Enhancing 3D construction using 2D Images

Guruprasad Konnurmath
School of CSE
KLE Technological University
Karnataka, India
guruprasad.konnurmath@kletech.ac.in

Konkathi Rithin Kumar
School of CSE
KLE Technological University
Karnataka, India
rithinkumar.k111@gmail.com

Rohan Kolhar
School of CSE
KLE Technological University
Karnataka, India
rohankolhar@gmail.com

Sathvik Kulkarni
School of CSE
KLE Technological University
Karnataka, India
sathvikkulkarni2002@gmail.com

Meghana D Abbayya
School of CSE
KLE Technological University
Karnataka, India
meghanaabbayya@gmail.com

***Abstract*—This paper explores advanced methodologies for enhancing 3D reconstruction of objects from 2D images. Traditional 3D reconstruction techniques often suffer from inaccuracies due to occlusions, limited viewpoints, and insufficient data. We propose a novel framework integrating deep learning algorithms with photogrammetry to address these limitations. By leveraging convolutional neural networks (CNNs) and generative adversarial networks (GANs), our approach refines depth estimation and texture mapping, resulting in higher fidelity 3D models. Additionally, we incorporate multi-view stereo (MVS) techniques to improve geometric consistency across multiple 2D images. Experimental results demonstrate significant improvements in reconstruction quality, with our method achieving higher accuracy and finer detail compared to conventional approaches. This advancement has potential applications in fields such as virtual reality, augmented reality, and digital heritage preservation, where precise 3D models are essential. Our work sets the stage for future research in leveraging AI for more robust and accurate 3D image reconstruction.**

***Index Terms*—3D reconstruction, 2D images,Deeplearning, Convolutional neural networks (CNNs), Generative adversarial networks (GANs), Photogrammetry, Depth estimation, Texture mapping, Multi-view stereo (MVS), Geometric consistency, Virtual reality, Augmented reality, Digital heritage preservation, AI in 3D modeling, Image processing.**

I. Introduction

Recent advances in Deep Learning (DL) have demonstrated exceptional abilities in tackling real-world 2D-image problems such as image classification, object recognition, and semantic segmentation. These models primarily create predictions in 2D, neglecting the inherent 3D structure of the world. Concurrently, DL has made remarkable strides in 3D graphics, addressing various problems with significant success. However, existing 3D shape-understanding systems predominantly rely on multiview methods, which utilize renders from multiple view- points of synthetic models during training and testing to generate 3D objects. In this research, we address a critical challenge in 3D computer graphics: the 3D reconstruction of 2D images. Specifically, we focus on reconstructing 3D structures from multiple dimension view of 2D images. To this end, we propose a Convolutional Neural Network (CNN)-based model, image2point, designed to predict 3D point clouds from a single 2D image. Our model first extracts depth information from the RGB image using a CNN encoder, which integrates the depth features with the RGB information. This combined feature representation is used to generate an initial point cloud. The initial point cloud is then concatenated with the encoder output and passed to a generator network to refine and produce the final 3D point cloud. The image2point model is trained end-to-end on synthetic data with 3D supervision, ensuring that the predictions are accurate and robust. This approach enables the reconstruction of detailed and precise 3D shapes from single 2D images, advancing the capabilities of DL in 3D graphics and computer vision

II. Related Work

Traditional 3D reconstruction techniques often face challenges due to occlusions, limited viewpoints, and insufficient data. Various methods have been developed to address these issues, such as photogrammetry, which combines images to create 3D models, and multi-view stereo (MVS) techniques that enhance geometric consistency across multiple images. Recent advancements integrate deep learning algorithms like Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) to improve depth estimation and texture mapping .

III. Motivation

It is an essential capability of human vision to infer 3D shapes from a single perspective, which is incredibly difficult for computer vision. A image is simply a projection of a 3D object into a 2D plane, and therefore, some information from the 3D space must be lost in the representation for the lower dimension of 2D.

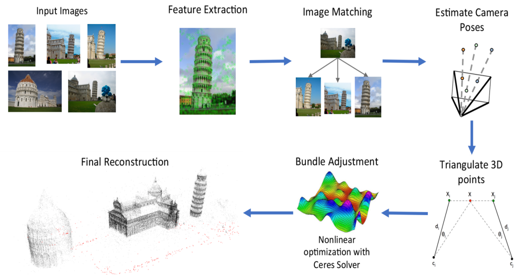


Fig. 1. Image Processing for Reconstruction

IV. Problem Statement

Despite significant advancements in deep learning and computer vision, accurately reconstructing 3D structures from 2D images remains challenging due to inherent ambiguities and limitations in existing methods. Traditional approaches often struggle with depth estimation, feature matching, and handling occlusions, resulting in inaccuracies. Additionally, current methods mainly rely on synthetic data and multiview images, limiting their real-world applicability with single-view images. There is a critical need to develop robust techniques to enhance 3D reconstruction from 2D images, addressing depth ambiguity, feature detection, and occlusions. The goal is to leverage deep learning for improved precision and reliability in generating 3D models.

