

A Hybrid Deep Learning Model For 3D Scene Alignment And Reconstruction With 3D U-Net Network

Naveen Kumar N¹, GiribabuKande ², and B Prabhakar Rao³

¹ Research Scholar, Department of ECE, JNTUK, Kakinada, 533003, Andra Pradesh, India

²Professor, ECE Department, VVIT, Nambur, 522509, Andra Pradesh, India

³Professor, ECE Department, JNTUK, Kakinada, 533003, Andra Pradesh, India

Abstract

Reconstructing 3D geometry from multi-view 2D RGB imagery has emerged as a key research focus due to its wide applicability in areas such as virtual reality, robotic perception, and human-computer interaction. This work presents a hybrid deep learning framework that integrates 3D U-Net with transformer-based attention mechanisms. The architecture consists of a visual autoencoder and a dedicated 3D refinement network. Feature representations extracted from unordered image sets are aggregated through a self-attention module to produce a rich global feature map, which is decoded into an initial volumetric 3D representation and subsequently refined for improved structural accuracy. The proposed model supports 3D reconstruction from RGB inputs captured from arbitrary viewpoints. Experimental validation on ShapeNet and PASCAL datasets demonstrates that the proposed method consistently surpasses existing state-of-the-art approaches, achieving superior mean Intersection over Union (IoU) performance.

Keywords: U-Net, Encoder, Decoder, 3D Reconstruction, IoU, ShapeNet

1. Introduction

Image-based 3D reconstruction aims to determine the geometry and structure of objects and scenes from 2D photographs. Many applications, including robot navigation, object recognition, 3D modeling, industrial control, and medical diagnostics, rely on this unsolved problem. The purpose of traditional Multiview stereo and shape-from-X techniques, which have been thoroughly studied for many decades, has been to recover the lost dimension from simply 2D images. In order to develop mathematical or algorithmic solutions to the ill-posed inverse problem, the initial generation of approaches used a geometric approach to the problem and concentrated on comprehending and formally formalizing the 3D to 2D projection process.

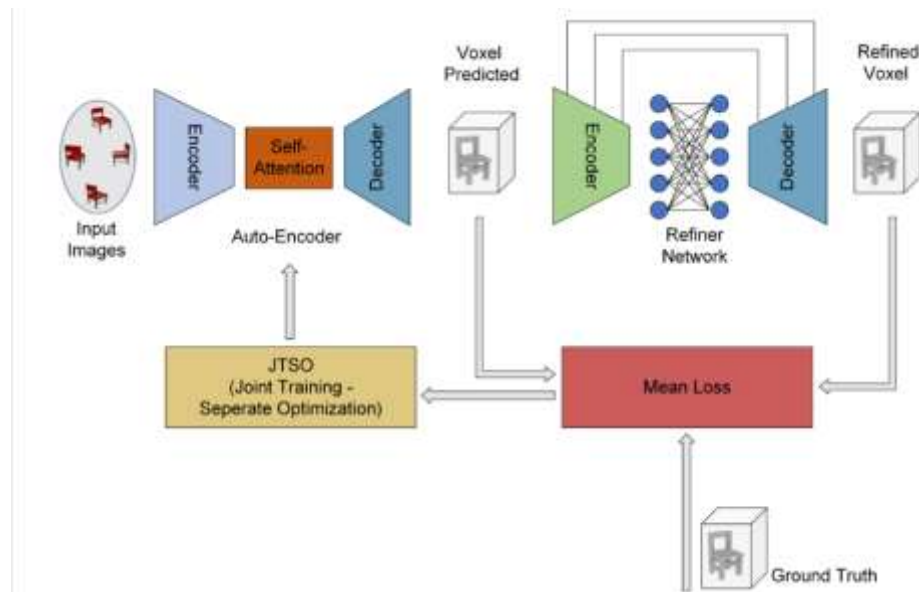
Several images taken with precisely calibrated cameras are usually needed for effective solutions. For instance, stereo-based methods [1] use the triangulation concept to reconstruct the 3D coordinates of the picture pixels after matching characteristics between images taken from slightly different viewing angles. Accurately segmented 2D silhouettes are necessary for shape-from-silhouette and shape-by-space-carving techniques [2]. It's interesting to note that humans are adept at using past information to solve such ill-posed inverse problems. With just one eye, they are able to deduce an object's approximate size and general geometry. They are even able to speculate on how it would appear from a different perspective. We are able to do this because all of the scenes and objects we have already seen have helped us create mental models of what items look like and establish prior knowledge. By framing the 3D reconstruction problem as a recognition problem, the second generation of 3D reconstruction techniques attempted to take advantage of this existing knowledge. A new generation of techniques that can recover the 3D geometry and structure of objects from one or more RGB images without the complicated camera calibration process has been made possible by deep learning techniques and, more significantly, by the growing availability of large training data sets. Even though these techniques are new, they have shown intriguing and encouraging outcomes on a variety of computer vision and graphics-related tasks.

