The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W9-2025
ISPRS Intl. Workshop “Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine”
PSBB25 , 9–11 June 2025, Moscow, Russia

Generative 3D Inpainting of Scene Using Diffusion Model

Vladimir V. Kniaz[1,2], Tatyana N. Skrypitsyna[3], Petr V. Moshkantsev[1], Vladislav D. Pashtanov[1],
Artyom N. Bordodymov[1], Maxim A. Karpov[1]

[1] Moscow Institute of Physics and Technology (MIPT), Moscow, Russia - (kniaz.vv)@mipt.ru
[2] State Research Institute of Aviation Systems (GosNIIAS), Moscow, Russia -
(vl.kniaz,petr-mosh,pashtanov,bordodymov, karpov-ma)@gosniias.ru
[3] Moscow State University of Geodesy and Cartography (MIIGAiK), Moscow, Russia - fot@miigaik.ru

Keywords: cultural heritage, neural networks, diffuse network model, generative adversarial networks.

Abstract

The reconstruction of partially destroyed buildings is a critical aspect of architectural preservation, disaster recovery, and urban planning. Accurate reconstruction not only aids in restoring the historical and cultural significance of structures but also plays a vital role in ensuring safety and functionality in urban environments. Traditional methods often struggle with incomplete data, necessitating the exploration of advanced techniques that can improve reconstruction accuracy and efficiency. In this paper, we propose a novel approach for 3D inpainting of partially destroyed models using a diffusion neural network, termed `Restore3D`. This method leverages the principles of diffusion processes to iteratively refine and reconstruct missing sections of 3D wireframe structures. By integrating temporal features into the inpainting process, `Restore3D` effectively captures the intricate details and spatial relationships within the models, providing a more holistic reconstruction compared to conventional techniques. Our experimental results demonstrate that `Restore3D` not only competes with but also outperforms modern baselines in 3D model inpainting tasks. The evaluation metrics indicate significant improvements in reconstruction fidelity and detail preservation, showcasing the potential of our approach in practical applications. The results highlight the effectiveness of leveraging deep learning techniques for complex reconstruction challenges. In conclusion, this study presents `Restore3D` as a promising method for the reconstruction of partially destroyed 3D models. The encouraging results underscore the potential of deep learning in enhancing reconstruction accuracy, paving the way for future research and application in architectural restoration and urban planning.

1. Introduction

The reconstruction of partially destroyed objects of cultural heritage is a crucial endeavor in preserving the historical and cultural narratives embedded in these sites. Such efforts are particularly vital for ancient cities in the Middle East and abandoned churches in central Russia, where the ravages of time and conflict have left significant portions of these invaluable treasures in decay. The task of reconstructing these culturally significant structures involves filling in the gaps that time or destruction has created, allowing us to appreciate and study these heritage sites in their intended grandeur.

Traditionally, 3D object completion methods have been employed to undertake this monumental task. These methods range from manual reconstructions based on historical records and photographs, to more sophisticated computational approaches using 3D scanning and modeling technologies. Recent advancements have introduced machine learning techniques, such as deep neural networks, which have shown promise in automating parts of the reconstruction process by predicting missing geometries and textures. However, while these methods have improved the efficiency and accuracy of 3D reconstructions, they often fall short in generating the high-fidelity results needed to restore cultural heritage artifacts that possess intricate details and unique characteristics.

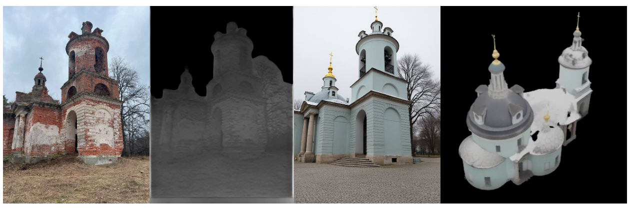

Figure 1. Our `Restore3D` framework is focused on the prediction of the 3D model of partially destroyed cultural heritage.

Recent advancements in neural methods for 3D reconstruction have significantly pushed the boundaries of what is achievable in the field of cultural heritage restoration. Among the state-of-the-art techniques, RenderDiffusion (Anciukevicius et al., 2022a) and NeRFiller (Weber et al., 2024) have emerged as prominent approaches. RenderDiffusion leverages image diffusion techniques to enhance the quality and resolution of 3D reconstructions, inpainting missing parts of models with remarkable accuracy. Similarly, NeRFiller employs generative 3D inpainting to complete scenes, utilizing neural radiance fields to synthesize missing geometric details and textures. These methods have demonstrated impressive capabilities in completing and refining existing 3D models, contributing substantially to the automation of restoration processes.

Despite these advancements, a critical challenge remains in the realm of cultural heritage preservation: generating a complete 3D model of a building from a single photograph of its partially destroyed state. Existing methods like RenderDiffusion and NeRFiller (Weber et al., 2024) primarily focus on the completion of pre-existing 3D models and do not address the unique challenge of reconstructing a full model from a limited, singular viewpoint. This gap in capability requires novel approaches that can infer complex structures and details from minimal input data, thereby enabling the reconstruction of heritage sites that

Publisher's note: Copernicus Publications has not received any payments from Russian or Belarusian institutions for this paper.
This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-151-2025

lack extensive photographic documentation. The development of such technologies would represent a significant leap forward, facilitating the safeguarding of historical sites with limited existing records.

