BetterScene: 3D Scene Synthesis with Representation-Aligned Generative Model

Yuci Han[1], Charles Toth[2], Alper Yilmaz[2], John E. Anderson[3], William J. Shuart[3]

[1] Dept. of Electrical and Computer Engineering, The Ohio State University
han.1489@osu.edu
[2] Dept. of Civil, Environmental and Geodetic Engineering, The Ohio State University
{toth.2, yilmaz.15}@osu.edu
[3] USACE ERDC GRL
{john.e.anderson, william.j.shuart}@usace.army.mil

Keywords: 3D Gaussian Splatting, Video Diffusion Model, Novel View Synthesis.

Abstract

We present BetterScene, an approach to enhance novel view synthesis (NVS) quality for diverse real-world scenes, using extremely sparse unconstrained photos. BetterScene leverages the production-ready Stable Video Diffusion (SVD) model pretrained on billions of frames as a strong backbone, aiming to mitigate artifacts and recovering view-consistent details at inference time. Conventional methods have developed similar diffusion-based solutions to address these challenges of novel view synthesis. Despite significant improvements, these methods typically rely on off-the-shelf pretrained diffusion priors and fine-tune only the UNet module while keeping other components frozen, which still leads to inconsistent details and artifacts even when incorporating geometry-aware regularizations like depth or semantic conditions. To address this, we investigate the latent space of the diffusion model and introduce two components: (1) temporal equivariance regularization and (2) vision foundation model-aligned representation, both applied to the variational autoencoder (VAE) module within the SVD pipeline. BetterScene integrates a feed-forward 3D Gaussian Splatting (3DGS) model to render features as inputs for the SVD enhancer and generate continuous, artifacts-free, consistent novel views. We perform evaluation using the challenging DL3DV-10K dataset, demonstrating significant visual quality improvements over previous state-of-the-art diffusion-based methods on NVS tasks.

1. Introduction

Novel View Synthesis (NVS) plays a critical role in recovering 3D scenes. With the advent of Neural Radiance Fields (NeRF) Mildenhall et al. (2020) and 3D Gaussian Splatting (3DGS) Kerbl et al. (2023), we can now render photorealistic views of complex scenes efficiently. Yet, both NeRF and 3DGS suffer from performance degradation in sparse-view settings, particularly in under-observed areas for scene-level view synthesis, which hampers their practical applicability in real-world scenarios.

To tackle this ill-posed challenge, many methods incorporate additional regularizations during the training of NeRF or 3DGS, such as cost volumes Chen et al. (2024a), depth priors Xu et al. (2025) Deng et al. (2021) Li et al. (2024)Wang et al. (2023) Roessle et al. (2021) or visibility Somraj and Soundararajan (2023) Kwak et al. (2023). Despite their improvements in rendering quality of NVS, these methods still exhibit significant artifacts including spurious geometry and missing regions. Fortunately, recent advancements in video generative models pretrained on internet-scale datasets demonstrate promising capabilities in generating sequences with plausible 3D structure Blattmann et al. (2023a)Blattmann et al. (2023b). Researchers Liu et al. (2024b) Luo et al. (2024)Wu et al. (2025)Wang et al. (2025)Wu et al. (2023)Chen et al. (2024b) have employed diffusion models as effective enhancers for NVS from sparse views, capable of “imagining” unobserved regions and mitigating artifacts. Despite these improvements, these methods have limitations, particularly in two aspects: (1) lack of shift stability, and (2) limited ability to hallucinate plausible detailed appearance in underconstrained regions. Meanwhile, it is worth noting that most contemporary diffusion-based NVS enhancement methods primarily focus on optimizing solely the denoising module, specifically the U-Net denoiser architecture in video diffusion pipelines. However, the potential of diffusion models’ latent representations for NVS enhancement remains unexplored.

In this work, we exploit the capabilities of unconstrained high-dimensional latent space for enhancing 3D scene synthesis. Several influential works Blattmann et al. (2023a) Dai et al. (2023) have demonstrated that under the same spatial compression rate (or "down-sampling rate"), increasing the dimension of latent visual tokens leads to better reconstruction quality (see Fig. 2). This plays a key role in maintaining scene realism when using generative models as enhancers for NVS, avoiding over-hallucination while achieving higher-quality detail reconstruction. However, research Blattmann et al. (2023a) Xie et al. (2024) also revealed an optimization dilemma: while increasing token feature dimensions improves reconstruction, it significantly degrades generation performance. Common strategies to address this issue include either scaling up model parameters as demonstrated by Stable Diffusion 3 Blattmann et al. (2023a) or sacrificing reconstruction quality with limited token dimensions. However, neither approach is suitable for NVS tasks. We argue that both the reconstruction and generative capability of latent diffusion models (LDMs) are crucial for tackling the aforementioned limitations of conventional NVS methods. Moreover, the video diffusion backbone inherently constrains model scaling.

