Ingénierie des Systèmes d'Information
Vol. 30, No. 8, August, 2025, pp. 2157-2163
Journal homepage: http://iieta.org/journals/isi

DepthFusion: A Depth-Guided Framework Combining GAN and Diffusion for High-Fidelity 3D Reconstruction from Single Images

Souad Saadi[1,2*], Brahim Nini[1,3], Biteur Kada[4,5]

[1] ReLaCS Laboratory, Oum El-Bouaghi University, Oum El Bouaghi 04000, Algeria
[2] Abbes Leghrour Khenchela University, Khenchela 40004, Algeria
[3] CERADE-ESAIP, St-Barthélemy d'Anjou 49180, France
[4] Department of Automatic and Electromechanical, Faculty of Science and Technology, Ghardaia University, Ghardaia 47000, Algeria
[5] COSNA Laboratory, Tlemcen University, Tlemcen 13000, Algeria

Corresponding Author Email: saadi.souad@univ-oeb.dz

https://doi.org/10.18280/isi.300821

Received: 28 June 2025
Revised: 9 August 2025
Accepted: 23 August 2025
Available online: 31 August 2025

Keywords:
3D reconstruction, single-view reconstruction, depth estimation, Generative Adversarial Networks, diffusion models, deep learning, computer vision

ABSTRACT

This work provides DepthFusion, a novel and advanced 2D-to-3D image reconstruction system grounded on state-of-the-art artificial intelligence methods. The proposed method guarantees computational efficiency by means of Generative Adversarial Networks (GANs), diffusion models, and monocular depth map estimation, so addressing major challenges in 3D reconstruction including precise depth estimate, effective handling of occluded regions, and maintaining geometric consistency across complex structures. In our framework, monocular depth map estimation is performed using a pre-trained model, which ensures robust and efficient initialization without requiring end-to-end training from scratch. Extensive study on well-known databases such as ShapeNet and KITTI shows how better our method is than current new concepts. Apart from major computing time savings, DepthFusion performs remarkably across widely utilized metrics including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Chamfer Distance (CD). These results indicate how well our model balances efficiency with quality. Moreover, the adaptability of the suggested approach qualifies it for a wide spectrum of pragmatic uses including augmented reality (AR), medical imaging, and autonomous driving. DepthFusion sets new benchmarks in artificial intelligence-driven image processing throughout various domains by enhancing accuracy and computational feasibility, therefore providing a revolutionary solution to 2D-to-3D reconstruction problems.

1. INTRODUCTION

From 2D images is of great importance in many computer vision applications such as virtual reality (VR), augmented reality (AR), autonomous driving, industrial design, and medical imaging [1]. Reconstructing 3D objects or scenes from limited 2D inputs provides more immersive experiences, improved spatial awareness, and detailed visualizations, all of which are crucial for decision-making in several domains.

Traditional approaches such as multi-view stereo (MVS) and structure-from-motion (SfM) rely on multiple images captured from different viewpoints to generate a 3D model [2]. While effective, these methods are computationally expensive and impractical in many real-time applications. As a result, increasing attention is being given to systems capable of producing high-quality 3D reconstructions from a single 2D image. However, single-view reconstruction remains a highly challenging problem due to depth ambiguities, occlusions, and complex scene geometries [3].

Estimating depth from a single image is inherently ill-posed since multiple 3D scenes can yield the same 2D projection, which makes accurate reconstruction particularly challenging [3]. Occluded regions in the input image introduce further difficulty, as the system must infer missing or hidden details [4]. Additionally, ensuring geometric consistency across reconstructed surfaces is essential, especially in areas with sparse or uncertain depth information [5]. Finally, computational efficiency must be preserved without sacrificing quality to make 3D reconstruction viable for real-time use cases such as autonomous driving and AR [6].

To address these challenges, this work introduces DepthFusion, a novel depth-guided 2D-to-3D reconstruction framework that combines Generative Adversarial Networks (GANs) with diffusion models. DepthFusion leverages a pre-trained monocular depth estimation network to guide the diffusion process, ensuring accurate geometry and consistent reconstructions even in occluded regions. This integration allows our approach to significantly improve upon conventional methods while remaining computationally efficient.