The primary objective of this paper is to develop a neural model capable of reconstructing the original appearance and complete 3D model of partially destroyed objects of cultural heritage using a single photograph. This ambitious goal addresses the critical gap in current methodologies, which lack the ability to generate detailed and accurate 3D reconstructions from minimal input data. By harnessing the power of diffusion models, our approach aims to infer and restore the intricate geometries and textures that define these cultural artifacts, ensuring that their historical and aesthetic values are preserved for future generations. Through this research, we aim to provide a transformative tool that not only aids in the preservation of cultural heritage but also expands the capabilities of current generative methods in 3D reconstruction.

In this paper, we present significant contributions to the field of 3D inpainting through the development of novel methodologies and resources. First, we introduce the `Restore3D` model, a pioneering approach specifically designed for single-image 3D object reconstruction and the inpainting of missing parts. This model leverages the power of diffusion processes to generate detailed and accurate 3D representations from limited visual input, addressing the complex challenge of reconstructing objects from incomplete data. Second, we provide a new Destroyed2Restored dataset, which serves as a rich resource for research by offering an unpaired collection of images capturing objects of cultural heritage that are labeled as either partially destroyed or restored. This dataset is instrumental in understanding the nuances of cultural heritage preservation and provides a diverse range of examples for testing and validation. Lastly, we conduct a comprehensive evaluation of our `Restore3D` model and compare it with existing baselines using the Destroyed2Restored dataset, demonstrating the effectiveness and robustness of our approach in achieving high-quality 3D reconstructions and inpainting results. Through these contributions, we aim to advance the capabilities of generative modeling for 3D scenes, particularly in the context of cultural heritage objects.

The aim of this work is to develop a novel neural model, `Restore3D`, specifically designed to enable the rapid reconstruction of destroyed buildings from a single image. This approach addresses the urgent need for efficient restoration techniques in cultural preservation and disaster recovery scenarios, where detailed architectural data is often sparse or nonexistent. By leveraging the capabilities of diffusion models, our method seeks to reconstruct detailed 3D models of buildings using minimal visual input, allowing for precise inpainting of missing or damaged portions. Through this work, we strive to provide a powerful toolset for architects, historians, and preservationists, facilitating the digital restoration of invaluable cultural heritage and enabling the visualization of structures that might otherwise remain lost to time.

In this paper, we introduce the `Restore3D` model, an innovative approach for 3D inpainting and reconstruction of cultural heritage objects from a single image. Our model builds upon the foundation of the Stable diffusion model and integrates with the SSZ model, an Image-to-Voxel Model Translation framework designed for 3D scene reconstruction and segmentation. The `Restore3D` model works by processing an image of a partially destroyed building, where we identify and mask the destroyed regions, similar to the SmartBrush model, and fill these areas with Gaussian noise. The core strength of our model lies in providing a 'restored' prompt to the diffusion model, which facilitates the reverse diffusion process, effectively recovering and reconstructing the masked regions with high fidelity. Subsequently, the inpainted 2D image serves as an input to the SSZ model, which constructs a detailed 3D voxel model of the object, capturing its intricate structure and historical essence. Through this integrated framework, presented in Figure 1, we offer a powerful method for digitizing and preserving cultural heritage, enabling a seamless transition from 2D inpainting to comprehensive 3D reconstruction.

The main contributions of the study are the following:

- the `Restore3D` model, a pioneering approach specifically designed for single-image 3D object reconstruction and the inpainting of missing parts
- new *Destroyed2Restored* dataset, which serves as a rich resource for research by offering an unpaired collection of images capturing objects of cultural heritage that are labeled as either partially destroyed or restored
- evaluation of the `Restore3D` model and baselines on Destroyed2Restored dataset

The experimental evaluation of our `Restore3D` model reveals promising results, showcasing its effectiveness in the realm of 3D model inpainting when benchmarked against two modern state-of-the-art models. Our approach demonstrates competitive performance, particularly excelling in scenarios where reconstructing missing sections of buildings is crucial. Unique to the `Restore3D` model is its ability to generate style-consistent reconstructions, seamlessly integrating the restored parts with the original architectural aesthetics. This capability is pivotal in preserving the integrity and cultural significance of heritage structures. The findings underscore the potential of our model as a valuable tool for cultural preservation initiatives, highlighting its superior ability to maintain stylistic coherence in reconstructed imagery, a feat unmatched by existing methodologies.

The potential future implications of our `Restore3D` model are significant, particularly in the preservation and restoration of cultural heritage. The model's ability to rapidly reconstruct objects from a single image makes it invaluable during archaeological expeditions, where time and resources are often limited. By enabling quick and accurate 3D reconstructions, our model can assist researchers in documenting and analyzing artifacts in situ before they degrade further. Additionally, the `Restore3D` model's capacity to generate comprehensive 3D reconstructions from a limited number of archived photos opens new avenues for revitalizing historical records and reconstructing lost or damaged heritage sites. This capability not only aids in academic research but also enhances public engagement and education by providing immersive and interactive representations of cultural heritage, allowing broader access to our shared history. Through these applications, the `Restore3D` model stands to make a lasting impact in fields such as archaeology, history, and digital preservation.

