

A Decade of Deep Learning-Based 2D-to-3D Image Reconstruction: Taxonomy, Challenges, and Emerging Paradigms

Suchismita Mohapatra¹, Mohit Ranjan Panda²

¹*School of Computer Engineering, Kalinga Institute of Industrial Technology, Deemed to be University, 751024 Bhubaneswar, India, Email: 2081037@kiit.ac.in*

²*School of Computer Engineering, Kalinga Institute of Industrial Technology, Deemed to be University, 751024 Bhubaneswar, India Email: mohit.pandafcs@kiit.ac.in*

ABSTRACT

In recent years, two-dimensional to three-dimensional (2D-to-3D) image reconstruction has become one of the main research problems in the field of computer vision due to the development of deep learning. The survey will overview the development of key learning paradigms in the reconstruction of 2D-to-3D (convolutional neural networks (CNNs), generative adversarial networks (GANs), support vector machines (SVMs), and how they are used in hybrid reconstruction models). The methods reviewed are discussed concerning the ability to extract features, the ability to perceive and the ability to compute. Special focus is given to hybrid systems which comprise CNN-based learning of spatial representations and GAN-based learning of generative improvement and SVM-based learning of discriminative decision support to enhance fidelity to replication. Such important issues as depth ambiguity, scalability, and cross dataset generalization are discussed systematically. The survey also identifies the emerging paradigms, such as Neural Radiance Fields

(NeRFs), Vision Transformers (ViTs), and diffusion-based models that have a high potential to represent the implicit aspects of scenes and generate high-fidelity 3D scenes. In general, this review seeks to offer a systematic reference to any future research on interpretable, scalable and robust 3D vision systems.

Keywords: 2D-to-3D Image Conversion; Hybrid Deep Learning; Convolutional Neural Networks (CNN); Generative Adversarial Networks (GAN); Support Vector Machines (SVM); Neural Radiance Fields (NeRF); Vision Transformers (ViT); Diffusion Models; 3D Reconstruction; Explainable AI (XAI).

INTRODUCTION

Two-dimensional (2D) image reconstitution into three-dimensional (3D) models is one of the most complicated and dynamic problems of contemporary computer vision [34,66,95, 10]. The importance of this task is general to autonomous robotics [105,104], medical imaging [69,26,97,49], virtual and augmented reality [51,42,24], computer-aided design [73,70], and digital entertainment [96]. Basically, it involves the retrieval of lost depth data and spatial structure of ostensibly two-dimensional visual data that necessitates algorithms that can extrapolate geometry and structure in the absence of abundant visual evidence. There have been multiple detailed surveys that have studied deep learning-based 3D vision under varying viewpoints. Zhang et al. [66] reviewed a reliable number of 3D object detections as a result of deep learning based on a variety of sensor modalities, and Samavati and Soryani [59] reviewed 3D reconstruction with deep neural networks. Muzahid et al. [37] surveyed the deep learning technique in 3D object recognition, and Hussain et al. [65] focused on the domain of Generative Adversarial Networks (GANs) in medical image reconstruction. Nonetheless, these reviews discuss either 3D reconstruction or classification separately, or they focus on domain-specific implementations. The recent advancement of learning-based 3D vision methods is also exemplified by complementary surveys by Han et al. [34] and Xu et al. [95]. This review addresses this gap by presenting a coherent analysis of hybrid CNN-GAN-SVM

architectures for 2D-to-3D reconstruction. The paper compares and contrasts the approaches of reconstruction and classification, although it does not confuse the conceptual differences between the two. One of the biggest problems in 2D-to-3D conversion is the accurate inference of 3D geometry and spatial consistency [62,47,56,76,1] with low-resolution 2D input data. This issue has a number of subproblems:

- Monocular or limited-view images, in which depth cues are typically incomplete or ambiguous: depth estimation, as well as single-view reconstruction has been a fundamentally ill-posed problem [35,34,56,99];
 - Third-dimensional geometry reconstruction maintaining the surface continuity and structure;
 - Texturing and detail conservation of the dimensional transformation process; and
 - Dealing with occlusions and viewpoint changes in the complicated real world scenes.
- Despite the fact that the field of deep learning has shown a profound improvement in performance in these aspects, current models have certain challenges in domain generalization, computational cost of real-time processing, and stability across different environmental conditions [27,4]

In order to counter these shortcomings, the survey offers a systematic analysis of hybrid deep learning systems comprising of Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Support Vector Machines (SVMs). The idea of using their complementary advantages where CNNs extract the features, GANs generate realistic images, and SVMs classify and refine them has been used to show how the hybrid systems can enhance preciseness, interpretability, and efficiency in the reconstruction of 2D-to-3D images.

0.1 Contributions of This Survey

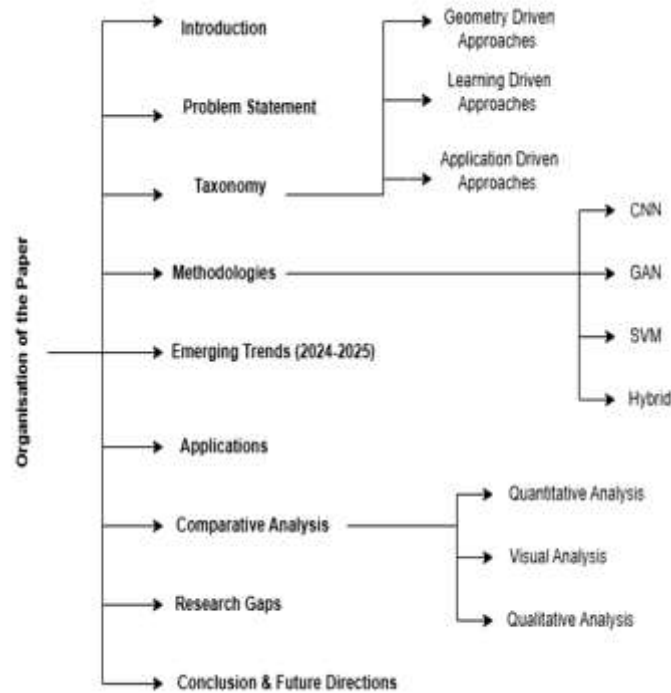
The current paper provides a unified and more modern view of the hybrid deep learning systems of 2D-to-3D conversion. Its major contributions may be summed up as the following:

- Comprehensive hybrid analysis: This paper critically analyzes CNN, GAN, and SVM architecture and its combination into hybrid models to perform 2D-to-3D reconstruction.
- Comparative evaluation: The paper determines the relative merits and drawbacks of individual and hybrid models according to the following criteria: reconstruction accuracy, the complexity of computation, and generalization ability.
- Identification of research gaps: The survey shows that there are still enduring problems concerning the occlusion handling, depth ambiguity, data scarcity, and interpretability of the model.
- Emerging paradigm integration: The review relates the hybrid architectures to the latest paradigm advances like Neural Radiance Fields (NeRFs), Vision Transformers (ViTs), and Diffusion Models and their role in changing the direction of research on 3D reconstruction [57].
- Future-oriented guidance: The paper ends by providing future research directions that could advance to energy-efficient, explainable, and real-time 3D vision systems that could be implemented in edge devices.

These submissions put this survey in the position of a complete resource to be used by researchers and practitioners in the fields of deep learning, computer vision, and 3D modeling [78].

0.2 Organization of the Paper

Figure 1. Hierarchical taxonomy of 2D-to-3D conversion approaches illustrating geometry-driven, learning-driven, and hybrid paradigms



The overall structure of this review is illustrated in Figure 1. The paper begins with Section 1, which introduces the motivation, background, and objectives of the study, followed by Section 2, where the research problem is defined and the scope of the survey is clearly established. Section 3 presents a detailed taxonomy of 2D-to-3D image conversion approaches, categorizing them into geometry-driven, learning-driven, and application-driven methodologies. The main architectures and hybrid models (convolutional neural networks (CNNs), generative adversarial networks (GANs), and support vector machines (SVMs)) are also described in detail in Section 4. Section 5 uses the 3D reconstruction application of deep learning to discuss its practical application in medical imaging, robotics, virtual reality, and digital modeling. Section 6 provides a comparative performance analysis, which consists of quantitative, qualitative, and visual analysis. Section 7 outlines open challenges and gaps in present research, which constrain contemporary methods, whereas the major contributions and the main conclusions of the review are presented in Section 8. Lastly, Section 9 presents the future research directions, which include building real-time, interpretable, and scalable hybrid models of next-generation 3D vision systems.

1. PROBLEM STATEMENT

Although astonishing progress has been achieved in 2D-to-3D image conversion with the advent of deep learning, the existing techniques still have multiple limitations that hinder their performance in terms of accuracy, scaling, and applicability in different real-world situations. The classic geometry-based methods tend to have issues in the areas of depth ambiguity, occlusion management, and light variability, which lead to partial or flawed 3D reconstructions. More expansive methods centered around learning are more flexible,

though they typically overfit and have poor generalization to environments that were not seen during training.

Convolutional Neural Network (CNN) models are themselves especially powerful at learning spatial and structural aspects but tend to produce over-smoothed results and cannot make depth work on large scales. Generative Adversarial Networks (GANs) are based on perceptually trained networks and vulnerable to instability mode collapse and geometric artifacts. SVMs are interpretable and computationally efficient but strongly rely on the quality of features extracted and cannot completely represent complicated spatial associations on their own.

The latest hybrid CNN-GAN-SVM models are still faced with choices between accuracy, cost of computation and robustness. Those systems usually require a huge amount of training and adjustment to ensure consistency in datasets of different visual and structural features.

Considering such difficulties, the urgent need is to have an all-encompassing survey on the study of the development of deep learning methods of 2D-to-3D conversion that is the most critical evaluation. This work attempts to answer that requirement by reviewing (in a systematic manner) the strengths of the existing architectures, their limitations, and how they interrelate in a systematic way. By doing this, it creates a structured taxonomy as well as offers analytical information that can be used to design the next-generation hybrid models that could realize real-time, high-fidelity, and interpretable 3D reconstruction.

2. TAXONOMY OF 2D-TO-3D CONVERSION APPROACHES

To offer a concise and systematized perspective of the techniques employed in the process of 2D-to-3D image conversion, the existing methods can be crudely categorized into three major groups, namely, geometry-driven approaches, learning-driven approaches, and hybrid approaches. The sublists where synthetic morphology fits are affected by the historical and methodological advancement of the field, moving away to modern data-intensive and integrative deep learning models. This classification is in line with recent surveys, as it is based on widely accepted classifications [34,66,95,10] and allows one to make a systematic comparison between the strengths, limitations, and application domains of each method.

2.1. Geometry-Driven Approaches

Geometry-based methods base their procedures on predicting depth and three-dimensional content, using explicit physical light sources of disparity, illumination, and motion. Early methods in 3D reconstruction methods are based on classical approaches such as stereo vision, structure-from-shading, and shape-from-motion [75,34,12]. These algorithms make use of geometrical concepts to deduce spatiality on the pixel of the image, usually resulting in a reliable result when imaging variables are held under strict control.

Nevertheless, they effectively deteriorate with the occurrence of scenes with occlusions, textless areas or depth discontinuities. Finally, lighting and surface reflectance variations also increase the difficulty in estimating depth to incomplete or distorted reconstructions. Furthermore, since such algorithms are constructed based on inflexible geometrical and photometric models, their flexibility to dynamically changing or complicated environments in the real world situation is restricted [35,47,56]. These restrictions notwithstanding, geometry-based approaches remain a nice baseline to assess

contemporary learning-based systems, offering interpretable metrics and geometrical information concerning the mechanics of the 3D reconstruction process.

2.2. Learning-Driven Approaches

The approaches based on learning use machine learning models to predict depth or create 3D structure directly on 2D images [106] .

CNN-based models, which produce hierarchical spatial features and depth using high structural precision [89,44,93,7,77] ;

GAN-based models, which improve the level of realism and refinement of texture with the help of adversarial learning [55,102,108,18,3,30] ;

SVM-based models, which can be used to research classification, refinement, and decision-level optimization in the 2D-to-3D pipeline [28,16,67,41] .

Generalization and reconstructive fidelity are enhanced by learning-based approaches, which need very high annotated datasets and considerable computational power [34,12,14].

2.3. Application-Driven Approaches

Application-oriented techniques focus on the application of 2D-to-3D conversion algorithms to particular real-world areas, including medical imaging, virtual reality (VR), robotics, and computer-aided design (CAD) [39] . In medical practice, an example is Shao et al. (2022), who applied a CNN-GAN hybrid model to classify lung tumors through CT scans and achieved a significant increase in the accuracy of the diagnosis [90] . Similarly, Xie et al. (2021) created a GAN-trained MRI super-resolution system that improved the clarity of the anatomies and facilitated the clinical interpretation [17] .

In addition to healthcare, the use of hybrid and learning models has been intensified in virtual and augmented reality realization, where 3D reconstruction of single images can be realized in an immersive and photorealistic manner [51,107,24,42] . Youm et al. (2023) showed how hybrid networks are capable of making realistic 3D worlds out of insufficient views of input but lower perception and simulation to the realm of interactivity. Such frameworks have been used to help depth perception, object manipulation, and object model generation tasks in robotics and CAD to design more adaptive and context-sensitive autonomous systems [55,58,53] .

On the whole, these domain-based implementations not only highlight the flexibility of hybrid deep learning models in structure but also necessitate a pragmatic view of deep learning with regard to 2D-to-3D reconstruction. Simultaneously, they demonstrate enduring limitations associated with real-time performance, explainability, and dataset specialization to which their large-scale use is confined. All in all, application-based research indicates that, although hybrid deep learning algorithms have been effectively scaled to real-world systems, the increase in computational efficiency, interpretability, and cross-domain approach has been one of the primary areas of exploration in the future.

