Single Image Depth Estimation using SparK Encoder Decoder

Ramsha Imran
phdcs22003@itu.edu.pk

Sadia Saeed
msds22012@itu.edu.pk

Salwa Batool
msds22022@itu.edu.pk

Momna Ibrahim
msds22054@itu.edu.pk

Abstract

We address the problem of Depth Estimation. Where we are taking an image and conducting depth estimation on that image using convolutional U-NET++ architecture. So the original paper [3] from where we inspired used random masking and hierarchical convolutional. But due to the system limitation, we are only able to work on U-NET++ architecture. The results shown are taken from our U-NET++ architecture which we train on the NYS Depth V2 dataset.

1. Introduction

Masked image modeling (MIM) [3] is a technique used in computer vision and image processing to replace or repair damaged or missing portions of an image. The fundamental goal of MIM is to develop a model that, given the information at hand, can forecast which portions of the image are missing or distorted. This can be helpful in a number of situations, including image in-painting, where a portion of an image is unintentionally or purposefully deleted, and image denoising, where an image is distorted by noise.

Depth estimation is a fundamental task in computer vision that involves inferring the distance or depth information from 2D images or image sequences. It plays a crucial role in understanding the three-dimensional structure of a scene and has a wide range of applications in various fields, including robotics, autonomous driving, augmented reality, virtual reality, and more.

The ability to estimate depth from images enables machines to perceive the world in a manner similar to humans, where depth perception is essential for spatial understanding and interaction. By accurately estimating depth, machines can infer the relative distances between objects, identify their positions in the scene, and reconstruct the 3D geometry of the environment.

Depth estimation algorithms leverage various techniques and technologies to infer depth information from images. Traditional approaches often rely on stereo vision, where depth is computed by analyzing the disparities between corresponding points in a pair of stereo images. These methods typically require calibrated camera setups and make assumptions about the scene geometry.

In recent years, deep learning-based approaches have revolutionized depth estimation. Convolutional neural networks (CNNs) have shown remarkable performance in learning depth representations directly from images. By training on large-scale datasets with ground truth depth information, CNNs can capture complex depth cues and generalize well to unseen images.

Depth estimation has numerous practical applications. In robotics, it enables accurate perception for tasks such as object manipulation, scene understanding, and navigation. In autonomous driving, depth estimation helps in scene reconstruction, obstacle detection, and path planning. Augmented reality and virtual reality systems benefit from depth estimation for realistic virtual object placement, environment reconstruction, and immersive user experiences.

Overall, depth estimation is a fundamental task in computer vision that empowers machines with the ability to perceive depth and understand the 3D structure of the world. It continues to be an active area of research, driving advancements in algorithms, sensors, and applications, with the goal of achieving more accurate and robust depth estimation in various real-world scenarios.

Many neural network architectures, including autoencoders, generative adversarial networks (GANs), and convolutional neural networks (CNNs), can be used for this operation.

1.1. Objective

A computer vision problem called depth estimation seeks to infer distance or depth information from 2D images. It is crucial for comprehending a scene's 3D structure and has uses in augmented reality, robotics, virtual reality, and other fields. It makes it possible to do tasks like object identification, obstacle detection, placement of realistic virtual objects, and immersive virtual experiences by inferring depth. In many fields, depth estimation is essential for

improving perception and interaction with the environment.

1.2. Motivation

The motivation behind depth estimation lies in enhancing computer vision applications, such as robotics, where accurate depth perception is vital for object recognition, scene understanding, and navigation. Depth estimation also plays a significant role in autonomous driving systems, enabling vehicles to perceive obstacles, detect road structures, and estimate scene depth for improved safety and efficiency. Additionally, depth estimation finds applications in augmented reality (AR) and virtual reality (VR), enhancing the realistic placement of virtual objects and creating immersive virtual environments. Overall, depth estimation is driven by the desire to enable machines to perceive depth like humans, opening up possibilities for advancements in robotics, autonomous driving, AR, and VR, and enabling machines to interact with the world intelligently and intuitively.

2. Literature Review

2.1. Masked Image Modeling

The paper [6] introduces a system where they input the random, clockwise, and square-masked images and extract the features using encoders of the Vit and Swim transformer. By using that features they predict they predict each pixel value using a one-layer prediction head. The datasets they have used are ImageNet-22K

The paper [5] introduces the Masked Feature Prediction (MaskFeat) approach, which computes five distinct features and classifies the extraction methods into two categories: one-stage and two-stage extractors. For the one-stage method, the authors extracted hand-crafted features such as the histogram of gradient (HOG) and pixel colors. For the two-stage method, they utilized feature extraction techniques such as the Discrete Variational Autoencoder (dVAE), Convolutional Neural Network (CNN), and Video Transformer (ViT). The proposed system was pretrained on unlabeled data and demonstrated superior performance compared to state-of-the-art systems. The authors conducted experiments using the Kinetics and ImageNet-21K datasets.