In this paper, building on representation-aligned LDM Yu et al. (2025) Yao et al. (2025), we propose BetterScene, a novel view synthesis framework that incorporates feed-forward Gaussian Splatting with a representation-aligned and equivariance-regularized video diffusion model Kouzelis et al. (2025) Zhou et al. (2025). Our key idea is to leverage high-dimensional

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-269-2026 | © Author(s) 2026. CC BY 4.0 License.

Figure 1. We demonstrate our **BetterScene** approach on diverse in-the-wild scenes. Given sparse inputs, recent novel view synthesis methods suffer from performance degradation due to insufficient visual information. BetterScene enhances novel view rendering quality by mitigating artifacts and recovering view-consistent details at inference time with an alias-free, representation-aligned video diffusion model.

equivariant latent representations for video LDM, achieving both superior reconstruction and generation quality to enable enhanced novel view synthesis while addressing the aforementioned limitations. Specifically, we first train a variational autoencoder (VAE) guided by vision foundation models using both an alignment loss Yao et al. (2025) and an equivariance loss that penalizes discrepancies between reconstructions of transformed latent representations and the corresponding input image transformations Kouzelis et al. (2025). We choose Stable Video Diffusion (SVD) Blattmann et al. (2023b) as the enhancer backbone, integrating our pretrained VAE module and fine-tuning the denoising UNet in the second stage. Furthermore, we leverage the feed-forward 3DGS model, MVSplat Chen et al. (2024a), to generate coarse novel views as SVD conditioning frames, bypassing the computationally expensive per-scene optimization required by conventional 3DGS approaches.

We evaluate our BetterScene on the real-world scene-level DL3DV-10K dataset. Extensive results demonstrate that BetterScene surpasses existing LDM-based NVS baselines in both fidelity and visual quality, yielding more photorealistic rendering outputs. Our main contributions can be summarized as follows.

- We propose an effective framework that combines feed-forward 3D Gaussian Splatting with a representation-aligned, equivariance-regularized video LDM for novel view synthesis.
- We exploit the capabilities of unconstrained high-dimensional latent spaces by training a variational autoencoder under the guidance of vision foundation models with both alignment and equivariance losses. By integrating our VAE with the SVD refinement module, we achieve enhanced reconstruction and generation quality while addressing limitations of traditional NVS methods.
- We conduct extensive experiments on the large-scale DL3DV-10K dataset, which contains unbounded real scenes. Results demonstrate our method’s superiority over existing state-of-the-art diffusion-based NVS approaches.

2. Related Work

Radiance fields novel view synthesis. Two standard techniques that revolutionized the field of novel view synthesis are NeRF Mildenhall et al. (2020) and 3D Gaussian Splatting Kerbl et al. (2023). NeRF utilizes an MLP to implicitly model the scene as a function and leverages volume rendering to generate novel views. Despite its high rendering quality, NeRF suffers from long training and inference times compared to 3DGS. In contrast to NeRF, 3DGS explicitly represents scenes as a set of Gaussian primitives, which are rendered to screen space through splatting-based rasterization. 3DGS offers significantly higher efficiency and competitive rendering quality compared to NeRF. However, all of these methods require high-quality, dense input views to optimize the model representation, introducing limitations in many situations. To address this, various regularization terms have been introduced to per-scene optimization Niemeyer et al. (2021); Li et al. (2024); Yu et al. (2022), while others focus on speeding up the optimization process or proposing effective scene representations Chen et al. (2022); Yu et al. (2021a,b). However, despite these improvements, these methods still lack generalization ability to unseen data.

Generalizable novel view synthesis. To avoid expensive per-scene optimization, feed-forward methods have been proposed to generate 3D representations directly from only a few input images Chen et al. (2024a); Charatan et al. (2023); Wewer et al. (2024). PixelSplat Charatan et al. (2023) predicts a dense probability distribution over 3D and generates Gaussian features from that probability distribution for scene rendering. LatentSplat Wewer et al. (2024) predicts semantic 3D Gaussians in latent space, which are decoded through a lightweight generative 2D architecture. MVSplat Chen et al. (2024a) introduces a cost volume as a geometric constraint to enhance multi-view feature extraction, while effectively capturing cross-view feature correlations for robust depth estimation. Splatt3r Smart et al. (2024) utilizes the foundation 3D geometry reconstruction method, MASt3R, to predict 3D Gaussian Splats without requiring any camera parameters or depth information. While these models generate photorealistic results for observed viewpoints,

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-269-2026 | © Author(s) 2026. CC BY 4.0 License.

