International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 5, Issue 3, 118-128, 2024
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V5I3P112

Original Article

Multiview Diffusion Models for High-Resolution Image Synthesis

Sajud Hamza Elinjulliparambil
Pace University, United States.

Abstract *- Multiview image synthesis aims to generate multiple coherent images of a scene from different viewpoints, a capability that is essential for applications such as 3D reconstruction, virtual reality, medical imaging, and autonomous systems. While recent advances in diffusion-based generative models have significantly improved image fidelity and training stability, ensuring geometric, photometric, and semantic consistency across multiple high-resolution views remains a fundamental challenge. This paper presents a comprehensive review of multiview diffusion models for high-resolution image synthesis, focusing on methods that explicitly or implicitly enforce cross-view consistency. We first introduce the theoretical foundations of diffusion models, including denoising diffusion probabilistic models, score-based generative frameworks, and latent diffusion. We then systematically analyze the core challenges of multiview high-resolution synthesis and propose a structured taxonomy of existing approaches based on conditioning strategies, architectural designs, and geometry-aware modeling mechanisms. Furthermore, we review resolution-scaling and computational optimization techniques that enable diffusion models to operate effectively at high resolutions. Widely used datasets and evaluation metrics are discussed, highlighting current limitations in benchmarking multiview consistency. Finally, we survey key application domains and identify open research challenges and future directions. This review provides a unified perspective on the intersection of multiview learning and diffusion-based generation, serving as a valuable reference for researchers and practitioners in generative modeling and 3D vision.*

Keywords *- Multiview Image Synthesis, Diffusion Models, High-Resolution Generation, Latent Diffusion, Geometric Consistency, View-Consistent Generation, 3D-Aware Generative Models, Novel View Syntheis.*

1. Introduction

The synthesis of high-resolution images has now emerged as an essential research problem in computer vision because it has a wide range of applications in graphics, medical image science, remote sensing, robotics, and the creation of immersive 3D content [1]. The fact that it is possible to produce visually realistic and high-resolution images allows many downstream tasks, such as reconstruction of scenes, augmentation of data, simulation worlds, and diagnostic imaging. The latest breakthroughs in generative modeling have greatly enhanced the quality of image fidelity and synthetic data are becoming more and more difficult to differentiate between real-world data and a generated one [2]. Irrespective of this development, a majority of the initial procedures of generative image synthesis are single-view image synthesis in which a single image is produced at a time. These single-view approaches have natural limitations such as ambiguity of viewpoints, artifacts of occlusion and geometric inconsistency in the cases where two or more views of the same scene are needed. In applications like 3D reconstruction, novel view synthesis and multi-camera In independently generated views, distinctly generated views typically result in inconsistent object structure, incorrectly oriented textures and physically unrealistic scene layouts [3].

In order to overcome these issues, multiview learning has been confirmed as an efficient paradigm, which utilizes the correlations among multiple viewpoints on the same scene. Multiview methods attempt to impose spatial, photometric and semantic consistency on the multiple views that have been bodily modeled in order to minimize ambiguity and enhance the physical plausibility of the generated images. Multiview image synthesis is especially essential in high-resolution cases, and inconsistencies are more apparent and harmful to downstream activities. Simultaneously with the development of multiview modeling, diffusion models have recently become a formidable type of generative models. Developed as an extension of denoising diffusion probabilistic models (DDPMs), diffusion-based methods produce samples by a denoising autoencoder, and thus provide better training stability and highly realistic image ideas than adversarial ones [4]. Diffusion models can form a promising base to high-resolution multiview image generation because, unlike a set of independent images, multiview image synthesis targets the creation of multiple consistent views of a scene as illustrated in Fig. 1.

This figure 1 illustrates the fundamental difference between traditional single-view image generation, where each image is synthesized independently, and multiview image synthesis, where multiple views of the same scene are generated with enforced geometric and semantic consistency across viewpoints. Generative image model development has evolved all the way through multiple distinct paradigms in the last decade [5]. Variational autoencoders (VAEs) were some of the earlier probabilistic models used to generate images, which introduced the concept of latent variables but tended to generate smooth images, because of limiting assumptions of likelihood models. Thereafter, generative adversarial networks (GANs)

demonstrated much higher levels of visual realism with the use of adversarial training, but GANs have been reported to be susceptible to training instability, mode collapse, and lack multiview-consistent generation.

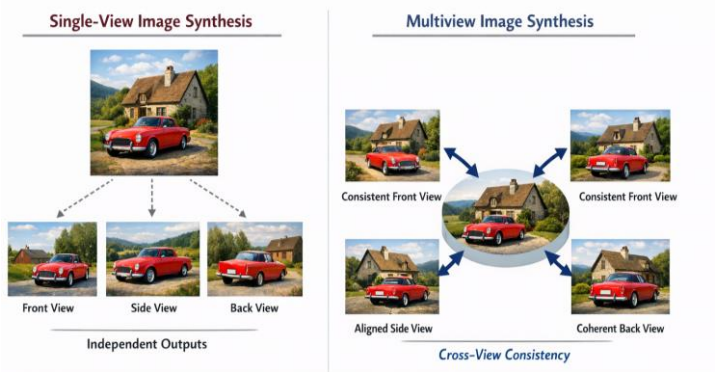

Fig 1: Conceptual Comparison between Single-View and Multiview Image Synthesis

Autoregressive models also enhanced likelihood evaluation and sample heterogeneity by conditioning images as chains of conditional distributions. Although useful in modeling intricate dependencies, the autoregressive models are computationally costly to infer and practically infeasible to apply in high-resolution multi-view synthesis since they are sequential [6]. Diffusion models have become the new trend in generative modeling. By describing image generation as a denoising process step by step under Gaussian noise, diffusion models have been found to have increased training stability, scalability and sample fidelity [7]. In comparison to GANs diffusion models do not need adversarial optimization and are associated with more stable convergence patterns. Such properties have also contributed to diffusion models being especially useful in multiview image synthesis, where views are to be generated reliably by joint optimization and expressive generative capability. Table 1 provides a high-level summary of key generative paradigms and the multiview synthesis-suitable ones.