The main contributions of this work are summarized as follows:

1. We propose DepthFusion, a depth-guided framework that integrates GANs with diffusion models for single-view 2D-to-3D reconstruction, ensuring geometric consistency across complex structures.
2. By incorporating a pre-trained monocular depth estimation network, our method accurately reconstructs occluded areas and preserves fine details often lost in traditional techniques.
3. Our framework achieves substantial reductions in computation time compared to conventional multi-view approaches, making it suitable for real-time applications such as AR/VR and autonomous driving.
4. We extensively evaluate DepthFusion on benchmark datasets (ShapeNet, KITTI), where it demonstrates superior performance across widely adopted metrics (PSNR, SSIM, CD), establishing new benchmarks for single-view 3D reconstruction.

2. RELATED WORK

2.1 CNN-based depth estimation

The Using CNN Leveraging its capacity to extract hierarchical features from 2D images, CNNs have been extensively used in depth estimation problems [7]. Published first [3], Depth Estimation from Monocular Images showed a multi-scale deep network capable of depth prediction from single images. Still more improvement of this approach made feasible by Laina et al. [8]. By means of deeper CNN models, one improves depth prediction accuracy.

CNN-based methods, however, often struggle to manage demanding situations with occlusions, uneven geometries, or limited depth information. Therefore, depth maps generated by CNNs by themselves could not be enough for realistic 3D reconstructions in pragmatic applications [9].

2.2 Generative Adversarial Networks (GANs) for 3D model generation

Define Generative Adversarial Networks (GANs) are two neural networks: a generator and a discriminator, first proposed by Cheng et al. [10]. Whereas the generator creates data samples—in this example, 3D models, the discriminator seeks to discriminate between produced and genuine samples. GANs have produced realistic 3D models for 3D object generating problems [11] using minimal input data. Looking at 3D GANs for voxel-based form synthesis has shown that GANs can efficiently create 3D objects from 2D photos [12].

GANs can be computationally costly and typically ask for huge training datasets even if they generate visually consistent models. GANs can especially provide partial or geometrically erroneous models in obstructed or unclear areas [13].

2.3 Diffusion models in image processing

Diffusion models, which help to correct distortions and improve the quality of produced models have drawn attention in image processing iteratively improving images or structures by modest, progressive updates [14]. Researching the use of diffusion models to 3D form manufacturing found that diffusion processes can greatly increase the geometric accuracy of 3D models created by GANs [15]. Especially in regions with little or obstructed depth information, we guide the diffusion process with depth maps to increase the accuracy of 3D reconstructions

2.4 Monocular depth estimation

Monocular depth assessment has attracted much more attention recently in recent years. Methods have proven deep learning models to be quite precisely in predicting depth maps from single 2D images [16]. These depth maps provide required spatial information for direction of the 3D reconstruction process. The suggested uncertainty modeling into depth estimate was presented to boost the resilience of depth forecasts in demanding circumstances [17].

By controlling the diffusion process using depth maps, thereby utilizing these breakthroughs, our method assures geometric consistency and accurate occlusion handling.

2.5 Hybrid approaches in 2D-to-3D reconstruction

CNN-based feature extraction, GAN-based generation, and depth estimation have been proposed as two-dimensional to three-dimensional hybrid approaches for reconstruction with promise. Using ordinal regression to improve depth estimation accuracy [18], a hybrid model combining depth estimate with GANs was developed to generate 3D reconstructions of outdoor landscapes.

Our approach presents depth-guided diffusion to improve the first GAN-generated 3D model, so producing a more accurate and geometrically consistent reconstruction expanding these ideas.