2. Related Work

Our model uses two kind of neural models: 2D Inpainting models and single image 3D reconstruction models. The following

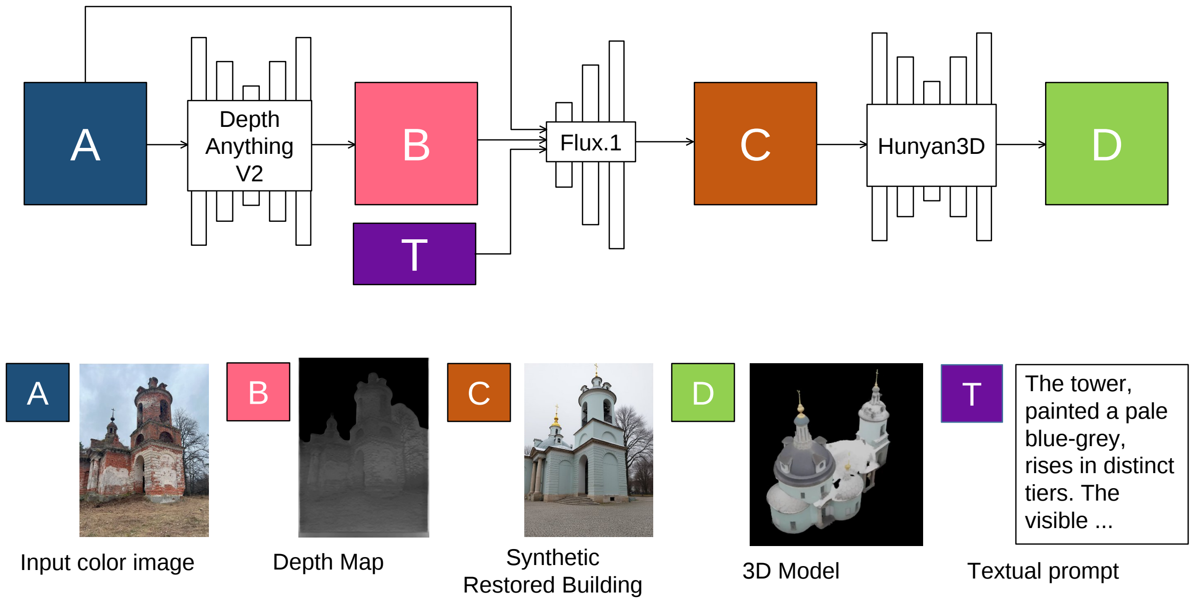

Figure 2. Overview of the proposed `Restore3D` framework.

section provides review of the related work on these families of neural models.

2.1 2D Inpainting

Image inpainting is a crucial process in digital image restoration, aimed at filling in missing pixels in a manner that appears plausible and seamless. This task is essential for repairing damaged images, removing unwanted objects, and enhancing visual content. The methodologies employed for image inpainting are broadly categorized into traditional techniques and those leveraging deep learning. Traditional methods rely heavily on mathematical reasoning and utilize existing information within the image to infer and reconstruct missing areas. Notable examples include the model proposed by Bertalmio (BSCB) (Bertalmío et al., 2000) as well as the Criminisi model (Criminisi et al., 2004). These foundational models have inspired subsequent developments such as the Total Variation (TV) model (Shen and Chan, 2002) and the PatchMatch model (Barnes et al., 2009). While traditional algorithms have demonstrated effectiveness in addressing minor defects, they often fall short when confronted with extensive damage or complex textures.

In contrast, the field of image inpainting has experienced a transformative shift with the advent of deep learning techniques, driven by advancements in computational power and hardware capabilities. The introduction of architectures like AlexNet (Krizhevsky et al., 2012, Kniaz et al., 2021a, Knyaz and Kniaz, 2020), VGG networks (Simonyan and Zisserman, 2015), ResNet (He et al., 2016), and Generative Adversarial Networks (GAN) (Goodfellow et al., 2014, Mizginov et al., 2021, Kniaz and Bordodymov, 2019) has significantly propelled the development of sophisticated image inpainting algorithms. Early deep learning models such as the Context Encoder (Pathak et al., 2016), Globally and Locally Consistent Image Completion (GLCIC) (Iizuka et al., 2017), Generative Multicolumn Convolutional Neural Networks (GMCNN) (Wang et al., 2018), Generative Image Inpainting with Contextual Attention (CA) (Yu et al., 2018), and Edge Connect (Nazeri et al., 2019) have employed encoder-decoder structures combined with GAN discriminators to optimize generative adversarial loss between inpainted and real images. These innovations have led to more realistic restorations by leveraging both global context understanding and local detail refinement.