Table 1: Evolution of Generative Models and Their Suitability for Multiview Image Synthesis.

Generative Model Type	Training Stability	Scalability to High Resolution	Multiview Consistency Capability	Key Limitations
Variational Autoencoders (VAEs)	High	Moderate	Limited	Over-smoothed outputs, limited detail
Generative Adversarial Networks (GANs)	Low–Moderate	High	Limited–Moderate	Mode collapse, unstable training
Autoregressive Models	High	Low	Moderate	Slow inference, high computational cost
Diffusion Models	High	High	High	Computationally intensive sampling

The review is based on multiview diffusion models of high-resolution image synthesis including all research works published. It has a tightly focused scope on diffusion-based generative models which either explicitly or implicitly deal with multiview consistency, geometric consistency, and high-resolution image generation. In contrast to the current surveys which present a general picture of the diffusion models of generic image synthesis, this article focuses on the area of intersection of multiview learning and diffusion-based generation, which is an issue that has been getting growing attention over the past few years. This review has three major contributions. First, it offers a structured taxonomy of multiview diffusion models, classifying the existing methods according to conditioning strategies, architecture and geometry awareness. Second, it discusses the major limitations and obstacles, such as scalability, computing efficiency and limits on consistency in high-resolution multiview synthesis. Third, it examines popular datasets and evaluation measures, outlines vulnerabilities in benchmarking behaviour, and reveals the gaps in the research. By conducting this systematic study, the review will provide a complete resource to researchers and practitioners in the multiview generative modeling field via diffusion techniques.

2. Fundamentals of Diffusion Models

Diffusion models constitute a class of probabilistic generative models that learn to synthesize data by iteratively transforming random noise into structured samples [8]. Their formulation is rooted in nonequilibrium thermodynamics and stochastic processes, enabling stable training and high-quality sample generation. This section introduces the core principles of diffusion-based generation and establishes the foundation for understanding their extension to multiview and high-resolution image synthesis.

2.1. Denoising Diffusion Probabilistic Models (DDPMs)

Denoising Diffusion Probabilistic Models (DDPMs) describe data generation as a two-stage stochastic procedure, which is a forward noising process and a reverse denoising process [9]. In the forward process, a clean sample of data is corrupted by slowly introducing Gaussian noise in a series of time steps. This process is specified so that, with a large number of steps, the data distribution is drawn to simple isotropic Gaussian distribution. Notably, the forward diffusion process is deterministic and does not have to be learnt. DDPMs are trained to learn the reverse process, which makes them have a generative capability whereby a neural network is trained to incrementally remove noise and restore the original data distribution [10]. The model does not directly predict the denoised sample, instead, it usually learns to approximate the noise in the sample added at each diffusion step. This re-formulation makes optimization easier and results in stable training dynamics. Its overall intuition is that complex data distributions can be trained by breaking the process of generation down into a series of simpler denoising tasks.

As a model, DDPMs may be introduced on a pixel space or on a latent space. The pixel representation in pixel-space diffusion directly acts on the high dimensional image pixels providing the capability of fine-grained detail modeling at a high computational and memory cost, especially with high-resolution images. Conversely, latent-space diffusion encodes the images into a lower-dimensional latent value and uses the diffusion process, which is much more efficient, and the perceptual quality of the images does not significantly decrease [11].The relationship between the forward noise injection and the learned reverse denoising process is illustrated in Fig. 2.

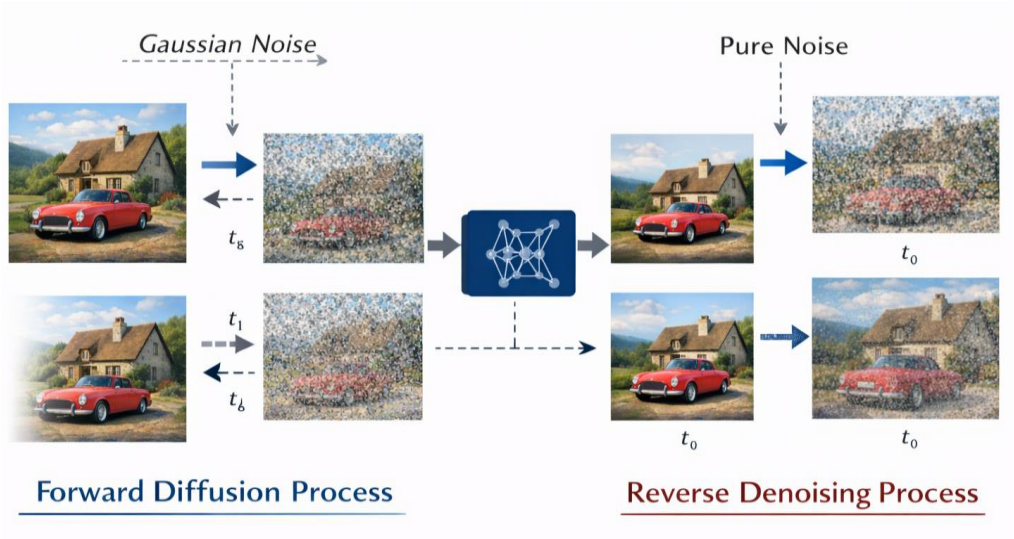

Fig 2: Schematic Illustration of the Diffusion Process

This figure depicts the forward diffusion process, where Gaussian noise is progressively added to a clean image over multiple time steps, and the reverse denoising process, where a neural network learns to iteratively remove noise to reconstruct a high-quality image sample from pure noise.