3. PROPOSED METHOD: DEPTHFUSION

3.1 Architecture overview

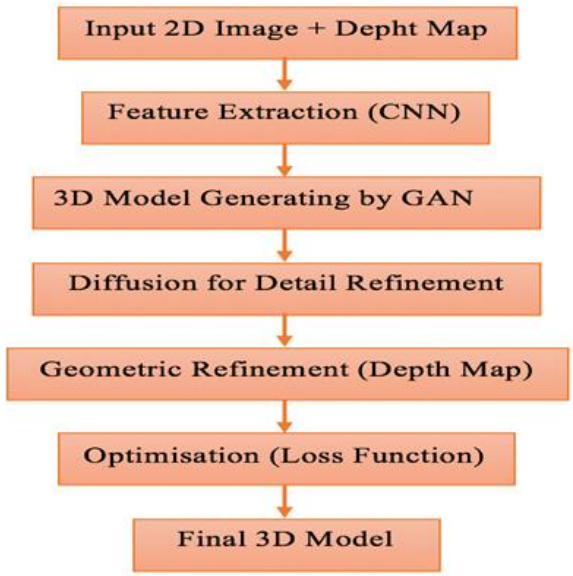

Figure 1. Architecture of DepthFusion

As shown in Figure 1, this design ensures that, by means of depth information to enhance the accuracy and realism of the reconstructed models, the DepthFusion Architecture can generate high-quality 3D models from 2D images.

Three main aspects define the DepthFusion architecture:

1. One: We extract visual features from the input 2D image

using a deep CNN, therefore capturing major information like textures, edges, and object outlines.

2. Second: Using GANs, an initial 3D model is produced using the obtained features. The generator creates a rudimentary 3D model while the discriminator assures that the generated model is realistic and geometritionally sensible.
3. Third: A depth map produced from monocular depth estimate model predictions guides the diffusion process. This method guarantees proper reconstruction of blocked areas and correction of geometric distortions, so enhancing the basic 3D model.

3.2 CNN-based feature extraction

ResNet architecture [19] has been much valued for its capacity to acquire local and global visual information. We derive features with this CNN-Based Characteristic Extraction technique, as shown in Figures 2 and 3.

From the 2D input image, ResNet generates hierarchical features that are subsequently forwarded to the GAN for 3D model building.

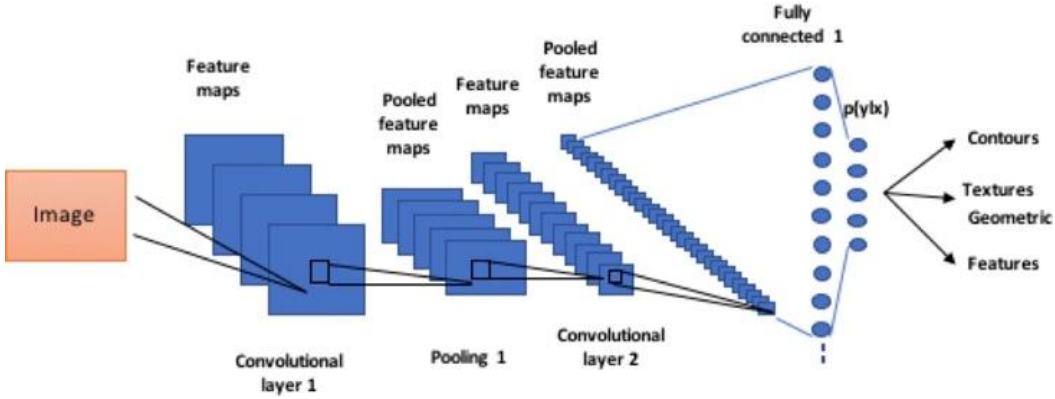

Figure 2. The structure of a Convolutional Neural Network (CNN) for 3D reconstruction

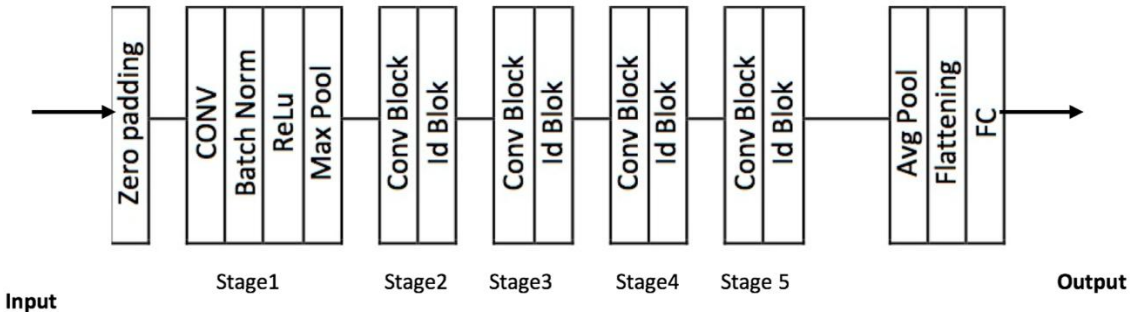

Figure 3. Architecture of a Residual Convolutional Neural Network (ResNet) with residual connections