3. DEEP LEARNING FRAMEWORKS FOR 2D-TO-3D CONVERSION

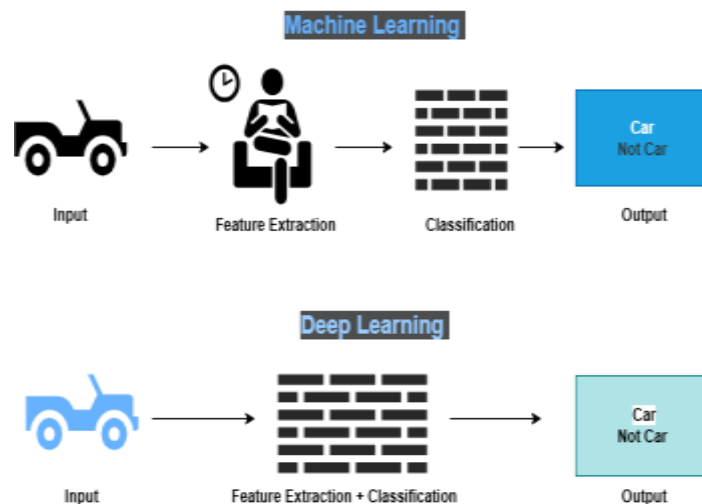
This part provides a summary of the most prominent deep learning networks—Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs) , and Support Vector Machines (SVMs), which constitute the party of modern 2D-to-3D image conversion and classification systems. The advantages and disadvantages of each approach in terms of methodology, comparative strengths, and internal limitations of each approach are presented to develop an understanding of future hybrid architectures. Machine Learning (ML) is an important subfield of artificial intelligence that

deals with empowering machines to learn after exposure to data and make a prediction without a clear set of instructions. Deep Learning (DL) has become a specialized subdiscipline in this area attempting to automate feature extraction and representation learning through hierarchical neural networks. Contrary to the traditional ML, in which handwritten features are defined by humans, DL models inherently explore multi-level features based on raw data, minimizing human input to a significant degree.

Deep learning has its conceptual origins in, verbatim, the initial work in the field of neural network research [23], specifically in the 1980s, the discovery of Geoffrey Hinton and his co-authors that learning was possible by spacing neurons out and interconnecting them to form networked structures. That is because each neuron can perform basic mathematical operations but, when combined together, can allow a network to design complex patterns by means of training and weight optimization.

Depending on the availability of data and the complexity of the task being undertaken, deep learning frameworks can be trained in a supervised, unsupervised, or semi-supervised learning configuration. The models have proven to be amazing in their ability to process various types of data, such as images, video, and volumetric data, and have been shown to be more accurate and scalable than conventional ML algorithms. When considering 2D-to-3D image conversion, noticeable among performance and recognized methods, the use of neural networks nowadays is essential, and it has the potential to predict the geometric depth, rebuild space relationships, and create photorealistic three-dimensional output with little human correction.

Figure 2. Comparison between Machine Learning and Deep Learning



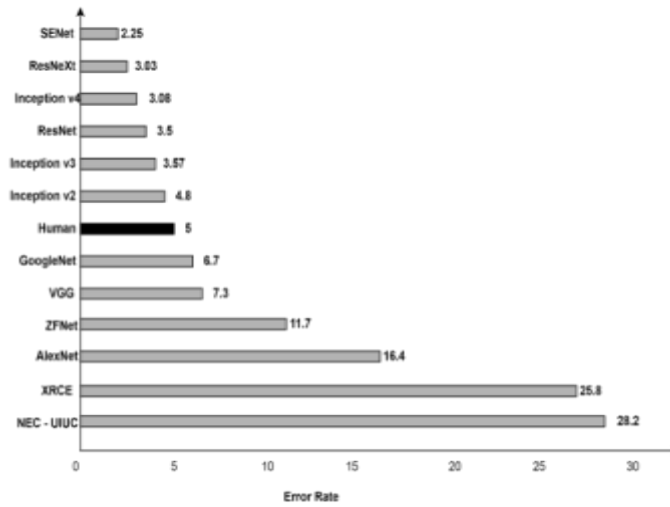
Consider building a system to classify images as either "Car" or "Not Car," as illustrated in Figure 2. In classical machine learning (ML), the system uses data to learn this classification. A human might label specific features indicative of a "car" (e.g., wheels) to aid in assessment. In contrast, a deep learning (DL) solution also attempts to autonomously identify which parts of the image constitute a car, such as wheels, wing mirrors, headlamps, and the windscreen. Consequently, DL reduces the amount of manual feature definition required, streamlining the dataset preparation process.

Table 1. Comparison between Machine Learning (ML) and Deep Learning (DL).

Feature	Machine Learning (ML)	Deep Learning (DL)
Definition	Uses algorithms to evaluate and learn from data in order to make predictions.	Uses neural networks to extract intricate patterns from vast amounts of data.
Data Requirement	Works well with structured, smaller datasets.	Requires massive datasets for high performance.
Feature Engineering	Manual feature selection required.	Automatically extracts features via neural networks.
Model Complexity	Uses simpler models like decision trees and SVM.	Uses deep architectures like CNNs and GANs.
Computational Power	Runs efficiently on CPUs.	Needs high-performance GPUs/TPUs.
Training Time	Trains faster due to lower complexity.	Training takes longer due to multiple layers.
Interpretability	More interpretable and explainable.	Harder to interpret because of its deep architecture.
Performance on Unstructured Data	Less effective when applied to images, video, or speech data.	Excels in image, video, and speech processing.
Best Use Cases	Fraud detection, recommendation systems, medical diagnostics.	Autonomous driving, medical imaging, natural language processing.
Example Models Used	SVM, Random Forest, Decision Trees, KNN.	CNN, GAN, RNN, Transformers, LSTMs.

From Table 1, we observe that ML models are more interpretable and computationally efficient, making them suitable for structured datasets with small-to-medium size. In contrast, DL models leverage deep neural architectures, excelling in unstructured data processing (e.g., image, speech, and video recognition) but requiring larger datasets and significant computing power.

Figure 3. Deep learning performance compared to human—ImageNet Top 5 Error Rate [87]



The ImageNet Top-5 Error Rates for several deep learning models and human performance are shown in Figure 3. SENet outperforms other models such as ResNeXt (3.03%) and Inception v4 (3.08%), achieving the lowest error rate (2.25%). The 5% human error rate is emphasized as a standard. The occurrence of errors in older models, including NEC-UIUC (28.2%) and AlexNet (16.4%), is significantly higher, which also proves the progress in deep learning structures. The data illustrated in Figure 3 illustrates that modern computer vision models are more accurate than human beings in terms of the modern versions provided [87].

Working of Deep Neural Network (DNN): The DNN classifier takes the form of a layered architecture of the interrelated neurons, with each having its own activation function and rectified linear unit (ReLU). The activation mechanism assists the network to determine complex trends in the input data through non-linearity. The DNN can be mathematically expressed as given below when it has SCMR-SVM input. The output layer is optimized using a cost function, often implemented via cross-entropy and a SoftMax function together. The ReLU activation function is characterized as follows:

$$f(x) = \max(0, x)$$

or equivalently,

$$f(x) = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

where x , acting as a ramp function, is the neuron input from FSVM. A smooth approximation of the ReLU is obtained by applying the *softplus* function:

$$f(x) = \ln(1 + e^x),$$

which acts as a continuously differentiable stand-in for the ReLU operator $\max(0, x)$. After the phase of prediction, the hidden layer representations are iteratively updated through the nonlinear transformation:

$$H^{(l)} = f(W^{(l)}H^{(l-1)} + B^{(l)}),$$

where H represents the activation function, $B^{(l)}$ indicates the bias of the l -th hidden layer, and $W^{(l)}$ represents the weight matrix. The ReLU activation function is employed to determine the parameters [52]. Critically, the FSVM preprocesses the input for the hidden

units instead of using the raw training or testing dataset directly. This preprocessing step allows the DNN to focus on higher-level feature extraction.

Table 2. Summary of Deep Learning Approaches and Solutions for 2D-to-3D Conversion and Image Analysis.

Sl. No.	Survey	Authors	Scheme	Issues	Solutions
1	The use of Deep CNN and Deep GAN in computer vision-driven analysis of images [2]	Usman Tariq, Kathiravam Sorinivasan, Chuan-Yu Chang, P.M. Durai Raj Vincent, R. Nandhini Abirami	DCNN & DGAN	Insufficient data and poor image quality generation.	Dropout and data augmentation to improve quality.
2	Automatic subcentimeter lung cancer classification using a combination of GANs and CNNs [90]	Cheng Shao, Lili Shi, Shuyi Yang, Zhiyong Zhang, Fei Shan, Lei Liu, Yunpeng Wang, Lingxiao Zhou, and Mingming Wang	TORGB (feature map transfer)	Performance limitations and high design complexity.	Design of WGAN with synthetic image generation.
3	DeepMIB: User-friendly, open-source program for training deep learning networks for biological image segmentation [11]	Eija Jokitalo, Ilya Belevich et al.	Preprocessing, Training, Prediction, Data Augmentation	Deep learning complexity in biological imaging.	DeepMIB framework simplifying image segmentation tasks.
4	3D multilevel DenseNet and GAN-based MRI super-resolution [17]	Yibin Xie, Debiao Li, Feng Shi, Zhengwei Zhou, Yuhua Chen, Anthony G. Christodoulou, et al.	MDCSRN (Multilevel Dense Super-Resolution Network)	Long scan time, high cost, low signal-to-noise ratio.	Deep CNN with GAN improves quality and reduces acquisition time.
5	Synergizing Medical Imaging & Radiotherapy with Deep Learning [69]	Ge Wang, Yunyao Li, Harald Paganetti, Pingkun Yan, Xun Jia, and Hongming Shan	NLP & GNN-based semantic analysis	Low diagnostic accuracy and slow image validation.	Integration of NLP and Knowledge Graphs for deep learning advancement.

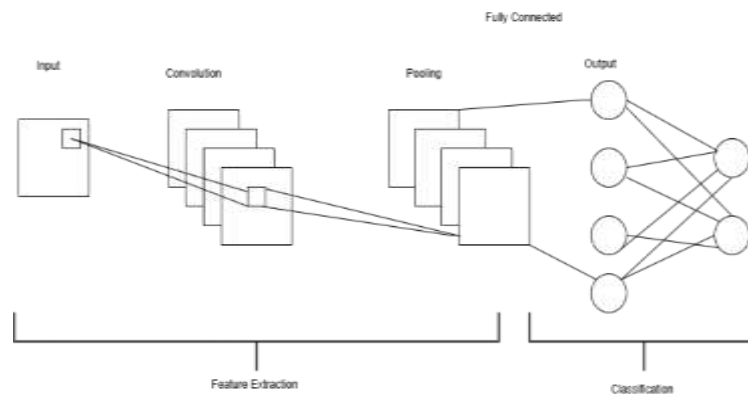
6	Color Deep Learning and Depth 3D Conversion for Human Height Estimation [43]	Soon Kak Kwon, Seok Chan Jeong, Jong Soo Kim, and Dong Seok Lee	Mask R-CNN & Depth Frames	Inaccurate height estimation under motion or occlusion.	Combine color and depth data using Mask R-CNN for precision.
7	Non-Contact Sensor System for 3D Body Reconstruction from 2D Photos [45]	Sekyoung Youm, Jiyong Lee, Kihyeon Kwon, and Changgyun Kim	Concurrent Neural Network, Point Cloud GAN, SMPL	Difficulties in 3D anatomy creation from 2D images.	Novel hybrid network achieving 93% accuracy using phone photos.
8	R-3D-YOLOv3 Image Classification and Filtration for Animal Study [43]	Subramaniaswamy Vairavasundaram, Priya Velayutham, Indragandhi Vairavasundaram, Sudhakar Sengan, Ketan Kotecha, Vijayakumar Varadarajan, and Logesh Ravi	R-3D-YOLOv3	Poor accuracy in dynamic environments.	Multi-scalar 3D object view with embedded real-time systems.
9	Size-Invariant 3D Creation from a Single 2D Picture of Rock [68]	Frank Lindseth, Gabriel Kiss, Johan Phan, and Leonardo Ruspini	VQ-VAE & GAN	High cost and inefficiency of 3D rock imaging.	GAN-based size-invariant synthesis improving speed and cost.
10	Computer-aided detection of brain tumors using DNNs [15]	Maibam Manga, Khelchandra Thongam, and Leibi Chanu	2D CNN with Keras	Low diagnostic accuracy in MRI.	Improved accuracy and speed using DNN classification.

Table 2 emphasizes the primary topics of deep learning methodologies, including CNNs and GANs, employed for 2D-to-3D transformation and image analysis in many fields such as medical imaging, object recognition, and reconstruction. It also underscores the solutions proposed to the problem. R. Nandhini Abirami et al. [2] presented the work on deep GAN and deep CNN image processing for computational visual perception. This work faces the issue of inadequate data and incorrectly generated picture quality. The author uses dropout and data augmentation to solve the problem. Yunpeng Wang and Lingxiao Zhou et al. [90] published a survey on automatic subcentimeter pulmonary adenocarcinoma classification using GAN and CNN. The objective of this survey is to create artificial medical imaging and design the progressively growing WGAN, which improves the performance and accuracy. Ilya Belevich and Eija Jokitalo et al. [11] studied deepMIB,

which is user-friendly open-source software used for segmenting biological images using DNN training. This study deals with various methods, such as preprocessing, training, predicting, and data argumentation. This survey states deep learning is still considered complex, which needs experts to understand, but with deep MIB, the author addresses the problems and provides the solution by training CNN and applying them to separate multicolor or grayscale 2D and 3D files. Yuhua Chen and Anthony G. Christodouleu et al. [17] published their work on MRI super resolution using GAN and 3D multilevel DenseNet, which is more accurate, quicker, and reduces size. High-resolution MRI diagnosis anatomy information, which is critical to understand, takes high cost, time, low signal-to-noise ratio, and small spatial coverage. In order to overcome this issue, the author comes up with deep CNN, which enables the extraction of generic high-quality images from inputs with low resolution with GAN-guided training, which recovers fine details and improves perceptual quality with less time. It fits for real-world clinical applications. The article by Hongming Shan and Xun Jia et al. [69] can give a contribution to the research of the opportunities of deep learning to transform medical imaging and radiotherapy. Although major improvements have been realized, there are still challenges to deal with: low diagnostic accuracy, slow image interpretation rates, validation and standardization problems, costliness, and the inability to interpret large amounts of data. One way forward to fighting these limitations is to integrate NLP and knowledge graphs that will bring the benefits of both connectionism and computationism together to boost the power of the deep learning models. More studies and advancements in these spheres are needed to reap all the benefits of using AI in clinical practice. The height of a person is hard to determine due to the complexity in the backgrounds and changes in poses. To be more precise, Dong Seok Lee et al. [43] propose to use Mask R-CNN with depth 3D conversion and color deep learning. This method enhances depth video body area extraction and segmentation, which are necessary in correct height measurement. They add backdrop analysis and the Euclidean distance measures to the usefulness of their method in gesture recognition and estimating the height of animals. To predict body form and obesity in individuals aged 20 and 30, Lee et al. [45] offer an innovative method of non-invasive sensor, whose application relies on the 2D images to create 3D images of humans. Their method, which integrates 3D Recurrent Reconstruction (3D-R2N2), Skinned Multiperson Linear Model (SMPL), Point Cloud Generative Adversarial Network (PointGAN), and Pixel Aligned Impact Function (PIF) with 93% accuracy, is presented. They have a higher grade compared to the earlier phones, including PIFuHD, and make use of normal photographs taken using a mobile phone to deliver proper and cost-effective body measurements. The improvement of BMI and WHtR assessments by this technology allows continued health monitoring without using specialist 3D scanners. The next step will involve making it more accessible and more accurate to allow comprehensive 2D picture-based body examination in any perspective. The R-3D-YOLOv3 model that Sengan et al. [43] work with is used to develop a deep learning and 3D construction ability to identify animals automatically on Indian roadways in real time. Their method, confirmed on nuScenes datasets on day and night scenes, enhances the accuracy of detection of multi-scalar views of 3D objects. Its embedded system optimization makes the model contribute significantly to road safety and wildlife management, as it provides the ability to ensure a high level of real-time monitoring of the wildlife on the road without consuming too much of the processing power. Phan et al. [68] develop a method that employs size-invariant GAN, image transformer, and VQ-VAE technologies to generate large 3D images based on the image of one 2D rock shot. In cases of hydrology, oil and gas, and CO₂ collection and storage, among other industries, this technology serves as a faster and cheaper approach than the customary 3D imaging technologies, as they are inefficient and expensive. The

characteristics of the method in speeding up the determination of rock features in various companies are illustrated by its validation on the basis of effective comparisons of actual rock photographs. Although early studies of Chanu et al.[15] inform about the application of combined 2D CNNs together with deep learning to support the computer-aided detection of brain cancers based on MRI . With regard to traditional MRI practices, their methodology that uses Keras enhances the diagnostic rate and accuracy. It is a giant leap in medical diagnosis and has some significant ramifications in improving clinical decision-making in neuroimaging.