2.2. Score-Based Generative Models

An equally close yet very different formulation of diffusion-based generation is score based generative models. These models are not only implicit models of the reverse diffusion transitions, but also learn the score function, the gradient of the log-probability density with respect to the data. Score matching allows the model to approximate how to modify a noisy sample so as to move the sample towards areas that are more likely to have data. An important step forward in score-based models is that they are defined by stochastic differential equations (SDEs), which are equations of motion of the diffusion process in time. Through the definition of forward and reverse-time SDEs, score-based models bring discrete diffusion models and continuous stochastic processes together into the same theoretical framework [12]. The sampling is achieved through interpretation of the reverse-time SDE by numerically solving with the learned score function. Although DDPMs and score-based models are not the same at the mathematical level (discrete-time Markov chains versus continuous-time SDEs), their expressive power is essentially the same. The two methods are based on a progressive noise perturbation and trained denoising dynamics. In reality, DDPMs are commonly favored due to their ease of implementation, whereas more flexible schedules of noise and sampling plans are available in score-based models. This has allowed cross-fertilization of ideas between the two paradigms, and has guided the design of multiview diffusion models.

2.3. Latent Diffusion Models

Despite outstanding generative results, diffusion models are facing a challenge in their ability to generate large-resolution images because the cost of interacting directly in pixel space cannot be readily reduced to allow scaling to large images. Latent diffusion models overcome this weakness by running the diffusion model on a learned latent image (instead of on image

pixels). The heart of the matter is to encode an image with an autoencoder into a small latent space and to generate with the use of diffusion in this small space. Latent diffusion models are typically based on an encoder-decoder architecture, in which the encoder processes images of high resolution into a continuous or discrete latent representation, and the decoder is used to recover images by taking latent samples generated by the encoder [13]. Continuous autoencoders or vector-quantized variational autoencoders (VQ-VAEs) are often inspired by architecture. Latent diffusion has created a desirable trade-off between efficiency and visual quality by the decoupling of perceptual compression and generative modeling. Latent diffusion models are especially effective in multiview and high-resolution image generation, as they greatly lower the amount of memory needed, and allow large datasets to be trained. Moreover, the use of latent space helps to combine the conditioning signals, e.g., camera viewpoints or reference images, required to impose the consistency between the views. Consequently, many developed multiview diffusion frameworks have latent diffusion as a core in them.

3. Multiview Image Synthesis: Concepts and Challenges

Multiview Image synthesis is aimed at producing various images of the same image at varying viewpoints and ensuring consistency between the views. Multiview generation, unlike single-view generation, explicitly models the geometric relationships and common place scene structure and is a core element of 3D perception, reconstruction, and scene interpretation. This section makes this concept of multiview consistency more formal, explains the hardest problems of high-resolution multiview synthesis, and surveys the classic generation methods invented before diffusion models.

3.1. Definition of Multiview Consistency

Multiview consistency can be defined as the fact that images obtained at various perspectives of the same scene can be coherent with each other at various dimensions [14]. This consistency may be divided into geometric, photometric and semantic.

3.2. Geometric Consistency

Geometric consistency can be used to make sure that the spatial structure of objects is consistent with a common underlying 3D structure in all the views. Connected locations in various images are expected to obey epipolar conditions and have constant depth relationships. Geometric consistency violations tend to be in the form of distorted shapes of objects, wrong relative positioning, or depth differences, which impact 3D reconstruction and view synthesis performance in a very harmful way.

3.3. Photometric Consistency

Photometric consistency It must be that appearance properties of color, texture, and shading are consistent across views [15]. Although the viewpoint-dependent illumination variations are not surprising, significant differences in the texture patterns or surface appearance are the signs of the inability to reproduce the common scene identity. High-resolution synthesis Photometric consistency Photometric attack Photometric detection Photometric glucometry Photometric AFM Photometric scanning AFM Photometric cell culture

3.4. Semantic Consistency

Semantic consistency is used to maintain high level meaning between views. The semantic labels, categories, and identities of objects and other elements of the scene should be the same irrespective of viewpoint. As an illustration, an object that is known to be a particular anatomical structure or vehicle in one perspective should be the same semantically in all views generated. This detail is paramount to downstream activities of understanding the scene and reasoning. Taken together, these varieties of consistency are needed to understand the 3D, since the multiview observations give a complementary information that eliminates the depth ambiguity, occlusion, and view-dependent distortions.

3.5. Challenges in Multiview High-Resolution Image Synthesis

The requirement of multiview consistency is even more difficult to attain when high resolution images are being synthesized. A number of basic challenges are present in this environment.

3.5.1. Viewpoint Ambiguity

The reason is that the ambiguity of viewpoint is due to the fact that a single image in 2D may have many possible interpretations of the 3D scene. According to the case of many different views being produced separately, these ambiguities are likely to produce conflicting scene structures and inconsistent geometry across views.