(a) SD-VAE c=4, rFID=13.8 (b) BetterScene-VAE c=32, rFID=7.74 (c) BetterScene-VAE c=64, rFID=4.75 (d) GT

Figure 2. The visual quality and reconstruction FID score (rFID) for autoencoders with different channel sizes. We trained all the autoencoders on the DL3DV-10K Ling et al. (2024) dataset. Results show that the original 4-channel autoencoder design Rombach et al. (2021), which is widely used in diffusion models is unable to reconstruct fine details. Moreover, as shown in (b) and (c), increasing channel size leads to much better reconstructions. We choose to use a 64-channel BetterScene autoencoder for our video diffusion model.

their ability to reconstruct high-fidelity details in occluded or unobserved regions remains limited.

Novel view synthesis with diffusion priors. Recently, leveraging diffusion priors for aiding or enhancing novel view synthesis has proven to be an effective approach to improving rendering quality. By mitigating artifacts and hallucinating missing details, these methods significantly enhance the quality of synthesized views Chen et al. (2024b); Wang et al. (2025); Liu et al. (2024a,b); Wu et al. (2023). ReconFusion Wu et al. (2023) fine-tunes a diffusion model on a mixture of real-world and synthetic multi-view image datasets and employs it to regularize a standard NeRF reconstruction process in a manner akin to Score Distillation Sampling. VideoScene Wang et al. (2025) introduces a 3D-aware leapflow distillation strategy to bypass low-information diffusion steps. Their method enables single-step 3D scene generation. DIFIX3D+ Wu et al. (2025) also allows one-step scene generation with the benefit of a consistent generative model. Furthermore, it progressively refines the 3D representation by distilling back the enhanced views to achieve significant results. 3DGSEnhancer Liu et al. (2024b) employs video diffusion to restore view-consistent novel view renderings, then utilizes these refined views to optimize the initial 3DGS model. MVSplat360Chen et al. (2024b) leverages a feed-forward 3DGS model to directly generate coarse geometric features in the latent space of a pre-trained SVD model, enabling efficient synthesis of photorealistic, wide-sweeping novel views. While our approach builds upon MVSplat360’s pipeline, we introduce an innovative representation-aligned, equivariance-regularized high-dimensional latent feature representation instead of using an off-the-shelf pretrained SVD. Our experiments demonstrate superior fidelity and visual quality compared to baseline methods.

3. Methodology

3.1 BetterScene Overview

BetterScene consists of a feed-forward 3DGS reconstruction module, MVSplat Chen et al. (2024a), and a refinement module based on a stable video diffusion Blattmann et al. (2023b) backbone. Specifically, given N sparse-view inputs $\mathcal{I} = \{\boldsymbol{I}^i\}_{i=1}^N$, our goal is to synthesize realistic images from novel viewpoints in an end-to-end manner. The framework of our BetterScene is illustrated in Fig. 3. The training of our BetterScene consists of two stages. In the first stage, we train an autoencoder using a representation-aligned and equivariance-regularized objective function. In the second stage, we freeze the pretrained BetterScene-VAE and fine-tune the denoiser U-Net within the SVD framework. As shown in Fig. 3, we leverage a feed-forward 3DGS rendering module, MVSplat, to generate both coarse synthesized views and corresponding Gaussian feature latents $\hat{\boldsymbol{f}}_i$. The SVD module then processes these coarse features to decode enhanced high-quality images. Further details are discussed in subsequent sections.

3.2 Representation-aligned Equivariance-regularized VAE

In this section, we introduce the representation-aligned and equivariance-regularized variational autoencoder for achieving superior quality in both reconstruction and generation. This optimization improves both synthesis fidelity and visual quality of NVS by incorporating unconstrained high-dimensional latent representations into the SVD pipeline. Specifically, we scale the original SD-VAE architecture which uses 8× spatial downsampling and 4 latent channels, to 16× downsampling with 64 latent channels that maintain a comparable model scale. This modification triggers the aforementioned optimization dilemma: while reconstruction quality improves, generation performance degrades. This phenomenon likely stems from the Gaussian prior assumption in the VAE’s KL divergence loss. The objective function of the original VAE Kingma and Welling (2022) derived from maximizing the evidence lower bound (ELBO), can be expressed as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = &-D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) \\ &+\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right] \end{aligned} \quad (1)$$

where the posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})$ is constrained to match a standard Gaussian distribution. This inherently restricts latent embedding expressiveness, especially in high-dimensional spaces. Another observation is that increasing the latent dimension leads to underutilization of the feature space, which is also observed in autoregressive generation with codebook embeddings. Bo and Liu (2024); Zhu et al. (2024).