Figure 4. Working diagram of a Convolutional Neural Network (CNN)



3.1. Convolutional Neural Networks (CNNs)

CNNs are webs with one or more convolutional layers that are mostly used for image processing, segmentation, classification, and alteration [44,93,7] . A neural network with more than four layers is called a DNN [7] . The three layers that comprise a neural network are the input layer, hidden layer, and output layer. In CNNs, mostly for processing, classification, segmentation, resizing, etc., different layers are used. Additionally, the input and output layers are interchangeable. In CNN, the system learns the pattern, shape, appearance, and features of an image [89,7] , so it may be in any shape or position it will identify. In ANN, the system will only learn the position and pattern. If we change the position or the size, it may be difficult for the computer to identify an image. Mostly for proper detection, CNN is the best. The primary function of CNNs is to analyze and interpret data with higher accuracy [7,103] than the human brain. CNNs are a kind of DNN that was created using a model inspired by biology [7] . Researchers found how humans perceive the image in the brain through different layers. As shown in Figure 4. Suppose you consider an input image of a cat. As you know, a cat has various attributes like shape, color, appearance, attitude, and so on, which all will be passed through a convolutional layer [93,7] where each feature will be assigned to a separate convolutional layer using the ReLU activation function. The image will be transferred through patches; in fig. 4, the small squares indicate the patches. Different layers will be for different patches, and then the same patches after filtration will be transferred to the pooling layer. The pooling layer is the type of non-linear downsampling layer, and it has various layers to extract the best features to identify the image. Mostly we use the max pooling layer [7] for image classification. After this, the process of feature extraction from the input image is ended, and it moves to the flattening phase, where it converts the two-dimensional array to a one-dimensional [7] array, or we can say it is converted to a vector. Ultimately, the classification phase began with a neural network that included a fully connected layer that produced

the output in the form of classes. The model outputs either a single class or a probability distribution over classes [7]. Input images are in the form of an array of pixel values in RGB format that are red, green, and blue. CNN has three fundamental features that reduce parameters in neural networks. First, in spatial layer interaction: Each neuron in a layer is interconnected with every other layer, resulting in a vast number of parameters (training data) that the network must learn. This might lead to additional issues for one layer. Additionally, convergence time lengthens, and an overfitting model might result. CNN can use indirect interactions to lower the number of parameters. Second, just like it does with sparse interactions, the parameter-sharing principle also affects the lowered parameter. Prior to that, you must understand how a convolutional network generates face characteristics.

Assuming that an image has a volume after going through the convolutional layer, a portion of this volume taken via the depth will reflect the characteristics of that same area of the picture. Additionally, the same filter is used to construct each feature in the same depth layer as convolve X . It indicates that each point in the feature map at the same depth is the output 3D volume produced by the same kernel. Lastly, the equivalent representation [7,52], which can be explained mathematically. Let us consider an example where two functions, F and G , are said to be *equivalent* if, for a given input X , the following condition holds:

$$F(G(X)) = G(F(X)).$$

This relationship implies that the order in which the two functions are applied does not affect the final outcome. To illustrate this concept, consider the image translation operation g and the convolution operation f . In the context of convolutional neural networks, convolution is said to be *translation variant*, meaning that:

$$f(g(X)) \neq g(f(X)).$$

This indicates that when an image is convolved and then translated, the result differs from first translating the image and then performing convolution. In other words, the convolution operation is sensitive to shifts in the input image, leading to different feature activations after translation. This property has important implications for feature extraction in CNNs, motivating the use of techniques such as pooling and data augmentation to mitigate translation variance.

Table 3. Summary of key CNN-based and hybrid machine learning approaches for 2D-to-3D image conversion.

Sl. No.	Survey Title	Authors	Techniques	Issues	Solutions
1	2D to 3D image conversion using a machine learning approach [20].	Pranjali Joshi, Atul Gunjal, Shreya Jadhav, and Rameshwar Dhondge.	Deep Map.	Difficult to predict the disparity map in blurred images.	More complex ML approach needed.
2	Simplified 2D-3D CNN architecture for spatial-	Meiping Song, Caiyu Liu, Rui Han, Chunyan Yu, and Chein I. Chang.	Fusion of 2D and 3D convolutional	Overfitting and high training cost.	Combines 2D and 3D layers to improve classification accuracy.

	spectral fusion-based hyperspectral categorization [100].		layers; 2D for spatial features, 3D for band relationships.		
3	Adaptive O-CNN: A deep representation of 3D forms based on patches [89].	Chunyusun, Peng Shuai Wang, et al.	3D shape adaptive O-CNN encoder and decoder.	Adaptive patches are not seamless; planar patches fail to approximate curves.	Apply surface reconstruction and fitting metrics for smoothness.
4	Knowledge-based 3D building reconstruction from 2D images (3DBR) employing CNNs [5].	Federico Tombari, Fatemeh Alidost, and Hossein Arfi.	Multiscale convolutional-deconvolutional network (MSCDN).	Complex dataset comparisons (Potsdam, Zeebrugge).	Deep CNN improves reconstruction speed.
5	Automated 2D images to 3D model: A survey [19].	Sanket Kua Dale, Sukhada Dole, Mahipal Mehta, and Shweta Kothawade.	CNN encoder-decoder, graph convolution, MaskRCNN.	3D modeling with 3D supervision requires high resources.	2D supervision balances accuracy and resource use.
6	Deep3D: Fully automatic 2D-to-3D video conversion with deep CNN [93].	Ali Farhadi, Ross Girshick, and Jun Yuan Xie.	Deep neural network (DNN).	Requires large groundtruth datasets.	Uses stereo pairings and temporal information for training efficiency.
7	LonchaNet: CNN architecture for real-time 3D object recognition [29].	S. Orts-Escalano, M. Cazorla, J. Garcia Rodriguez, A. Garcia Garcia, and F. Gomez Donoso.	ModelNet projection from point cloud slices.	Deep models slow and error-prone.	Reduces computation cost and improves accuracy.

8	Hyperspectral image classification using 3D-2D CNN [21].	Fu Xiaping, Gui Jiangsheng, and Alou Diakite.	3D-2D CNN on multiple datasets (Indian Pines, KSC, Pavia, Salinas).	Long training times.	Combines 3D and 2D CNNs for faster, accurate classification.
9	Sign Language Recognition System using CNN and computer vision [88].	Hurroo Mehreen and Walizad Mohammad Elham.	HSV color algorithm, dilation, and mask operations.	Requires high computational power.	CNN simplifies recognition.
10	CNNs for automated 2D-3D tracking of submicron-scale particles [60].	Phoebe Lee, M. Gregory Forest, Jay M. Newby, Alison M. Schaefer, Samuel K. Lai, et al.	CNN localization with machine learning training.	User bias and inefficiency.	Neural networks improve automation and accuracy.

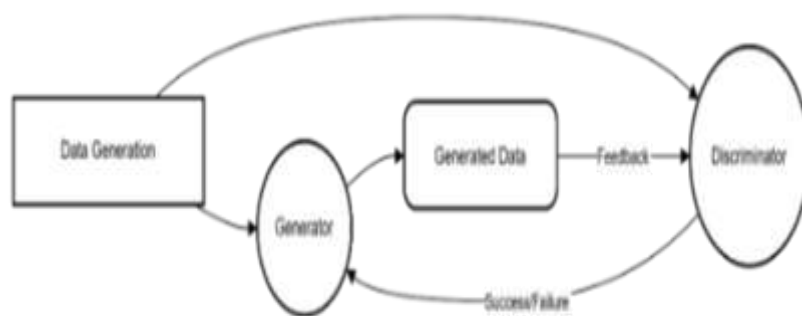
Moreover, Table 3 summarizes studies using CNNs for 2D-to-3D conversion and image analysis, detailing the techniques, issues, and solutions in various applications. Rameshwar Dhondge et al. [20] presented a survey on 2D to 3D image conversion using CNNs, which states obtaining stereo pictures from 3D movies and training the CNN to generate 3D images. The author uses CNN to extract 3D images from blurred images. Similarly, in another report, Chunyan Yu and Rui Han et al. [100] come up with a spatial spectral fusion method for converting 2D to 3D by the use of CNN, which helps to overcome the problem of overfitting. CNN improves the classification accuracy with less training cost. Peng Shuai Wang et al. [89] survey a paper on Adaptive O-CNN, a patch-based 3D form representation. Adaptive OCNN is an encoder and decoder. The work states the lack of approximate features and seamlessness patches. The authors use the reconstruction method and other quality metrics to overcome the issue. Fatemeh Alidost et al. [5] published the work on knowledge-based 3D building reconstruction (3DBR) with CNNs and a single aerial picture to predict the output from two methods, RMSE and NMAD, by using two different datasets, which are the Potsdam and Zeebrugge test data. Mahipal Mehta and Shweta Kothawade et al. [19] reported the work on automated generation of 3D models from 2D images by using the PyTorch 3D library, which speeds up the conversion, reconstruction, and development techniques of 3D models. The study uses the encoder-decoder graph and mask RCNN technique for convolutions of 3D models. Junyuan Xie et al. [93] have studied deep 3D, which states the automatic conversion of 2D to 3D video by using CNN. This study helps to reduce cost and time and improve performance in quantitative evaluation. F. Gomez Donoso et al. [29] presented a work on

the loncha net (slice-oriented architecture) used for the identification of 3D objects in real time. This method helps in point cloud classification to achieve high accuracy with low computation cost by preserving the 3D information. Alou Diakite et al. [21] demonstrated a study on hyperspectral image classification using a CNN and computer vision, where CNN reduces training time and improves accuracy during converting 2D to 3D images, whereas computer vision takes a high training time for good accuracy. Mehreen Hurroo et al. [88] studied a sign language recognition system using CNN and computer vision to achieve low computation power with high accuracy, making the system robust. In Ref. [60], the author published a paper on automatic detection for monitoring particles at submicron sizes in 2D and 3D images using CNNs. This technique uses a tracking method to track bright objects and manage the range of spatiotemporal heterogeneity and use submicron species to detect low signal-to-noise ratios in a complicated biological setting.

3.2. Generative Adversarial Networks (GANs)

GAN is a machine learning framework that comes from the model called the generative adversarial model, which is developed by Ian Goodfellow [31,30]. Three guiding concepts underlie GAN's operation. GAN first aids in learning and generating probabilistic data [31,3] for the generative model. Using DNNs and artificial intelligence algorithms, it trains the model in the second step and then the complete system. Despite being primarily used for unsupervised learning, GAN deployment networks have also proven to be highly effective for semi-supervised and reinforcement learning [3,30,42]. GAN has two parts, such as a discriminator and generator [31]. The generator generates fake images from sample images, whereas the discriminator identifies fake images by comparing them to the real images. The generator always tries to fool the discriminator [31] and continues to correct data till the discriminator is unable to identify fake data. For different applications, different types of GANs and architectures are used. GANs are widely used for 3D generation, texture synthesis, and image-to-image translation [74,55,102,108,18,51,3,42,6]. GANs have expanded significantly into medical imaging, geological modeling, and avatar generation [98,26,97,18,42,8].

Figure 5. Schematic Diagram of Generative Adversarial Network.



GAN works in two steps, as shown in Figure 5 [31,30]. First, the discriminator operates while the generator is not in use. During this step, the network is forward propagated, and the discriminators are trained on both actual data, or data sets, and false data generated by a generator to see whether they can properly identify the data. Second, the discriminator will be idle, and the generator will be operated and trained to fool the discriminator correctly and generate fake data till it confuses the discriminator. This continues till the result is genuine; when it is acceptable, then the training will be stopped.

GANs are designed as a minimum-maximum challenge where the generator aims to optimize the discriminator's loss, or decrease its reward $V(D, G)$, while attempting to reduce the incentive for the discriminator. It is described mathematically by the following formula [31,30] : Generator (G):

$$G : z \sim P(z) \rightarrow G(z) \in \mathbb{R}^n \quad (1)$$

where z is the latent vector and $P(z)$ is a prior noise distribution (such as Gaussian). Discriminator (D):

$$D : x \rightarrow D(x) \in [0,1] \quad (2)$$

where $D(x)$ computes the likelihood that the variable x , taken from either $P_{\text{data}}(x)$ or $G(z)$, is authentic. The core adversarial objective[31] is defined as:

$$\min_G \max_D \mathbb{E}_{x \sim P_{\text{data}}} [\ln D(x)] + \mathbb{E}_{z \sim P(z)} [\ln (1 - D(G(z)))] \quad (3)$$

where:

- G : A generator network that converts synthetic samples $G(z)$ to noise $\sim P(z)$.
- D : A discriminator network that, given input x , returns a probability $D(x) \in [0,1]$.
- $P_{\text{data}}(x)$: The distribution of actual data.
- $P(z)$: The prior distribution of latent noise vectors (e.g., Gaussian).