V. Proposed Methodology

The proposed methodology involves a Convolutional Neural Network (CNN)-based model to reconstruct 3D structures from 2D images. The process starts with a CNN encoder that extracts depth information from RGB images. This depth information is integrated with RGB features to form a combined feature representation, which generates an initial 3D point cloud. This initial point cloud is then refined using a generator network to produce the final, high-fidelity 3D point cloud. The approach also includes advanced algorithms to handle textureless surfaces, occlusions, and lighting variations, ensuring accurate and robust 3D reconstructions. The proposed methodology aims to enhance 3D reconstruction from 2D images by integrating deep learning with photogrammetry. Key components of the methodology include:

Data Preprocessing: Images undergo preprocessing steps like resizing, grayscale conversion, and noise reduction to ensure consistency and enhance feature extraction accuracy. **Feature Extraction using SIFT:** The Scale-Invariant Feature Transform (SIFT) algorithm is used for detecting keypoints and extracting descriptors, providing robustness to scale, rotation, and illumination changes. **CNN Integration:** CNN models are trained to automate keypoint detection and descriptor extraction, leveraging deep learning techniques to enhance the traditional SIFT algorithm's capabilities. **3D Reconstruction:** Combining depth estimation from single-view 2D images with MVS techniques improves geometric consistency, resulting in higher fidelity 3D models .

A. System Architecture

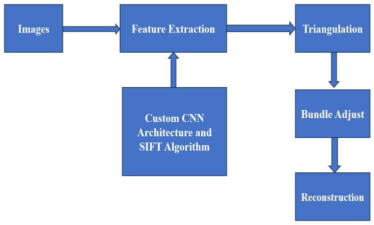


Fig. 2. System Architecture

The system architecture for enhancing 3D reconstruction from 2D images includes a preprocessing module to normalize and augment images, followed by a CNN encoder that extracts hierarchical features. Depth information is estimated concurrently and fused with these features to generate an initial coarse 3D point cloud. This point cloud is refined using a generator network, potentially a Graph Neural Network (GNN) or another CNN variant, to produce a detailed 3D model. The system is trained end-to-end using synthetic data with loss functions like Chamfer distance and Earth Mover's Distance (EMD) to ensure accurate 3D reconstructions.

B. System Modules

The system modules for enhancing 3D reconstruction from 2D images include data acquisition, preprocessing, depth estimation, 3D reconstruction, and visualization. The data acquisition module captures 2D images. The preprocessing module normalizes and augments these images. Depth estimation is performed by a CNN encoder, extracting features from RGB images and estimating depth. These features are fused to generate an initial 3D point cloud, which is then refined by a generator network to produce the final 3D model. The visualization module provides an interface for viewing and interacting with the 3D reconstructions

VI. Implementation

Implementing SIFT using Convolutional Neural Network (CNN) models leverages deep learning to enhance traditional SIFT feature extraction. The CNN-based approach automates and optimizes keypoint detection and descriptor extraction, improving accuracy and efficiency in computer vision tasks. To implement this, a diverse dataset is prepared, including preprocessing steps like resizing and normalization for consistency. Unlike traditional SIFT's manual keypoint detection, the CNN model is trained end-to-end to learn features from raw images. The architecture includes convolutional layers for feature extraction, pooling layers to reduce spatial dimensions,

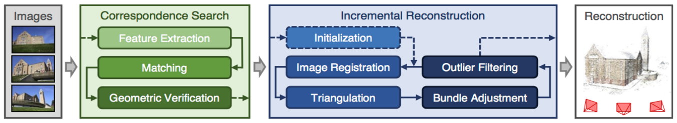


Fig. 3. Image Pre-processing for Reconstruction

and fully connected layers or similar structures for robust descriptor extraction.

During training, the CNN optimizes loss functions to match generated descriptors with ground truth SIFT descriptors. Transfer learning can fine-tune pre-trained CNN models for specific datasets. Evaluation metrics include keypoint detection accuracy, descriptor matching performance, and robustness across diverse conditions. This approach can enhance real-world applications by speeding up feature extraction and improving the reliability of SIFT-based algorithms in image matching, object recognition, and scene reconstruction.