Representation alignment loss. To generate a high-dimensional latent space for enhanced novel view synthesis (NVS), we introduce a vision foundation model alignment loss Yu et al. (2025); Yao et al. (2025) to optimize the VAE component within the original SVD framework. The key idea involves constraining the latent space by leveraging the vision foundation model’s feature space. This enables a flexible feature distribution that improves feature utilization while escaping the limitations of the standard Gaussian distribution assumption.

Specifically, given an input image I, we process it through both our modified VAE with 64 latent channels and DINOv2 Oquab et al. (2023), a vision foundation model that extracts robust visual

features. The resulting image latents are denoted as Z_V and F_D. Z_V is projected to match the dimensionality of F_D through a linear transformation: $Z' = \mathrm{W}\, Z_V$. We employ the cosine similarity loss Yao et al. (2025) to minimize the discrepancy between corresponding feature representations with margin m_1.

$$\mathcal{L}_{\text{cos-align}} = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} \text{ReLU}\left(1 - m_1 - \frac{z'_{ij} \cdot f_{ij}}{\|z'_{ij}\| \|f_{ij}\|}\right) \tag{2}$$

Furthermore, a distance similarity loss is employed as a complementary objective to align the internal distributions of Z_V and F_D with margin m_2.

$$\mathcal{L}_{\text{dist-align}} = \frac{1}{N^2} \sum_{i,j} \text{ReLU}\left(\left|\frac{z_i \cdot z_j}{\|z_i\| \|z_j\|} - \frac{f_i \cdot f_j}{\|f_i\| \|f_j\|}\right| - m_2\right) \tag{3}$$

Equivariance regularization. Recent research reveals that the SD-VAE latent representations lack equivariance under spatial transformations. Specifically, given an input image I and its corresponding VAE latent $Z_V(I)$, if we apply a transformation τ to both I and $Z_V(I)$, the latent representation of the transformed image should satisfy Kouzelis et al. (2025):

$$\forall \mathbf{I} \in \mathcal{I}: \quad \mathcal{Z}(\tau \circ \mathbf{I}) = \tau \circ \mathcal{Z}(\mathbf{I}). \tag{4}$$

Violating this property implicitly leads to temporal inconsistency in video LDMs, as the noise patterns between frames lack transformation consistency. Consequently, the decoded images cannot form an equivariant frame sequence, resulting in sudden scene shifts or inconsistent content across consecutive frames. This creates fundamental limitations for using video LDMs to enhance novel view synthesis (NVS), which requires strict temporal consistency.

Therefore, we directly enforce latent equivariance by incorporating the constraint from (4) as a regularization term during autoencoder training with augmented transformations.

$$\mathcal{L}_{\text{latent-equivariance}}(\mathbf{I}) = \|\tau \circ \mathcal{Z}(\mathbf{I}) - \mathcal{Z}(\tau \circ \mathbf{I})\|_2^2, \tag{5}$$

where τ represents a set of spatial transformations. In addition to the latent equivariance loss, we employ a reconstruction equivariance loss to align the reconstructions of transformed latent features ($\mathcal{D}(\tau \circ \mathcal{Z}(\mathbf{I}))$) with the corresponding transformed inputs ($\tau \circ \mathbf{I}$). The reconstruction equivariance objective is as follows:

$$\begin{aligned}\mathcal{L}_{\text{recon-equivariance}}(\mathbf{I}, \tau) = \mathcal{L}_{rec}\Big(\tau \circ \mathbf{I}, \mathcal{D}\big(\tau \circ \mathcal{Z}(\mathbf{I})\big)\Big) + \\ \lambda_{gan}\mathcal{L}_{gan}\Big(\mathcal{D}\big(\tau \circ \mathcal{Z}(\mathbf{I})\big)\Big) + \lambda_{reg}\mathcal{L}_{reg}\end{aligned} \tag{6}$$

BetterScene-VAE. We train our autoencoder on the DL3DV-10K dataset with the objective function:

$$\begin{aligned}\mathcal{L}_{\text{BetterScene-VAE}} = w_{\text{align}} * (\mathcal{L}_{\text{dist-align}} + \mathcal{L}_{\text{cos-align}}) + \\ w_{\text{equi}} * (\mathcal{L}_{\text{latent-equivariance}} + \mathcal{L}_{\text{recon-equivariance}})\end{aligned} \tag{7}$$

By leveraging representation alignment and equivariance regularization, our autoencoder achieves both superior reconstruction fidelity and generation capability while enabling transformation equivariance in the latent space. By integrating this high-dimensional latent representation into SVD refinement modules, we produce enhanced visual fidelity and rendering quality for NVS.

3.3 BetterScene: Video LDM NVS Enhancer

Given coarse rendered novel views with artifacts $\tilde{\mathcal{I}}$, our model generates a sequence of cleaned predictions. We build our pipeline upon a feed-forward 3DGS reconstruction model, MVSplat, and a pretrained SVD backbone. We introduce the details of each module in the following parts.