Richard T. Marriott et al. [55] presented a survey on 3D GAN in order to improve largepose facial recognition, such as hair, glasses, teeth, lips, eyes, and so on. The survey uses a 3D morphable model incorporated in the generator to generate synthetic images with fully disentangled poses, which is lacking in 2D GAN methods. This survey shows it learns about lighting effects and capturing shadows, which enables it to generate realistic images. Similarly, in another report, Paul Henderson and Vittorio Ferrari [35] presented a 3D reconstruction image from a single image that states on various meshes as a way of representation, uses a shading method to improve performance, and white light techniques to exploit clues. Meng Zhang et al. [102] come up with a paper that helps to restore 3D hair structure while preserving clarity, resolution, and uniqueness from a single photograph that depicts the evolution of hairstyles from various perspectives. Aamir Khan et al. [40] demonstrated a deep GAN holography framework for hologram construction, which improves the time complexity and performance of the image that was lacking in the computer-generated hologram. The authors use two techniques for this work: firstly, they use a Fresnel-based method to extract data samples or a data set, and secondly, they train both the holographic label image data and the raw input data. Jun Yan Zhu et al. [108] utilize cycle-consistent adversarial networks to study unpaired image-to-image translation, which states the translation of an image without labeled data. Guillaume Coiffier et al. [18] published the paper on the conversion of 2D example data to synthesized 3D geological images by use of Dimension Augmentor (DIGAN). These techniques convert the 2D to 3D without labeled data but enable the incorporation of some inbuilt data in the generator and discriminator of GAN. Xingde Ying et al. [98] have studied X2CT GAN, which states about the reconstruction of CT scans from bipolar X-ray reports. This study shows it reduces the price of X-ray and radiation treatment incubated to the patients. The author uses CNN and GAN techniques to guide the generator in order to produce high-quality CT volume, both statistically and visually. Jie Lu and Yang Yang et al. [51] presented about converting 2D to stereo panoramas with concentric mosaics and GANs. These techniques generate high-quality 3D stereoscopic panoramas. The author comes up with two techniques: firstly, using GAN architecture for adding a selection layer to the generator and adding a loss term for pseudo depth peeling, and secondly, using

concentric mosaics to analyze the panoramic stereo from shifted views. Dule Shu and James Cunningham et al. [73] reported a survey on 3D design using GAN and physics-based validation, which states that the repetitive generation and validation of design to achieve a model called the aircraft model helps to create high-quality performance and design without ruining its original quality. Michael Varga and Jan Jadlovisky et al. [86] have tackled the challenge of noise in the training data, which usually affects the efficiency of 3D image classifiers. Typically, noise affects non-generative classifiers that operate poorly. In an effort to lessen the effects of noise, Varga et al. (2023) explore generative classifiers, which can create fresh samples of data that accurately reflect the initial training data. As demonstrated by their research, generative classifiers are an improved choice for noisy data situations simply because they substantially enhance performance and accuracy. In an effort to attain higher levels of accuracy and precision in 3D picture classification, this study advocates merging generative models. This Table 4 summarizes the above surveys on 3D generation and related GAN techniques, outlining the authors, methods used, issues encountered, and proposed solutions for improvement:

Table 4. Summary of GAN-based techniques for 2D-to-3D image conversion.

Sl. No.	Survey Title	Authors	Techniques	Issues	Solutions
1	A 3D GAN for better facial recognition in large poses [55].	Liming Chen, Sami Romdhani, Richard T. Marriott, et al.	3D Morphable Model; Augmentation datasets.	Disentanglement issues; misidentification of hair, glasses, and teeth.	3D GAN improves facial pose handling and identity consistency.
2	Learning 3D reconstruction of a single image by generative modeling of shading, posture, and form [35].	Vittorio Ferrari and Paul Henderson.	Trained in 2D techniques; meshes; shading under single light source.	Shape identification errors; low-quality geometry.	Shading improves performance; GANs capture fine structural details and smooth surfaces.
3	HairGAN: Recovering 3D hair structure from a single image using GANs [102].	Meng Zhang and Youyi Zheng.	2D CNNs for 3D cue learning; dimensional expansion layer; GAN-based training.	Complex hairstyles challenge training; tracking issues.	Increased resolution and larger volumetric field size improve final output.

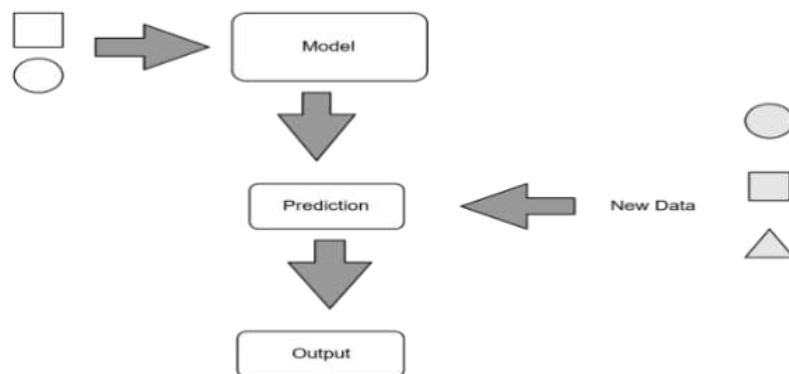
4	GANHolo: GAN-based generated holography using deep learning [40].	Muhammad Amir Khan, Ketao Yan, Khizar Aziz, Yingjie Yu, Aamir Khan, and Zhang Zhijiang.	Fresnel-based GAN holography framework.	Limited real-time performance and accuracy.	GAN holography improves reconstruction stability and performance.
5	Translation of unpaired images using cycle-consistent adversarial networks [108].	Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros.	Adversarial and cycleconsistency loss functions.	Requires paired datasets for translation tasks.	Combines cycle and adversarial losses to enable unpaired translation.
6	3D geological image synthesis from 2D examples using GANs [18].	Guillaume Coiffier, Philippe Renaud, and Sylvain Lefebvre.	DiGAN (Dimension Augmenter).	Requires extensive 2D training samples.	DiGAN improves 3D geological structure generation from limited 2D data.
7	X2CT-GAN: Reconstructing CT from biplanar X-rays with GANs [98].	Xingde Ying, Heng Guo, Kai Ma, Yefeng Zheng, and Jian Wu.	CNN and GAN combined with reconstruction, projection, and adversarial losses.	High radiation exposure; costly CT/MRI imaging.	Reconstructing CT scans from X-rays reduces radiation and cost.
8	Converting 2D to stereo panoramas with concentric mosaics and GANs [51].	Jie Lu, Ruiyang Liu, Sing Bing Kang, and Jingi Yu.	Pseudo-depth peeling loss; concentric mosaic generation.	Low-quality panoramas.	GANs with CMs generate high-quality stereoscopic panoramas.
9	3D design using GAN and physics-based validation [73].	Conrad S. Tucker, Timothy W. Simpson, Michael A. Yukish, Simon W. Miller, James Cunningham, Gary Stump,	Repetitive generation-validation method.	Loss of quality after repeated training cycles.	Physics-based validation preserves original data integrity.

		and Dule Shu.			
10	Generative enhancement of 3D image classifiers [86].	Michael Varga, Jan Jadlovska, and Slavka Jadlovska.	Non-generative classifier combined with generative model.	Noise during training disrupts classifier performance.	Generative classifiers deliver higher accuracy and robustness.

3.3. Support Vector Machines (SVMs)

SVM is a potent technique for data analysis tasks, including regression, classification, and outlier identification [81,82] in machine learning under supervision. The mechanics of the SVM may be applied to a wide range of applications, such as gene expression analysis, text classification, picture classification, face recognition, anomaly detection, handwriting identification, spam detection, and more. SVMs are flexible and useful in many different domains because of their capacity to manage high-dimensional data and nonlinear relationships. SVMs were formulated by Vladimir Vapnik and colleagues at Bell Laboratories [13,33,81], AT&T, such as Cortes and Vapnik (1995), Boser et al. (1992), and Guyon et al. (1993). These works established SVMs as a robust classification method. SVMs are theoretically grounded in Vapnik-Chervonenkis (VC) theory, which was put out by Chervonenkis (1974) and Vapnik (1982, 1995), which provides principles for understanding the generalization capabilities [83,85] of learning algorithms. This statistical learning framework has been pivotal in the evolution of SVMs [81,83,85], making them among the most researched and applied models in the field of machine learning (Vapnik et al., 1997). As shown in figure 6, The data is utilized to train the

Figure 6. Workflow of Machine Learning Model for Forecasting and Output Generation.



model, which is subsequently tested in order to know whether it is working properly or not. In order to check, new data is given. If the output is a circle, then it is classified as a circle class, and if it is identified as a triangle, then it belongs to the triangle class, and so on. It follows the classification algorithm and classifies the data into different classes. The classes can be represented in both a linear and a nonlinear manner. Let us consider an xy representation graph as shown in Figure 7, where we have a bunch of circles and squares.

Figure 7. Linearly Separable Data Representation.

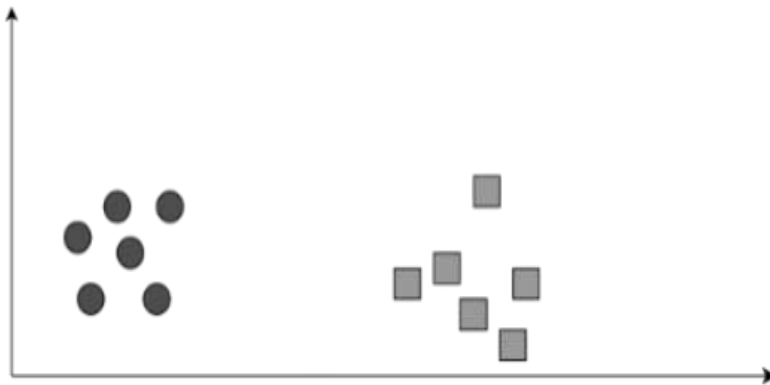


Figure 8. Linear Separation in a Binary Classification Problem.

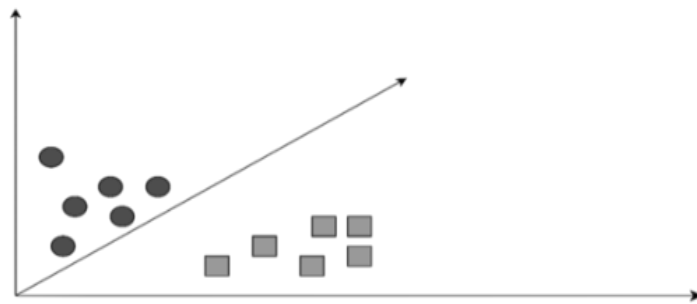
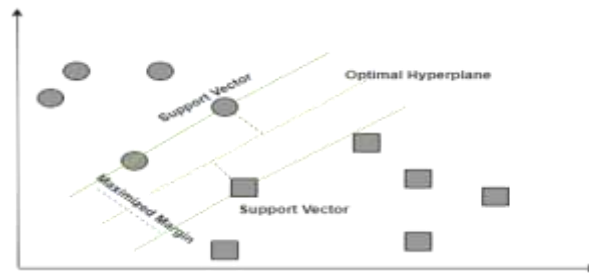


Figure 9. Support Vector Machine with Optimal Hyperplane and Margin.



The sum of the distances between the positive and negative points is the margin. D_+ represents the distance to the positive point. D_- stands for the distance to the negative point. Margin decides which type of hyperplane to be used For example, linear or nonlinear [81]

The greater the margin, the more precise the result with fewer errors as illustrated in the

Figure 10. Support Vector Machine : Separating Hyperplanes and Optimal Hyperplane.

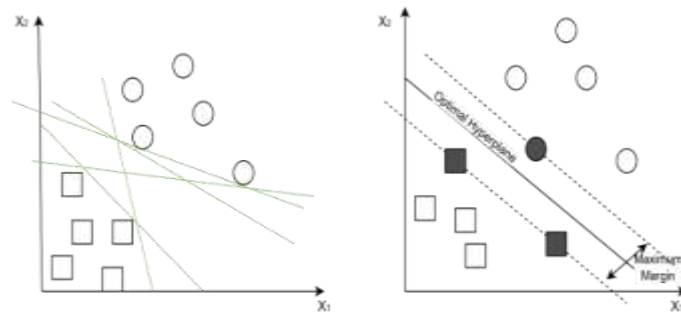


figure 10 . With the vertical direction of separation, we get less margin, so it is not fit for the model to provide better accuracy. Therefore, for better results, models always prefer a maximal margin hyperplane. In non-linear separation, the data are converted from low dimension to high dimensions with the help of the SVM kernel [33,81] . The decision boundary is circular; in order to convert to linear separation, the input data are passed through a 3-dimensional plane representation shown in figure 11.

Figure 11. Support Vector Machine with Optimal Hyperplane.

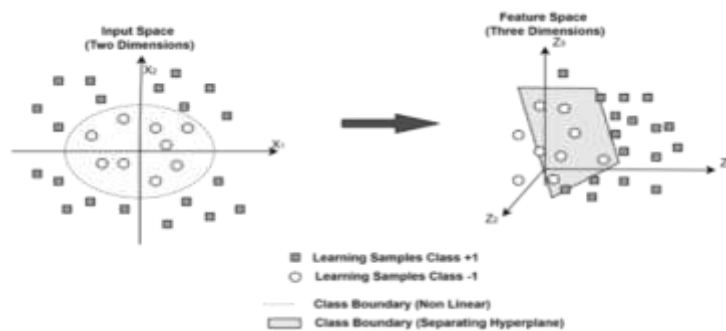


Figure 12. SVM Process Flow

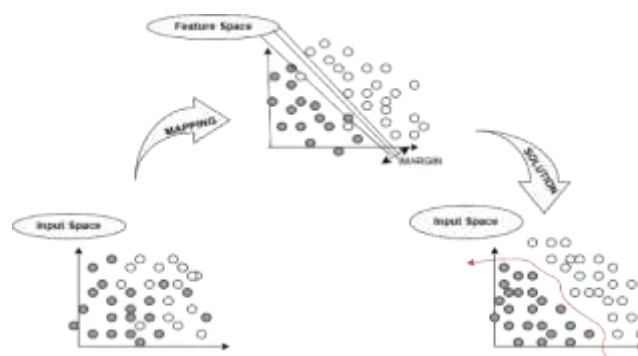


Figure 12 shows how the SVM method works. It starts with the input space, which is where raw data points are first represented. Usually, they cannot be separated linearly in a lower-dimensional space. Separability is made possible by using a kernel function to transform these data points into a feature space with more dimensions through a procedure known

as mapping. Support vectors are the informational values from different classes that are in closest proximity to one another, and the SVM finds the best hyperplane by maximizing the distance between them in the feature space. This hyperplane is then translated back into the input space as a decision boundary, which, as the curving boundary in the picture illustrates, can manage both linear and nonlinear separations. This method underscores the efficacy of SVM in accurately identifying intricate datasets [28,67,41] .

