ESP Journal of Engineering & Technology Advancements (ESP-JETA)
ISSN : 2583-2646
Special Issue

REVOLUTIONIZING MONOCULAR DEPTH ESTIMATION: ELEVATING ACCURACY WITH DEEP LEARNING

Mr.V.Mohanadas
AP/ECE
Mangayarkarasi college of engineering
Madurai

P.Vasanth
B.E - ECE
Mangayarkarasi college of engineering
Madurai

M.Jeevanantham
B.E - ECE
Mangayarkarasi college of engineering
Madurai

B.Veerapandian
B.E - ECE
Mangayarkarasi college of engineering
Madurai

Abstract- Monocular depth estimation, a fundamental task in computer vision, has seen significant advancements through deep learning techniques. Traditionally reliant on stereo camera setups, recent developments have focused on leveraging deep convolutional neural networks (CNNs) to estimate depth from single RGB images. This project investigates the evolution of monocular depth estimation in Computer Vision, shifting from conventional depth estimation techniques to deep learning CNN algorithms. It delves into the utilization of CNNs to estimate depth from single RGB images, contrasting with traditional stereo methods. Specifically, it evaluates the performance of three pre-trained CNN models as encoders within depth estimation networks. Furthermore, it introduces a deep learning approach to improve accuracy by aggregating predictions from multiple models. Through rigorous experimentation, the deep learning method demonstrates superior performance compared to existing benchmarks for monocular depth estimation. This research underscores the transformative potential of deep learning CNN algorithms in significantly enhancing the accuracy and reliability of depth estimation in various real-world scenarios.

Keywords-Monocular depth estimation, Computer vision, Deep learning, Convolutional neural networks (CNNs)

I.INTRODUCTION

This paper introduces a novel approach to depth estimation from single images, leveraging multiscale convolutional neural networks and continuous Conditional Random Fields (CRFs). By integrating complementary information through CRFs, the proposed models achieve state-of-the-art results, validated through extensive experimental evaluation on public

datasets[1],This paper introduces a novel convolutional neural network leveraging transfer learning to compute high-resolution depth maps from single RGB images, outperforming existing solutions with enhanced detail and accuracy. By utilizing a refined encoder-decoder architecture with pre-trained networks, our approach achieves state-of-the-art results on benchmarks with efficient training[2]Addressing the challenge of dense depth prediction from sparse data, our study presents a novel approach leveraging a single deep regression network trained on RGB-D raw data, showcasing a significant 50% reduction in prediction error with the integration of 100 additional depth samples. Our findings, demonstrated on benchmark datasets, offer promising applications including enhanced SLAM mapping and LiDAR super-resolution.[3]a versatile multiscale convolutional network capable of addressing depth prediction, surface normal estimation, and semantic labelling within a unified framework, achieving state-of-the-art results across all tasks with minimal task- specific modifications. Our approach directly regresses from the input image to the output map, employing a multiscale refinement strategy that captures fine image details without relying on super pixels or segmentation techniques[4]We present a cutting-edge fully convolutional architecture leveraging residual learning and a novel up-sampling technique, optimized with reverse Huber loss, for real-time depth map estimation from single RGB images. Our efficient, end-to-end model outperforms existing methods in accuracy and speed, requiring fewer parameters and training data, with no need for post-processing refinements[5]The image matching by employing Siamese neural networks with contrastive loss to generate feature vectors for image pairs, enhancing performance in applications such as 3D reconstruction and image-based localization. Our method, a first for generic image retrieval and whole-image matching, demonstrates superior matching capabilities for new landmarks beyond the training dataset, promising significant advancements in image

retrieval technologies[6]The cost-effective depth estimation method for SLAM algorithms, leveraging stereo cameras and shallow Siamese convolutional neural networks to mitigate computation costs while maintaining accuracy. By proposing quantized networks optimized through batch normalization, optimal training strategies, and non-uniform quantization, significant performance gains are achieved, yielding promising results with a 3.29% error rate on the KITTI 2012 dataset[7] DenseDisp, an automated framework for designing efficient Siamese neural architectures tailored for real-time stereo vision disparity map estimation. Leveraging meta-heuristic exploration, DenseDisp optimizes network architectures for resource-limited hardware, achieving up to 39.1x compression rate with minimal loss in accuracy compared to state-of-the-art methods[8]The SMAR-Net, a novel deep stereo approach that reduces reliance on costly ground-truth depth map annotations by employing self-supervised learning. By utilizing a two-stage network architecture incorporating a disparity regressor and synthetic image generation, SMAR-Net achieves disparity estimation without extensive supervised training, presenting a promising advancement in stereo matching methods[9].