Moreover, several image inpainting models have adopted the U-Net architecture (Ronneberger et al., 2015) due to its efficacy in handling various scales of defects through its unique encoder-decoder structure with skip connections. Models such as Shift-Net (Yan et al., 2018), Deep Fusion Network (DFNet) (Hong et al., 2019), Partial Convolutions (Liu et al., 2018) designed for random defects, and Pyramid-Context Encoder Network (PEN-Net) (Zeng et al., 2019) exemplify this approach. These U-Net-based models often incorporate GAN's generative adversarial loss to enhance the realism of restored images further. Consequently, many studies categorize these models under both U-Net class and GAN-based image inpainting models due to their dual reliance on structural encoding-decoding mechanisms and adversarial training paradigms. This dual classification underscores their capability to produce credible and visually coherent results even when faced with complex inpainting challenges.

2.2 Single Image 3D Reconstruction and Inpainting

The field of 3D model inpainting has evolved considerably over the years, beginning with traditional handcrafted photogrammetry methods that leverage stereo pairs and manual reconstruction techniques. These early approaches relied heavily on skilled human intervention and detailed measurements to generate 3D models, often involving complex processes to match and blend images for reconstructing missing parts of an object or scene. As technology advanced, the introduction of the structure-from-motion (SfM) pipeline (Remondino and El-Hakim, 2006) marked a significant improvement by automating the process of 3D reconstruction. SfM utilizes a series of overlapping images to estimate camera positions and recover dense

Figure 3. Example images from the dataset used for fine-tuning the `Flux.1` model.

point clouds, eventually constructing a coherent 3D model of the scene. However, these methods are limited by the quality and number of input images available. In recent years, neural-based techniques like RenderDiffusion (Anciukevicius et al., 2022b) and NeRFiller (Weber et al., 2023) have further revolutionized 3D inpainting. RenderDiffusion, for instance, applies diffusion models to enhance and complete 3D reconstructions, refining details and textures with high fidelity. NeRFiller, on the other hand, employs generative neural radiance fields to effectively synthesize missing components in 3D space, achieving impressive results in scene completion. These modern approaches mark a significant shift toward using artificial intelligence to tackle the challenges of 3D model inpainting with reduced dependency on extensive input data.

The rapidly evolving landscape of 3D inpainting has seen significant advancements through the integration of neural radiance fields and diffusion models. One noteworthy contribution is the SPIn-NeRF dataset (Mirzaei et al., 2023), which centers on multiview segmentation and perceptual inpainting using neural radiance fields. This dataset facilitates the exploration of robust methods to segment and reconstruct missing portions of a scene, providing a pivotal resource for training and evaluating advanced inpainting models. Another important development is the Instruct-NeRF2NeRF method (Haque et al., 2023), which introduces an innovative approach for instruction-based editing of NeRFs utilizing a 2D diffusion model. This enables intuitive modifications to existing NeRFs, offering users a powerful tool to edit and enhance 3D scenes with minimal manual intervention. Extending these capabilities further, InNeRF360 (Wang et al., 2023) presents a text-guided framework for 3D-consistent object inpainting across 360-degree neural radiance fields. By incorporating text inputs, InNeRF360 allows precise and coherent inpainting over complete panoramic views, ensuring high levels of detail and continuity. These advancements underscore a growing trend towards using machine learning to achieve intricate and contextually aware reconstructions in the realm of 3D inpainting.

Other approaches for 3D reconstruction take in account reinforcement learning (Kniaz et al., 2021b, Kniaz, 2015, Kniaz, 2014), structured light techniques (Knyaz, 2012) and deep volumetric U-nets with skip connections (Knyaz et al., 2019, Kniaz et al., 2020). The important task in the field of 3D reconstruction is a development of a holistic validation technique that compares the ground truth model with respect to the predicted model (Mizginov and Kniaz, 2019, Knyaz and Moshkantsev, 2019).

3. Method

The aim of our `Restore3D` framework is simulatenous estimation of the original appearance of a partially destroyed object of cultural heritage and synthesis of its 3D model. Our framework operates in four domains. The input image domain $\mathcal{A} \in \mathbb{R}^{w \times h \times 3}$, the depth map domain $\mathcal{B} \in \mathbb{R}^{w \times h}$, the synthetically restored image domain $\mathcal{C} \in \mathbb{R}^{w \times h \times 3}$, and the implicit 3D model domain $\mathcal{D} = \{\mathrm{x} \in \mathbb{R}^3 | f(\mathrm{x}) = 0\}$, where $\mathcal{D}$ is the surface of a given 3D object represented by an implicit function $f(\mathrm{x})$ that represents a signed distance function from the object's surface to a given point in a 3D volume x (Wang et al., 2021, Yang et al., 2024c). We use the marching cube algorithm to transform the implicit 3D model representation to the explicit 3D mesh that can be utilized by 3D artist and scientists in the field of history of architecture. The overview of our framework is presented in Figure 2.

The rest of this Section presents details on the architecture of our framework and provides a brief description of fine-tuning procedure leveraging the Low-rank adaptation approach.