3.5.2. Occlusion Handling

The issue of occlusions is a significant problem in multiview synthesis since objects that are visible in one view are not always visible in the other. A generative model should be able to reconstruct plausible occluded content, as well as maintain geometric plausibility among views. Poor occlusion reasoning may often result in missing structures, redundant objects, or wrong object boundaries [16].

3.5.3. Coherence across Views at High Resolutions.

The issue of keeping cross-view coherence would become more challenging with an increase in the image resolution. The amplification of small irregularities in texture matching, object outlines, and minute structural features is made by high-resolution synthesis. Minor aberrations may produce artifacts that are aesthetically unpleasing and thus ruin the effect of the realism of the produced multiview images.

3.5.4. Computational Cost

Multiview synthesis can raise a considerable level of computational complexity because the relation has to be modeled in multiple images of high dimensionality. The costs of training and inference increase exponentially with the resolution of the images and the number of perspectives, which puts scalability restrictions across most domains in practice. These difficulties are graphically outlined in Fig. 3.

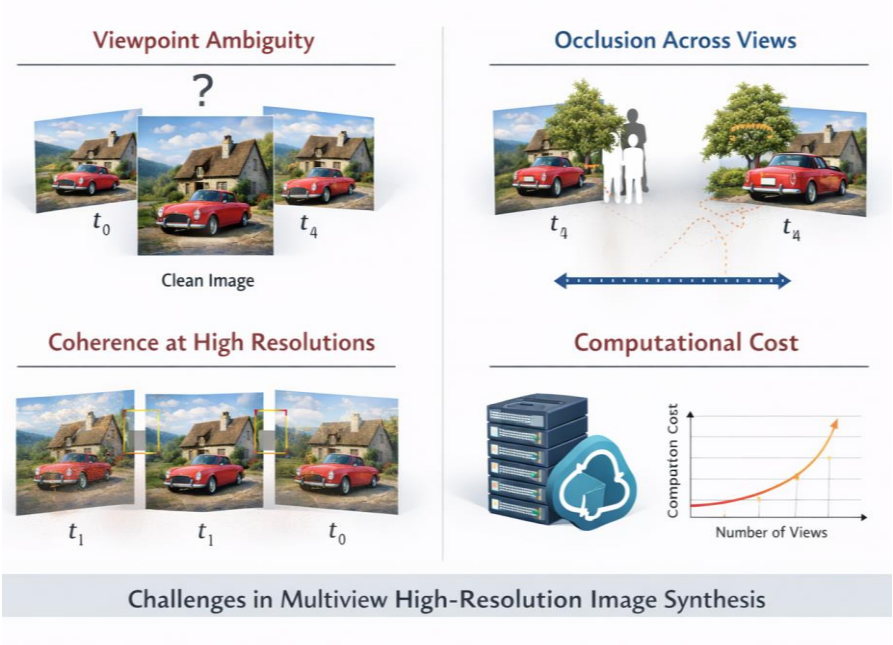

Fig 3: Key Challenges in Multiview High-Resolution Image Synthesis

This figure illustrates major challenges, including viewpoint ambiguity, occlusion across views, loss of cross-view coherence at high resolutions, and increased computational complexity when jointly modeling multiple viewpoints.

3.6. Traditional Multiview Generative Approaches (Pre-Diffusion)

Before the adoption of diffusion models, multiview image synthesis was predominantly addressed using GAN-based frameworks and NeRF-related generative pipelines.

3.6.1. GAN-Based Multiview Synthesis

Multiview based on GAN methods were generally based on conditional adversarial training where the generator was conditioned on viewpoint parameters, camera pose or reference images. These approaches were visually appealing but had problems of training instability, mode collapse and bad enforcement of geometric consistency across views. Therefore, it was challenging to scale GANs to high-resolution multi-view synthesis [17].

3.6.2. Generative Pipelines related to NeRF.

Neural Radiance Field (NeRF)-based methods trained scenes in the form of continuous volumetric representations and produced images at arbitrary viewpoints [18]. These techniques had powerful geometric priors and state of the art view-consistent rendering. Nevertheless, the classic NeRF pipelines are computationally intensive and they need dense multiview supervision, and are not naturally suited to a variety of high-quality image generation.

3.6.3. Motivations Limitations Motivating Diffusion-Based Models.

The GAN-based and NeRF-related models demonstrated serious drawbacks. GANs do not have a stable optimization and global consistency, and NeRF-based approaches have scalability and flexibility problems. These drawbacks encouraged the development of diffusion models, which provide stable training dynamics, high sample fidelity and allowable conditioning mechanisms, which is why they are suitable for high-resolution multiview image synthesis.

4. Taxonomy of Multiview Diffusion Models

Multiview diffusion models can be systematically categorized based on how multiview information is incorporated into the diffusion process. This section presents a taxonomy along three principal dimensions: conditioning strategies, architectural designs, and geometry-aware modeling mechanisms. Such a classification provides a structured understanding of existing approaches and highlights design trade-offs in achieving multiview consistency and high-resolution synthesis.

4.1. Conditioning Strategies for Multiview Diffusion

Conditioning strategies characterize the external multiview information injection into diffusion process. Conditional proper enforcement is the only way of maintaining consistency among the views as generative diversity is maintained.