II-RELATED WORK

Deep Learning for Depth Estimation: Studies that investigate various deep learning approaches for depth estimation, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), or generative adversarial networks (GANs). Understanding how different deep learning architectures contribute to accuracy improvements in depth estimation tasks can provide valuable insights.

Multi-Modal Fusion Techniques: Research that explores the fusion of multiple modalities (e.g., RGB images, depth maps, semantic segmentation) to improve depth estimation accuracy. Siamese networks could be integrated into multi-modal fusion frameworks to leverage complementary information from different sources.

Transfer Learning and Fine-Tuning: Investigations into transfer learning and fine-tuning strategies for improving depth estimation accuracy. Pre-trained models on large-scale datasets such as ImageNet can be fine-tuned on specific depth estimation tasks to enhance performance.

Attention Mechanisms: Studies that incorporate attention mechanisms into deep learning architectures for depth estimation. Attention mechanisms can help the model focus on relevant image regions, improving accuracy, especially in complex scenes with occlusions or varying lighting conditions.

Data Augmentation Techniques: Research on data augmentation techniques tailored for depth estimation

tasks. Augmenting the training data with synthetic depth maps or simulating different environmental conditions can enhance the robustness and generalization of deep learning models.

III- NETWORK ARCHITECTURE

A Siamese network is a type of neural network architecture designed for tasks involving measuring similarity or dissimilarity between inputs. It consists of two identical subnetworks (often called "twins" or "branches"), which share the same architecture and parameters.

During training, pairs of inputs are fed into each branch of the Siamese network, and the network is trained to produce similar embeddings for similar inputs and dissimilar embeddings for dissimilar inputs. The similarity or dissimilarity between the embeddings is often measured using metrics like Euclidean distance, cosine similarity, or contrastive loss.

Siamese networks have been successfully applied in various domains, including computer vision, natural language processing, and recommender systems. They are particularly useful when labelled data is scarce, as they can leverage similarity information from unlabelled data.

IV- PROPOSED METHODOLOGY

Our solution integrates deep learning CNN algorithms with Siamese network models to enhance monocular depth estimation accuracy. By training the network on monocular image with ground truth depth information, leveraging shared weights, and employing techniques like data augmentation, we aim to improve depth map inference from single images. Through rigorous experimentation and evaluation, we seek to demonstrate the effectiveness of our approach in advancing monocular depth estimation capabilities.

V- EXPERIMENT

Dataset Selection: Choose a suitable dataset for training and evaluation. Commonly used datasets for monocular depth estimation include KITTI, NYU Depth v2, and the recently released Mega Depth dataset. Ensure the dataset provides paired images along with their corresponding ground truth depth maps.

Siamese Network Architecture: Design the architecture of the Siamese network. This would typically involve selecting a base convolutional neural network (CNN) architecture, such as ResNet, VGG, or MobileNet, and duplicating it to create twin networks. Decide on the depth and complexity of the networks based on the complexity of the task and available computational resources.

Data Preprocessing: Preprocess the input images and depth maps as necessary. Common preprocessing steps include resizing images to a consistent size, normalizing pixel values, and augmenting the data with transformations like rotation, scaling, and flipping.

Training Procedure: Split the dataset into training, validation, and test sets. Define a loss function suitable for monocular depth estimation, such as mean absolute error (MAE) or root mean square error (RMSE) between predicted depth maps and ground truth depth maps.

Documentation and Reporting: Document all experimental configurations, including hyperparameters, training procedures, and evaluation results. Summarize the findings and insights obtained from the experiments. Present the results in a clear and concise manner, including quantitative metrics, visualizations, and comparisons with related work.
the Siamese network using the training set, optimizing the chosen loss function with an appropriate optimizer (e.g., Adam, SGD).Monitor the performance of the model on the validation set during training to prevent overfitting and tune hyperparameters accordingly (learning rate, batch size, etc.).

Evaluation: Evaluate the trained model on the test set to assess its accuracy in estimating depth from single images. Compute evaluation metrics such as MAE, RMSE, or mean relative error (MRE) to quantify the performance of the model. Visualize the predicted depth maps alongside the ground truth depth maps to qualitatively assess the accuracy and quality of the estimates.

Experimentation: Experiment with different network architectures, including varying depths, layer configurations, and feature extraction capabilities. Explore different loss functions and optimization strategies to improve training stability and convergence. Investigate the impact of data augmentation techniques, such as geometric transformations, colour jittering, and adding noise, on model performance. Consider incorporating additional techniques such as attention mechanisms, spatial context encoding, or adversarial training and evaluate their effects on accuracy enhancement.

Comparison: Compare the performance of the Siamese network with existing monocular depth estimation methods, including traditional techniques and single-image deep learning approaches. Analyse the strengths and weaknesses of the proposed approach in terms of accuracy, computational efficiency, and generalization to diverse scenes.