3.1 Framework Overview

Our framework operates by receiving a single color image $A \in \mathcal{A}$ presenting a partially destroyed object of cultural heritage. Our approach is generation of an synthetically restored image $C \in \mathcal{C}$ using a diffusion model. Following the recent research (Margaryan et al., 2024) on the depth-guided image synthesis with diffusion models, we assume that an additional input channel representing the depth map B for an image A will improve the stability of prediction of the restored image C.

Hence, we firstly estimate a depth map B for an image A using the `DepthAnythingV2` model (Yang et al., 2024a, Yang et al., 2024b).

After that, we leverage the `Flux.1` (Labs, 2024) model to perform depth-guided image synthesis using the original image A, and its depth map B as an input. We additionally provide a textual prompt T that indicates that a translation from the destroyed to the restored state is required. The `Flux.1` predicts the restored image C that is processed by the `Hunyan3D` model (Yang et al., 2024c) to predict the implicit representation of a 3D model as signed distance function $f(\mathrm{x})$. Processing of the object volume with a marching cubes algorithms produce the required explicit 3D model representation of the restored state of the given object of cultural heritage.

3.2 Low Rank Adaptation

Fine-tuning large diffusion models is a resource-intensive process, often requiring significant computational power and memory resources. The Low-Rank Adaptation (LoRA) technique offers a solution by enabling the fine-tuning of these models with a substantially reduced number of parameters. This is achieved by incorporating smaller matrices into the attention weights of the model, which typically results in a reduction of trainable parameters by approximately 90

LoRA operates by freezing the pre-trained model weights and introducing trainable rank decomposition matrices into various layers of the model. Unlike traditional fine-tuning methods that require updating all model parameters, LoRA employs low-rank decomposition to break down weight updates into smaller, more manageable matrices. This strategy significantly diminishes the number of parameters that need to be trained while preserving the model’s efficacy. For instance, when applied to GPT-3 175B, LoRA achieved a reduction in trainable parameters by a factor of 10,000 and decreased GPU memory requirements by threefold compared to conventional full fine-tuning methods.

The mechanism of LoRA involves adding pairs of rank decomposition matrices specifically to transformer layers, with a primary focus on attention weights. During inference, these adapter weights can be seamlessly integrated with the base model without incurring additional latency overheads. This makes LoRA particularly advantageous for adapting large language models to specific tasks or domains while keeping resource demands within practical limits. The ability to merge adapter weights back into the base model ensures that there is no increase in computational delay during inference.

LoRA presents several key advantages, particularly in terms of memory efficiency and training features. By storing only adapter parameters in GPU memory and keeping base model weights frozen—potentially in lower precision—LoRA facilitates the fine-tuning of large models even on consumer-grade GPUs. Furthermore, it supports native integration with minimal setup and offers compatibility with QLoRA (Quantized LoRA) for enhanced memory efficiency. Adapter management is streamlined through features that allow for saving adapter weights during checkpoints and merging them back into the base model as needed. In our work, we leverage LoRA to fine-tune the Flux model for 2D generative inpainting tasks involving partially destroyed objects of cultural heritage, demonstrating its practical applicability and effectiveness in specialized domains.

4. Evaluation

We evaluated our `Restore3D` framework and two baselines using the collected *Destroyed2Restored*. The following section presents details on the evaluation protocol and qualitative and quantitative results for the task of the prediction on the restored 3D model of a partially destroyed object of cultural heritage. Finally, we present a brief ablation study proving the necessity of all components of our `Restore3D` framework.

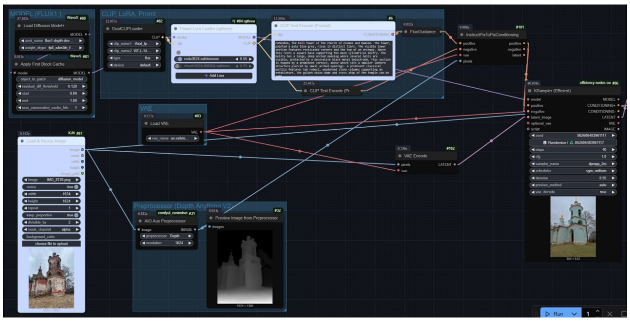

Figure 4. Implementation of the proposed `Restore3D` framework using ComfyUI.

4.1 Evaluation Protocol

We use the following evaluation protocol. For each model we use the training split of the *Destroyed2Restored* dataset including 5k images of 5 objects of cultural heritage to train the model. We use the test split including 200 images of two other objects of cultural heritage for the evaluation.

We perform evaluation in terms of 3D Intersection over Union and Frechet Inception Distance (Heusel et al., 2017) between the real images of object and synthetic renderings of its digitally restored 3D model.

4.2 Qualitative Evaluation

We evaluate our `Restore3D` and baselines qualitatively on the task of reconstruction of a restored 3D model of a partially destroyed object of cultural heritage. Qualitative results for our framework and baselines are presented in Figure 5.

Figure 5. Qualitative results for our `Restore3D` framework and baselines.