In paper [28] , curvelets, invariant moments, and SVM are investigated by Ghazal and Abdullah for face recognition. SVMs offer precise classification, curvelets capture fine features, and invariant moments manage changes in size and rotation. This technique successfully handles problems like illumination and position fluctuations. SVM is used by Lei Xiao and Jingyi Wang in paper [92] to optimize 3D animation design and enhance feature selection for analysis of ink painting. Compared to genetic algorithms and particle swarm optimization, their SVM technique achieves 94.89% accuracy, a 0.27% error rate, and shorter response times (0.37 seconds), leading to more accurate 3D models and higher overall performance. The paper [16] , 2-D-3-D CNNs in Parallel for Hyperspectral Pictures Categorization Using Hyperspectral Target Detection by Chen et al. discusses the main issues with hyperspectral image classification. To improve spectral characteristics and lessen noise, they employ HTD. 3D-CNNs incorporate both spectral and spatial information, whereas 2D-CNNs only record spatial data. By efficiently combining these features, the classification accuracy is increased when HTD-2D-CNN and PCA-3D-CNN are combined. Sanil et al., in paper [67] , present a thorough investigation on enhancing machine learning, geodesic distance computation, and 3D landmarks for monozygotic twin face recognition. Their work demonstrates important breakthroughs in overcoming difficulties with recognition, increasing accuracy, and bolstering security. But there are still issues with expression variability and real-time computing that need to be worked out further. Rivera-Romero et al., in paper [64] , increase image quality for classifying posture in bed using histogram equalization and median filtering. To select the most effective RGB texture descriptors for training SVMs, they employ ANOVA and Tukey testing, which allows them to achieve high accuracy with little preprocessing. They do, however, draw attention to difficulties in identifying posture and the effect that extensive pre-processing has on ongoing monitoring. Mansour et al. in Paper [54] investigate the use of hierarchical SVM in infrastructure scene segmentation for 3D point clouds [32]. They tackle problems such as low SVM performance, high dimensionality, and missing texture characteristics. Transfer learning for improved generalization, deep learning combined with SVM optimization, and parallel processing optimization are some of the solutions. Additionally, they recommend dimensionality reduction and texture features for enhanced feature extraction. Islam et al. (2022), in paper [38] , address the shortcoming of existing SER models that only account for emotion classification, ignoring intensity. Using a deep learning framework that includes 3D CNN, TDF layer, and Bi-LSTM, their method combines MFCC, STFT, and Chroma STFT into a 3D feature set. Their approach achieves high accuracy on the RAVDESS dataset by breaking down emotion and intensity categorization into discrete steps, improving comprehension of emotional intensity as well as practical applicability. In order to achieve high accuracy in facial expression detection in Western databases, Kim et al. (2022), in paper [41] , offer a hybrid model that combines CNNs and SVMs. The usage of solely Western databases, differences between emotions shown on the face and in real life, and the requirement for improved CNN weighting and classification are some of the challenges. The use of various datasets, enhanced emotion inference, CNN and weight optimization, and benchmarking against novel techniques are suggested remedies. In paper [61] , Pokorny et al. (2024) showed

that SVMs can successfully classify brain strokes using artificial data and a single frequency (1 GHz). While Vrba et al. (2024) emphasized difficulties in generalizing stroke identification across various head positions, Tosi et al. (2024) showed that smaller stroke data improved SVM accuracy. While this research collectively improves SVMs and microwave imaging for the categorization of strokes, it also highlights persistent difficulties with data generalization. In the paper [63], in order to identify brain cancers in MRI images, Rasheed et al. (2023) developed a 98% accurate approach combining adaptive noise filtering with SVM classification. Their method effectively handles noise and improves segmentation over current approaches. Further developments might concentrate on integrating deep learning in future work. As per the above discussion, here is Table 5, which summarizes various surveys and studies, detailing the techniques used, problems faced, and proposed solutions:

Table 5. Summary of SVM Techniques for Image Recognition, Classification, and Segmentation.

Sl.No	Survey	Authors	Techniques	Issues / Problems	Solutions
1	Recognition of face using curvelets, SVM, and invariant moments [28].	Karam Abdullah and Muhammad Talal Ghazal.	FDCvT, Invariant Moments, SVM.	Variations in illumination, pose, and expression; difficulty depicting edges.	High accuracy, reduced error rate, effective face and side-face detection.
2	Optimization of 3D animation design based on SVM algorithm [92].	Lei Xiao and Jingyi Wang.	Feature Selection, Defect Identification, 3D Virtual Reconstruction.	Difficult to reconstruct precise 3D virtual ink paintings; low accuracy in traditional GA/PSO.	SVM achieves 94.89% accuracy, 0.27 error rate; fast reconstruction.
3	2D–3D Parallel CNNs for hyperspectral image classification [16].	Tzu-Hsien Kuo, KaiHsun Hsu, and Shih-Yu Chen.	HTD, 2D CNN, 3D CNN, Feature Map Combination.	Noise in background, poor spatial-spectral feature fusion.	Combined HTD–2D–3D CNN improves classification accuracy.
4	Facial similarity measure using 3D facial landmarks [67].	Vinod Nayak, Srikanth Prabhu, Krishna Prakash, and Gangothri Sanil.	Geodesic Distance Calculation, ML-based Feature Extraction.	Hard to distinguish identical twins; accuracy drops with varied expressions.	Real-time computation, higher accuracy, robust to expression change.

5	In-bed posture classification using SVM [64].	MisaellLopez-Ramirez,Carlos Lastre-Dominguez, Jorge Munoz-Minjares, and Claudia RiveraRomero.	SVM, Histogram Equalization, Median Filter, RGB/HSV Color Models.	Hard to identify patient posture; heavy preprocessing.	Median filtering reduces noise; optimized SVM improves accuracy and efficiency.
6	Semantic segmentation of 3D point clouds [54].	Jorg Blankenbach, Jan Martens, Kou Yamada, Md. Abdus Samad Kamal, and Md. Riadul Islam.	SVM, Point-Cloud Segmentation, Optimization, Deep-Learning Integration.	Missing texture features; high-dimensional data; slow optimization.	Combine SVM with deep learning; parallel optimization boosts performance.
7	Emotion intensity recognition from speech [38].	Md. Abdus Samad Kamal, M.A.H. Akh, Md. Riadul Islam, and Kou Yamada.	MFCC, STFT, Chroma STFT, 3D CNN, BiLSTM, TDF Layer.	Conventional models ignore intensity; low interpretability.	Combines 3D features and Bi-LSTM for accurate emotion intensity recognition.
8	Hybrid method for facial expression recognition using SVM and CNN [41].	Muhammad Tahir Naseem,Chan-Su Lee,Han-Enul Suh, Jin-Chul Kim, and Min-Hyun Kim.	CNN,SVM, Dense Facial Motion Flows, Landmark Flows.	Limited datasets; CNN weight optimization needed.	CNN-SVM hybrid enhances classification consistency and emotion inference.
9	SVM-based classification and detection of microwave brain strokes [61].	Ondrej Fiser, David Vrba, Alessandro Polo, Jan Vrba, Luca Tosi, TomasPokorny, and Marco Salucci.	SVM, Synthetic Data Generation, Antenna Array, Frequency Analysis.	Dataset variability affects classification; difficult stroke localization.	Synthetic data and 1 GHz training improve accuracy and robustness.
10	Adaptive noise filtering and statistical	Adel Bahaddad, Khalid	Adaptive Noise Filtering,	MRI noise, high false detections,	Adaptive filtering + GLCM features

	features for brain tumor recognition [63].	Almarhabi, Ahmed Alghamdi, Muhammad Ashraf, Arfan Jaffar, Mehwish Rasheed, and Muhammad Iqbal.	Statistical Feature Extraction, SVM.	segmentation errors.	enhance tumor segmentation accuracy.
--	--	--	--------------------------------------	----------------------	--------------------------------------

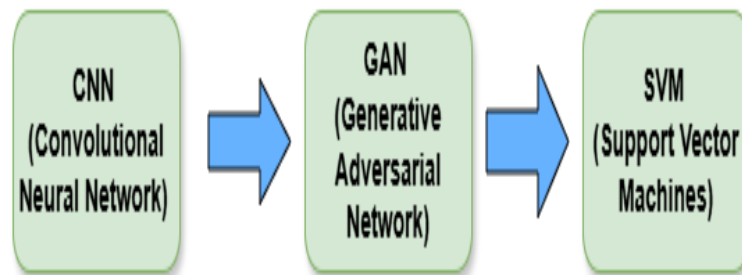
3.4. Comprehensive Hybrid Architectures for 3D Reconstruction

Hybrid deep learning models represent the current state-of-the-art in 2D-to-3D image conversion by combining the complementary strengths of Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Support Vector Machines (SVMs) [55,58,53,10,74,9,72]. This section provides an in-depth overview of how these architectures are integrated to address the fundamental challenges [44,93,35,56,99] of 3D reconstruction, with particular attention to recent advances documented in the literature [2,90,17,68,55,58].

3.4.1. CNN-Based Feature Extraction in Hybrid Models

CNNs represent the basic building block of hybrid architectures that offer powerful features extraction [89,7] that are critical in learning 2D image semantics and geometric features. The recent works by Nandhini Abirami et al. [2] and Alzubaidi et al. [7] emphasize

Figure 13. Hybrid CNN-GAN-SVM architecture for 2D-to-3D image reconstruction.



the multitask of deep CNNs in terms of visual perception and learning representations. The CNNs are usually useful in:

Depth Estimation : CNNs are used to estimate depth maps using hierarchical spatial features. U-Net and its variants are encoder-decoder networks that have been popular in dense depth prediction [44,93,62,47]. Lee et al. [44] show the automatic 2D-to-3D conversion with multi-scale deep neural networks, which proves that CNNs can learn fine-grained depth information.

Feature Representation: CNNs are trained to acquire geometric features and contextual features that are important in reconstruction accuracy [89,5,34,12]. Wang et al. [89] introduce adaptive octree-based CNN representations of 3D shapes whereas Alidoost et al. [5] use CNNs to reconstruct 3D buildings using aerial images.

Geometric Understanding: Deep CNNs have the ability to implicitly learn perspective and occlusion constraint, thus realistically perform spatial reasoning in the context of reconstruction [29,21] .

Overall, CNNs serve as the skeleton of the hybrid pipelines by producing semantically rich features that control the further refinement steps using GAN- and SVM-based methods.

3.4.2. GAN-Based Realistic 3D Synthesis

Generative Adversarial Networks (GANs) take the time to improve hybrid models with the use of adversarial learning [55,102,108,51,3,30] to enhance the realism and fidelity of the reconstructed 3D output. There are several extensions to 3D synthesis and rendering which have developed since the original GAN framework proposed by Goodfellow et al. [31,30] :

Realistic Texture Generation: GAN generators synthesize fine surface textures [102,18,107] missing in the original 2D images, while discriminators enforce photorealistic outputs [102,18,51] . Hair reconstruction is realizable with GANs as demonstrated by Zhang et al. [102] , and Coiffier & Renard [18] use the same architecture to synthesize geological images.

Shape Refinement: GAN refiners maximally refine coarse 3D geometries [55,98,51] output by CNN-based depth estimators. In conditional GANs (cGANs), the conditioning information is structural, which is 2D in nature [55,35] . Marriott et al. [55] use 3D GANs in huge pose facial recognition.

Multi-View Consistency: GANs enhance consistency of the images at different viewpoints (3D) [51,56,99] . Lu et al. [51] accomplish 2D-stereo panorama conversion with GANs and concentric mosaics.

Adaptation: GANs enable cross-domain transfer by learning from unpaired datasets through adversarial domain adaptation (e.g., CycleGAN [108]), facilitating generalization to real-world imagery [3] .

Applying CNN-based structural insight alongside adversarial realism, GANs increase the photo-realism of texture, as well as, the visual consistency of 3D-generated images.

3.4.3. SVM-Based Classification and Refinement

Despite being classical, Support Vector Machines (SVMs) are also very useful in a hybrid system due to their ability to be interpreted and due to their high accuracy [81, 28,41] . Vapnik et al. [81,84,85] have introduced strong classifiers in SVMs and the recently developed studies indicate that SVMs are useful in 3D reconstruction and quality evaluation:

Geometric Primitive Classification: SVMs classify geometric primitives such as planes, edges, or surfaces, providing structural labels that guide CNN-based reconstruction [28,54] .

Quality Assessment: SVMs categorize reconstructed 3D models based on fidelity metrics, enabling targeted refinement of lower-quality regions [16,41] .

Feature Selection: Integrating SVMs with CNN feature extraction reduces dimensionality and enhances computational focus [92,67] .

Post-Processing Refinement: SVMs can refine ambiguous boundaries [28,67] and provide interpretable classification for uncertain regions [64,61] .

By offering interpretability and decision-level refinement, SVMs complement the representational depth of CNNs and the generative power of GANs, forming an essential bridge between feature extraction and model validation.

3.4.4. Integrated Hybrid Architectures

The collaboration of CNNs, GANs, SVMs, and single systems is the character of the recent progress in the image conversion 2D-3D. Figure 13 shows a three-way hybrid CNNGANSVM systems to convert 2D images to 3D, which consists of three key processes: CNN-based feature extraction and initial depth estimation;

GAN-based refinement for texture generation and geometric realism; and

SVM-based quality assessment and adaptive post-processing.

In this structure, CNN encoder is used to extract multi scale features and coarse 3D geometry is developed by the decoder. These predictions are further refined by the GAN generator, as they are made real by the discriminator, whereas the quality of the output is determined by the SVM classifier and causes localized refinement to be improved. This feedback loop produces interpretable high-fidelity 3D reconstructions that are obtained through an iterative feedback loop [2,90,58,80]. The benefits of these combined strategies are proved by recent research:

CNN-GAN Hybrids: Two layers can either be fixed or trained either way. Shao and Wang et al. [90] indicate that better diagnostic accuracy in medical imaging is achieved under CNN -GAN incorporation and pulse nodule detection respectively.

CNN-SVM Hybrids: Use SVM with CNN-extracted features of classification and of geometric quality assessment with the ability to refine where they structure a loop of refinement [16,41].

CNN-GAN-SVM Architectures: This is the most sophisticated type of architecture, which uses features, adversarial reinforcement, and explanatory measures. Mirzaee et al. [58] reveal the state of art in the porous media reconstruction and Mamand and Abdulla [80] introduce a full 2D-to-3D framework with all of the three building blocks.

All these hybrid architectures underscore the effective combination of deep learning and classical machine learning methods and represent a new paradigm of effective, accurate, and explainable 3D reconstructions of 2D images [72].