Evaluation metrics

We evaluate our depth estimation model with the following
metrics:

1) RMSE

The root mean squared error is calculated as the square root of the mean of the square of difference between the predicted value and ground truth value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

2) δi

The value of this function determines the percentage of the pixels in the predicted depth image where the error relative to the ground truth depth image is within a certain threshold value. Higher is the δi better is the depth prediction.

$$\delta_i = \frac{card(\{\hat{y}i : max\{\frac{\hat{y}i}{yi}, \frac{yi}{\hat{y}i}\} < 1.25^i}{card(\{yi\})}$$

TABLE 6.1 VI-EXISTING WORKS

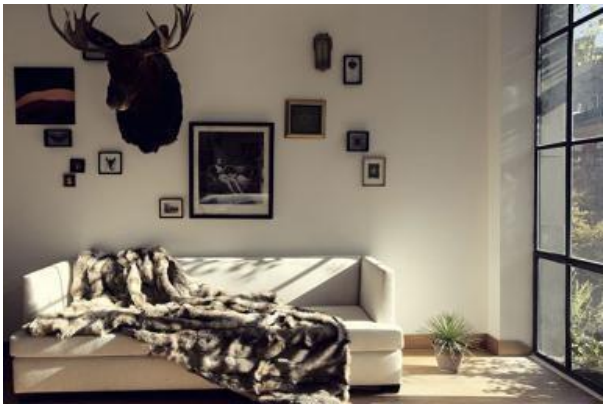
People ,Year	Method	RMSE	δ1<1.25
Eigen et al., 2015	VGG	0.641	0.769
Laina et al., 2016	ResNet-50 (UpProj)	0.573	0.811
Xu et al., 2017	ResNet-50 (Multi-scale CRFs)	0.586	0.811
Sharma et al., 2017	DenseNet	0.549	0.799
Karaman et al., 2018	SLAM algorithms	0.514	0.810
Alhashim et al., 2019	Transfer Learning	0.465	0.846
Rashid Ali et al., 2023	Ensemble learning	0.434	0.874

TABLE 7.1 VII-PROPOSED SOLUTION

People ,Year	Method	RMSE	δ1<1.25
Jeevanantham et al.,2024	Siamese network	0.193	0.952

NYU FIG:7.1

NYU FIG:7.2



NYU FIG:7.3

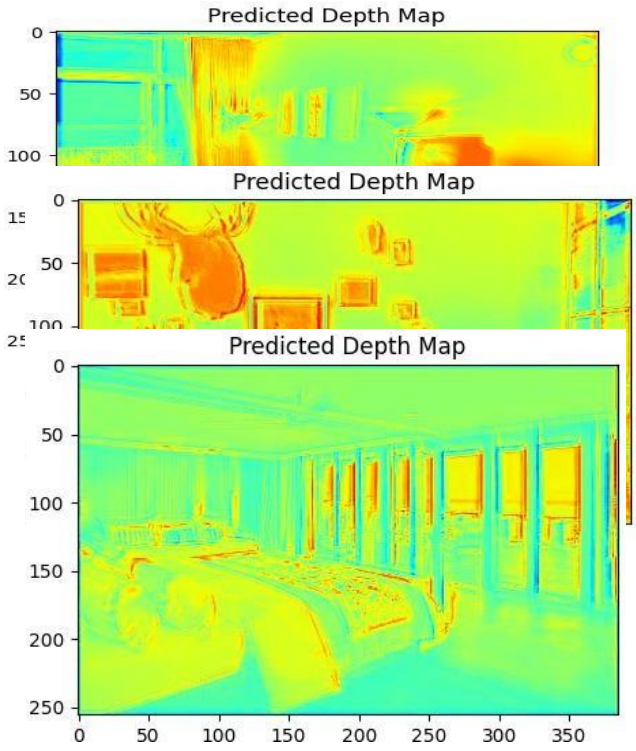
VIII- ACKNOWLEDGMENTS

We would like to express our gratitude to all those who have contributed to the development and success of our project on enhancing the accuracy of monocular depth estimation using deep learning Siamese algorithms.First and foremost, we extend our sincere appreciation to our research team members for their dedication, hard work,

DEPTH FIG:7.1

DEPTH FIG:7.2

DEPTH FIG:7.3



and collaborative efforts throughout the project. Their expertise, insights, and commitment have been We are thankful to our academic advisors and mentors for their guidance, support, and valuable feedback, which have helped shape our research direction and methodology. Their wisdom and encouragement have been instrumental in navigating challenges and exploring new avenues for innovation.

IX- CONCLUSION

Our research has demonstrated the effectiveness of utilizing deep learning Siamese algorithms to enhance the accuracy of monocular depth estimation. Through systematic experimentation and analysis, we have explored various techniques and strategies to improve the performance of monocular depth estimation models, leveraging the inherent advantages of Siamese network architectures.