4.2. Viewpoint-Conditioned Diffusion

Viewpoint-conditioned diffusion models directly use camera parameter or viewpoint information as conditioning of the generative process. The conditioning signals that are commonly used are represented by camera pose, which can be extrinsic and intrinsic parameters or view relative transformations. Positional embeddings or learned projection layers are incorporated into the diffusion network to allow the model to learn to match specific viewpoints with image structures. These models are trained to produce geometrical aligned images with the viewing direction by conditioning the diffusion process on camera viewpoints. This strategy is especially efficient in controlled view synthesis and novel view generation, in which there is a necessity to align the viewpoints correctly. Viewpoint-conditioned models can however be problematic with noisy camera parameters or unavailable parameters, and in an unconstrained environment.

4.3. Image-Conditioned Multiview Diffusion

Image-conditioned multiview diffusion models are based on one or more reference images, used to inform the creation of more views. These methods do not rely on explicit camera parameters, rather they rely on visual cues obtained by using reference images to determine the underlying structure and appearance of a scene. Conditioning is normally done via feature extraction networks which encode reference images into latent representations [19]. Image-conditioned diffusion models heavily rely on the processes of cross-attention whereby the generative mechanism chooses the most pertinent spatial and semantic information of the reference views. Cross-attention facilitates the preservation of identity, texture continuity, and semantic consistency of objects by dynamically combining information across views. Image-conditioned methods come in handy especially in cases where there is no metadata in the camera, but they can be problematic when trying to generalize over the perspectives of the reference images.

4.4. Architectural Designs

In addition to the use of conditioning approaches, architectural design options also play a major role in determining the level of efficiency and consistency of the multiview diffusion models. The current methods mainly are based on shared-backbone structures or multi-branch diffusion models.

4.5. Shared Backbone Architectures.

Shared backbone architectures use one diffusion model to create multiple views, by sharing parameters on all views. The similarity of representations and unified denoising dynamics are promoted in this design to achieve multiview consistency. By sharing parameters, the complexity of models is reduced, as is the memory usage, and sharing features encourages the ability to learn features across views. This design is especially beneficial to high-resolution synthesis because a high-resolution synthesis can be scaled more effectively with this design by learning to use previously learned representations. Shared backbones can however be limited when it comes to capturing highly view-specific information particularly where the viewpoints are quite different.

4.5.1. Multi-Branch and Coupled Diffusion Models

Multi-branch diffusion models have different diffusion processes per view, and the views are explicitly linked to make the processes consistent. Such coupling methods can be common latent variables, cross-view attention modules, or terms of consistency loss that discourage errors between generated viewpoints. Multi-branch architectures are more flexible in modeling variations depending on view because each branch is allowed to specialize in one particular view. Nonetheless, a higher model complexity and cost of computation in the case of multiple diffusion branches is challenging to scale. Table 3 gives a comparative account of shared and multi-branch architectural designs.

Table 2: Comparison of Architectural Designs for Multiview Diffusion Models

Architecture Type	Parameter Sharing	Computational Cost	Multiview Consistency	Scalability
Shared Backbone	High	Low–Moderate	Moderate–High	High
Multi-Branch / Coupled	Low–Moderate	High	High	Moderate

4.6. Geometry-Aware Diffusion Models

Geometry-aware diffusion models The geometric priors of 3D geometry are explicitly placed in the diffusion process in order to provide better multiview consistency and structural correctness. The incorporation of Epipolar Geometry is implemented since a light beam can interact with a moving object, or radiation is capable of moving through a time-dependent medium, both of which cannot be conveniently represented by geometric models.

4.6.1. Introduction of Epipolar

Certain multiview diffusion models use epipolar constraints to ensure that the views are geometrically aligned. These strategies can be used to reduce geometrical ambiguity by basing the correspondence of points between views on known camera geometry. Losses in projection consistency are additional punishments to the differences between features projected or points reconstructed in different perspectives [20]. This kind of geometry-aware constraints enhances the physical realism of the images being generated, and has proved useful in tasks that demand high accuracy in spatial correspondence, including 3D reconstruction and novel view synthesis. Nevertheless, such techniques tend to be based on a proper calibration of the camera, and such calibration might be unavailable in some cases.

4.6.2. 3D-Conscious Latent Representations.

The other way of learning 3D-aware diffusion is to learn 3D-aware latent representations that implicitly capture scene geometry. These models are learned on latent space rather than explicit geometric constraints, and are view-consistent with underlying 3D geometry. To encode spatial relationships, implicit 3D representations are commonly used e.g. neural fields or volumetric feature grids. These models can be used to generate coherent multi-view images without direct geometric guidance by working in view-consistent latent spaces. Multiview diffusion models that use geometry knowledge to encode camera geometry and 3D prior are shown in Fig. 4.

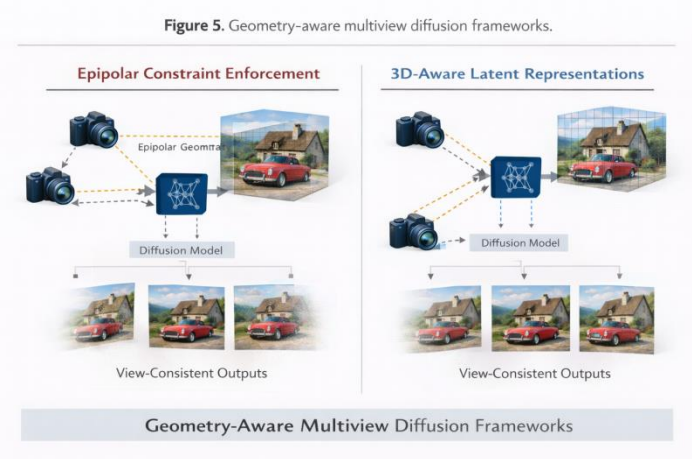

Fig 3: Geometry-Aware Multiview Diffusion Frameworks

Figure 4 illustrates representative approaches that integrate camera geometry and 3D priors into diffusion models, including epipolar constraint enforcement and 3D-aware latent representations for achieving view-consistent image synthesis.