Coarse feature generation. We bypass expensive per-scene optimization Barron et al. (2021a,b); Gao et al. (2024), and adopt MVSplat, a generalizable feed-forward 3DGS generation model capable of synthesizing novel views for unseen scenes from sparse-view inputs. Specifically, MVSplat first fuses multi-view information and obtains cross-view aware features $\mathcal{F} = \{\boldsymbol{F}^i\}_{i=1}^N$ given sparse-view observations $\mathcal{I} = \{\boldsymbol{I}^i\}_{i=1}^N$ and their corresponding camera poses $\mathcal{P} = \{\boldsymbol{P}^i\}_{i=1}^N$. Then, N cost volumes $\mathcal{C} = \{\boldsymbol{C}^i\}_{i=1}^N$ are constructed through cross-view feature correlation matching, enabling per-view depth estimation. Finally, we compute the Gaussian parameters: mean $\boldsymbol{\mu}$, covariance $\Sigma \in \mathbb{R}^{3\times3}$, and spherical harmonic coefficients $\boldsymbol{c} \in \mathbb{R}^{3(S+1)^2}$ where S is the order. The target view $\tilde{\mathcal{I}}$ can be rendered through rasterization.

Gaussian feature conditioning. In the original SVD framework, the first ground truth frame usually serves as the conditioning input, which is concatenated with Gaussian noise during the denoising process. In our framework, we leverage coarse rendered priors by directly concatenating rasterized features $\hat{\mathcal{F}}$ as conditioning with the latent space noise in SVD. Notably, the coarse Gaussian priors are directly combined with noise latents, bypassing the encoding step, similar to the method described in Chen et al. (2024b). The key advantage of this operation is its ability to leverage supervision from ground truth frame VAE embeddings, which simultaneously optimizes both the conditioning Gaussian features and the MVSplat modules. This is where our optimized VAE plays a critical role in the pipeline. The high-dimensional expressive latent representations of target frames provide strong supervision for the latent conditioning, thereby offering more effective guidance for the generation process.

Moreover, similar to the original SVD, we leverage CLIP Radford et al. (2021) to generate conditioning embeddings from the input views $\mathcal{I}$. These embeddings serve as global semantic cues that are injected into the denoising process via cross-attention operations, helping the model maintain both semantic coherence and fidelity.

Fine-tuning and losses. We fine-tune the denoiser U-Net of SVD while keeping our pretrained VAE encoder and decoder frozen. The model takes sparse context images as input and generates refined target images as output through our BetterScene framework. The entire system is trained end-to-end. We supervise our SVD model with: (1) the standard v-prediction formulation as the diffusion loss, and (2) a linear combination of ℓ_2 and LPIPS Zhang et al. (2018) discrepancies between the predicted outputs $\tilde{\mathcal{I}}^{\text{pred}}$ and the corresponding ground truth $\mathcal{I}^{\text{gt}}$ as the reconstruction loss. Additionally, as previously mentioned, we align the conditioning Gaussian features with the high-dimensional latent representations of target images encoded by our pretrained VAE with a latent feature loss $\min_{g_\theta} \mathbb{E}_{\hat{\mathcal{Z}} \sim g(\mathcal{I})} \|\mathcal{E}(\mathcal{I}^{\text{gt}}) - \hat{\mathcal{Z}}^{\text{gs}}\|_2^2$.