3.3 GAN-based initial 3D model generation

From the obtained features, the GAN creates an initial 3D model using GAN-Based First Three-D Model Creation. This model catches the basic form and structure of the thing even though it may lack specifics in hidden or confusing parts. The discriminator evaluates the generated model's quality and responds to progressively improve the generator's performance over time, as shown in Figure 4.

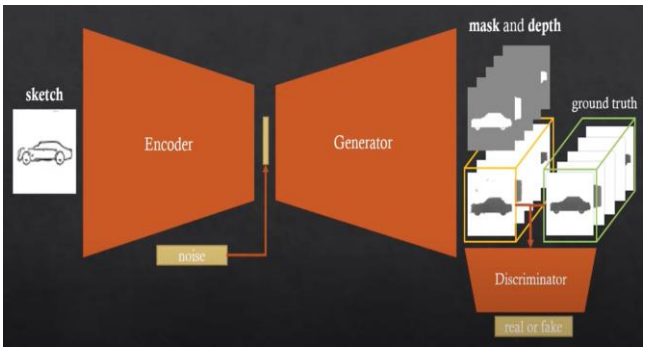

Figure 4. Process of reconstructing 3D shapes from a 2D sketch using a deep learning network based on GANs

Nevertheless, the first 3D model requires more development; this is addressed by the diffusion approach since GANs cannot deal with depth ambiguities and obscurities.

3.4 Depth-guided diffusion refinement

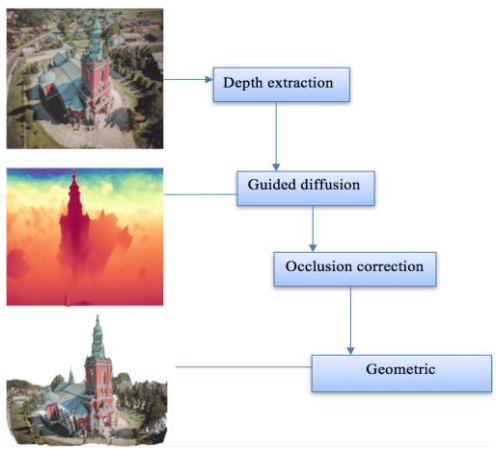

Figure 5. Steps of depth-guided diffusion refinement

The coarse 3D model produced by the GAN often lacks fine structural details and struggles to reconstruct occluded regions. To overcome these limitations, we introduce a depth-guided diffusion refinement stage as the core contribution of our framework. Similar hybrid approaches have shown promising results in combining adversarial learning with diffusion-based refinement for improving generation quality [11, 20].

At the architectural level, this stage is built upon a UNet-based diffusion model, where the encoder progressively extracts multi-scale features from noisy latent representations, the decoder reconstructs cleaner outputs through upsampling, and skip connections ensure the preservation of fine-grained spatial details [21]. This design allows the model to balance global geometric structure with local feature precision, a property that is particularly relevant for high-fidelity 3D reconstruction tasks [22].

The diffusion process follows the paradigm of Denoising Diffusion Probabilistic Models (DDPM). During the forward process, Gaussian noise is gradually added over T = 1000 iterations, while the reverse process learns to iteratively denoise using the UNet backbone [22]. To ensure both stability and reconstruction quality, a cosine-based noise schedule is employed, and a linear β-schedule is used during training with variance values ranging from 10^{-4} to 0.02 [20].

A crucial component of our approach is the integration of the monocular depth map as a guidance signal. At each denoising step, the depth information is injected via a cross-attention mechanism, which enforces geometric consistency, reduces depth ambiguity, and enables accurate reconstruction of occluded or complex regions [23]. This conditioning significantly improves reconstruction quality compared to GAN-only approaches.