2. Related Work

Previous research in the field of 3D model reconstruction has explored various methodologies aimed at enhancing the accuracy, efficiency, and inference time of generated 3D model representations. These approaches range from single-view reconstruction to more complex multi-view methods. In this section, we delve into the developments in this area, focusing on the different methodologies and the challenges they address. One classical computer-vision challenge is extracting three-dimensional information from images. Conventional techniques include visual Simultaneous Localization and Mapping (vSLAM) [3] and Structure from Motion (SfM) [4] to match manually crafted features across several views and estimate the camera pose for each image. In response to the limitations of traditional methods, deep learning approaches have emerged, utilizing extensive datasets like ShapeNet [5] and Pascal 3D+ [6]. Based on these datasets, several efficient deep-learning methods for 3D reconstruction have been developed. Most of these techniques utilize a shared encoder-decoder design. GRU units are used in LSM [7] techniques to fuse deep image features and create 3D models, which introduces permutation variance and gradient issues. Besides the encoder-decoder structure, LSM [7] includes a 3D Grid Reasoning module for generating final 3D models and depth maps. Methods like DeepMVS [8] and 3DensNet use pooling operations to fuse image features extracted by the shared encoder. However, pooling operations capture only first-order information, resulting in the loss of valuable information. Some of the learning-based works try to encode the shape prior into a feature vector but stick to the traditional representations by decoding the vector into mesh [9-13], voxel [14-18] or point cloud [19, 20]. Others try to learn structured representations which decompose the shape into simple shapes [21, 22]. Recently, implicit surface function [23-26] has been widely used as a new representation method to overcome the disadvantages of traditional methods. We propose an innovative CNN architecture for 3D model reconstruction capable of generating 3D voxels from one or multiple 2D images captured from various viewpoints. Our comprehensive evaluation using the ShapeNet Core dataset demonstrates that the methods proposed in this study surpass state-of-the-art results, particularly with a small number of input images.

3. Methods

Our method introduces a technique to generate 3D models from RGB images, both single and multi-view, utilizing a predicted grid of voxel occupancy in three dimensions. The architecture of the network is depicted in Fig. 1. Our method introduces a novel network architecture for 3D model reconstruction, complemented by a dedicated training algorithm to improve efficiency. The architecture begins with a shared CNN encoder-decoder framework, followed by a 3D U-Net for refining features. Addressing feature fusion challenges, we incorporate an attention network inspired by Transformer models to effectively transform latent space features. For robust feature extraction and object generation, we employ a ResNet18-based autoencoder architecture. Additionally, refiner networks based on 3D U-Net optimize and enhance the 3D structures generated by decoders. Mean cross-entropy loss from intermediate and refined 3D models is used to update the encoder decoder and attention module. We used the Adam optimizer with a step-wise decreasing learning rate. Detailed descriptions of each module used in our architecture are provided in the following sections.