4. EMERGING TRENDS (2024-2025): TRANSFORMERS, NERFS, AND DIFFUSION MODELS

The more recent developments in 3D vision have brought forward excellent architectures that can be extended to the classical CNNGAN designs. Neural Radiance Fields (NeRFs) to Vision Transformers (ViTs) and Diffusion Models represent not the last innovations that transform the way 2D images are transformed into 3D scaffolds.

Neural Radiance Fields (NeRFs): NeRFs are 2D image-based learners of continuous volumetric representations of scenes in 3D by learning to model radiance and density functions. Different approaches like Instant-NGP (2022) or SparseNeRF (2023) require much less time to train, and NeRF-Edge [91] is capable of training on edge devices in real time and 3D to rebuild objects. Recent Hybrid NeRF-GANs [48] use top-notch volumetric rendering and adversarial texture synthesis to achieve the photorealistic output [36,94].

Vision Transformers (ViTs) : ViTs utilize self-attention to capture long-range dependencies to enhance depth estimation and understanding of the scene. These works as Trans3DNet [50] and PointViT [101] demonstrate that transformers-based models have superiority over CNNs when it comes to dealing with occlusion and structural consistency.

Diffusion Models: The diffusion models have also revolutionized the generative 3D synthesis. The development of works such as DreamFusion [22], Magic3D [71], and Zero123++ [25] has shown a strong ability to generate scenes in 3D using single input 2D

images or textual instructions, and the techniques of DiffusionDepth shown further can produce high-quality monocular depth completion.

4.1. Comparative Summary

A comparative summary of current state-of-the-art models that are used to outline the 2024-2025 research frontier in 2D-to-3D image conversion is provided in Table 6. The three prevailing paradigms, including Neural Radiance Fields (NeRF), Vision Transformers (ViT), and Diffusion Models, are compared in terms of the mechanisms, strengths of their models, limitations grounded in the inherent flaws, and exemplary works.

Table 6. Comparative analysis of emerging 2D-to-3D model paradigms (2024-2025).

Model Type	Core Mechanism	Strength	Limitation / Key References
NeRF	Volumetric scene modeling; ray sampling for geometry learning	Precise geometry and realistic lighting	High computational cost [91,48]
ViT	Attention-based feature extraction; global context reasoning	Excellent scalability and representation learning	Large data requirement [50,101]
Diffusion	Noise-to-signal generation; probabilistic synthesis	High-quality sample realism	Slow inference; resource intensive [71,22]

The details of the NeRF-based architecture [91,48] are based on volumetric scene modeling used to learn model geometry and realism of lighting, but are still costly to run in a large-scale or real-time application. Transformer based systems [50,101] utilize attention based learning of features to obtain better reasoning in the global context and cross-view consistency, but with large labeled datasets and large amount of computation. Diffusionbased models [71,22] use iterative probabilistic synthesis to produce stable training of generative work, and photorealistic 3D reconstruction, at the cost of frequently having high computational requirements. In general, the table highlights a noticeable tendency of research to hybrid models to unify the accuracy of NeRF with the contextual understanding of ViT coupled with the stability of Diffusion Models to achieve interpretable and effective 3D reconstruction systems.

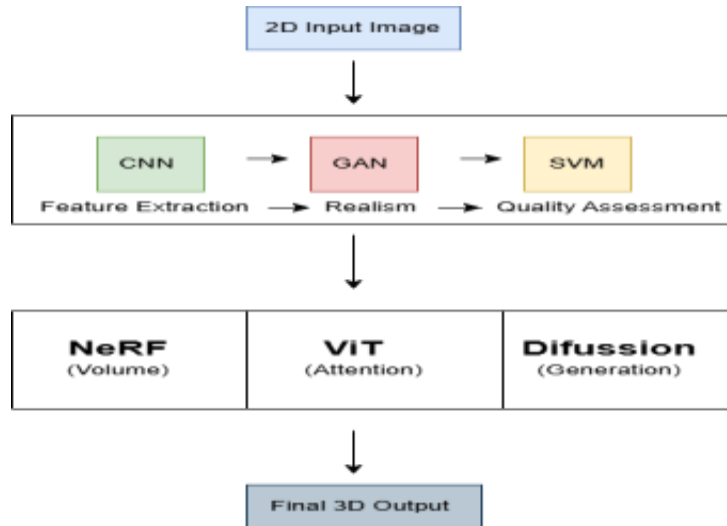
4.2. Future Integration

The combination of NeRFs, ViTs, and diffusion models with hybrid CNNGANsVM models is an approach that would provide intelligible, efficient and real time 3D reconstruction frameworks that can operate on embedded and edge computing systems [46].

Figures 14 illustrates the conceptual framework of the proposed unified hybrid 2D-to-3D reconstruction model that integrates traditional deep learning and emerging AI paradigms. The framework combines the representational power of Convolutional Neural Networks (CNNs), the generative realism of Generative Adversarial Networks (GANs), and the interpretability of Support Vector Machines (SVMs) with modern architectures such as Vision Transformers (ViTs), Neural Radiance Fields (NeRFs), and Diffusion Models. The process begins with CNN-based feature extraction, where spatial and depth features are

learned from 2D images. These features are refined through a GAN, which enhances texture and structural realism by learning photometric consistency. The SVM module subsequently performs quality assessment and classification to ensure geometric and perceptual accuracy. The resulting output is then optimized by the emerging models: NeRF captures volumetric radiance information for realistic 3D representation, ViT enhances global contextual reasoning through self-attention, and Diffusion Models progressively refine the final 3D reconstruction through probabilistic denoising and texture synthesis. Collectively, this multi-level integration provides a robust and interpretable framework that supports high-fidelity, generalizable, and computationally efficient 3D reconstruction suitable for next-generation vision systems [78].

Figure 14. Conceptual framework of the unified next-generation hybrid 2D-to-3D reconstruction architecture integrating CNN–GAN–SVM with Transformers (ViT), Neural Radiance Fields (NeRF), and Diffusion Models. The framework leverages CNNs for spatial feature extraction, GANs for texture realism, SVMs for interpretability, ViTs for contextual attention, NeRFs for volumetric scene synthesis, and Diffusion Models for stable 3D generation.



5. APPLICATIONS

The 2D to 3D conversion has numerous applications in the areas:

1. *Robotics*: 2D-to-3D conversion is capable of producing 3D models of objects and environments, which can be used for activities including gripping, object identification, and self-navigating [21], [59], [72], [89], [90], [91].
2. *Computer-Aided Design (CAD)*: 2D-to-3D conversion can be able to create 3D models of objects and environment, which can be utilized to accomplish tasks such as gripping, object recognition, and self-navigating [35,105,104,56,99,95] .
3. *Virtual Reality (VR)*: 3D model can be obtained through 2D-to-3D conversion technique of making use of 2D photos to generate immersive VR experiences in games, education, training simulation, and virtual tours [51,107,24,42] .

Table 7. Comparative Analysis of 2D-to-3D Conversion Applications Based on Depth Accuracy, Real-Time Processing, Computational Complexity, and Industry Impact.

Application	Depth Accuracy (1-5)	Real-Time Processing (Yes/No)	Computational Complexity	Industry Impact (1-5)
Robotics	5	Yes	High	5
Computer-Aided Design (CAD)	4	No	Medium	4
Virtual Reality (VR)	5	Yes	High	5
Medical Imaging	5	No	High	5
Facial Recognition	4	Yes	Medium	4
Geospatial Analysis	5	No	High	4
Entertainment	4	Yes	Medium	5

1. *Medical Imaging*: 2D to 3D conversion, which can be used to deform 2D scans such as X-rays, CT, and MRI into 3D, supports the creation of 3D anatomical structures, supporting diagnosis, surgical and treatment planning [17,69,98,26,97,49] .

2. *Facial Recognition and Biometrics*: 3D facial models based on 2D images enhance accuracy of the recognition system at different levels of lighting, angles and facial expression [93,67,41,24,42] .

3. *Geospatial Analysis*: 2D satellite and aerial images can be transformed to 3D ground maps to be used in urban planning, disaster management and environmental monitoring [5,75,34,18,58] .

4. *Entertainment Industry*: 2D-3D conversion is applied in the production of films, animation and visual effects to introduce level of depth and realism to digital content [102,73,107,70,42,77].

As seen in Table 7, applications such as Robotics and Medical Imaging demand the most of depth accuracy and the highest level of computational issues whereas CAD and Facial Recognition are at the brink between performance and real-time demands. In order to visualize these performance indicators better. The chart shows clearly that the cumulative performance requirements of applications like VR and Robotics are more demanding in terms of real-time processing and depth accuracy unlike other applications like Geospatial Analysis whose applications are more concerned with depth accuracy with no real-time requirements.

6. COMPARATIVE ANALYSIS

The survey will evaluate the performance of various deep-learning models (2D-to-3D conversion) to assess the performance of the models based on the quantitative performance (accuracy), visual performance (reality) and qualitative performance (scalability, cost and usability) in order to understand the performance of each model regarding these properties.

The comparison will involve works published in 2019-2025 [34,66,95,10] that provided the examples of such architectures.

6.1. Quantitative Comparison

Table 8. Comparative evaluation of representative 2D-to-3D conversion models (2019–2025).

Model Type	Representative Study	Dataset / Domain	Accuracy (%)	Time (s/frame)	Params	Remarks
CNN	Lee et al. (2020) [43]	NYU Depth v2	89.5	0.32	22M	Effective for structured depth estimation but limited texture realism.
GAN	Marriott et al. (2021) [55]	CelebA / 3D Faces	93.2	0.41	38M	High realism, prone to training instability.
SVM	Kim et al. (2022) [41]	FER / Expression	86.7	0.12	3M	Low complexity, interpretable but limited scalability.
CNN–GAN Hybrid	Shao et al. (2020) [90]	CT / MRI	95.6	0.39	42M	High realism and structural consistency.
CNN–SVM Hybrid	Chen et al. (2020) [17]	Hyperspectral	92.4	0.28	24M	Improved discrimination and robustness.
CNN–GAN–SVM Hybrid	Mirzaee et al. (2023) [58]	Porous Media / General	96.8	0.36	45M	State-of-the-art accuracy and generalization.
Transformer–NeRF Hybrid	Thalmann et al. (2024) [79]	ShapeNet / Synthetic	95.4	0.48	52M	Integrates selfattention and implicit NeRF representation for high-fidelity 3D reconstruction.

Table 8 gives an overview of the quantitative summary of representative 2D-to-3D conversion models in the period between 2019 and 2025. It is found that the analysis has been shifting to more advanced hybrid and transformer-powered models combining spatial, adversarial, and implicit concepts of learning that started with the basic convolution-based depth estimation models. Architectures based on CNN like that of Lee et al. (2020) [43] showed good feature extraction and stability but had low texture realism. GAN-based designs [55] were more real, but less training-stable and computationally efficient, and SVM-based systems [41] were less accurate and had lightweight choices.

CNN to hybrid models such as CNN-GAN, CNN-SVM and CNN-GAN-SVM has significantly enhanced accuracy and generalization, because Mirzaee et al. (2023) [58] have demonstrated that a hybrid architecture achieved accuracy of 96.8 per cent with balanced performance and realism. Most recently, transformer-enhanced systems have changed the 3D reconstruction paradigm. Transformer-NeRF hybrid presented by Thalmann et al. [79]

performed with an accuracy of 95.4 percent and provided better photorealistic and structural coherence based on attention-based feature encoding and implicit representation of volumes.

Altogether, the numerical data emphasizes the regular tendency towards the integration of multi-architectures, where the hybridization of the convolutional, adversarial, and transformer based methods allow the higher precision of depth, texture, and cross-domain flexibility. All these findings point to the fact that future 2D-to-3D conversion studies will still be biased toward modular, data-saving and transformer-based hybrid designs that are equally realistic and computationally efficient.

6.2. Visual Comparison of Model Performance

Figure 15. Accuracy comparison among 2D-to-3D conversion models.

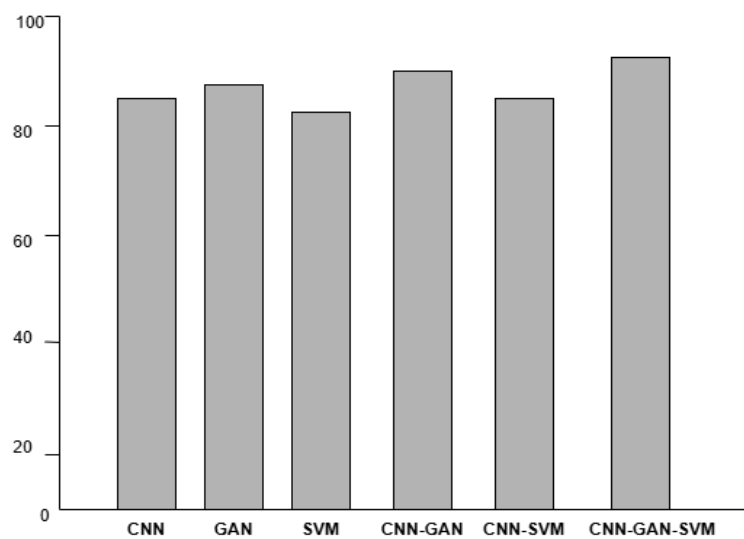


Figure 16. Processing time versus model complexity among deep learning models.