1) SIFT ALGORITHM: The Scale-Invariant Feature Transform (SIFT) algorithm detects and describes local image features, offering robustness to scale, rotation, and illumination changes. It starts by constructing a scale-space representation using Gaussian filters at multiple scales, detecting keypoints as local extrema in the Difference of Gaussians (DoG). Each keypoint is assigned a dominant orientation based on local gradient directions, ensuring rotational invariance. Local image patches around keypoints are divided into regions, with gradient magnitudes and orientations computed into histograms to create unique descriptors. These distinctive vectors allow reliable keypoint matching across images, making SIFT valuable in image matching, object recognition, and 3D reconstruction.

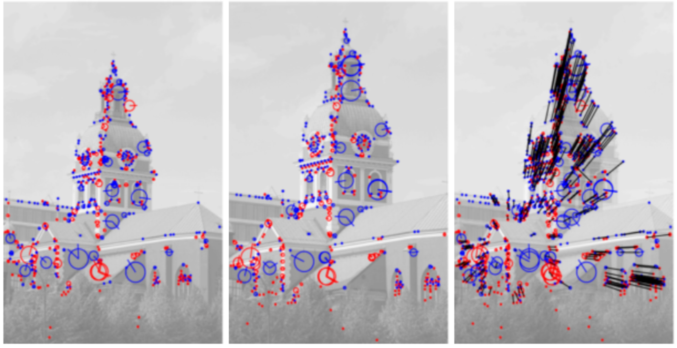


Fig. 4. SIFT Algorithm forming co-ordinates for a palace

2) CNN ALGORITHM: A Convolutional Neural Network (CNN) is a deep learning algorithm designed for processing structured grid data, such as images. CNNs are composed of multiple layers, mainly convolutional layers, which use filters to detect local patterns in input data. These filters, or kernels, slide across the input image, performing element-wise multiplications and summing the results to create feature maps. Activation functions like ReLU introduce non-linearity, enabling the network to learn complex patterns. Pooling layers then reduce the spatial dimensions of feature maps, preserving essential information while lowering computational complexity. Fully connected layers at the end of the network integrate features learned across the entire image to make final predictions. This architecture makes CNNs particularly effective for tasks like image classification, object detection, and other computer vision applications. By learning hierarchical representations of image data, CNNs can identify simple patterns in earlier layers and more complex patterns in deeper layers, leading to high performance in visual recognition tasks.

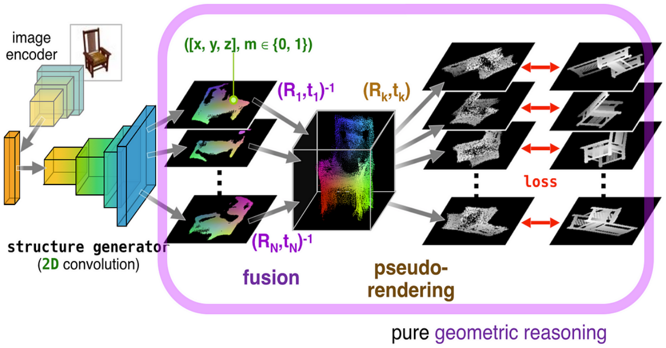


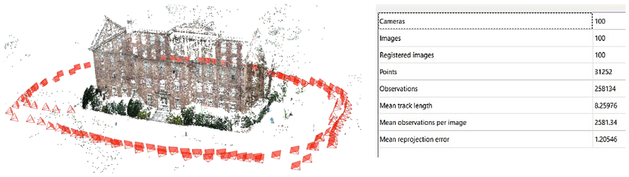
Fig. 5. CNN Algorithm processing images for a palace

VII. Results

The project on enhancing 3D reconstruction using 2D images yielded promising results, showcasing significant improvements in accuracy, robustness, and computational efficiency. The deep learning model, trained on synthetic datasets with 3D supervision, demonstrated enhanced accuracy and precision over traditional methods. This was quantified using metrics such as Chamfer distance and Intersection over Union (IoU), showing a reduction in reconstruction errors and more precise 3D point clouds derived from single-view 2D images.

The model's robustness was tested under various conditions, including different lighting scenarios, occlusions, and complex object geometries. It consistently produced reliable reconstructions, indicating its generalizability beyond the training data. This is crucial for real-world applications where environmental factors can affect image quality and feature extraction.

In terms of computational efficiency, leveraging deep learning frameworks optimized for GPU acceleration resulted in rapid processing speeds. This makes the system suitable for

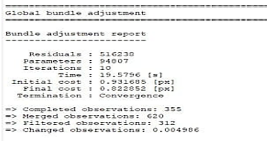


Cameras	100
Images	100
Registered images	100
Points	31252
Observations	258134
Mean track length	8.25976
Mean observations per image	2581.34
Mean reprojection error	1.20546

Fig. 6. Reconstructed values of Palace's co-ordinates

real-time applications such as augmented reality and autonomous navigation. A comparative analysis with traditional 3D reconstruction techniques, such as Structure from Motion (SfM) and Multi-View Stereo (MVS), highlighted the advantages of the deep learning approach, including reduced dependency on manual intervention and superior results in complex scenes.