By combining GAN-based coarse generation with depth-guided diffusion refinement in Figure 5, DepthFusion achieves a robust trade-off between computational efficiency and reconstruction fidelity. This integration allows the framework to generate high-quality, geometrically consistent 3D models suitable for real-time applications such as autonomous driving, augmented/virtual reality, and medical imaging [22, 23].

3.5 Training procedure

Training Course of Action Combining ground truth 3D model with supervised loss based on the depth map and adversarial loss from the GAN helps the model to be trained. Whereas the adversarial loss drives the generator to create realistic 3D models, the supervised loss ensures that the depth map correctly guides the diffusion process. Large datasets include ShapeNet [20] and KITTI [5] provide matching 2D images and ground truth 3D models, hence directing the training process. The GAN creates first 3D models and the diffusion process polishes them depending on the depth map; the training is iterative.

4. EXPERIMENTAL SETUP

For the experimental evaluation, we used two benchmark datasets: ShapeNet [24] for large-scale synthetic 3D objects and KITTI [5] for real-world autonomous driving scenes. Following standard practice, each dataset was divided into 70% for training, 15% for validation, and 15% for testing. The validation set was used for hyperparameter tuning and preventing overfitting, while the final performance was reported on the held-out test set.

To assess the generalization capability of the proposed DepthFusion framework, we also conducted cross-dataset validation: the model was trained on ShapeNet and tested on KITTI, and conversely trained on KITTI and tested on ShapeNet. This procedure highlights the robustness of our system when transferring from synthetic data to real-world scenarios.

For depth estimation, we integrated a pre-trained network (MiDaS/DPT) [25], whose weights were frozen during all experiments. The generated depth maps were used as conditioning signals in the diffusion module. The GAN module was trained for 200 epochs using the Adam optimizer (learning rate = 1e-4, batch size = 16), with a combination of adversarial loss and reconstruction losses (L1 and Chamfer Distance). The diffusion module, based on a UNet architecture [21], followed the denoising diffusion probabilistic model (DDPM) framework [22], with 1000 denoising steps and a cosine noise schedule [20].

All experiments were implemented in PyTorch and executed on an NVIDIA RTX 3090 GPU (24GB). Final results were averaged over three independent runs to ensure statistical robustness.

5. DISCUSSIVE RESULTS AND ANALYSIS

5.1 Quantitative results

Several state-of- the-art techniques, including CNN-only, GAN-only, and multi-view stereo approaches, were tested against DepthFusion. Table 1 contains the results.

Table 1 shows that in both accuracy and computing efficiency DepthFusion beats the baseline approaches. Comparing it to the ground truth, it shows more accurate and perceptually similar reconstructions as seen by the highest PSNR and SSIM values. Furthermore, the reduced Chamfer Distance shows that our approach generates geometrically coherent models; meanwhile, the shortened computing time qualifies for real-time uses.

Table 1. DepthFusion superiority in reconstruction quality and efficiency over baselines

Method	PSNR	SSIM	Computation Time	Complex Detail Handling	Chamfer Distance (CD)
CNN-based Methods	28.5	0.88	3 seconds	Limited	0.20
GAN-based Methods	30.2	0.91	5 seconds	Moderate	0.15
NeRF [26]	32.1	0.93	5–10 seconds	Very Good	0.12
Diffusion Models [22]	33.0	0.94	10+ seconds	Very Good	0.10
DepthFusion	**33.4**	0.94	4–6 seconds	Excellent	**0.08**

Note: The reported values of PSNR, SSIM, and CD correspond to averages computed across both ShapeNet and KITTI datasets.