Fig1. Proposed Architecture

The input images undergo processing through an encoder network designed to extract a specific set of deep image features using 2D convolution layers. We utilize the ResNet-V1 architecture comprising 12 layers, to extract 1024 feature vectors from images sized $127 \times 127 \times 3$. The encoder network consists of 12 convolution layers organized into 6 residual blocks. Each residual block includes 2 convolutional blocks with the Leaky ReLU activation function, followed by a max-pool layer and an identity connection. Identity connections incorporate a 2D convolution layer with a 1×1 kernel size. The output channel sizes of the residual blocks are 96, 128, 256, 256, 256, and 256 respectively. All max-pool layers have a kernel size and stride of 2, and identity connections follow each residual block except for the 4th block. Ultimately, the encoder network produces flattened feature vectors consisting of 1024 dimensions. The decoder generates 3D models from the aggregated feature output produced by the Attention-Net block. It takes as input a 1024-dimensional vector obtained from the self-attention layer. The decoder consists of five residual blocks with output filter sizes of 128, 128, 128, 64, 64, and 32, respectively. Each transpose convolution layer in the decoder has a kernel size of 3.

3.1 Datasets

We use different datasets as shown in fig 2,3 to train each module individually and jointly. i) ShapeNet: The ShapeNet [7] dataset is a collection of 3D CAD models that are organized according to the WordNet hierarchy. We used a subset of the ShapeNet dataset which consists of 50,000 models and 13 major categories.

Table 1. Overview of the dataset

S.No	Datasets	Number of Frames	Number of Labels	Object	5 Common Classes
1.	PASCAL3D+ [22]	30,899	12	3D CAD Scans	Boat, Bus, Car, Chair, Sofa

2.	Shape Net [10]	220,000	3135	Scans of Artefact,	Plant, Person, Table, Car, Chair, Sofa, Rifle
----	----------------	---------	------	--------------------	---

Table 1. Gives an overview of the dataset used. We split the dataset into training and testing sets, with 4/5 for training and the remaining 1/5 for testing. We refer to these two datasets as the ShapeNet [7] training set and testing set throughout the experiments section. ii) PASCAL 3D: The PASCAL 3D dataset is composed of PASCAL 2012 detection images augmented with 3D CAD model alignment [8]. Online Products: The dataset contains images of 23,000 items sold online. MVS and SFM methods fail on these images due to ultra-wide baselines. Since the dataset does not have the ground-truth 3D CAD models, we only used the dataset for qualitative evaluation.

Fig 2. Sample images from Shape Net



Fig 3. Sample images from PASCAL 3D



To assess the performance of the 3D reconstruction, we employed the Intersection over Union (IoU) metric. This metric is a standard for evaluating the accuracy of object reconstruction by comparing the predicted voxel occupancy with the ground truth voxel occupancy. Specifically, we binarized the predicted probabilities using a fixed threshold of 0.25 to determine voxel occupancy. The IoU is calculated as follows:

$$IOU = \frac{\sum_{i,j,k} I(p_{(i,j,k)} > t) I(gt_{(i,j,k)})}{\sum_{i,j,k} I[I(p_{(i,j,k)} > t) I(gt_{(i,j,k)})]} \dots\dots\dots [1]$$

In this equation [1], (i,j,k) represents the predicted probability of occupancy at the voxel coordinate (i, j, k) , and $gt(i,j,k)$ denotes the ground truth occupancy at the same coordinate.

