Zero Shot Novel View Synthesis in the Wild

Tim Tomov
Technical University of Munich
tim.tomov@tum.de

Michael Neumayr
Technical University of Munich
michael.neumayr@tum.de

Abstract

Recent advancements in novel view synthesis (NVS) have shown remarkable success in generating images from novel viewpoints, particularly for objects in a centered position. However, the application of such techniques to scene-level datasets introduces exponential growth in the degrees of freedom as it encompasses a more complex combination of objects. Thus posing unique challenges to the novel view synthesis task.
We build on methods from 3D content creation that show promising novel view synthesis results on a single image and extend them to a complex scene dataset. We combine and adapt semantic global conditioning modules of the object-centric approaches with a geometric module that can provide pixel-level partial depth or RGB image conditioning for the generation process.
Qualitative as well as quantitative results show superior performance compared to our baseline Zero123 and confirm our intuition that more rigorous geometric conditioning is needed in a scene-setting compared to the object-centric setting.
The code is available at https://github.com/michaelnoi/scene_nvs.

1. Introduction

The domain of novel view synthesis (NVS) has been a focal point of recent research endeavors within the field of computer vision and graphics, primarily due to its vast potential in applications ranging from virtual reality (VR) to autonomous navigation. NVS techniques are designed to generate images of a scene or object from viewpoints different from those of the original images.
More specifically, given a single RGBD image of an unseen scene and a target pose, the task is to generate a realistic image at the target pose. We want to generalize to previously unseen scenes in a zero-shot fashion and constrain the generation process given only a single image of the scene.
The advent of highly available, powerful text-to-image models [12] opened exciting research directions leveraging their generative capabilities and priors about the world to tackle such an underconstrained task. NVS techniques building on top of generative models like Stable Diffusion have shown remarkable success when applied to 3D objects positioned centrally within a scene [1, 5]. This success, however, is often confined to scenarios involving singular, well-defined objects in a controlled environment of rotating poses in a half-sphere around the object.
Unique challenges arise when applying NVS to scene-level datasets. Unlike object-centric NVS, scene-level synthesis must account for an exponentially greater number of combinations of real-world objects for multi-view consistent NVS. Recognizing the limitations of existing NVS approaches in dealing with scene-level complexity, our research seeks to bridge this gap. As a baseline, we build on 3D content creation methods Zero-1-to-3 and its Zero123++ [13] evolution. We extend those to tackle the complexities of scene-level datasets, aiming to synthesize multi-view consistent views.
We propose a framework that synergizes the strengths of object-centric approaches with an additional geometry module to locally enforce correct geometric constraints from the source image. Like the recent work, GeNVS [1], we aim to bridge the gap by re-projecting meaningful information to the target pose for conditioning the generation process. Our re-projection includes a depth map or colored image, and we observe that it is a crucial local conditioning signal that considerably improves our results in the experiments. This coincides with the intuition that even partial pixel-wise conditioning of our generative model helps constrain the generation to not violate basic scene geometry, which is the case when trying to extend the vanilla Zero-1-to-3 approach to the more challenging scene setting.
At the core of our methodology is the integration of a geometric conditioning module, which has been pivotal in object-centric NVS, with our proposed re-projection module. Our contributions include:

1. We extend an object-centric approach to a complex real-world scene dataset
2. Add re-projected depth map and RGB image conditioning with geometry module

2. Related Work

Novel view synthesis (NVS) research intersects with 3D reconstruction, generative models, and deep learning, aiming to generate photorealistic images from sparse viewpoints. Regression-based deep learning methods utilize 3D scene representations and differentiable neural rendering to predict unseen views. These approaches typically require optimization for individual scenes or adapt to new scenes in a few-shot learning manner, facilitating the generation of new views with limited input images [6, 17].

In tackling zero-shot NVS, generative models emerge as a novel frontier, with recent efforts including the optimization of NeRF using 2D diffusion models as priors [7, 14] and the fine-tuning of diffusion models with various constraints for multi-view consistency [1, 5, 13].