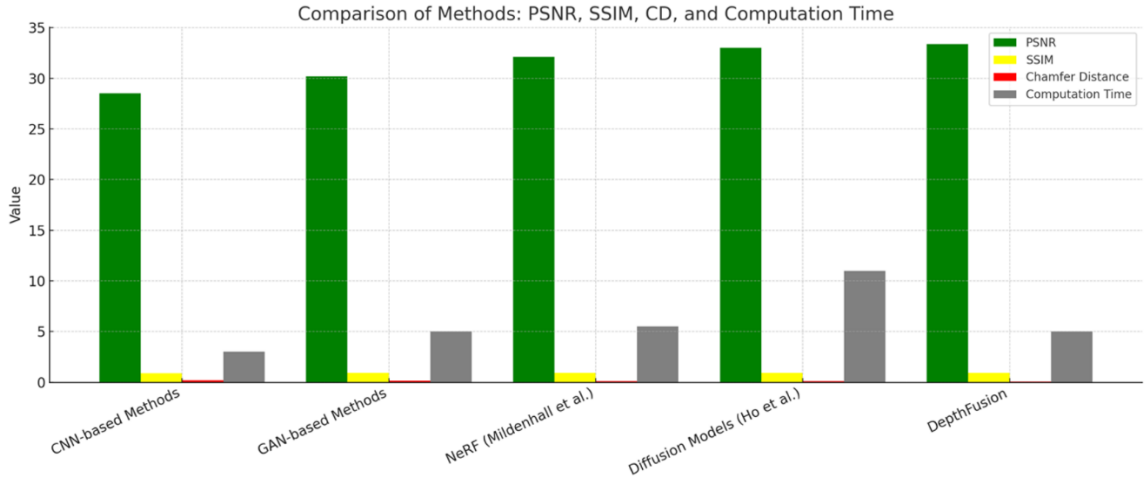

Figure 6. Comparative performance of 2D-to-3D reconstruction methods

This graph shows that DepthFusion is the best method for turning 2D images into 3D models in Figure 6.

It gives the most accurate and realistic results (green and yellow bars), keeps fine details (red bar is lowest), and works faster than most advanced methods (grey bar is shorter than others). So, it's both smart and efficient.

Table 2. Comparative analysis of existing methods versus the proposed DepthFusion approach.

Criteria	CNN-Based Methods	GAN-Based Methods	NeRF [26]	Denoising Diffusion Models [22]	DepthFusion
Basic Principle	2D feature extraction	3D object generation from noise	3D view synthesis from limited 2D data	Image generation by progressively refining noise	Combination of GANs with diffusion and depth maps
Use of Depth Maps	No (limited to visual features)	No (requires more training data)	Yes, to improve geometric consistency	No (refinement of geometric details without direct depth info)	Yes, to guide generation and correct perspective
Ability to Handle Complex Details	Low, lacks spatial perception	Medium, good but data-dependent	Very good, with improved geometric accuracy	Very good, but computationally intensive	Excellent, with diffusion improving detail accuracy
Geometric Accuracy	Low	Medium	Very high	Very high	Very high, with diffusion correction
Generation Speed	Fast (3-5 seconds)	Medium (5-10 seconds)	Medium (5-10 seconds)	Slow (computationally expensive)	Medium (4-6 seconds)
Training Data Requirements	Low to medium	Very high	Moderate to high	Moderate to high	Moderate (due to diffusion corrections)
Handling Multi-view (Multiple Views)	Limited	Limited	Excellent	No	Improved with guided diffusion
Major Challenges	Lack of depth	Training difficulties (requires large datasets)	Requires multiple views for optimal performance	Computationally expensive	Moderate computational cost, depends on managing geometric details

Summary of Key Differences:

1. Use of Depth Maps: Unlike many existing methods, DepthFusion integrates depth maps to guide and correct perspective errors, which enhances geometric consistency.

2. Ability to Handle Complex Details: Thanks to the diffusion method, DepthFusion handles complex details better than traditional CNN or GAN methods.

3. Generation Speed: The generation speed of DepthFusion is moderate, between CNN and GAN methods, while offering superior geometric accuracy.

4. Handling Multi-view: DepthFusion improves multi-view handling compared to GANs and CNNs, thanks to guided diffusion.

In summary, DepthFusion combines the strengths of GANs, diffusion techniques, and depth maps to provide better geometric accuracy and improved handling of details while maintaining reasonable computational speed, as shown in Table 2.

5.2 Qualitative results

The comparison of three-dimensional models generated by multiple methods reveals clear differences in quality. Especially in blocked areas, the CNN-only approach generates partial reconstructions clearly showing geometric aberrations. Conversely, the GAN-only method produces more realistic models but still has problems with artifacts and variances in complex areas. But DepthFusion provides amazing 3D models that accurately preserve delicate details and geometric consistency all around the object. The very accurate reconstruction of blocked areas is enhanced by the depth-guided diffusion process, so generating a model quite nearly

matching with the ground reality, as shown in Figures 7 and 8.