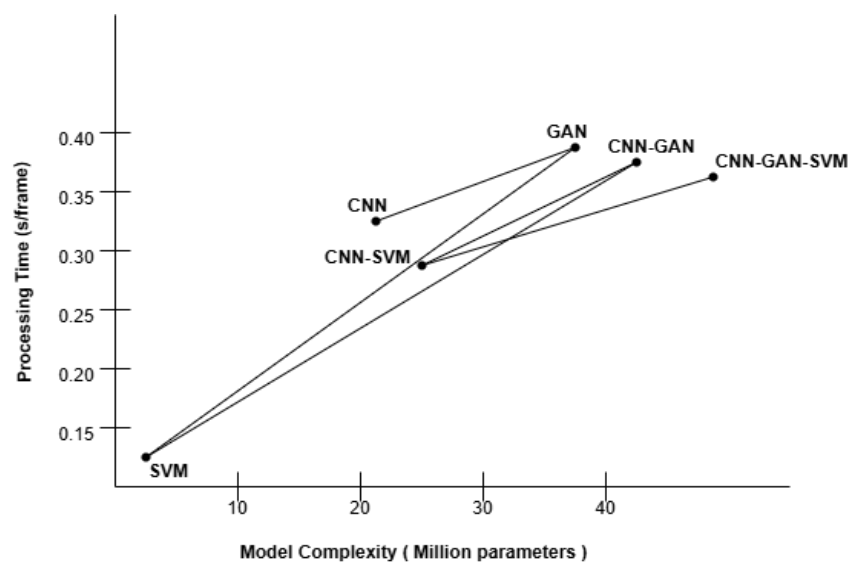
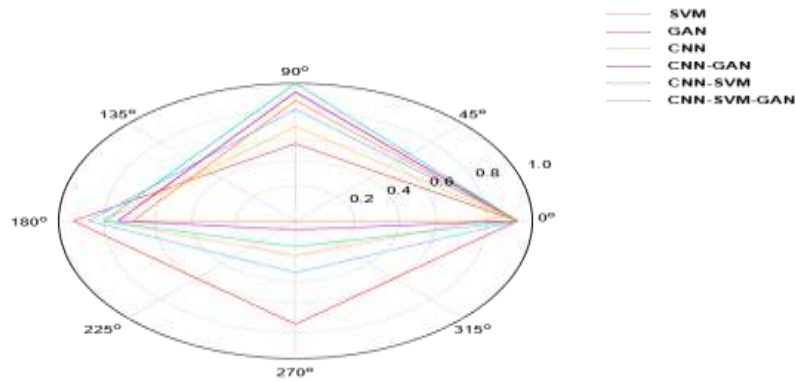


Figure 17. Multi-metric radar comparison of 2D-to-3D conversion models.

Visual comparison gives a feeling of what is going on with various 2D-to-3D conversion models that do not compromise geometric integrity, textual fidelity, and perceptual realism.

Hybrid frameworks, as shown in Figure 15, are always the most effective in reconstruction due to their ability to integrate both the geometric feature extraction ability of CNNs and the textual refinement ability of GANs. Among these, the CNN-GAN-SVM hybrid model is the most accurate and realistic; as it manages to retain the detailed structure of the structure and provide a smooth surface at the same time.

The way between model complexity and computational efficiency is pointed out In Figure 16. Though the SVM-based systems present the quickest inference time coupled with the minimum amount of parameters, they have limited representational ability. By contrast, the hybrid architectures (which incur higher parameter loads (about 45M) and middle-of-the-road inference times (0.36 s/frame)) have much higher visual realism without the prohibitive cost of full realism. According to this trade-off, it is possible to have high fidelity and achievable efficiency on real-time 3D applications through well-planned hybrid models.

Figure 17 a radar plot, summarizes the comparison on the four metrics of evaluation which include accuracy, realism, scalability and efficiency. CNNs are efficient and have a high edge clarity whereas GAN-based models are more realistic in textures, but cannot be trained with less computational resources. Both are outperformed by hybrid systems that produce structurally accurate systems as well as visually coherent depth estimation. It is worth noting that transformer-augmented hybrids, like the Transformer-NeRF model presented by Thalmann et al. (2024) also have much more detailed textures and geometric consistency, as a result of a combination of attention-based feature encoding, resulting in perceptually better 3D reconstructions.

On the whole, the visual evaluation supports the quantitative results that hybrid and transformer-integrated models provide the most balanced results in the realism, scalability, and accuracy facets, which can be considered a significant step towards attaining photorealistic and computational efficient 3D reconstruction.

6.3. Qualitative Comparison

Qualitative assessment is a complement of quantitative and visual evaluation because it emphasizes the perceptual, structural, and contextual realism of reconstructed 3D scenes. As in Figures 15–17, the CNN-based models are very geometrically consistent and define their edges sharply, but the results tend not to have the richness of surface texture and fineness of realism. Architectures based on GANs, conversely, generate detailed and

appealing surfaces and tend to have artifacts and to be unstable when dealing with complicated illumination or occlusion scenarios.

These shortcomings can be addressed successfully by the combination of geometric accuracy and discriminative and generative learning parts via the use of hybrid solutions like CNNGAN and CNNSVM. The CNNGANSVM hybrid is one of them, with a good trade-off between spatial coherence and texture fidelity, it gives a very realistic depth reconstruction with minimal perception errors. Moreover, the new Transformer-NeRF hybrid framework has all these benefits, yet it adds additional possibilities to encode specific feature via attention and the implicit volumetric representation, which allows a more natural interpretation of the setting and transitions of depths across diverse surfaces.

In general, the qualitative evaluation confirms that architectures based on the hybrid and transformer augmentation not only deliver a higher numerical precision but also yield perceptually and structurally convincing 3D results. They are therefore a good way forward in 2D-to-3D conversion studies in the future and especially in fields that demand high spatial realism (high-fidelity medicine imaging, autonomous perception, and content synthesis of digital content).

7. RESEARCH GAPS

The above discussion has outlined the main advances of 2D-to-3D image conversion with deep learning and the comparative advances in CNN-, GAN-, and SVM-based networks. Although great progress has been made in reconstruction accuracy, visual realism and hybrid model design, there are still a number of challenges not addressed, which limit scalability, interpretability and cross-domain performance [34,66,95,10]. It is important to identify these limitations in order to develop a direction of the future studies that will develop a stronger and more generalized type of hybrid frameworks that may be utilized at in reflex mode in a real-life setting.

Although recent breakthroughs take place, the methods available are continuing to struggle with the challenge of achieving the ballast between precision, computational efficiency, and model generalization. The research gaps discussed below identify the main areas that need to be probed further and innovated in the coming years.

7.1. Precise Depth Approximation

The single 2D image estimation of depth is still one of the most longstanding and technologically challenging problems in three-dimensional reconstruction [44,93,62,47, 12,35,56,99]. The present CNN-based models are known to be effective in order setups but are unable to generalize when faced with irregular depth variation, occlusion or very limited contextual information. It could be beneficial to integrate with GAN-based refiners (and self-supervised depth learning) to allow networks to make better predictions on depth and ensure spatial consistency across a wide range of visual and photing conditions.

7.2. Handling Missing Data and Occlusions

The real-world images often contain places where visibility is not total or lacks visual data [35,51,56,99] which is a significant setback to the problem of accurate 3D reconstruction. The patch-filling or interpolation-based methods usually cause structural anomalies and artifacts in space [44,93,47]. Further development of hybrid CNN-SVM or GAN-based reconstruction models [90,28,16,58,53,80] that use context-aware completion and geometric completion arguments can have a significant positive effect on the performance of occlusion recovery and structural continuity in the reconstruction model.

7.3. Texture Preservation and Detail Integrity

Fine surface textures and material realism throughout 2D-to-3D transformation continue to be a serious unresolved problem [55,102,51,42]. Though algorithmic methods can now recreate figure geometrically correct conveying 3D shapes, they very often lose small scale visual details, giving it blurred or naturalistic textures. In future studies, it could be useful to combine texture synthesis with the GAN and SVM-directed feature refinement to provide consistent specification of the surface during its production [102,28,16,41] to produce a more realistic perceptual experience upon reconstruction.

7.4. Scalability and Computational Efficiency

The majority of deep learning models that have been applied to 2D-to-3D reconstruction, especially CNNs and GANs, are computationally expensive and require large amounts of memory [51,62,107,66]. This complexity provides a disadvantage to be used in embedded systems and even mobile devices. In order to alleviate these shortcomings, researchers are considering lightweight models, pruning and quantization schemes and distributed learning pipelines [10]. These methods have potentials of being able to realize a scalable, low-energy and real-time 3D reconstruction across edge devices.

Overall, the given research gaps indicate that the evolution of the hybrid deep learning models is still required to be made more accurate, flexible, and efficient. By addressing such obstacles, the emerging generation of 3D vision systems will attain real-time, interpretable or cross-domain functionality across a broad field of application, such as robotics, medical imaging or virtual reality [58,53,10].

8. CONCLUSION AND FUTURE DIRECTIONS

In this survey we provided a detailed overview of the deep learning-based methodologies to 2D-to-3D image reconstruction, specifically, convolutional neural networks, generative adversarial networks, support vector machines, and combinations of these. The systematic study of geometry-driven, learning-driven, and hybrid paradigm provided an insight into the effectiveness of hybrid CNN GAN SVM architectures to integrate spatial feature learning, generative realism and discriminative refinement to enhance reconstruction fidelity and strength. Comparative study of different application areas showed that CNNbased models are more effective in feature extraction and depth estimation, GAN-based models are more suitable in perceptual realism and SVMs provide interpretability and decision level optimization.

Although significant advances have been made, there are a few problems that are yet to be solved, such as depth ambiguity in monocular reconstruction, a weak extrapolation of datasets, excessive computational demands, and the inability to explain deep reconstruction models. These issues limit the utilization of the available methods in resource-based and real-time set-ups.

Future studies ought to seek to come up with scalable and explainable hybrid architectures that offer a balance between reconstruction accuracy and computational efficiency.

The combination of new paradigms like Neural Radiance Fields, Vision Transformers, and diffusion-based generative models should deliver avenues of improving implicit 3D scene representation and realism. Moreover, uncertainty modeling, self-supervised learning, and edge-sensitive optimization are another way to make the system more resilient and applicable to real-world systems. On the whole, the purpose of this survey is to become a structured reference and research roadmap towards the next-generation 2D-to-3D reconstruction system development in the field of artificial intelligence.

List of Acronyms

2D Two-Dimensional
3D Three-Dimensional
AI Artificial Intelligence
ANN Artificial Neural Network
AR Augmented Reality
CAD Computer-Aided Design
CNN Convolutional Neural Network
CPU Central Processing Unit
DL Deep Learning
DNN Deep Neural Network
GAN Generative Adversarial Network
GPU Graphics Processing Unit
MRI Magnetic Resonance Imaging
NeRF Neural Radiance Fields
ReLU Rectified Linear Unit
RGB Red–Green–Blue
SfM Structure from Motion
SLAM Simultaneous Localization and Mapping
SVM Support Vector Machine
ViT Vision Transformer
VR Virtual Reality

Acknowledgement

The authors sincerely thank Dr. Mohit Ranjan Panda for his invaluable advice and guidance during the course of this research. The authors also acknowledge the School of Computer Engineering, Kalinga Institute of Industrial Technology (Deemed to be University), Bhubaneswar, India, for providing academic support and facilities essential to completing this work.

Data Availability: The data used in this study are openly accessible.

Funding: No funding was received for conducting this study.

Conflicts of Interest: The authors declare no conflicts of interest.

REFERENCES

1. S. Abdulwahab, H. A. Rashwan, M. A. Garcia, M. Jabreel, S. Chambon, and D. Puig. Adversarial learning for depth and viewpoint estimation from a single image. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2947–2958, 2020.
2. N. R. Abirami, P. M. D. R. Vincent, K. Srinivasan, U. Tariq, and C. Y. Chang. Deep cnn and deep gan in computational visual perception-driven image analysis. *Complexity*, 2021(1):5541134, 2021.
3. Aggarwal, M. Mittal, and G. Battineni. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, 1(1):100004, 2021.
4. F. Alanazi. Comparative analysis of deep fake detection techniques. In *Proceedings of the 14th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 119–124, 2022.
5. F. Alidoost, H. Arefi, and F. Tombari. 2d image-to-3d model: Knowledge-based 3d building reconstruction using single aerial images and convolutional neural networks. *Remote Sensing*, 11(19):2219, 2019.

6. Alotaibi. Deep generative adversarial networks for image-to-image translation: A review. *Symmetry*, 12(10):1705, 2020.
7. L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, and future directions. *Journal of Big Data*, 8:1–74, 2021.
8. T. I. Anderson, B. Vega, J. McKinzie, S. A. Aryana, and A. R. Kovscek. 2d-to-3d image translation of complex nanoporous volumes using generative networks. *Scientific Reports*, 11(1):20768, 2021.
9. T. E. Andreassen, D. R. Hume, L. D. Hamilton, S. E. Higinbotham, and K. B. Shelburne. Automated 2d and 3d finite element overclosure adjustment and mesh morphing using generalized regression neural networks. *Medical Engineering and Physics*, 126:104136, 2024.
10. K. Bayoudh. A survey of multimodal hybrid deep learning for computer vision: Architectures, applications, trends, and challenges. *Information Fusion*, 105:102217, 2024.
11. Belevich and E. Jokitalo. Deepmib: User-friendly and open-source software for training of deep learning network for biological image segmentation. *PLoS Computational Biology*, 17(3):e1008374, 2021.
12. Bhoi. Monocular depth estimation: A survey. *arXiv preprint arXiv:1901.09402*, 2019.
13. Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.
14. M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.
15. M. M. Chanu and K. Thongam. Retracted article: Computer-aided detection of brain tumor from magnetic resonance images using deep learning network. *Journal of Ambient Intelligence and Humanized Computing*, 12(7):6911–6922, 2021. Retracted.
16. S. Y. Chen, K. H. Hsu, and T. H. Kuo. Hyperspectral target detection-based 2d-3d parallel convolutional neural networks for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
17. Y. Chen, A. G. Christodoulou, Z. Zhou, F. Shi, Y. Xie, and D. Li. Mri super-resolution with gan and 3d multi-level densenet: Smaller, faster, and better. *arXiv preprint arXiv:2003.01217*, 2020.
18. G. Coiffier and P. Renard. 3d geological image synthesis from 2d examples using generative adversarial networks. In *Petroleum Geostatistics 2019*, pages 1–5. European Association of Geoscientists & Engineers, 2019.
19. Sanket Kua Dale, Sukhada Dole, Mahipal Mehta, and Shweta Kothawade. Automated 2d images to
20. 3d model: Construction—a survey. *International Journal of Engineering Research and Technology (IJERT)*, 10(6), 2021.
21. Rameshwar Dhondge, Shreya Jadhav, Atul Gunjal, and Pranjali Joshi. 2d to 3D image conversion using machine learning approach. *International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)*, 8(3), 2019.
22. Diakite, G. Jiangsheng, and F. Xiaping. Hyperspectral image classification using 3d–2d cnn. *IET Image Processing*, 15(5):1083–1092, 2021.