While GeNVS leverages a re-projected feature from a feature volume in conditioning the diffusion model, Zero-1-to-3 solely conditions the generative process on the input view and the relative pose information. Our work builds on the Zero-1-to-3 family, adapting their fine-tuning setup and introducing more direct re-projected geometric constraints than the GeNVS setup to successfully extend the object-centric model to scenes.

3. Method

To create a realistic novel view in the scene setting, our model requires both additional geometric constraints and semantic information. We achieve this by introducing corresponding modules (3.2, 3.3) and also incorporating the relative pose into the semantic encoding (3.1).

3.1. Pose Module

In object-centric frameworks, pose variance is often represented through differences in polar coordinates. This method is particularly effective when an object is centrally positioned within a spherical domain, with pose variations corresponding to different locations on the sphere's surface. However, this polar coordinate approach proves less effective in scene setups. To address this limitation, our architecture adopts a quaternion-based approach for encoding rotational differences. Quaternions offer a more robust and efficient method for representing 3D rotations. Alongside this, we capture translational differences using straightforward translation vectors.

The resulting combined representation is a concatenation of quaternion and translation vector components, yielding a seven-dimensional vector:

$$c(R, T) \in \mathbb{R}^7$$

.

Furthermore, to enhance the model's ability to utilize this pose information effectively, we incorporate a sinusoidal encoding scheme. This approach, also employed in Neural Radiance Fields (NeRF) [6], facilitates a more detailed interpretation of pose data.

3.2. Geometric Module

Understanding the geometric structure is vital for our model to grasp the scene's spatial relations. Utilizing monocular inverse depth maps can provide geometric information by aiding the model in understanding the proximity of objects in the scene [10].

Our approach involves back-projection of an image with the given depth map into a point cloud. Formally, we convert image pixel coordinates (u, v) and depth values Z into 3D space coordinates (X_c, Y_c, Z_c) using the intrinsic camera matrix $\mathbf{K}$, and subsequently into world coordinates (X_w, Y_w, Z_w) using the source pose. This process is described by:

$$\begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = \mathbf{T}_{cw} \cdot Z \cdot \mathbf{K}^{-1} \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix},$$

where $\mathbf{T}_{cw} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix}$ combines the rotation and translation into a single operation.

We can then rasterize the point cloud to obtain a depth map or render an RGB image leveraging the PyTorch3D library [11, 15], given there exists an overlap between the source and target pose, which we made sure of in the image pair selection. Through this process, we acquire partial depth information that serves as a crucial constraint for the generation process. Specifically, the depth map is integrated into the network by first downsampling it to match the latent resolution. Subsequently, it is concatenated along the channel with the noised latent representation as illustrated in Figure 1.

When using RGB image conditioning, the model gains access to high-frequency details and local structural information. Such details, often elusive in semantic representations, are essential for a comprehensive scene understanding. In this case, the conditioning image is input to the original autoencoder as the RGB image mostly matches the image distribution of the training images and is then again concatenated to the latent vector in the generation process. It is worth noting that this RGB-based approach introduces similarities with the task of outpainting.

To maintain training stability, the channels newly introduced for this depth information are initialized with zero weights following [18]. This allows the model to gradually adjust and incorporate the geometric data effectively.