4.3 Quantitative Evaluation

We evaluate our `Restore3D` and baselines qualitatively in terms of 3D intersection over union and Frechet Inception Distance (FID) (Heusel et al., 2017). Results are presented in table 1. The usage of a modern diffusion architecture and an implicit 3D model representation allows our Recover3D model surpas modern 3D Inpainting models.

Table 1. Quantitative comparison between our Recover3D framework and baselines.

Model	3D IoU	FID
Recover3D	0.371	23
NeRFiller (Weber et al., 2024)	0.261	65
RenderDiffusion (Anciukevicius et al., 2022b)	0.201	45

4.4 Ablation Studies

We evaluate the necessity of all components of our model by performing 3D model reconstructions using an ablated version of our model.

We compare the performance of our `Restore3D` using empty or wrong inputs for the depth map input B and the reference image A. Results are presented in Figure 6. The different reference images A inputs are given for different columns. The different depth maps B are given for different rows. The resulting image C is located on the intersection of a given row and a given column.

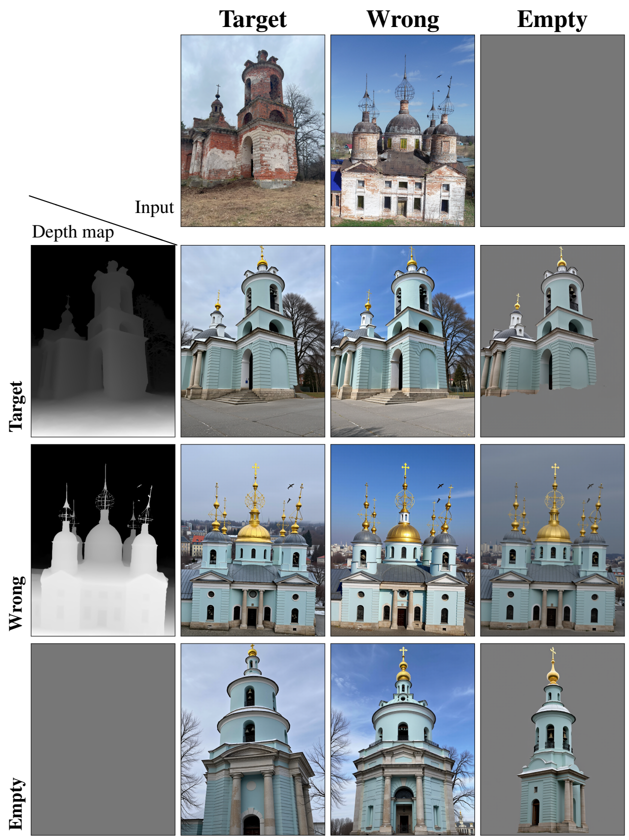

Figure 6. Evaluation of ablated versions of our `Restore3D` framework.

It's obvious that the depth map B defines the shape of the object in the output image. The reference image A provides the reference style for the output image. The empty reference image A causes the background to be empty. While the empty depth map B completely removes any relation with the shape of the original object.

The ablation study proves the necessity of all components of our `Restore3D` model. Only the combination of the input reference image and depth map allows independent control of the shape and style of the object in the output image.

4.5 Discussion

The main application of our `Recover3D` framework is a synthesis of a draft prototype of a reconstruction of a partially destroyed 3D object. Still, the consistency with the architectural style of a given epoch is an important requirement for scientists in the field of history of architecture.

The models predicted from a single image in Figure 5 and Figure 6 are highly realistic and have elaborate architectural elements. However, it can be noted that probabilistic nature of a neural diffusion model took some liberties in modeling details. That does not allow us to speak about the completeness of the development of the method for obtaining a historically sound three-dimensional models from a single image of a partially destroyed object.

This is expressed in changes in the shape and number of windows on the facades of buildings, the appearance of columns where they should not be.

Such a freedom in the historical style appearing in the reconstruction can be explained if we take into account that the time of construction of the building is not provided to the model in the input prompt. Also no differentiation into architectural styles was provided in the training dataset (Figre 3) that was used to fine-tune the `Flux.1` model. The dataset was collected by crawling 20k of open source images of churches from the internet. Hence even some images of churches with a modern architecture were used for training.

It is obvious that the reason for these errors is the imperfection of the training methodology. Firstly, for architecture, where architectural details are significant because they are a kind of style markers, the training samples and the modeled objects must belong to the same styles. Secondly, it is necessary to take into account the time of construction of the building and do not include the training samples with a modern architecture for a model designed for 3D reconstruction of old buildings. Such a differentiation should be strict even if architectural styles are similar.

Finally, the images in the training dataset must have sufficient spatial resolution so that it is possible to unambiguously determine small but key architectural elements for the style and the time of construction.