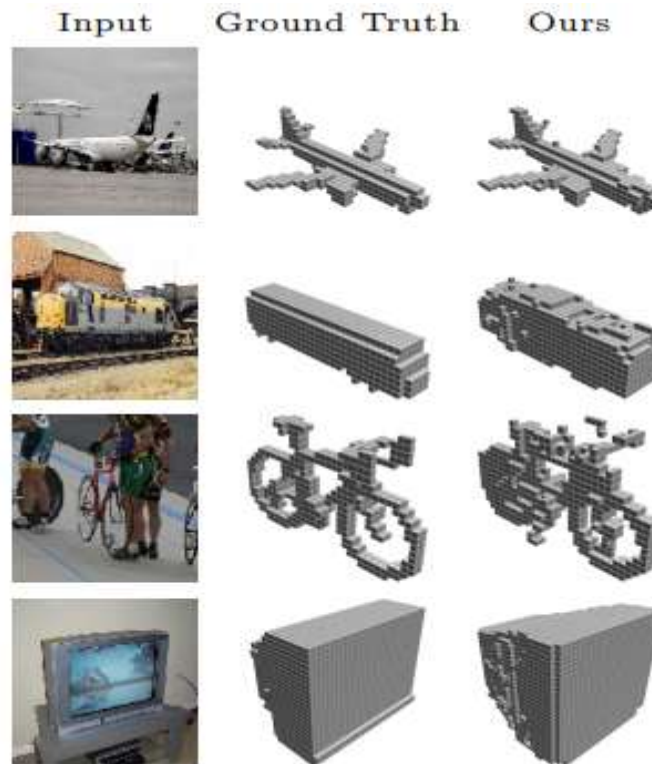
3.2 Implementation

We utilize 127×127 RGB images as input to train the proposed network, with a batch size of 4. The voxelized output model is produced with a resolution of 32^3 . Our implementation leverages TensorFlow 2.8.0 and Python 3.9.8, alongside CUDA 12.2 as the backend driver and library, respectively.

4. Experimental Results

We evaluated the performance of our network in single-view reconstruction using real-world images. To make a quantitative comparison, we used images from the Shape Net dataset [7] and its corresponding 3D models from the PASCAL 3D+ dataset [8]. The 3D reconstructed images are as shown in figure 4. We first investigate the quality of the reconstructed models under different numbers of views. Fig. 5(a) and (b) show that reconstruction quality improves as the number of views increases. We also report the reconstruction IoUs on each of the 13 major categories in the testing set in Table 2.

Fig 4. 3D Reconstructed images PASCAL dataset





We observed that the reconstruction quality improved for every category as the number of views increased, but the quality varied depending on the category. Cabinets, cars, and speakers had the highest reconstruction performance since the objects are bulky-shaped and have less (shape) variance compared to other classes. The network performed worse on the lamp, bench, and table categories. These classes have much higher shape variation than the other classes. For example, a lamp can have a slim arm or a large lampshade which may move around, and chairs and tables have various types of supporting structures.

Fig 5(a). Cross Entropy Loss

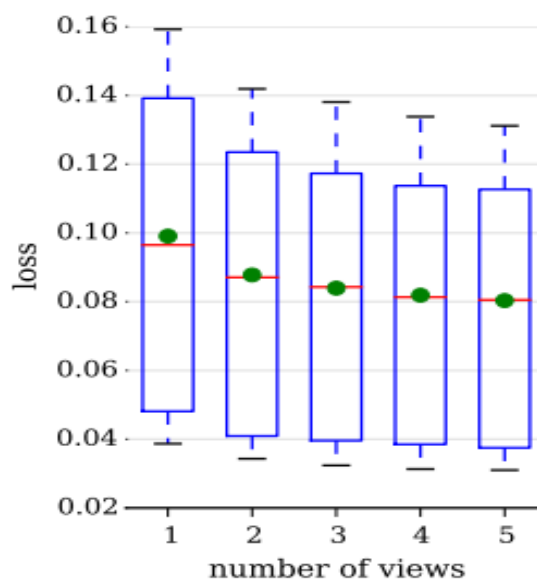
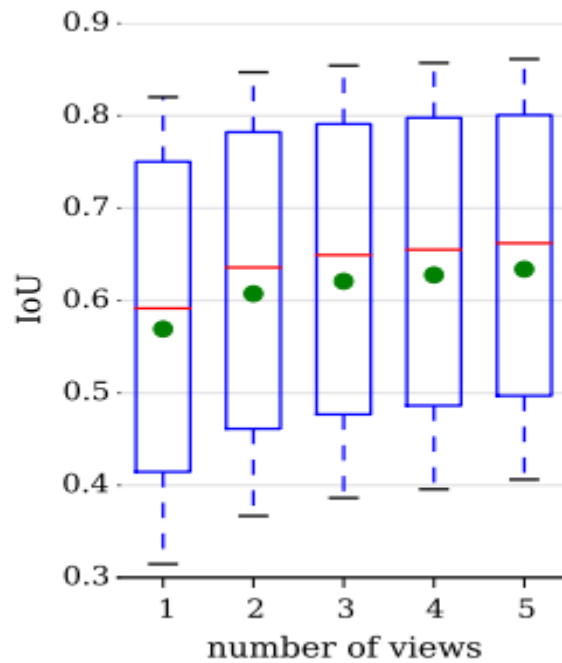