23. Y. Duan, X. Guo, and Z. Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 432–449, 2024.
24. 2024.
25. C. Dugas, Y. Bengio, F. Belisle, C. Nadeau, and R. Garcia. Incorporating second-order functional knowledge for better option pricing. In Advances in Neural Information Processing Systems, volume 13, 2000.
26. D. Esan, P. A. Owolawi, and C. Tu. 3d image generation using generative adversarial network for virtual art gallery. The Indonesian Journal of Computer Science, 13(6), 2024.
27. T. Feng, W. Wang, R. Quan, and Y. Yang. Shape2scene: 3d scene representation learning through pre-training on shape data. In Proceedings of the European Conference on Computer Vision (ECCV), pages 73–91, 2024.
28. Ferreira, J. Li, K. L. Pomykala, J. Kleesiek, V. Alves, and J. Egger. Gan-based generation of realistic 3d volumetric data: A systematic review and taxonomy. Medical Image Analysis, 93:103100, 2024.
29. E. Fourati, W. Elloumi, and A. Chetouani. Anti-spoofing in face recognition-based biometric authentication using image quality assessment. Multimedia Tools and Applications, 79(1):865–889, 2020.
30. M. T. Ghazal and K. Abdullah. Face recognition based on curvelets, invariant moments features and svm. TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(2):733–739, 2020. [29] Francisco Gomez-Donoso, Alberto Garc'ia-Garc'ia, Jorge Garc'ia-Rodr'iguez, Sergio Orts-Escolano, and Miguel Cazorla. Lonchanet: A slice-based cnn architecture for real-time 3d object recognition. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), pages 412–419, 2017.
31. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020.
32. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2014.
33. Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun. Deep learning for 3d point clouds: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(12):4338–4364, 2020.
34. Isabelle Guyon, Bernhard Boser, and Vladimir Vapnik. Automatic capacity tuning of very large vc-dimension classifiers. In Advances in Neural Information Processing Systems, volume 5, 1992.
35. X. F. Han, H. Laga, and M. Bennamoun. Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(5):1578–1604, 2019.
36. P. Henderson and V. Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. International Journal of Computer Vision, 128(4):835–854, 2020.
37. W. Hu, Y. Wang, L. Ma, B. Yang, L. Gao, X. Liu, and Y. Ma. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 19774–19783, 2023.

39. J. Hussain, M. Bath, and J. Ivarsson. Generative adversarial networks in medical image reconstruction:° A systematic literature review. *Computers in Biology and Medicine*, 191:110094, 2025.
40. M. R. Islam, M. A. Akhand, M. A. Kamal, and K. Yamada. Recognition of emotion with intensity from speech signal using 3d transformed feature and deep learning. *Electronics*, 11(15):2362, 2022.
41. K. S. Joga, K. K. Sree, N. Spandana, and G. G. Pushpa. 3d reconstruction of regular objects from multiple 2d images using a reference object. *International Journal of Engineering Research and Technology*, 2023.
42. Khan, Z. Zhijiang, Y. Yu, M. A. Khan, K. Yan, and K. Aziz. Gan-holo: Generative adversarial networks-based generated holography using deep learning. *Complexity*, 2021(1):6662161, 2021.
43. J. C. Kim, M. H. Kim, H. E. Suh, M. T. Naseem, and C. S. Lee. Hybrid approach for facial expression recognition using convolutional neural networks and svm. *Applied Sciences*, 12(11):5493, 2022.
44. J. H. Koh, S. Y. Tan, and M. F. Nasrudin. A systematic literature review of generative adversarial networks (gans) in 3d avatar reconstruction from 2d images. *Multimedia Tools and Applications*, 83(26):68813–68853, 2024.
45. D. S. Lee, J. S. Kim, S. C. Jeong, and S. K. Kwon. Human height estimation by color deep learning and depth 3d conversion. *Applied Sciences*, 10(16):5531, 2020.
46. J. Lee, H. Jung, Y. Kim, and K. Sohn. Automatic 2d-to-3d conversion using multi-scale deep neural network. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 730–734, 2017.
47. J. Y. Lee, K. Kwon, C. Kim, and S. Youm. Development of a non-contact sensor system for converting 2d images into 3d body data: A deep learning approach to monitor obesity and body shape in individuals
48. in their 20s and 30s. *Sensors*, 24(1):270, 2024.
49. Y. Li, H. Zhang, K. Xu, and J. Han. Hybrid neural radiance fields for multi-view 3d reconstruction: Integrating cnns, gans, and transformer representations. *IEEE Transactions on Image Processing*, 33:1452–1466, 2024.
50. Y. Z. Li, S. J. Zheng, Z. X. Tan, T. Cao, F. Luo, and C. X. Xiao. Self-supervised monocular depth estimation by digging into uncertainty quantification. *Journal of Computer Science and Technology*, 38(3):510–525, 2023.
51. Cheng-Han Lin, Jun Gao, Linjie Tang, Towaki Takikawa, Xiaohang Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In
52. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2023.
53. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van der Laak, B. Van Ginneken, and C. I. Sanchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.
54. T. Liu, S. Geng, Y. Fu, Z. Lei, Y. Huo, X. Zhang, F. Wang, B. Han, M. Sha, and Z. Wu. Triplane generator-based nerf-gan framework for single-view ship reconstruction. *International Journal of Digital Earth*, 18(1):2496406, 2025.
55. J. Lu, Y. Yang, R. Liu, S. B. Kang, and J. Yu. 2d-to-stereo panorama conversion using gan and concentric mosaics. *IEEE Access*, 7:23187–23196, 2019.
56. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 30, page 3, 2013.

57. S. S. Mamand and A. I. Abdulla. 2d to 3d image conversion algorithms. In ITM Web of Conferences, volume 64, page 01010, 2024.
58. M. Mansour, J. Martens, and J. Blankenbach. Hierarchical svm for semantic segmentation of 3d point clouds for infrastructure scenes. *Infrastructures*, 9(5):83, 2024.
59. R. T. Marriott, S. Romdhani, and L. Chen. A 3d gan for improved large-pose facial recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13445–13455, 2021.
60. R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7210–7219, 2021.
61. B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
62. H. Mirzaee, S. Kamrava, and P. Tahmasebi. Minireview on porous media and microstructure reconstruction using machine learning techniques: Recent advances and outlook. *Energy & Fuels*, 37(20):15348–15372, 2023.
63. M. Muzahid, H. Han, Y. Zhang, D. Li, Y. Zhang, J. Jamshid, and F. Sohel. Deep learning for 3d object recognition: A survey. *Neurocomputing*, 608:128436, 2024.
64. J. M. Newby, A. M. Schaefer, P. T. Lee, M. G. Forest, and S. K. Lai. Convolutional neural networks automate detection for tracking of submicron-scale particles in 2d and 3d. *Proceedings of the National Academy of Sciences*, 115(36):9026–9031, 2018.
65. T. Pokorny, J. Vrba, O. Fiser, D. Vrba, T. Drizdal, M. Novak, L. Tosi, A. Polo, and M. Salucci. On the role of training data for svm-based microwave brain stroke detection and classification. *Sensors*, 23(4):2031, 2023.
66. R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021.
67. M. Rasheed, M. W. Iqbal, A. Jaffar, M. U. Ashraf, K. A. Almarhabi, A. M. Alghamdi, and A. A.
68. Bahaddad. Recognizing brain tumors using adaptive noise filtering and statistical features. *Diagnostics*, 13(8):1451, 2023.
69. C. A. Rivera-Romero, J. U. Munoz-Minjares, C. Lastre-Dominguez, and M. Lopez-Ramirez. Optimal~ image characterization for in-bed posture classification by using svm algorithm. *Big Data and Cognitive Computing*, 8(2):13, 2024.
70. S. Sabnam and S. Rajagopal. Application of generative adversarial networks in image, face reconstruction and medical imaging: challenges and the current progress. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 12(1):2330524, 2024.
71. T. Samavati and M. Soryani. Deep learning-based 3d reconstruction: A survey. *Artificial Intelligence Review*, 56(9):9175–9219, 2023.
72. G. Sanil, K. Prakasha, S. Prabhu, and V. Nayak. Facial similarity measure for recognizing monozygotic twins utilizing 3d facial landmarks, efficient geodesic distance computation, and machine learning algorithms. *IEEE Access*, 2024.
73. S. Sengan, K. Kotecha, I. Vairavasundaram, P. Velayutham, V. Varadarajan, L. Ravi, and S. Vairavasundaram. Real-time automatic investigation of indian roadway animals by 3d reconstruction detection using deep learning for r-3d-yolov3 image classification and filtering. *Electronics*, 10(24):3079, 2021.

74. H. Shan, X. Jia, P. Yan, Y. Li, H. Paganetti, and G. Wang. Synergizing medical imaging and radiotherapy with deep learning. *Machine Learning: Science and Technology*, 1(2):021001, 2020.
75. N. Sharma, D. Mishra, P. Yadav, M. Singh, and P. Mukherjee. 3d image modeling using data science. In
76. *Recent Advances in Computational Intelligence and Cyber Security*, pages 311–320. CRC Press, 2020.
77. R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su. Zero123++: A single image to consistent multi-view diffusion-based model. *arXiv preprint arXiv:2310.15110*, 2023.
78. M. A. Shoukat, A. B. Sargano, A. Malyshev, L. You, and Z. Habib. Ss3dnet-af: A single-stage, single-view 3d reconstruction network with attention-based fusion. *Applied Sciences*, 14(23), 2024.
79. D. Shu, J. Cunningham, G. Stump, S. W. Miller, M. A. Yukish, T. W. Simpson, and C. S. Tucker. 3d design using generative adversarial networks and physics-based validation. *Journal of Mechanical Design*, 142(7):071701, 2020.
80. N. Subramanian, O. Elharrouss, S. Al-Maadeed, and A. Bouridane. Image steganography: A review of the recent advances. *IEEE Access*, 9:23409–23423, 2021.
81. J. S. Suri, K. Liu, S. Singh, S. N. Laxminarayan, X. Zeng, and L. Reden. Shape recovery algorithms using level sets in 2-d/3-d medical imagery: A state-of-the-art review. *IEEE Transactions on Information Technology in Biomedicine*, 6(1):8–28, 2002.
82. Z. Tang, C. Xu, and S. Yan. An adaptive interpolation and 3d reconstruction algorithm for underwater images. *Machine Vision and Applications*, 35(2):33, 2024.
83. O. Taybi, T. Gadi, and R. Alaoui. 2dslicesnet: A 2d slice-based convolutional neural network for 3d object retrieval and classification. *IEEE Access*, 9:24041–24049, 2021.
84. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, Y. Wu, et al. Advances in neural rendering. *Computer Graphics Forum*, 41(2):703–735, 2022.
85. N.M. Thalmann, X. Hu, B. Sheng, D. Thalmann, and T. Peng. Transformer-nerf hybrid for implicit 3d reconstruction and realistic animation. In *Computer Animation and Social Agents (CASA 2024)*, 37th International Conference, Wuhan, China, June 5–7, 2024, Revised Selected Papers, Part II, *Lecture Notes in Computer Science*, pages 112–125. Springer, 2024.
86. Hsiao-Yu Tung, Adam W. Harley, Wei Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4364–4372, 2017. [81] Vladimir Vapnik. The nature of statistical learning theory. *Machine Learning*, 20:273–297, 1995.
87. Vladimir Vapnik, Steven Golowich, and Alex Smola. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems*, volume 9, 1996.
88. Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1998.
89. Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, 2006.
90. Vladimir N. Vapnik and Alexey Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, 1982.
91. M. Varga, J. Jadlovsky, and S. Jadlovsk´ a. Generative enhancement of 3d image classifiers. *Applied Sciences*, 10(21):7433, 2020.

92. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis. Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018:7068349, 2018.
93. M. E. Walizad and M. Hurroo. Sign language recognition system using convolutional neural network and computer vision. *International Journal of Engineering Research and Technology*, 9(12):59–64, 2020.
94. P. S. Wang, C. Y. Sun, Y. Liu, and X. Tong. Adaptive o-cnn: A patch-based deep representation of 3d shapes. *ACM Transactions on Graphics*, 37(6):1–13, 2018.
95. Y. Wang, L. Zhou, M. Wang, C. Shao, L. Shi, S. Yang, Z. Zhang, M. Feng, F. Shan, and L. Liu. Combination of generative adversarial network and convolutional neural network for automatic subcentimeter pulmonary adenocarcinoma classification. *Quantitative Imaging in Medicine and Surgery*, 10(6):1249–1264, 2020.
96. Z. Wang, L. T. Yang, B. Ren, J. Zhao, Z. Li, and G. Zeng. Capturing detail variations for lightweight neural radiance fields. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3935–3939, 2024.
97. L. Xiao and J. Wang. Optimization of 3d animation design based on support vector machine algorithm.
98. *International Journal of Multimedia and Ubiquitous Engineering*, 10(6):225–234, 2015.
99. J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 842–857, 2016.
100. M. Xing, S. Qu, and Q. Zhang. Exploration of neural radiance fields: A review. In *Proceedings of the International Conference on Applied Intelligence*, pages 311–322, 2024.
101. Q. C. Xu, T. J. Mu, and Y. L. Yang. A survey of deep learning-based 3d shape generation. *Computational Visual Media*, 9(3):407–442, 2023.
102. P. S. Yeoh, K. W. Lai, S. L. Goh, K. Hasikin, Y. C. Hum, Y. K. Tee, and S. Dhanalakshmi. Emergence of deep learning in knee osteoarthritis diagnosis. *Computational Intelligence and Neuroscience*, 2021:4931437, 2021.
103. X. Yi, E. Walia, and P. Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.
104. X. Ying, H. Guo, K. Ma, J. Wu, Z. Weng, and Y. Zheng. X2ct-gan: Reconstructing ct from biplanar x-rays with generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
105. Yu, V. Ye, M. Tancik, and A. Kanazawa. Pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021.
106. Chunyan Yu, Rui Han, Meiping Song, Caiyu Liu, and Chein I. Chang. A simplified 2d–3d cnn architecture for hyperspectral image classification based on spatial–spectral fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020.
107. J. Zhang, Y. Wu, and H. Jiang. Survey on monocular metric depth estimation. *Computers*, 14(11):502, 2025.
108. M. Zhang and Y. Zheng. Hair-gan: Recovering 3d hair structure from a single image using generative adversarial networks. *Visual Informatics*, 3(2):102–112, 2019.

109. X. Zhang, H. Wang, and H. Dong. A survey of deep learning-driven 3d object detection: Sensor modalities, technical architectures, and applications. *Sensors*, 25(12):3668, 2025.
110. Y. Zhao, H. Qin, L. Xu, H. Yu, and Y. Chen. A review of deep learning-based stereo vision techniques for phenotype feature and behavioral analysis of fish in aquaculture. *Artificial Intelligence Review*, 58(1):1–61, 2025.
111. D. Zhou, X. Song, J. Fang, Y. Dai, H. Li, and L. Zhang. Context-aware 3d object detection from a single image in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):18568–18580, 2022.
112. Y. Zhou, Y. Jin, A. Luo, S. Chan, X. Xiao, and X. Yang. Toonnet: A cartoon image dataset and a dnn-based semantic classification system. In *Proceedings of the 16th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, pages 1–8, 2018.
113. J. Zhu, J. Xie, and Y. Fang. Learning adversarial 3d model generation with 2d image enhancer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
114. J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.