Thus, it can be said that with the proposed efficient algorithm, an accurately annotated dataset, and a properly trained model, it will be possible to obtain plausible three-dimensional models of partially destroyed objects. In further research we are going to focus on semi-automatic analysis of the training dataset that will divide it into splits for different historical epochs. Also we are going to annotate images with textual prompts indicating the historical epoch and style of a depicted object. This will allow us to develop a controllable robust model for single image 3D reconstruction

5. Conclusion

We developed the `Restore3D` generative diffusion model for 3D inpainting and reconstruction. This model's ability to simultaneously perform single image 3D reconstruction and restoration offers a robust solution for addressing the challenges posed by digital restoration of partially destroyed architectural structures. By harnessing the power of diffusion processes, `Restore3D` effectively reconstructs missing sections with high fidelity, preserving intricate details and spatial relationships that are often lost in traditional methods. Our comprehensive dataset collection facilitated rigorous training and testing of the framework, ensuring its robustness and generalizability across diverse scenarios.

The empirical results substantiate the efficacy of `Restore3D`, as it consistently outperforms two modern baselines by achieving an 11% improvement in terms of 3D Intersection over Union (IoU) and a notable enhancement of 22 points in the Frechet Inception Distance (FID) metric. Such advancements not only demonstrate the superiority of our approach but also highlight its applicability for scientists working in the history of architecture, where accurate reconstructions are paramount. The findings from this study pave the way for future research endeavors aimed at further refining generative models for architectural preservation, disaster recovery, and urban planning, underscoring the transformative potential of integrating deep learning techniques into these critical domains.

Acknowledgements

The research was carried out at the expense of a grant from the Russian Science Foundation No. 24-21-00314, https://rscf.ru/project/24-21-00314/

References

Anciukevicius, T., Xu, Z., Fisher, M., Henderson, P., Bilen, H., Mitra, N. J., Guerrero, P., 2022a. RenderDiffusion: Image Diffusion for 3D Reconstruction, Inpainting and Generation. *arXiv*.

Anciukevicius, T., Xu, Z., Fisher, M., Henderson, P., Bilen, H., Mitra, N. J., Guerrero, P., 2022b. RenderDiffusion: Image Diffusion for 3D Reconstruction, Inpainting and Generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12608-12618. https://api.semanticscholar.org/CorpusID:253708307.

Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D. B., 2009. PatchMatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3), 24. https://doi.org/10.1145/1531326.1531330.

Bertalmío, M., Sapiro, G., Caselles, V., Ballester, C., 2000. Image inpainting. J. R. Brown, K. Akeley (eds), *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000*, ACM, 417–424.

Criminisi, A., Pérez, P., Toyama, K., 2004. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.*, 13(9), 1200–1212. https://doi.org/10.1109/TIP.2004.833105.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., Bengio, Y., 2014. Generative adversarial nets. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (eds), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2672–2680.

Haque, A., Tancik, M., Efros, A. A., Holynski, A., Kanazawa, A., 2023. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 19683-19693. https://api.semanticscholar.org/CorpusID:257663414.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, 770–778.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (eds), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 6626–6637.

Hong, X., Xiong, P., Ji, R., Fan, H., 2019. Deep fusion network for image completion. L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, W. T. Ooi (eds), *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, ACM, 2033–2042.

Iizuka, S., Simo-Serra, E., Ishikawa, H., 2017. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4), 107:1–107:14. https://doi.org/10.1145/3072959.3073659.

Kniaz, V. V., 2014. A Fast Recognition Algorithm for Detection of Foreign 3D Objects on a Runway. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-3, 151–156. https://isprs-archives.copernicus.org/articles/XL-3/151/2014/.

Kniaz, V. V., 2015. Fast instantaneous center of rotation estimation algorithm for a skied-steered robot. F. Remondino, M. R. Shortis (eds), *Videometrics, Range Imaging, and Applications XIII*, 9528, International Society for Optics and Photonics, SPIE, 95280L.

Kniaz, V. V., Bordodymov, A. N., 2019. LONG WAVE INFRARED IMAGE COLORIZATION FOR PERSON RE-IDENTIFICATION. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W12, 111–116. https://isprs-archives.copernicus.org/articles/XLII-2-W12/111/2019/.

Kniaz, V. V., Grodzitskiy, L., Knyaz, V. A., 2021a. DEEP LEARNING FOR CODED TARGET DETECTION. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIV-2/W1-2021, 125–130. https://isprs-archives.copernicus.org/articles/XLIV-2-W1-2021/125/2021/.

Kniaz, V. V., Knyaz, V. A., Mizginov, V., Papazyan, A., Fomin, N., Grodzitsky, L., 2021b. Adversarial dataset augmentation using reinforcement learning and 3d modeling. B. Kryzhanovsky, W. Dunin-Barkowski, V. Redko, Y. Tiumentsev (eds), *Advances* *in Neural Computation, Machine Learning, and Cognitive Research IV*, Springer International Publishing, Cham, 316–329.

Kniaz, V. V., Knyaz, V. A., Remondino, F., Bordodymov, A., Moshkantsev, P., 2020. Image-to-voxel model translation for 3d scene reconstruction and segmentation. A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (eds), *Computer Vision – ECCV 2020*, Springer International Publishing, Cham, 105–124.

Knyaz, V. A., 2012. IMAGE-BASED 3D RECONSTRUCTION AND ANALYSIS FOR ORTHODONTIA. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXIX-B3, 585–589. https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XXXIX-B3/585/2012/.