Fig 5(b). IoU Values.**Table 2. Mean IoU Performance vs Number of Views**

Category	1 View	2 Views	3 Views	4 Views	5 Views
plane	0.513	0.536	0.549	0.556	0.561
bench	0.421	0.484	0.502	0.516	0.527
cabinet	0.716	0.746	0.763	0.767	0.772
car	0.798	0.821	0.829	0.833	0.836
chair	0.466	0.515	0.533	0.541	0.55
monitor	0.468	0.527	0.545	0.558	0.565
lamp	0.381	0.406	0.415	0.416	0.421
speaker	0.662	0.696	0.708	0.714	0.717
firearm	0.544	0.582	0.593	0.595	0.6
couch	0.628	0.677	0.69	0.698	0.706
table	0.513	0.55	0.564	0.573	0.58
cellphone	0.661	0.717	0.732	0.738	0.754
watercraft	0.513	0.576	0.596	0.604	0.61

5. Conclusion

This work introduced a hybrid deep learning framework for reconstructing 3D geometry from multi-view 2D RGB images by effectively combining a 3D U-Net architecture with transformer-based attention mechanisms. By leveraging self-attention to aggregate features from unordered image sets, the proposed approach captures global contextual information and enables robust reconstruction from arbitrary viewpoints. The inclusion of a dedicated 3D refinement network further improves structural coherence and geometric accuracy of the reconstructed volumes. Experimental results on the ShapeNet and PASCAL datasets demonstrate that the proposed method consistently outperforms existing state-of-the-art approaches, achieving higher mean Intersection over Union

(IoU) scores. These results validate the effectiveness and generalization capability of the proposed framework, making it a promising solution for real-world 3D reconstruction applications in areas such as virtual reality, robotics, and human-computer interaction.

References

1. R. Hartley and A. Zisserman, Multiple view geometry in computer vision. Cambridge university press, 2003.
2. A. Laurentini, "The visual hull concept for silhouette-based image understanding," IEEE TPAMI, vol. 16, no. 2, pp. 150–162, 1994.
3. C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard, "Past, Present, and Future of Simultaneous Localization and Mapping: Towards the Robust-Perception Age," IEEE Transactions on Robotics, 32(6):1309–1332, 2016.
4. O. Ozyesil, V. Voroninski, R. Basri, and A. Singer, "A Survey of Structure from Motion," Acta Numerica, 26:305–364, 2017.
5. Li Yi , Lin Shao , Manolis Savva , Haibin Huang , Yang Zhou , Qirui Wang , Benjamin Graham , Martin Engelcke , Roman Klokov , Victor Lempitsky , Yuan Gan ." arge Scale 3D Shape Reconstruction and Segmentation from ShapeNet Core55". ECCV 2017
6. Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A benchmark for 3D object detection in the wild. In WACV 2014".
7. Abhishek Kar, Christian Häne, Jitendra Malik, "Learning a Multi-View Stereo Machine", NIPS 2017.
8. Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, Jia-Bin Huang, "DeepMVS: Learning Multi-view Stereopsis", CVPR 2018.
9. Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry, "A papier-mâché approach to learning 3d surface generation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 216–224, 2018.
10. Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images. In Eur. Conf. Comput. Vis., 2018.
11. Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia, "Deep mesh reconstruction from single rgb images via topology modification networks. In Int. Conf. Comput. Vis., pages 9964–9973, 2019.
12. Edward J Smith, Scott Fujimoto, Adriana Romero, and David Meger, "Geometrics: Exploiting geometric structure for graph-encoded objects. arXiv preprint arXiv:1901.11461, 2019.
13. Georgia Gkioxari, Jitendra Malik, and Justin Johnson, "Mesh r-cnn. In Int. Conf. Comput. Vis., pages 9785–9795, 2019.
14. Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao, "3d shapenets: A deep representation for volumetric shapes. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1912–1920, 2015.
15. Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In Eur. Conf. Comput. Vis., pages 628–644. Springer, 2016.
16. Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston, "Generative and discriminative voxel modeling with convolutional neural networks. arXiv preprint arXiv:1608.04236, 2016.
17. Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In Adv. Neural Inform. Process. Syst., pages 82–90, 2016.