5. High-Resolution Image Synthesis with Multiview Diffusion

The high-resolution image synthesis makes the task of establishing multiview consistency even harder because small-scale spatial details should be consistent between a variety of perspectives. Diffusion models deal with these issues with specialized resolution scaling methods and computational optimization, which allow multiview generation to scale and be constant across runs. In this section, the review of the key techniques that are applied to facilitate the high-resolution synthesis is provided, and trade-offs between visual fidelity and cross-view coherence are discussed.

5.1. Strategies of Resolution scaling.

Resolution scaling techniques are used to scale the outputs of diffusion models, trained to produce low or medium resolution images, to high-resolution multi-view images without the views being consistent. A popular method is cascaded diffusion in which image generation is carried out in a series of steps at successively higher resolutions. A diffusion model is then used to produce a rough, low-resolution multiview representation, which embodies the structure of the scene globally, and later a series of diffusion models that are capable of super-resolution are then trained to improve the original representation. This hierarchical break-down provides stability to training and allows the control of the synthesis of details between global and local levels better, however, the errors made at early stages can be propagated through the cascade.

Progressive upsampling is another technique that is also commonly employed; here, the denoising process increases the spatial resolution gradually as the diffusion process advances. The diffusion network is not trained to use isolated super-resolution networks and instead trains to add finer details as the noise level drops. Progressive upsampling in multiview allows consistency constraints to be imposed at multiple views, enhancing coherence. The method however, is more complex in training and consumes more memory especially in synthesizing multiple high-resolution views at a time. Latent super-resolution has a computationally efficient variant, which applies resolution enhancement to a learned latent space. With diffusion models, it is possible to produce high-resolution outputs by training them using compact latent representations instead of pixel-space images, which leads to much less computational overhead. Latent super-resolution is particularly useful with multiview diffusion where view-consistent latent representations are useful in maintaining geometric structure and appearance across views and scaling to high resolutions.

5.2. Memory and Computational Optimization

A high-resolution multiview diffusion has computational requirements that require effective memory and optimization solutions to find [21]. Patch-based diffusion is one of the solutions to this issue in which images are divided into small spatial patches and diffusion is applied to these patches. This minimizes memory needs and makes it possible to train it on small hardware. Boundary artifacts Multiview Multiview systems are used to reduce the effects visible along the boundary between different views, which is a side effect of the latent context shared by the different views. In spite of these, patch based techniques can still fail to capture long-range dependencies. Latent compression also leads to efficiency by making latent representations of diffusion models lower dimensional [22]. Compressed latent spaces remove redundant content and enable the training and inference of images more quickly and thus they are suitable in large-scale multiview synthesis. Moreover, the information sharing across views can be done more efficiently with the help of latent compression and information sharing, which can, in turn, contribute to multiview consistency.

Such optimization methods provide trade-offs in quality and speed between images and computational speed. Crashy compression and patch-based processing may suffer fine-grained detail loss or cause visual artifacts, whereas high-quality results may demand more substantial models and more diffusion chains. Table 4 gives a comparative overview of resolution scaling and optimization strategies of high-resolution multiview diffusion models.

Table 3: Resolution scaling and optimization strategies for high-resolution multiview diffusion models.

Strategy	Core Idea	Computational Cost	Effect on Multiview Consistency	Limitations
Cascaded Diffusion	Multi-stage resolution refinement	High	Moderate–High	Error propagation
Progressive Upsampling	Gradual resolution increase	High	High	Training complexity
Latent Super-Resolution	Latent-space refinement	Low–Moderate	High	Decoder dependence
Patch-Based Diffusion	Spatial partitioning	Low	Moderate	Boundary artifacts
Latent Compression	Dimensionality reduction	Low	Moderate–High	Loss of fine details

5.3. Quality vs Consistency Trade-offs

Multiview diffusion models need to achieve visual fidelity and cross-view consistency, which are conflictual goals of high-resolution images[23]. The most common perceptual and statistical measures taken to assess image quality include FID, PSNR and SSIM, each of which is a measure of single-image realism. Nevertheless, maximization of these measures in itself does not ensure a consistency in a multitude of perspectives. Models with a high fidelity on single images can still produce a discrepancy in the geometry of objects, texture alignment, or semantic identity between different views.

The common failure modes of high-resolution multiview synthesis are the misalignment of textures, inhomogeneous boundary of objects, and drift in the semantics depending on the perspective of view. These problems only get more intense as the resolution grows and it is important to note that evaluation frameworks that can look at the two issues fidelity and multiview coherence are necessary. These trade-offs are one of the main barriers to the design of multiview diffusion models, and understanding and addressing them.

6. Datasets and Evaluation Metrics

The multiview diffusion models require high-quality datasets and necessary evaluation metrics to develop and benchmark them. In this section, the data sets that are typically utilized in the multiview image synthesis are reviewed and the measures that are typically used to measure the quality of the image and consistent results across different views are summarized.

6.1. Multiview Datasets

Multiview datasets Multiview datasets consist of images of the same object or scene taken with multiple viewpoints, allowing model learning to discover geometric and photometric consistency.