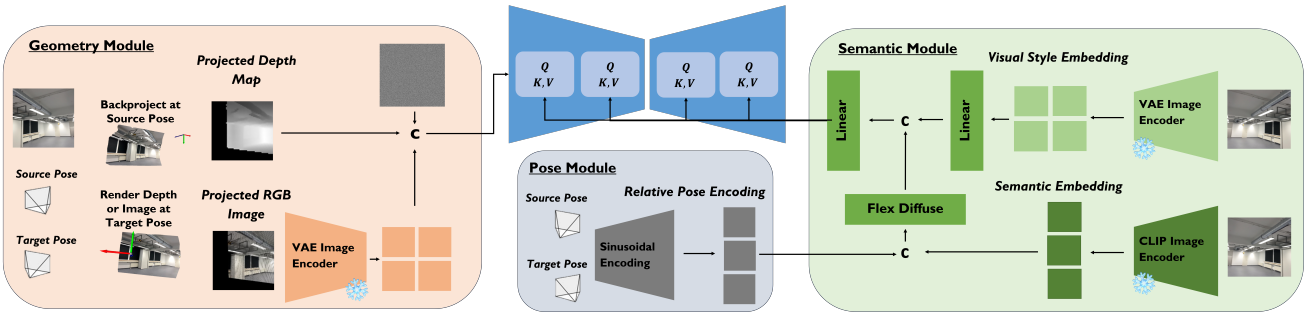

Figure 1. Overview of Architecture (1) *Geometry Module:* Using the source image and depth map as well as the camera poses, the content is first projected into a 3D point cloud of the scene. Afterward, we render the depth map / RGB image in the target pose. While the depth map is downsampled to the latent size, the RGB image is encoded using the variational autoencoder. (2) *Semantic Module:* (3) *Pose Module:* The relative poses are obtained using quaternions and translation vectors. Afterward, a sinusoidal embedding is used to project the poses in a higher dimensional space for better usage. During training, the variational autoencoder and CLIP encoder are frozen, while the cross-attention of the U-Net is trained using LoRA. All other proposed Layers are fully trained.

3.3. Semantic Module

By semantic conditioning, we refer to conditioning the diffusion model on ways that associate semantic information with the visual world. Stable Diffusion is pretrained to incorporate CLIP [8] text embeddings as a conditioning signal. To effectively use these pre-learned representations, we incorporate CLIP image embeddings, which we denote as *Semantic Embedding*. However, these embeddings are not directly useable. Therefore, we implement a linear flex diffuse layer similar to that in Zero123++. This layer serves as a mapping, converting the CLIP image domain representations into the CLIP text representation space.

Parallelly, inspired by [4], we have explored incorporating an additional embedding derived from the variational autoencoder of the conditioning image. This strategy aims to augment the representational richness, and we term this embedding as *Visual Style Embedding*.

Finally, we combined both signals through a linear mapping. The weights assigned to the visual style embedding are initially set to zero, following the same rationale as previously discussed.

4. Experiments

For our experiments, we used the ScanNet++ Dataset, which offers high-fidelity views of indoor scenes [16]. Our training process includes 218 distinct indoor scenes from this dataset. To evaluate the methods' generalizability, we tested them on 27 unseen scenes not included in the training set. To ensure moderate pose changes, we only selected image pairs under a certain threshold, which was defined as a combination of rotation and translation distance.

4.1. Implementation Details

As a diffusion model, we use Stable Diffusion 2 with a v-prediction target. Based on the findings of Zero123++, we adopt a linear noise schedule to ensure enough steps with low Signal-to-Noise Ratio (SNR). To efficiently train the model, we use LoRA [3] on the cross-attention layers while keeping the rest of the U-Net frozen. The variational autoencoder and CLIP encoder were frozen as well. To further make the training process more efficient, we also utilize the Min-SNR weighting strategy [2]. All training was done with a batch size of 8 and 4 gradient accumulation steps on 2 RTX A4000 using DeepSpeed Zero Stage 2 [9].

4.2. Baseline

As a baseline, we compare our approach to Zero-1-to-3. Our method demonstrates a significant improvement in accurately capturing the correct geometry and rotation of scenes compared to Zero-1-to-3, as can be seen in Figure 2. We attribute this enhanced performance to the main fact that the geometric information available in Zero-1-to-3 is misaligned with the target since the raw conditioning image is used. In contrast, our approach ensures a more accurate alignment with the target scene, thereby enhancing the geometrical adherence of the generated views to the source view.