18. David Stutz and Andreas Geiger. Learning 3d shape completion from laser scan data with weak supervision. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1955–1964, 2018.
19. Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. arXiv preprint arXiv:1706.07036, 2017.
20. Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In Int. Conf. Comput. Vis., pages 4541–4550, 2019.
21. Jun Li, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. Grass: Generative recursive autoencoders for shape structures. ACM Trans. Graph., 36(4):1–14, 2017.
22. Despoina Paschalidou, Ali Osman Ulusoy, and Andreas Geiger. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In IEEE Conf. Comput. Vis. Pattern Recog., pages 10344–10353, 2019.
23. Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4460–4470, 2019.
24. Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. arXiv preprint arXiv:1901.05103, 2019.
25. Qiangeng Xu, Weyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. arXiv preprint arXiv:1905.10711, 2019.
26. Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In Int. Conf. Comput. Vis., pages 2304–2314, 2019.

Author Bibliography

	Mr. Naveen Kumar Nekkalapu received the AMIETE degree in Electronics and Communication Engineering from IETE, New Delhi, in 2004, and the M.Tech degree in Electronics and Communication Engineering from Nagarjuna University, Guntur, in 2006. He is currently pursuing his Ph.D. in Electronics and Communication Engineering from Jawaharlal Nehru Technological University, Kakinada. Since 2006, he has been working as an Assistant Professor in the Department of Electronics and Communication Engineering at various institutions, including St. Mary's College of Engineering and Technology, SLC's Institute of Science and Technology, and Made Easy Institute Private Limited. His research interests include Image Processing and Computer Vision.
---	--

	Dr Giri Babu Kande (Senior Member, IEEE) received the B.Tech. degree in electronics and communication engineering from Nagarjuna University, Guntur, in 1996, the M.E. degree in electronics and communication engineering from Andhra University, Visakhapatnam, in 2000, and the Ph.D. degree in electronics and communication engineering from Jawaharlal Nehru Technological University, Hyderabad, in 2010. From 2008 to 2019, he was a Professor and the Head of the Department of Electronics and Communication Engineering Department. Since 2019, he has been a Professor and the Dean of Academics with the Vasireddy Venkatadri Institute of Technology, Nambur. He is the author of one book and more than 75 publications. His research interests include computer vision, statistical modeling and learning, and medical image analysis. He was a recipient of the Best Researcher by Jawaharlal Nehru Technological University, Kakinada, in 2018, and the ISTE AP Best Teacher Award, in 2021.
	Dr Rao has obtained his master's from SVU, Tirupati in 1982, PhD in Sonar Signal processing from Indian institute of Science, Bangalore in 1995. Dr. Bhima Prabhakara Rao is a Professor & Dean in the Department of Electronics and Communication Engineering at Godavari Global University Rajamahendravaram EGDistrict from March 2025. Prior to this he is having 43 years of Experience in teaching and administrative and Research experience. Dr Rao has contributed significantly in the areas of Optical Networks, Image Processing, Microwave Engineering, and emerging technologies such as Deep & Machine learning. He published 370 international And National journals and presented research papers in international And national conferences in India and Abroad like Singapore, Malaysia and USA. He served in different capacities starting from Hod, Vice Principal, Director of Evaluation, Rector and Vice Chancellor i/c.