6.1.1. Synthetic Datasets

Procedural modeling, which is often based on 3D engines, is often used to produce synthetic datasets with close multiview coverage and ground-truth geometry. Well-known synthetic datasets are ShapeNet, Objaverse as well as CLEVR-based multiview renderings. These datasets provide complete control of viewpoints, lighting and composition of the scene and this is helpful in the study of multiview consistency. Synthetic datasets can however not be as diverse in texture and have the characteristics of natural noise as real world images, and thus they have less generalization.

6.1.2. Multiview Datasets in the Real World.

Multiview datasets in the real world are recorded with the help of multiple cameras or controlled rigs. Some of them are DTU, Tanks and Temples, and the Stanford Multi-View Car/Chair datasets. These datasets offer quality real-world pictures having variable viewpoints and natural textures, which is useful to analyze realistic generative performance. They are limited by resolution, number of views, and coverage on large scale scenes which can be problematic in the high-resolution synthesis task of multiview diffusion models.Table 5 gives a summary of widely used multiview datasets used in training multiview diffusion models.

Table 4: Overview of Multiview Datasets Used In Diffusion-Based Image Synthesis.

Dataset	Type	Number of Views	Typical Resolution	Key Characteristics
ShapeNet	Synthetic	24–48 per object	128×128 to 256×256	Dense viewpoints, precise geometry
CLEVR Multiview	Synthetic	20–30 per scene	128×128	Controlled objects, compositional scenes
DTU	Real-world	49–64 per object	600×800	High-quality scans, varied lighting
Tanks & Temples	Real-world	20–60 per scene	1024×1024	Large-scale outdoor scenes
Stanford Car / Chair	Real-world	30–36 per object	512×512	Focused object category, multiple viewpoints

6.2. Evaluation Metrics

Evaluation of multiview diffusion models requires both image quality metrics and consistency metrics across views.

6.2.1. Image Quality Metrics

Common metrics for assessing visual fidelity include:

- FID (Fréchet Inception Distance): Measures the distance between the distributions of generated and real images in a feature space.
- IS (Inception Score): Evaluates image diversity and semantic meaningfulness.
- PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index): Quantify pixel-level fidelity between generated and reference images.

These metrics capture the perceptual quality of individual images but do not reflect geometric or semantic alignment across multiple views.

6.2.2. Multiview Consistency Metrics

To evaluate multiview consistency, specialized metrics have been proposed:

- Cross-view feature similarity: Measures the similarity of learned features (e.g., using a pretrained backbone) across generated views of the same scene.
- Geometry consistency errors: Compute discrepancies in depth, normals, or projections across views, often using ground-truth 3D data when available.

These metrics ensure that multiview diffusion models produce outputs that are not only visually plausible but also geometrically and semantically coherent across viewpoints.

7. Applications

Multiview diffusion models have allowed many applications in computer vision, graphics, robotics and in science. They are so useful in environments where views of the same image have to be taken in several different viewpoints due to their capability to produce high-resolution images without compromising geometric, photometric, and semantic consistency.

7.1. 3D Content Creation and View Synthesis

Multiview diffusion models, in high-resolution, have found application in content generation in 3D, such as asset generation in games, virtual reality and movie making. These models can generate novel view points by conditioning on only a small set of reference views or camera poses, meaning that they can generate a wide variety of 3D assets without necessarily having to manually model them exhaustively. This feature saves a lot of cost and time related to the production of content. Besides the creation of assets, multiview diffusion supports the construction of virtual environments, where various coherent views of complicated scenes are necessary in immersive applications. Such models have the capability of creating regular

textures and geometry at large scale surroundings, which make it smooth to transition between viewpoints and enhance the reality of a virtual simulation.

7.2. Medical and Scientific Imaging

Multiview consistency is essential in medical and scientific imaging as well as appropriate interpretation and analysis. Multiview scans Synthesis of missing views or improvement of resolution without artifact generation is especially useful in clinical and research practice. The medical imaging pipelines also have the support of these models in data augmentation. This diversification of datasets through the creation of more realistic images using the existing scans helps diffusion models to reduce overfitting in their downstream machine learning models, and enables robust training when there is limited annotated data.

7.3. Robotics and Autonomous Systems

In autonomous systems and robotics, multiview diffusion models make contributions to the understanding of scenes by synthesizing consistent high-resolution views of scenes with a limited amount of sensor input. As an example, a robot with a partial view of the camera can use multiview diffusion to predict the missing regions to improve perception and planning in a complicated environment. Simulation environments of autonomous agents can also be supported using these models. Synthetic data created with high-fidelity and view-consistency, with the help of diffusion models, allows the perception and control algorithms to be trained and tested in virtual worlds, instead of relying on expensive real-world data assembly and being safer to develop in the development phase.

8. Future Directions/Challenges in Multiview Diffusion.

The multiview diffusion models have a number of notable shortcomings and unsolved problems, despite making a considerable advancement. It is challenging to be scaled to very large view counts due to the high complexity of training, and the time required to infer an image is very high, especially when producing high-resolution results with strong cross-view consistency. There are also persistent physical and semantic consistency problems such as lighting inconsistencies between viewpoints and artifacts of deformed objects, which can ruin the realism and usefulness of generated scenes. Additionally, the discipline has a lack of data sets and testing processes and has no standardized benchmarks and uses subjective quality measurements that make it harder to compare and validate the models. In the future, research efforts focus on creating integrated 2D3D diffusion models that can learn geometry and appearance together to enhance the structural fidelity of the view and methods of efficient training and inference, such as model compression and knowledge distillation to make them less computationally expensive. Moreover, the development of physically grounded multiview diffusion based on physics-aware priors and illumination models will improve visual realism, as well as cross-view consistency, and find more applications in 3D content creation and medical imaging, along with robotics.