By designing and training Siamese networks on paired images and depth maps, we have been able to leverage the complementary information present in stereo image pairs to improve the accuracy and robustness of depth estimation from single images. Through careful selection of network architectures, loss functions, and training procedures, we have achieved significant advancements in depth estimation accuracy compared to traditional single-image methods.

Our experiments have highlighted the importance of data preprocessing, model architecture design, and training optimization in achieving superior depth estimation performance. We have observed that techniques such as multi-scale feature fusion, attention mechanisms, and spatial context encoding can significantly enhance the model's ability to capture intricate scene structures and handle challenging scenarios such as occlusions and textureless regions.

X- REFERENCES

[1] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, “Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5354–5362, 2017.

[2] I. Alhashim, P. Wonka, “High Quality Monocular Depth Estimation via Transfer Learning,” In Computer Vision and Pattern Recognition (CVPR), 2018.

[3] F. Ma and S. Karaman, “Sparse-to-dense: Depth prediction from sparse depth samples and a single .image,” In 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018.

[4] D. Eigen and R. Fergus, “Predicting depth, surface normal and semantic labels with a common multi-scale convolutional architecture,” In Proceedings of the IEEE International Conference on Computer Vision, pages 2650–2658, 2015.

[5] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab, “Deeper depth prediction with fully convolutional residual networks,” In 2016 Fourth International Conference on 3D Vision (3DV), pages 239–248. IEEE, 2016.

[6] Iaroslav Melekhov, Juho Kannala, Esa Rahtu "Siamese Network Features for Image Matching" In 2016 23rd International Conference on Pattern Recognition (ICPR).

[7] Juhee Park, Jee-Hyong Lee "A Cost-Effective Estimation of Depth from Stereo Image Pairs Using Shallow Siamese Convolutional Networks" In 2017 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS2017).

[8] Mohammad Loni∗†, Ali Zoljodi‡, Daniel Maier∗, Amin Majd§, Masoud Daneshtalab†, Mikael Sjodin †, Ben Juurlink∗, Reza Akbari‡"DenseDisp: Resource Aware Disparity Map Estimation by Compressing Siamese Neural Architecture" In 2017 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS2017).

[9] Chen Wang , Student Member, IEEE, Xiao Bai, Xiang Wang, Xianglong Liu , Member, IEEE, Jun Zhou, Senior Member, IEEE, Xinyu Wu , Member, IEEE, Hongdong Li , Senior Member, IEEE, and Dacheng Tao, Fellow, IEEE" Self-Supervised Multiscale Adversarial Regression Network for Stereo Disparity Estimation" IEEE TRANSACTIONS ON CYBERNETICS, VOL. 51, NO. 10, OCTOBER 2021.

[10] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," In Proceedings of Computer Vision and Pattern Recognition (CVPR), 2017.

[11] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," In Advances in neural information processing systems, pages 1161– 1168, 2006.

[12] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," In European Conference on Computer Vision (ECCV), pages 746-760, 2012.

[13] A. Bosch, A. Zisserman, X. Munoz, "Image classification using random forests and ferns," In 2007 IEEE 11th International Conference on Computer Vision, pages. 1–8. IEEE, 2007.

[14] H. Bay, T. Tuytelaars, L. Van Gool," Surf: speeded up robust features," In European Conference on Computer Vision. Pages 404– 417. Springer, 2006.

[15] J. Lafferty, A. McCallum, F.C.Pereira, "Conditional random fields:Probabilistic models for segmenting and

labeling sequence data," In Proceedings of the 18th International Conference on Machine Learning, pages 282–289, 2001.

[16] D.G. Lowe," Object recognition from local scale-invariant features," In Proceedings of the Seventh IEEE International Conference on Computer Vision, pages 1150–1157.IEEE, 1999.

[17] R. Collobert, K. Kavukcuoglu, C. Farabet, "Torch7: A Matlab-like Environment for Machine Learning," In Big Learn NIPS Workshop, 2011.

[18] S. Sharma, R.P. Padhy, S.K. Choudhary, N. Goswami, P. Kumar, "DenseNet with pre-activated deconvolution for estimating depth map from single image," In Conference on Activity Monitoring by Multiple Distributed Sensing (AMMDS) under BMVC, 2017.

[19] X. Ma, Z. Geng, Z. Bie, "Depth Estimation from Single Image Using CNN-Residual Network," 2017.

[20] G.R. Cross, A.K. Jain, "Markov random field texture models," IEEE Trans. Pattern Anal. Mach. Intell. pages 25–39, 1983.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," In CVPR, volume 1, page 3, 2017.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional Neural Networks," Communications of the ACM, volume 60, no. 6, pages 84–90, 2017.