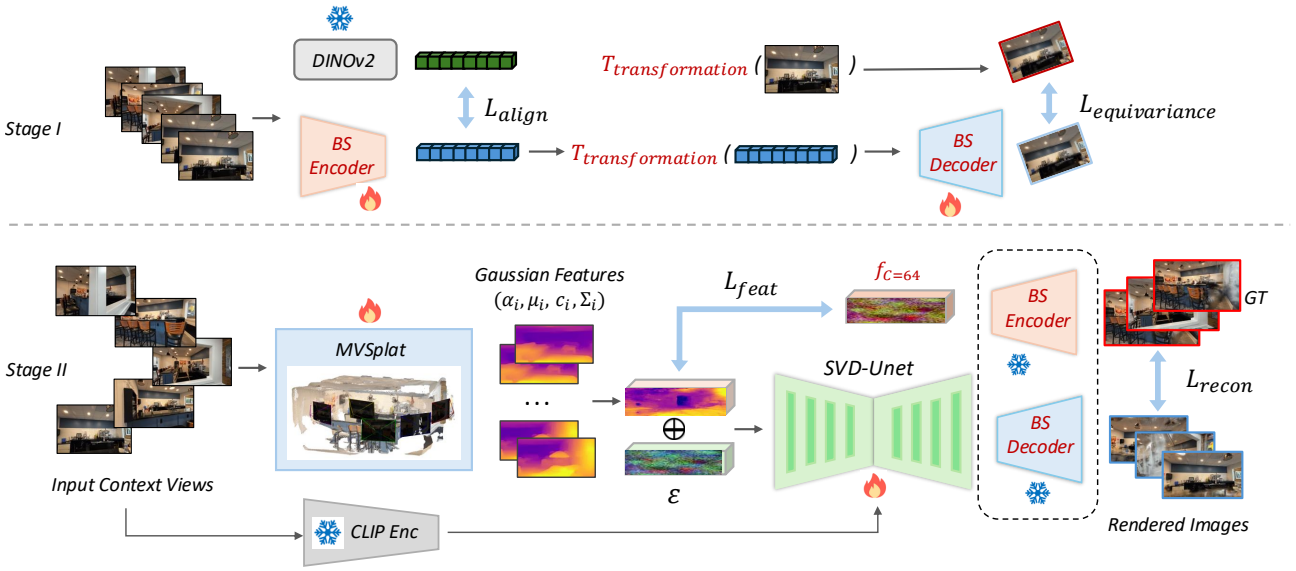

Figure 3. **Overview of our BetterScene**. The training process consists of two stages. In the first stage, we train an autoencoder using a representation-aligned and equivariance-regularized objective function. In the second stage, we freeze the pretrained BetterScene-VAE and fine-tune the denoiser U-Net within the SVD framework. We leverage a feed-forward 3DGS rendering module, MVSplat, to generate both coarse synthesized views and corresponding Gaussian feature latents. The SVD module then processes these coarse features to decode enhanced high-quality images.

Table 1. **A quantitative comparison of novel view synthesis performance using 5 input views**. Experiments on the DL3DV-10K dataset follow the setting of Chen et al. (2024b) (Note that since all experimental settings remain identical, we directly adopt the evaluation results for baseline methods reported in Chen et al. (2024b).)

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓
MVSplat Chen et al. (2024a)	17.05	0.499	0.435	61.92
latentSplat Wewer et al. (2024)	17.79	0.527	0.391	34.55
MVSplat360 Chen et al. (2024b)	**17.81**	0.562	0.352	18.89
BetterScene (ours)	**17.81**	**0.579**	**0.347**	**16.59**

4. Experiments

4.1 Implementation Details

To validate the efficacy of BetterScene, we conduct experiments on the challenging DL3DV-10K dataset, which contains 51.3 million frames from 10,510 real-world scenes. Our experiments follow the same benchmark settings as Chen et al. (2024b). The test partition contains 140 scenes and is filtered from the training set. We select 5 input views and evaluate 56 novel views, sampled uniformly from the remaining frames. We fine-tune the original SVD that generates 14 frames per sampling epoch. Our autoencoder architecture employs a 16× downsampling rate and a latent channel size of 64. The entire pipeline is trained on four NVIDIA H100 GPUs.

4.2 Comparison with State-of-the-Arts

The quantitative results with 5 input views on the DL3DV test set are shown in Table 1. Our approach achieves superior performance compared to all baseline methods in SSIM, LPIPS, and FID metrics, while maintaining PSNR scores comparable to MVSplat360.

The qualitative results on the DL3DV-10K benchmark are presented in Fig. 4. We compare our BetterScene with MVSplat Chen et al. (2024a) and its diffusion-enhanced variant, MVSplat360 Chen et al. (2024b). Without the diffusion-based refinement, MVSplat generates blurry novel views due to insufficient constraints from sparse input views. With the refinement from video diffusion, MVSplat360 demonstrates significant improvement, achieving remarkable visual quality through effective artifact removal. However, imperfections persist in both the reconstructed geometry and detail consistency. The first column in Fig. 4 demonstrates BetterScene’s capability for effectively removing artifacts. The second and third columns in Fig. 4 validate: (1) the efficacy of high-dimensional latent representations for improved recovery in images, such as text on the wall, and (2) the effectiveness of our representation-aligned, equivariance-regularized autoencoder design for maintaining detail consistency. Overall, our approach outperforms all baseline methods in both visual quality and detail consistency, demonstrating the capability to synthesize high-fidelity novel views.

4.3 Ablations Study

The core innovation of our pipeline is the high-dimensional latent feature representation. In this section, we present an ablation study exploring the impact of latent channel size on our BetterScene-VAE performance. Due to the prohibitive computational cost of training the complete BetterScene pipeline with SVD on the full DL3DV-10K dataset, we focus our evaluation on the reconstruction performance of BetterScene-VAE across three latent channel configurations: 16, 32, and 64 dimensions.