Knyaz, V. A., Kniaz, V. V., Remondino, F., 2019. Image-to-voxel model translation with conditional adversarial networks. L. Leal-Taixé, S. Roth (eds), *Computer Vision – ECCV 2018 Workshops*, Springer International Publishing, Cham, 601–618.

Knyaz, V. A., Moshkantsev, P. V., 2019. JOINT GEOMETRIC CALIBRATION OF COLOR AND THERMAL CAMERAS FOR SYNCHRONIZED MULTIMODAL DATASET CREATING. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W18, 79–84. https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W18/79/2019/.

Knyaz, V., Kniaz, V., 2020. Object recognition for UAV navigation in complex environment. L. Bruzzone, F. Bovolo, E. Santi (eds), *Image and Signal Processing for Remote Sensing XXVI*, 11533, International Society for Optics and Photonics, SPIE, 115330P.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (eds), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, 1106–1114.

Labs, B. F., 2024. Flux. `https://github.com/black-forest-labs/flux`.

Liu, G., Reda, F. A., Shih, K. J., Wang, T., Tao, A., Catanzaro, B., 2018. Image inpainting for irregular holes using partial convolutions. V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, Lecture Notes in Computer Science, 11215, Springer, 89–105.

Margaryan, H., Hayrapetyan, D., Cong, W., Wang, Z., Shi, H., 2024. DGBD: depth guided branched diffusion for comprehensive controllability in multi-view generation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, IEEE, 747–756.

Mirzaei, A., Aumentado-Armstrong, T., Derpanis, K. G., Kelly, J., Brubaker, M. A., Gilitschenski, I., Levinshtein, A., 2023. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20669–20679.

Mizginov, V. A., Kniaz, V. V., Fomin, N. A., 2021. A METHOD FOR SYNTHESIZING THERMAL IMAGES USING GAN MULTI-LAYERED APPROACH. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIV-2/W1-2021, 155–162. https://isprs-archives.copernicus.org/articles/XLIV-2-W1-2021/155/2021/.

Mizginov, V., Kniaz, V. V., 2019. EVALUATING THE ACCURACY OF 3D OBJECT RECONSTRUCTION FROM THERMAL IMAGES. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 129-134. https://api.semanticscholar.org/CorpusID:209468392.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., Ebrahimi, M., 2019. Edgeconnect: Structure guided image inpainting using edge prediction. *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, IEEE, 3265–3274.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A. A., 2016. Context encoders: Feature learning by inpainting. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, 2536–2544.

Remondino, F., El-Hakim, S., 2006. Image-based 3D modelling: a review. *The photogrammetric record*, 21(115), 269–291.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. N. Navab, J. Hornegger, W. M. W. III, A. F. Frangi (eds), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, Lecture Notes in Computer Science, 9351, Springer, 234–241.

Shen, J., Chan, T. F., 2002. Mathematical Models for Local Nontexture Inpaintings. *SIAM J. Appl. Math.*, 62(3), 1019–1043. https://doi.org/10.1137/S0036139900368844.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. Y. Bengio, Y. LeCun (eds), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Wang, D., Zhang, T., Abboud, A., Süsstrunk, S., 2023. InNeRF360: Text-Guided 3D-Consistent Object Inpainting on 360° Neural Radiance Fields. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12677-12686. https://api.semanticscholar.org/CorpusID:258865532.

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W., 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan (eds), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 27171–27183.

Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J., 2018. Image inpainting via generative multi-column convolutional neural networks. S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (eds), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 329–338.

Weber, E., Holy’nski, A., Jampani, V., Saxena, S., Snavely, N., Kar, A., Kanazawa, A., 2023. NeRFiller: Completing Scenes via Generative 3D Inpainting. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20731-20741. https://api.semanticscholar.org/CorpusID:266052569.

Weber, E., Holynski, A., Jampani, V., Saxena, S., Snavely, N., Kar, A., Kanazawa, A., 2024. Nerfiller: Completing scenes via generative 3d inpainting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20731–20741.

Yan, Z., Li, X., Li, M., Zuo, W., Shan, S., 2018. Shift-net: Image inpainting via deep feature rearrangement. V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, Lecture Notes in Computer Science, 11218, Springer, 3–19.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H., 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. *CVPR*.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024b. Depth Anything V2. *arXiv:2406.09414*.

Yang, X., Shi, H., Zhang, B., Yang, F., Wang, J., Zhao, H., Liu, X., Wang, X., Lin, Q., Yu, J., Wang, L., Chen, Z., Liu, S., Liu, Y., Yang, Y., Wang, D., Jiang, J., Guo, C., 2024c. Tencent Hunyuan3D-1.0: A Unified Framework for Text-to-3D and Image-to-3D Generation. *CoRR*, abs/2411.02293. https://doi.org/10.48550/arXiv.2411.02293.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T. S., 2018. Generative image inpainting with contextual attention. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, Computer Vision Foundation / IEEE Computer Society, 5505–5514.

Zeng, Y., Fu, J., Chao, H., Guo, B., 2019. Learning pyramid-context encoder network for high-quality image inpainting. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, Computer Vision Foundation / IEEE, 1486–1494.