The paper [7] introduces a new method called denoising contrast MIM (ConMIM) for self-supervised visual representation learning, which is a pure MIM method with advanced downstream performance. The authors analyze the key factors that make MIM more effective on vision Transformers (ViTs) and emphasize the potential broader impacts of their work on revitalizing contrastive learning and inspiring further research in NLP and multimodal domains.

2.2. Depth Estimation

For monocular depth estimation, Kim D. et al. [1] provide a structure and training method. They have used a hierarchical transformer encoder to record and transmit the overall situation and create a compact yet effective decoder to provide an approximated depth map while taking into account local connection. The network can merge both representations and recover fine details by building linked routes between multi-scale local features and the global decoding stream using our suggested selective feature fusion module. The suggested decoder also performs better than those that were previously offered while requiring much less computing complexity. In addition, we improve the depth-specific augmentation approach by enhancing the model using an essential discovery in depth estimation.

BinsFormer is a cutting-edge framework that Li Z. et al. [2] developed specifically for classification-regression-based depth estimation. It primarily concentrates on two essential elements of the particular task: 1) Proper adaptive bin creation; and 2) Enough interaction between predictions of the probability distribution and the adaptive bins. To be more precise, we use the Transformer decoder to create bins and, in a new way, see the issue as a direct set-to-set prediction challenge. To better grasp spatial geometry information and estimate depth maps in a coarse-to-fine way, we incorporate a multi-scale decoder structure. In order to increase estimation accuracy, an additional scene understanding question is also suggested. It turns out that models can implicitly acquire important information from an additional environment categorization task.

3. Masking Strategies

There are different masking strategies applied to images. One technique is to replace the pixel value with zero. but it leads to the zero-out problem. Secondly, the technique is to directly drop the pixel values. This technique is mainly used in transformers. Although the second technique has no side effects this technique is only used in transformers because they can deal with the variable length. The third technique is to sparsely drop the pixel values that can be used in convolutional networks. The different masking strategies are shown in the figure below.

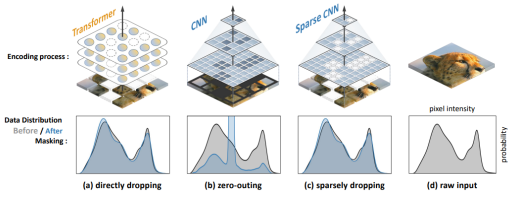

Figure 1. Different Masking Strategies [4]

4. Convolution VS Sparse Convolution

In Convolution, the mask is applied to the whole image while in sparse convolution majority of the image is sparse so convolution is applied to the unmasked areas only. So sparse convolution is more optimized for the computation while convolution networks are a better performer.

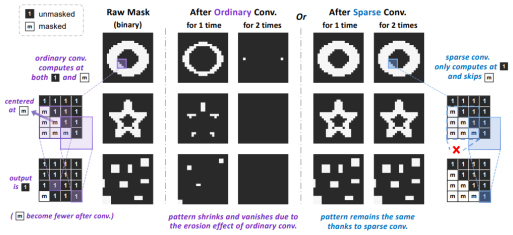

Figure 2. Convolution Vs Sparse Convolution [4]

5. Selected Paper

The paper [4] has applied masked image modeling or BERT-style pre-training to convolutional networks (convnets), they identify and remove two major barriers: Irregular, randomly masked input pictures that cannot be handled by convolution operation, and the single-scale nature of BERT pre-training is incompatible with the hierarchical structure of convnet. They employ sparse convolution to encode for (i) and regard unmasked pixels as sparse voxels of 3D point clouds. This is the first instance of 2D masked modeling using sparse convolution. In order to rebuild pictures from multi-scale encoded information for (ii), they create a hierarchical decoder. This approach, known as Sparse masKed modeling (SparK), is all-encompassing and can be used straight to any convolutional model without requiring changes to the model's backbone. the architecture diagram of the model is shown below:

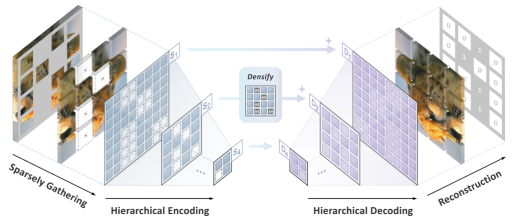

Figure 3. SparK Architecture Diagram [4]

5.1. Dataset

The NYU Depth V2 dataset is a popular dataset for depth estimation, which consists of RGB images and corresponding depth maps. The dataset contains a total of 1,449 pairs of RGB and depth images. Specifically, the dataset is divided into two subsets for training and testing: Training set: The NYU Depth V2 training set consists of 795 RGB-depth image pairs. Testing set: The NYU Depth V2 testing set contains 654 RGB-depth image pairs.

6. Experiments on Spark

Firstly we tried to run the Experiments on the SparK model but we faced system limitation issues. Here are some screenshots of our experiments. are given in the project folder.

7. Proposed Solution

In this project, we have used the UNet++ architecture for convolution to measure the depth of an image.

Training: For training, we have used the Nyu dataset training images and for validation, we have used the test images.

Testing: for testing, we used the testing images. Where we use the encoder part of the model.

Epochs: We have trained the model for 15 epochs.

Learning rate: 0.0001

Batch Size = 32

The model is also uploaded with the name beet_model.pkl in project folder.

7.1. Architectur Diagram

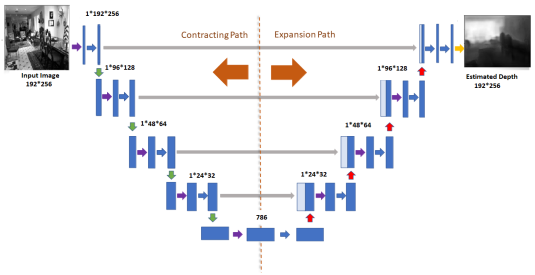

Figure 4. U-NET++Architecture Diagram

We take the Pre-train architecture of U-NET++ on ImageNet.

8. Results

Our results have been shown in Figure 4 and Figure 5. Where the left-most image is the input image. The center image is True Label and the right-most image is the predicted image.

9. Conclusion

SparK convolution [3] presents a good model that is less data-hungry and computationally less expensive than

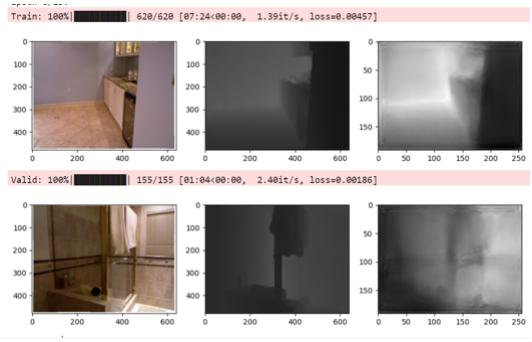

Figure 5. Right Image is input Image, Middle one is True Label and Right one is Predicted Image

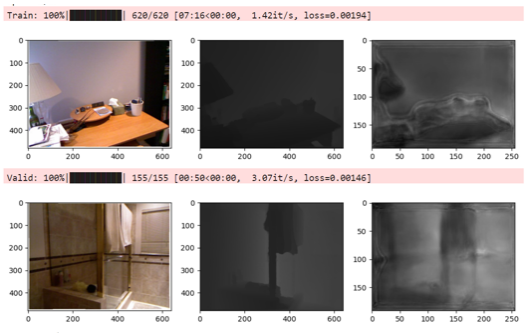

Figure 6. Right Image is input Image, Middle one is True Label and Right one is Predicted Image

the transformers but still computationally expensive. While UNET++ architecture is good to capture the information of objects near the camera and fails to capture the knowledge of small objects. So, It does not perform well on multi-scale. Also, it is unable to detect dark objects in darker regions that's why we can say that it has light effects. Also if the object and background color is similar then the model didn't differentiate well.

10. Contribution

Ramsha Imran (phdcs22003) and Sadia Saeed (msds22022) worked on coding. Salwa Batool and Momna Ibrahim worked on report writing and presentations. Also, all of our group members try to troubleshoot the errors, contribute to poster designing, and run the original Spark Model.

References

[1] Doyeon Kim, Woonghyun Ka, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022. 2

[2] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 2

[3] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 1, 3

[4] Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. *arXiv preprint arXiv:2301.03580*, 2023. 2, 3

[5] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 2

[6] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2

[7] Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast, 2023. 2