As demonstrated in Table 2, increasing the latent dimensionality yields significant improvements in reconstruction quality. The 64-channel configuration achieves superior performance across all metrics compared to lower-dimensional latent representations. Notably, higher-dimensional representations consistently produce superior reconstructions, with the 64-channel configuration achieving particularly robust detail consistency. These results suggest potential reasons for BetterScene’s superior performance in high-frequency detail and complex texture enhancement compared to existing approaches.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-269-2026 | © Author(s) 2026. CC BY 4.0 License.

Figure 4. A visual comparison of enhanced rendering results generated from 5 input views across scenes from the DL3DV benchmark test set. BetterScene demonstrates superior visual quality and enhanced detail consistency compared to existing state-of-the-art approaches.

Table 2. A quantitative comparison of reconstruction performance across latent channel sizes. The SD-VAE represents the original VAE architecture with 4 latent channels.

Method	PSNR↑	SSIM↑	LPIPS↓	rFID↓
SD-VAE	26.06	0.78	0.12	13.83
BetterScene-VAE(C=16)	25.14	0.758	0.12	13.41
BetterScene-VAE(C=32)	28.21	0.851	0.08	7.41
BetterScene-VAE(C=64)	**31.21**	**0.92**	**0.04**	**4.90**

5. Conclusion

We present BetterScene, an approach for enhancing novel view synthesis (NVS) quality from sparse and unconstrained photo collections. Unlike contemporary methods, we investigate the diffusion model’s latent space and introduce (1) equivariance regularization and (2) vision foundation model-aligned representations, both applied to the variational autoencoder (VAE) within the SVD pipeline. Our framework enhances NVS quality and generates artifact-free, temporally consistent novel views. We evaluated BetterScene on the challenging DL3DV-10K benchmark. Our method demonstrates significant visual quality improvements over baseline approaches. Our work may contribute insights for advancing 3D reconstruction and view generation in future research. However, the SVD model in the BetterScene framework requires computationally expensive training. Future work could explore replacing this pipeline with more efficient video diffusion architectures.

Acknowledgments

This work was supported in part by the U.S. Army Research Office under Grant AWD-110906.

References

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P. P., 2021a. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5835-5844. https://api.semanticscholar.org/CorpusID:232352655.

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., Hedman, P., 2021b. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5460-5469. https://api.semanticscholar.org/CorpusID:244488448.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R., 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., Kreis, K., 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bo, C., Liu, J., 2024. Enhancing Codebook Utilization in VQ Models via E-Greedy Strategy and Balance Loss. *2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 1-5. https://api.semanticscholar.org/CorpusID:276452211.

Charatan, D., Li, S. L., Tagliasacchi, A., Sitzmann, V., 2023. PixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19457-19467. https://api.semanticscholar.org/CorpusID:266362208.

Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H., 2022. TensoRF: Tensorial Radiance Fields. *ArXiv*, abs/2203.09517. https://api.semanticscholar.org/CorpusID:247519170.

Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.-J., Cai, J., 2024a. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *European Conference on Computer Vision*.

Chen, Y., Zheng, C., Xu, H., Zhuang, B., Vedaldi, A., Cham, T.-J., Cai, J., 2024b. MVSplat360: Feed-Forward 360 Scene Synthesis from Sparse Views. *ArXiv*, abs/2411.04924. https://api.semanticscholar.org/CorpusID:273877466.

Dai, X., Hou, J., Ma, C.-Y., Tsai, S. S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., Yu, M., Kadian, A., Radenovic, F., Mahajan, D. K., Li, K., Zhao, Y., Petrovic, V., Singh, M. K., Motwani, S., Wen, Y., Song, Y.-Z., Sumbaly, R., Ramanathan, V., He, Z., Vajda, P., Parikh, D., 2023. Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack. *ArXiv*, abs/2309.15807. https://api.semanticscholar.org/CorpusID:263151865.

Deng, K., Liu, A., Zhu, J.-Y., Ramanan, D., 2021. Depth-supervised NeRF: Fewer Views and Faster Training for Free. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12872-12881. https://api.semanticscholar.org/CorpusID:235743051.

Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P. P., Barron, J. T., Poole, B., 2024. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. *ArXiv*, abs/2405.10314. https://api.semanticscholar.org/CorpusID:269791465.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (TOG)*.

Kingma, D. P., Welling, M., 2022. Auto-encoding variational bayes.

Kouzelis, T., Kakogeorgiou, I., Gidaris, S., Komodakis, N., 2025. EQ-VAE: Equivariance Regularized Latent Space for Improved Generative Image Modeling. *ArXiv*, abs/2502.09509. https://api.semanticscholar.org/CorpusID:276317789.

Kwak, M., Song, J., Kim, S. W., 2023. GeCoNeRF: Few-shot Neural Radiance Fields via Geometric Consistency. *ArXiv*, abs/2301.10941. https://api.semanticscholar.org/CorpusID:256274740.

Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L., 2024. DNGaussian: Optimizing Sparse-View 3D Gaussian Radiance Fields with Global-Local Depth Normalization. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20775-20785. https://api.semanticscholar.org/CorpusID:268363574.

Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y. et al., 2024. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22160–22169.

Liu, F., Sun, W., Wang, H., Wang, Y., Sun, H., Ye, J., Zhang, J., Duan, Y., 2024a. ReconX: Reconstruct Any Scene from Sparse Views with Video Diffusion Model. *ArXiv*, abs/2408.16767. https://api.semanticscholar.org/CorpusID:272146325.

Liu, X., Zhou, C., Huang, S., 2024b. 3DGS-Enhancer: Enhancing Unbounded 3D Gaussian Splatting with View-consistent 2D Diffusion Priors. *ArXiv*, abs/2410.16266. https://api.semanticscholar.org/CorpusID:273508035.

Luo, Y., Zhou, S., Lan, Y., Pan, X., Loy, C. C., 2024. 3DEnhancer: Consistent Multi-View Diffusion for 3D Enhancement. *ArXiv*, abs/2412.18565. https://api.semanticscholar.org/CorpusID:274992751.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Niemeyer, M., Barron, J. T., Mildenhall, B., Sajjadi, M. S. M., Geiger, A., Radwan, N., 2021. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5470-5480. https://api.semanticscholar.org/CorpusID:244773517.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. Q., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y. B., Li, S.-W., Misra, I., Rabbat, M. G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. DINOv2: Learning Robust Visual Features without Supervision. *ArXiv*, abs/2304.07193. https://api.semanticscholar.org/CorpusID:258170077.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*.

Roessle, B., Barron, J. T., Mildenhall, B., Srinivasan, P. P., Nießner, M., 2021. Dense Depth Priors for Neural Radiance Fields from Sparse Input Views. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12882-12891. https://api.semanticscholar.org/CorpusID:244921004.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674-10685. https://api.semanticscholar.org/CorpusID:245335280.

Smart, B., Zheng, C., Laina, I., Prisacariu, V., 2024. Splatt3R: Zero-shot Gaussian Splatting from Uncalibrated Image Pairs. *ArXiv*, abs/2408.13912. https://api.semanticscholar.org/CorpusID:271957263.

Somraj, N., Soundararajan, R., 2023. ViP-NeRF: Visibility Prior for Sparse Input Neural Radiance Fields. *ACM SIGGRAPH 2023 Conference Proceedings*. https://api.semanticscholar.org/CorpusID:258426778.

Wang, G., Chen, Z., Loy, C. C., Liu, Z., 2023. SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9031-9042. https://api.semanticscholar.org/CorpusID:257771582.

Wang, H., Liu, F., Chi, J., Duan, Y., 2025. Videoscene: Distilling video diffusion model to generate 3d scenes in one step.

Wewer, C., Raj, K., Ilg, E., Schiele, B., Lenssen, J. E., 2024. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *European Conference on Computer Vision*.

Wu, J. Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M. Z., Fidler, S., Gojcic, Z., Ling, H., 2025. Difix3d+: Improving 3d reconstructions with single-step diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P. P., Verbin, D., Barron, J. T., Poole, B., Holynski, A., 2023. ReconFusion: 3D Reconstruction with Diffusion Priors. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21551-21561. https://api.semanticscholar.org/CorpusID:265659460.

Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., Han, S., 2024. SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers. *ArXiv*, abs/2410.10629. https://api.semanticscholar.org/CorpusID:273346094.

Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M., 2025. Depthsplat: Connecting gaussian splatting and depth. *CVPR*.

Yao, J., Yang, B., Wang, X., 2025. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., Kanazawa, A., 2021a. Plenoxels: Radiance Fields without Neural Networks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5491-5500. https://api.semanticscholar.org/CorpusID:245006364.

Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A., 2021b. PlenOctrees for Real-time Rendering of Neural Radiance Fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5732-5741. https://api.semanticscholar.org/CorpusID:232352425.

Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S., 2025. Representation alignment for generation: Training diffusion transformers is easier than you think. *International Conference on Learning Representations*.

Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A., 2022. MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. *ArXiv*, abs/2206.00665. https://api.semanticscholar.org/CorpusID:249240205.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586-595. https://api.semanticscholar.org/CorpusID:4766599.

Zhou, Y., Xiao, Z., Yang, S., Pan, X., 2025. Alias-free latent diffusion models: Improving fractional shift equivariance of diffusion latent space. *CVPR*.

Zhu, L., Wei, F., Lu, Y., Chen, D., 2024. Scaling the Codebook Size of VQGAN to 100,000 with a Utilization Rate of 99%. *ArXiv*, abs/2406.11837. https://api.semanticscholar.org/CorpusID:270560634.