Methods	**LPIPS** ↓	**SSIM** ↑
Baseline (Zero-1-to-3)	0.2539	0.1735
Ours (Depth)	0.2012	0.2759
Ours (Depth + RGB)	**0.1624**	**0.4297**

Table 1. **Comparison to Baseline (Zero123)**

4.3. Ablations

To evaluate the geometric and semantic module of our model, we conducted a series of ablation studies. It is im-

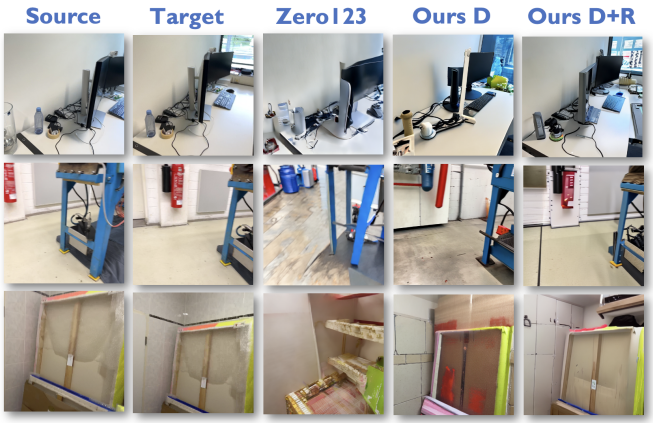

Figure 2. **Qualitative comparison to Baseline**. (D) Only Depth, (D+R) Depth + RGB. While Zero-1-to-3 understands the semantics of the scene, geometric information is not accurate. Our geometric module helps to solve this problem, resulting in geometrically consistent views.

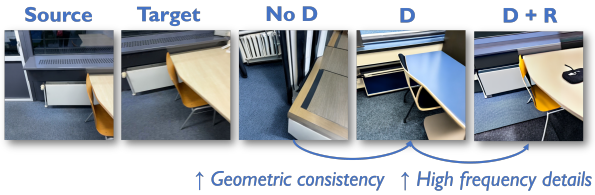

Figure 3. **Qualitative ablations on the geometry module**: (D) Only Depth Depth, (D+R) Depth + RGB). Information from the geometry module is critical to ensure geometric consistency. The use of RGB information enables to model to also capture correctly high-frequency details like patterns and color of individual small-scale objects

portant to note that due to computational constraints, these studies did not extend to full model convergence. Nonetheless, we posit that the trends observed are indicative and could be extrapolated to more extended training scenarios.

Depth	RGB	Visual Style	LPIPS ↓	SSIM ↑
		✓	0.2632	0.0811
✓		✓	0.2288	0.2036
✓			0.2245	0.2064
	✓	✓	0.1736	0.3984
✓	✓	✓	**0.1624**	**0.4297**

Table 2. **Ablations on geometric and semantic information**. Best performance is achieved using all modules.

For the geometric module, a key finding from our ablation tests is the critical role of local information in accurately positioning objects within a scene. The absence of depth or local RGB data notably hinders the model's ability to correctly localize objects, as can be seen in Figure 3. However, when provided with this information, our model adheres to the correct geometry in the generation process. Another observation is that the lack of color information significantly impairs the model's capacity to infer high-frequency details. This limitation likely arises from the inadequacy of the cross-attention conveying all the necessary visual information. As can be seen in Table 3, adding local RGB information significantly increases the correctness of high-frequency details when evaluated in terms of SSIM. In contrast, for the semantic module, we find that the added visual style embedding does not contribute to increased performance (Table 2).

5. Discussion

Our research introduces a method that extends object-centric approaches to scene data. By integrating local depth information, our method not only captures the overall semantic context but also ensures precise localization of objects within scenes. Additionally, we increased the accuracy of capturing high-frequency details by incorporating local RGB data.

In our study, we've started with the use of given low-resolution depth maps for novel view synthesis. A key area for future development involves leveraging methods to generate these depth maps and solving the challenge of accurately scaling them for projection. This advancement could make the approach less dependent on externally provided depth information and enable multiple new directions of research. Generating depth maps internally would e.g., enable the exploration of autoregressive techniques for producing consistent multi-view imagery across wider ranges of camera movement.

References

[1] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023. 1, 2

[2] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7441–7451, 2023. 3

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3

[4] Ting-Chun Wang Ira Kemelmacher-Shlizerman Johanna Karras, Aleksander Holynski. Dreampose: Fashion image-to-video synthesis via stable diffusion. 2023. 3

[5] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 1, 2

[6] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[7] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[9] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 3

[10] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 2

[11] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 2

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[13] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1, 2

[14] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2

[15] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 2

[16] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 3

[17] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2

[18] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2