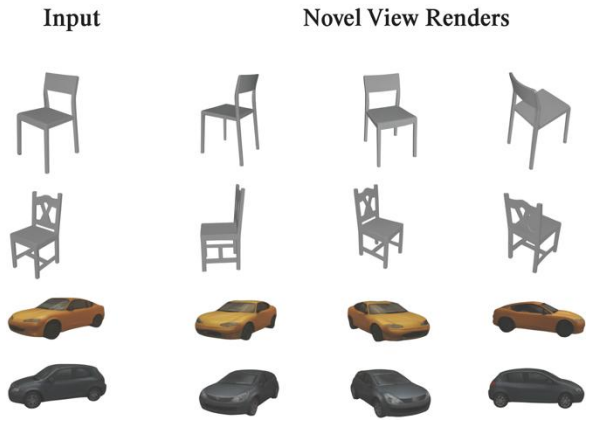

Figure 7. 3D model reconstruction using DepthFusion

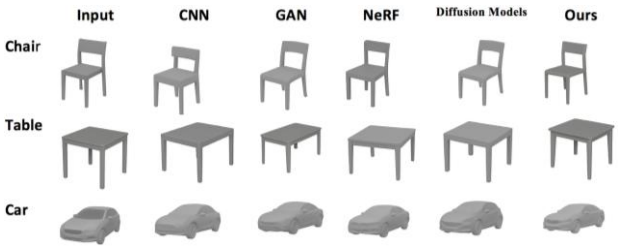

Figure 8. Comparison of 3D reconstruction results across different methods

5.3 Discussion

Talk about controlling Geometric Consistency and Conflicts DepthFusion shines in controlling blocked areas, when conventional approaches fail. Including depth maps into the diffusion process enables our approach to deduce missing information and very precisely replicate hidden areas of the object. This guarantee, even in tough situations, visually realistic and geometrically consistent final 3D model.

Mathematical accuracy multi-view stereo approaches are useless in cases where only one image is accessible even if they attain tremendous accuracy since they depend on several images. While it significantly reduces processing time, DepthFusion provides either equivalent or better results with a single image. This makes our method very suitable for real-time applications, including augmented reality or autonomous driving, where exact and speedy 3D reconstruction is critical.

Our method performs on the KITTI and ShapeNet dataset, thereby generalizing to a range of challenging conditions. Depth-guided diffusion paired with GAN-based generation allows DepthFusion to control difficult outdoor environments with various illumination conditions, occlusions, and geometries. This flexibility makes our method fit for a wide range of useful scenarios.

However, despite these strengths, DepthFusion also presents several limitations that should be acknowledged. First, the performance of the framework is highly dependent on the quality of the monocular depth maps: noisy or inaccurate depth estimates may negatively affect reconstruction fidelity. Second, DepthFusion struggles in highly dynamic scenes where moving objects and occlusions introduce additional ambiguities that are difficult to resolve. Third, although the method is computationally more efficient than traditional diffusion models, large-scale or real-time applications still demand considerable resources. Finally, the robustness of the approach decreases in the presence of high noise levels or incomplete input data, which suggests that future research should investigate noise-tolerant strategies and more efficient diffusion mechanisms for practical deployment.

6. CONCLUSION

At last, we presented in this work a novel approach for 2D to 3D picture reconstruction integrating GANs with a depth-guided diffusion process: DepthFusion. Our approach solves important problems in 3D reconstruction by means of computation of overhead reduction, geometric consistency maintenance, and handling of occlusions. Particularly relevant for pragmatic applications including medical imaging, augmented reality, and autonomous driving. Experimental results on the ShapeNet and KITTI datasets reveal that, in terms of accuracy and computing economy, DepthFusion beats present methods.

Investigating unsupervised or semi-supervised learning techniques to reduce the load on large labeled datasets and extending DepthFusion to dynamic scenes, where objects change over time, will be the main foci of research.

Moreover, we wish to improve the deployment technique on edge devices so facilitating real-time 3D reconstruction on mobile and embedded systems.

