space representations. This affects their utility in tasks involving data manipulation based on semantic representations. The challenge is compounded by the latent space often being as dimensionally extensive as the data space, which impacts sampling efficiency and the quality of representation learning.

Pitfalls in Diffusion Models: Despite their ability to generate high-quality synthetic images and scalability, diffusion models have several pitfalls, including the amplification of biases from training datasets, generation of inappropriate images, and privacy concerns. Addressing these challenges is imperative for future research, focusing on improving the models' robustness and ethical usage.

V. Conclusion

In this review, we explored the evolving landscape of Diffusion models in computer vision, a significant stride in generative modeling. These models, distinct in their forward diffusion and backward denoising processes, demonstrate versatility in applications ranging from image synthesis to video editing, distinguishing them from traditional methods like GANs and VAEs. Through our examination of various formulations such as DDPMs, LDMs, and SDEs, we showcased their strengths in balancing efficiency and quality. Despite remarkable achievements, challenges remain, including the need for improved sampling efficiency and theoretical understanding. Future research directions include refining efficiency, deeper theoretical insights, and responsible application, given their potential in AI-driven art and content creation. Ethical considerations and bias mitigation are paramount as we advance. Diffusion models hold immense promise for the future of generative modeling, with their continuous evolution marking a significant milestone in the realm of artificial intelligence.

REFERENCES

[1] Ho, J., Jain, A., Abbeel, P., & Srinivas, A. (2021). Denoising Diffusion Probabilistic Models. arXiv preprint arXiv:2102.09672.

[2] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., et al. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. arXiv preprint arXiv:2112.10752.

[3] Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., & Guo, B. (2021). Vector Quantized Diffusion Model for Text-to-Image Synthesis. arXiv preprint arXiv:2111.03857.

[4] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text driven editing of natural images," in Proc. Conf. Comput. Vis. Pattern Recognit., 2022, pp. 18208–18218.

[5] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022). Video diffusion models. arXiv preprint arXiv:2204.03458.

[6] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. (2022). Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303.

[7] Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. (2022). Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792.

[8] Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B., et al. (2022). Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2112.10752.