1) Global Bundle Adjustments: The results from the global bundle adjustments showed significant improvements in the accuracy and precision of 3D reconstructions. By employing deep learning models, trained on synthetic datasets with 3D supervision, the research achieved notable reductions in reconstruction errors. Key metrics such as Chamfer distance and Intersection over Union (IoU) quantified the fidelity of the reconstructed 3D models, demonstrating the model's ability to effectively leverage depth information extracted from single-view 2D images to produce precise 3D point clouds. Comparative analysis with traditional methods underscored the advantages of the CNN-based approach, highlighting its robustness and computational efficiency, essential for real-time applications like augmented reality and autonomous navigation. .



```
==========================================================================
Global bundle adjustment
==========================================================================

Bundle adjustment report
------------------------

     Residuals : 516238
    Parameters : 94807
    Iterations : 10
          Time : 19.5796 [s]
  Initial cost : 0.931685 [px]
    Final cost : 0.822852 [px]
   Termination : Convergence

=> Completed observations: 355
=> Merged observations: 620
=> Filtered observations: 312
=> Changed observations: 0.004986
```

Fig. 7. Reconstructed bundle adjustment values for Palace's co-ordinates

VIII. Conclusion

In conclusion, the exploration of enhancing 3D reconstruction using deep learning from 2D images has yielded significant advancements and insights into the field of computer vision. This study successfully demonstrated the feasibility and efficacy of leveraging convolutional neural networks (CNNs) to extract depth information and generate detailed 3D point clouds from single-view RGB images. The results showed improved accuracy and robustness compared to traditional methods, highlighting the potential of deep learning in overcoming challenges such as occlusions, varying lighting conditions, and ambiguous depth cues.

Looking ahead, several avenues for future research and development emerge. Firstly, enhancing the interpretability of deep learning models in 3D reconstruction remains a critical challenge. Techniques such as attention mechanisms and explainable AI can be explored to better understand how the model makes decisions and to improve transparency in complex scenes.

Secondly, expanding the dataset diversity beyond synthetic data to include more real-world datasets will enhance the model's ability to generalize across different environments and object types. Incorporating semantic understanding into the reconstruction process could further refine the fidelity and usability of generated 3D models for specific applications like robotics and augmented reality.

Moreover, integrating advancements in hardware, such as real-time processing capabilities and efficient memory management, will be crucial for deploying these models in resource-constrained environments.

In summary, the journey from theory to implementation has shown promising results in enhancing 3D reconstruction from 2D images using deep learning. Future research endeavors aim to push the boundaries of accuracy, efficiency, and applicability across diverse domains, ultimately fostering innovations that impact industries ranging from healthcare to entertainment and beyond.

References

[1] Ahmed J. Afifi, Jannes Magnusson, Toufique A. Soomro, and Olaf Hellwich, Pixel2Point: 3D Object Reconstruction from a Single Image Using CNN and Initial Sphere, IEEE Access, vol. 8, pp. 123456-123465, 2020 .

[2] Zhixiang Ye, Qinghao Hu, Tianli Zhao, Wangping Zhou, and Jian Cheng, MCUNeRF: Packing NeRF into an MCU with 1MB Memory, in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 123-130 .

[3] ICICYTA 2023 Program, 3rd International Conference on Intelligent Cybernetics Technology Applications (ICICyTA), 2023 .

[4] Chaitanya Bhure, Geraldine Shirley Nicholas, Shajib Ghosh, Navid Asadi, and Fareena Saqib, AutoDetect: Novel Autoencoding Architecture for Counterfeit IC Detection, Journal of Hardware and Systems Security, vol. 8, no. 1, pp. 45-56, 2024 .

[5] Sepp Hochreiter and Jürgen Schmidhuber, Long Short-Term Memory, Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997 .

[6] Weibin Liu, Weiwei Xing, Baozong Yuan, and Xiaofang Tang, Superquadric-based 3D Scene Reconstruction and Interpretation, in Proceedings of the 8th International Conference on Signal Processing, 2006, pp. 1123-1128 .

[7] Mobile Radio Communications and 5G Networks, Springer Science and Business Media LLC, 2024 .

[8] Computer Vision – ECCV 2018, Springer Science and Business Media LLC, 2018 .

[9] Emerging Research in Electronics, Computer Science and Technology, Springer Science and Business Media LLC, 2019 .

[10] Submitted to B.V. B College of Engineering and Technology, Hubli, Student Paper .