9. Conclusion

This research presented a comprehensive review of multiview diffusion models for high-resolution image synthesis, a rapidly emerging research area at the intersection of generative modeling and multiview learning. Unlike traditional single-view generative approaches, multiview diffusion models aim to produce multiple images of a scene that are not only visually realistic but also geometrically, photo metrically, and semantically consistent across viewpoints. Such consistency is critical for downstream tasks including 3D reconstruction, novel view synthesis, scientific imaging, and robotics. We first established the theoretical foundations of diffusion-based generative models and discussed why their stable training dynamics, scalability, and flexible conditioning mechanisms make them particularly suitable for multiview synthesis. A detailed taxonomy was then introduced, categorizing existing approaches according to conditioning strategies, architectural designs, and geometry-aware modeling techniques. This taxonomy highlights the design trade-offs between model complexity, scalability, and the degree of enforced cross-view consistency.

We further reviewed resolution-scaling strategies and computational optimizations that enable diffusion models to operate at high resolutions, while emphasizing the inherent trade-offs between visual fidelity and multiview coherence. In addition, we surveyed commonly used multiview datasets and evaluation metrics, revealing significant gaps in standardized benchmarks and consistency-aware evaluation protocols. Although impressive progress has been made, current models still struggle with scalability to large numbers of views, computational efficiency, and robust enforcement of physical and semantic consistency at very high resolutions. Looking forward, future research is expected to focus on unified 2D–3D diffusion frameworks, more efficient training and inference techniques, and physically grounded priors that better model geometry, lighting, and material properties. Addressing these challenges will be crucial for advancing multiview diffusion models toward practical, large-scale deployment. Overall, this review aims to consolidate existing knowledge, identify open problems, and provide clear research directions for the continued development of high-resolution multiview diffusion-based image synthesis.

References

[1] Z. Zhang, X. Zhang, et al., "Multi-view consistent generative adversarial networks for compositional 3d-aware image synthesis," International Journal of Computer Vision, vol. 131, no. 8, pp. 2219–2242, 2023.

[2] P. Eigenschink, et al., "Deep generative models for synthetic data: A survey," IEEE Access, vol. 11, pp. 47304–47320, 2023.

[3] N. A. Manap, "Multi-view image synthesis techniques for 3D vision and free-viewpoint applications," 2012.

[4] A. Kazerouni, et al., "Diffusion models for medical image analysis: A comprehensive survey," arXiv preprint arXiv:2211.07804, 2022.

[5] L. R. A. Wilde, "Generative imagery as media form and research field: Introduction to a new paradigm," 2023.

[6] D. Lee, "A comparison of conditional autoregressive models used in Bayesian disease mapping," Spatial and Spatio-temporal Epidemiology, vol. 2, no. 2, pp. 79–89, 2011.

[7] L. Yang, et al., "Diffusion models: A comprehensive survey of methods and applications," ACM Computing Surveys, vol. 56, no. 4, pp. 1–39, 2023.

[8] D. Watson, et al., "Learning to efficiently sample from diffusion probabilistic models," arXiv preprint arXiv:2106.03802, 2021.

[9] A. Alimanov and M. B. Islam, "Denoising diffusion probabilistic model for retinal image generation and segmentation," in 2023 IEEE International Conference on Computational Photography (ICCP), 2023.

[10] X. Wang, et al., "Efficient transfer learning in diffusion models via adversarial noise," arXiv preprint arXiv:2308.11948, 2023.

[11] S. Yu, et al., "Video probabilistic diffusion models in projected latent space," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[12] H. Sun, et al., "Score-based continuous-time discrete diffusion models," arXiv preprint arXiv:2211.16750, 2022.

[13] R. Rombach, et al., "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[14] T. Khot, et al., "Learning unsupervised multi-view stereopsis via robust photometric consistency," arXiv preprint arXiv:1905.02706, 2019.

[15] T. Kruisselbrink, R. Dangol, and A. Rosemann, "Photometric measurements of lighting quality: An overview," Building and Environment, vol. 138, pp. 42–52, 2018.

[16] K. Saleh, S. Szénási, and Z. Vámossy, "Generative adversarial network for overcoming occlusion in images: A survey," Algorithms, vol. 16, no. 3, p. 175, 2023.

[17] T. Mahmud, M. Billah, and A. K. Roy-Chowdhury, "Multi-view frame reconstruction with conditional gan," in 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2018.

[18] B. Kaya, et al., "Neural radiance fields approach to deep multi-view photometric stereo," in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision, 2022.

[19] Y. Shi, et al., "Mvdream: Multi-view diffusion for 3d generation," arXiv preprint arXiv:2308.16512, 2023.

[20] H.-Y. Tseng, et al., "Consistent view synthesis with pose-guided diffusion models," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[21] Z. Zhou and S. Tulsiani, "Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[22] Z. Pan, X. Zhou, and H. Tian, "Extreme generative image compression by learning text embedding from diffusion models," arXiv preprint arXiv:2211.07793, 2022.

[23] K. Nagano, Multi-scale Dynamic Capture for High Quality Digital Humans, Ph.D. dissertation, Univ. of Southern California, 2017.