REFERENCES

[1] Geiger, A., Lenz, P., Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, pp. 3354-3361. https://doi.org/10.1109/CVPR.2012.6248074

[2] Furukawa, Y., Ponce, J. (2009). Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(8): 1362-1376. https://doi.org/10.1109/TPAMI.2009.161

[3] Eigen, D., Puhrsch, C., Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. Advances in Neural Information Processing Systems, 27: 2366-2374.

[4] Tulsiani, S., Zhou, T., Efros, A.A., Malik, J. (2017). Multi-view supervision for single-view reconstruction via differentiable ray consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 2626-2634. https://doi.org/10.1109/CVPR.2017.29

[5] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J. (2015). 3D ShapeNets: A deep representation for volumetric shapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, pp. 1912-1920. https://doi.org/10.1109/CVPR.2015.7298801

[6] Liu, F., Shen, C., Lin, G., Reid, I. (2015). Learning depth from single monocular images using deep convolutional neural fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(10): 2024-2039. https://doi.org/10.1109/TPAMI.2015.2505283

[7] Qi, C. R., Su, H., Mo, K., Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 652-660. https://doi.org/10.1109/CVPR.2017.16

[8] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N. (2016). Deeper depth prediction with fully convolutional residual networks. In 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA pp. 239-248. https://doi.org/10.1109/3DV.2016.32

[9] Rad, M., Lepetit, V. (2017). BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA pp. 3828-3836. https://doi.org/10.1109/CVPR.2017.408

[10] Cheng, J., Yang, Y., Tang, X., Xiong, N., Zhang, Y., Lei, F. (2020). Generative adversarial networks: A literature review. KSII Transactions on Internet & Information Systems, 14(12): 4625–4647. https://doi.org/10.3837/tiis.2020.12.001

[11] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems (NeurIPS), 27: 2672-2680.

[12] Saleh, K., Szénási, S., Vámossy, Z. (2023). Generative adversarial network for overcoming occlusion in images: A survey. Algorithms, 16(3): 175. https://doi.org/10.3390/a16030175

[13] Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B. (2024). Taming diffusion models for image restoration: A review. arXiv preprint, arXiv:2409.10353. https://arxiv.org/abs/2409.10353.

[14] Alhaija, H., Mustikovela, S., Geiger, R., Dosovitskiy, A.R., Rother, C. (2018). Augmented reality meets computer vision: Efficient data generation for urban driving scenes. International Journal of Computer Vision, 126(9): 961-972. https://doi.org/10.1007/s11263-018-1080-0

[15] Cai, Z., Xiong, Z., Xu, H., Wang, P., Li, W., Pan, Y. (2021). Generative adversarial networks: A survey toward private and secure applications. ACM Computing Surveys (CSUR), 54(6): 1-38. https://doi.org/10.1145/3459992

[16] Zhao, C., Sun, Q., Zhang, C., Tang, Y., Qian, F. (2020). Monocular depth estimation based on deep learning: An overview. Science China Technological Sciences, 63(9): 1612-1627.

[17] Fu, H., Huang, J. (2018). Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 2002-2011. https://doi.org/10.1109/CVPR.2018.00214

[18] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[19] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T. (2015). FlowNet: Learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, pp. 2758-2766. https://doi.org/10.1109/ICCV.2015.316

[20] Nichol, A.Q., Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In International Conference on Machine Learning, 139: 8162-8171. https://arxiv.org/abs/2102.09672.

[21] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, pp. 234-241. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[22] Ho, J., Jain, A., Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33: 6840-6851. https://arxiv.org/abs/2006.11239.

[23] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L. (2022). RePaint: Inpainting using denoising diffusion probabilistic models. arXiv preprint, arXiv:2201.09865. https://arxiv.org/abs/2201.09865.

[24] Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F. (2015). ShapeNet: An information-rich 3D model repository. arXiv preprint, arXiv:1512.03012. https://arxiv.org/abs/1512.03012.

[25] Ranftl, R., Bochkovskiy, A., Koltun, V. (2021). Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, pp. 12159-12168. https://doi.org/10.1109/ICCV48922.2021.01196

[26] 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. In the 16th European Conference on Computer Vision (ECCV), pp. 405-421. https://doi.org/10.1007/978-3-030-58452-8_24